Prof. Dr. Michael Herrmann
Technische Universität Braunschweig
Mathematik – Institut iPDE
michael.herrmann@tu-braunschweig.de

Skript der Vorlesung

Analysis 2

für die Studiengänge der Mathematik und der Physik

im Sommersemester 2022

Version vom 2. Oktober 2022

Der Autor ist für Hinweise und Kommentare jederzeit dankbar.

Griechisches Alphabet

klein	*groß*	*Name*	*Laut*
α	A	alpha	a
β	B	beta	b
γ	Γ	gamma	g
δ	Δ	delta	d
ε, ϵ	E	epsilon	ĕ
ζ	Z	zeta	z
η	H	eta	ē
θ, ϑ	Θ	theta	th
ι	I	iota	i
κ	K	kappa	k
λ	Λ	lambda	l
μ	M	my	m

klein	*groß*	*Name*	*Laut*
ν	N	ny	n
ξ	Ξ	xi	x
o	O	omikron	ŏ
π	Π	pi	p
ϱ, ρ	P	rho	r
σ	Σ	sigma	s
τ	T	tau	t
υ	Υ	upsilon	y
φ, ϕ	Φ	phi	ph, f
χ	X	chi	ch
ψ	Ψ	psi	ps
ω	Ω	omega	ō

Literatur

Es gibt viele sehr gute Lehrbücher zur Analysis, zum Beispiel:

[For] Otto Forster: *Analysis 1/2*
12./10. Auflage, Vieweg 2016/2013

[Heu] Harro Heuser: *Lehrbuch der Analysis, Teil 1/2*
11./9. Auflage, Teubner 1990/1991

[Koe] Konrad Königsberger: *Analysis 1/2*
6./5. Auflage, Springer 2004/2004

Anmerkungen

Bei der Ausarbeitung dieses Skriptes hat der Autor die gelisteten Werke regelmäßig konsultiert und dabei viele Beweisstrategien, Präsentationsideen und Beispiele übernommen. Er hat sich darüber hinaus von anderen Quellen sowie den Vorlesungsskripten von Helga Baum (Humboldt-Universität zu Berlin), Martin Brokate/Johannes Zimmer (Technische Universität München), Dirk Lorenz (Technische Universität Braunschweig) und Barbara Niethammer (Universität Bonn) inspirieren lassen. Viele Bilder, Beispiele und Erklärungen wurden auch aus eigenen Ausarbeitungen zu früheren Vorlesungen übernommen.

Der Autor dankt den Studierenden des Kurses, die durch ihre Fragen, Kommentare und Hinweise dieses Skript verbessert haben. Ein besonderer Dank geht an Katia Kleine und Dirk Janßen für das regelmäßige und sehr sorgfältige Korrekturlesen sowie an Harald Löwe für die vielen inhaltlichen Anregungen.

Inhaltsverzeichnis

Kapitel 1

Metrische Räume

Vorlesungswoche 01

Bemerkungen zur Notation Im Folgenden werden wir oftmals die Menge $\mathbb{R}^m$ betrachten, deren Elemente (bzw. *Punkte*) wir mit x bezeichnen. Wir können jedes x via

$$x = (x_1, \ldots, x_m) \qquad \text{bzw.} \qquad x = \begin{pmatrix} x_1 & \ldots & x_m \end{pmatrix} \qquad \text{bzw.} \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

als *Tupel* bzw. *Zeilenvektor* bzw. *Spaltenvektor* schreiben, wobei x_j gerade die j-te Komponente ist.[1] In den Anwendungswissenschaften werden oftmals spezielle Notationen für vektorielle Variablen verwendet, d.h. man schreibt zum Beispiel

$$\boldsymbol{x} = (x_1, \ldots, x_m) \qquad \text{oder} \qquad \vec{x} = (x_1, \ldots, x_m)$$

in Tupelnotation. Obwohl diese Schreibweisen viele Vorteile haben, werden sie innerhalb der Mathematik in aller Regel nicht verwendet, da sie auch einige Nachteile mit sich bringen. Außerdem ist Abstraktion eine der Stärken der Mathematik. Wir werden zum Beispiel sehen, dass sowohl Zahlen, Vektoren und Funktionen jeweils als Punkte in einem abstrakten Raum verstanden und in einheitlicher Weise behandelt werden können.[2]

Für jeden Vektor $x \in \mathbb{R}^m$ ist sein *euklidischer* (oder *kartesischer*) Betrag durch

$$|x| = \left(\sum_{j=1}^{m} x_j^2 \right)^{1/2} = \sqrt{x_1^2 + \ldots + x_m^2}$$

gegeben und wir werden sehen, dass die Betragsfunktion $|\cdot| : \mathbb{R}^m \to \mathbb{R}$ als spezielle Norm auf $\mathbb{R}^m$ angesehen werden kann.

Eine weitere Besonderheit sind Folgen von vektoriellen Größen, da dann unterschiedliche Indizes verwendet werden müssen. Wir schreiben eine Folge im $\mathbb{R}^m$ meist als

[1] In einigen Bereichen der Mathematik muss man sauber zwischen Spalten- und Zeilenvektoren unterscheiden. Auch wir werden dies im Fortgang dieser Vorlesung gelegentlich tun, aber im Moment ist dies nicht wichtig.

[2] In der Mathematik ist *Raum* ein Synonym für *Menge mit Struktur*, wobei diese Struktur sehr unterschiedlicher Natur sein kann.

$(x_n)_{n\in\mathbb{N}}$, wobei die Komponenten von $x_n \in \mathbb{R}^m$ mit $x_{n,j}$ bezeichnet werden. Dies meint zum Beispiel

$$x_n = (x_{n,1},\, \ldots,\, x_{n,m})\,.$$

Eine alternative Notation für Folgen (die wir aber in der Regel nicht verwenden werden) ist $\left(x^{(n)}\right)_{n\in\mathbb{N}}$ mit

$$x^{(n)} = \left(x_1^{(n)},\, \ldots,\, x_m^{(n)}\right),$$

wobei der untere Index j die Komponenten und der obere Index n die Folgenglieder nummeriert.

Achtung Mathematische Notationen sind immer kontextabhängig! Sie müssen sich bei jeder Formel immer klarmachen, ob ein verwendetes Symbol (etwa x, u, α, oder B) nun eine reelle Zahl, ein Element des $\mathbb{R}^m$ oder etwas ganz anderes (Funktion, Menge, ...) bezeichnet.

1.1 Grundbegriffe

Vorbemerkung In diesem Abschnitt beschäftigen wir uns mit einem vergleichsweise abstrakten Gebiet der Analysis, das uns aber wichtige Hilfsmittel für die Differentialrechnung und die Theorie der Differentialgleichungen bereitstellen wird.

Definition Ein metrischer Raum $(X,\, d)$ besteht aus einer Menge X sowie einer Abbildung $d : X \times X \to \mathbb{R}$, sodass die folgenden Aussagen erfüllt sind.

1. *Positivität*: Es gilt

   $$d(x,\, x) = 0\,, \qquad d(x,\, y) > 0$$

   für alle $x \in X$ und alle $y \in X$ mit $y \neq x$.

2. *Symmetrie*: Die Gleichung

   $$d(x,\, y) = d(y,\, x)$$

   ist für alle $x,\, y \in X$ erfüllt.

3. *Dreiecksungleichung*: Es gilt

   $$d(x,\, z) \leq d(x,\, y) + d(y,\, z)$$

   für alle $x,\, y,\, z \in X$.

Die Funktion d wird dabei Metrik oder Abstandsfunktion (auf der Menge X) genannt.

Interpretation Die Funktion d ermöglicht es uns, in sinnvoller Weise den Abstand von zwei Elementen bzw. *Punkten* aus der Menge X zu quantifizieren, wobei es gewisse abstrakte Spielregeln gibt. Damit können wir (siehe weiter unten) Begriffe wie *Konvergenz*, *Stetigkeit*, *Kompaktheit* einführen, die die analogen Konzepte aus *Analysis 1* verallgemeinern. Dabei ist es nicht wichtig, ob die Elemente von X Zahlen, Vektoren oder andere mathematische Objekte sind.

Beispiele

1. Auf jeder Menge X wird durch
$$d(x,\, y) = \begin{cases} 0 & \text{falls } x = y \\ 1 & \text{sonst} \end{cases}$$
die sogenannte triviale Metrik definiert, wobei die geforderten Eigenschaften leicht überprüft werden können (Übungsaufgabe). Es handelt sich allerdings um ein sehr entartetes Beispiel.

2. Jede Norm auf einem Vektorraum induziert eine entsprechende Metrik (siehe unten), aber nicht jede Metrik auf einem Vektorraum wird durch eine Norm erzeugt. Außerdem können Metriken auf beliebigen Mengen existieren.

3. Ist $\eta : \mathbb{R} \to \mathbb{R}$ eine stetige und strikt monotone Funktion, so wird durch
$$d(x,\, y) = \bigl|\eta(x) - \eta(y)\bigr|$$
eine Metrik auf $X = \mathbb{R}$ definiert.

   Beweis: Die Positivität und die Symmetrie von d folgen unmittelbar aus der Definition, wobei die strikte Monotonie von η wichtig ist. Die Dreiecksungleichung für d ergibt sich via
$$d(x,\, z) = \bigl|\eta(x) - \eta(z)\bigr| \le \bigl|\eta(x) - \eta(y)\bigr| + \bigl|\eta(y) - \eta(z)\bigr| = d(x,\, y) + d(y,\, z)$$
direkt aus der normalen Dreiecksungleichung, d.h. den Eigenschaften der reellen Betragsfunktion. □

4. Wir betrachten $X = \mathbb{R}^2$ sowie einen festen Punkt $p \in \mathbb{R}^2$. Dann wird durch
$$d(x,\, y) = \begin{cases} \bigl|x - y\bigr| & \text{falls } x,\, y \text{ und } p \text{ alle auf einer Geraden liegen} \\ \bigl|x - p\bigr| + \bigl|y - p\bigr| & \text{sonst} \end{cases}$$
die französische Eisenbahnmetrik definiert.[3]

Definition Ein normierter Raum $(X,\, \|\cdot\|)$ besteht aus einem reellen Vektorraum X sowie einer Abbildung $\|\cdot\| : X \to \mathbb{R}$, die den folgenden Bedingungen genügt.[4]

1. *Positivität*: Es gilt
$$\|0\| = 0\,, \qquad \|x\| > 0$$
für alle $x \in X$ mit $x \neq 0$.

2. *Homogenität*: Für jedes $\lambda \in \mathbb{R}$ und jedes $x \in X$ gilt
$$\|\lambda\, x\| = |\lambda|\, \|x\|\ ,$$
wobei $|\lambda|$ der Betrag von λ ist.

3. *Subadditivität* bzw. *Dreiecksungleichung*: Die Abschätzung
$$\|x + y\| \le \|x\| + \|y\|$$
gilt für alle $x,\, y \in X$.

Die Funktion $\|\cdot\|$ wird dabei Norm (auf X) genannt.

[3] Der Name kommt daher, dass in früheren Zeiten jede Eisenbahnverbindung zwischen zwei französischen Städten (x und y) über Paris (p) lief. Siehe auch Wikipedia.

[4] Wir betrachten im Folgenden immer Vektorräume über dem Zahlenkörper $\mathbb{R}$, aber man kann Normen ganz analog auch auf komplexen Vektorräumen definieren. Auch in diesem Fall ist aber $\|x\|$ immer eine nichtnegative reelle Zahl.

Lemma (**normierte Räume sind auch metrisch**) Auf einem normierten Raum wird durch

$$d(x,\, y) := \|x - y\|$$

eine Metrik definiert.

Beweis Nachrechnen! □

Merkregel Normierte Räume sind spezielle Beispiele für metrische Räume. Zum einen ist die zugrunde liegende Menge X ein Vektorraum, zum anderen wird die Metrik durch eine Norm erzeugt.

Skalarprodukte* Ein Skalarprodukt auf einem endlich-dimensionalen Vektorraum X ist eine Abbildung $\langle \cdot,\, \cdot \rangle : X \times X \to \mathbb{R}$, die den folgenden Bedingungen genügt.[5]

1. *Positivität*: Es gilt $\langle 0,\, 0 \rangle = 0$ sowie

   $$\langle x,\, x \rangle > 0$$

   für jedes $x \in X$ mit $x \neq 0$.

2. *Bilinearität*: Die Formeln

   $$\big\langle \lambda\, x + \tilde{\lambda}\, \tilde{x},\, y \big\rangle = \lambda \big\langle x,\, y \big\rangle + \tilde{\lambda} \big\langle \tilde{x},\, y \big\rangle\,, \qquad \big\langle x,\, \mu\, y + \tilde{\mu}\, \tilde{y} \big\rangle = \mu \big\langle x,\, y \big\rangle + \tilde{\mu} \big\langle x,\, \tilde{y} \big\rangle$$

   sind für alle x, $\tilde{x}$, y, $\tilde{y}$ aus X und alle λ, $\tilde{\lambda}$, μ, $\tilde{\mu} \in \mathbb{R}$ erfüllt.

3. *Symmetrie*: Es gilt

   $$\big\langle x,\, y \big\rangle = \big\langle y,\, x \big\rangle$$

   für alle x, $y \in X$.

Lemma* (**Normen und Skalarprodukte**) Sei X ein reeller Vektorraum mit endlicher Dimension. Dann erzeugt jedes Skalarprodukt via

$$\|x\| := \sqrt{\langle x,\, x \rangle}$$

eine entsprechende Norm auf X, die außerdem der Parallelogramm-Identität

$$\tfrac{1}{2} \|x + y\|^2 + \tfrac{1}{2} \|x - y\|^2 = \|x\|^2 + \|y\|^2$$

genügt.

Beweis Mit den abstrakten Eigenschaften eines Skalarproduktes können wir leicht nachrechnen, dass für jedes gegebene Skalarprodukt durch die gegebene Formel eine Norm auf X definiert wird und dass diese auch die Parallelogramm-Identität erfüllt (Übungsaufgabe). □

[5] Skalarprodukte können analog auch für unendlich-dimensionale Vektorräume eingeführt werden, aber dann wird üblicherweise zusätzlich noch die Stetigkeit der Abbildung $\langle \cdot,\, \cdot \rangle$ gefordert.

Bemerkungen

1. Nicht für jede Norm gibt es ein entsprechendes Skalarprodukt. Als Beispiel sei auf die weiter unten eingeführten p-Normen im $\mathbb{R}^m$ verwiesen, für die es nur im Fall von $p = 2$ ein Skalarprodukt gibt.

2. *Ausblick*: Der *Satz von Jordan-von Neumann* garantiert, dass eine gegebene Norm genau dann durch ein Skalarprodukt erzeugt wird, wenn die Parallelogramm-Identität gilt, wobei dieses Skalarprodukt mithilfe der Polarisationsformel

$$\langle x, y \rangle = \tfrac{1}{4} \|x + y\|^2 - \tfrac{1}{4} \|x - y\|^2 \ .$$

   bestimmt werden kann. Für den erstaunlich schwierigen Beweis sei auf die Literatur verwiesen.

Konvergenz in metrischen Räumen

Definition (**Konvergenz von Folgen in metrischen Räumen**) Eine Folge $(x_n)_{n\in\mathbb{N}} \subset X$ konvergiert (bzgl. der Metrik d) gegen einen Grenzwert $x_\infty \in X$, falls zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ existiert, sodass

$$d(x_n, x_\infty) < \varepsilon \qquad \text{für alle} \quad n > N$$

gilt. In diesem Fall schreiben wir auch $x_\infty = \lim_{n\to\infty} x_n$ und nennen x_∞ den Grenzwert.

Bemerkungen

1. Die Definition verallgemeinert in natürlicher Weise den Konvergenzbegriff für Zahlenfolgen aus *Analysis 1*, wobei wir nun den Abstand von x_n und x_∞ nicht mehr mit der Betragsfunktion $|\cdot|$, sondern mithilfe der Metrik d quantifizieren.

2. *Äquivalente Charakterisierung*: Die Folge $(x_n)_{n\in\mathbb{N}} \subset X$ konvergiert genau dann für $n \to \infty$ gegen x_∞, wenn

$$d(x_n, x_\infty) \quad \xrightarrow{n\to\infty} \quad 0$$

   gilt, wobei letzteres eine Konvergenz im Sinne der reellen Zahlen ist und für normierte Räume auch als

$$\|x_n - x_\infty\| \quad \xrightarrow{n\to\infty} \quad 0$$

   geschrieben werden kann.

   Beweis: Übungsaufgabe. □

3. Wie schon in *Analysis 1* gilt: Nicht jede Folge besitzt einen Grenzwert, aber wenn sie konvergiert, so ist der Grenzwert eindeutig (Übungsaufgabe).

4. Besonders in normierten Räumen gibt es auch alternative, nicht äquivalente Konvergenzbegriffe, die nicht auf Metriken oder Normen, sondern auf anderen Konzepten beruhen. Die Konvergenz im Sinne der obigen Definition wird oftmals auch starke Konvergenz oder Normkonvergenz genannt.

5. *Achtung*: Der Konvergenzbegriff hängt von der Metrik bzw. der Norm ab. Siehe dazu auch die Diskussion weiter unten.

Lemma (**Konvergenz in Norm impliziert Konvergenz der Norm**) In jedem normierten Raum $(X, \|\cdot\|)$ gilt die Implikation

$$x_n \xrightarrow{n\to\infty} x_\infty \qquad \Longrightarrow \qquad \|x_n\| \xrightarrow{n\to\infty} \|x_\infty\| \,,$$

wobei links eine Konvergenz bzgl. der Norm $\|\cdot\|$ und rechts eine Konvergenz reeller Zahlen steht.

Beweis Aus der Subadditivität der Norm ergibt sich

$$\|x_n\| - \|x_\infty\| \le \|x_n - x_\infty\| \qquad \text{sowie} \qquad \|x_\infty\| - \|x_n\| \le \|x_\infty - x_n\| = \|x_n - x_\infty\|$$

und insgesamt

$$\big|\|x_n\| - \|x_\infty\|\big| \le \|x_n - x_\infty\| \,.$$

Die Behauptung folgt nun mithilfe der vorangegangenen Bemerkungen. □

Achtung Die umgekehrte Aussage ist falsch, wobei dies leicht mit Gegenbeispielen aus dem $\mathbb{R}^2$ begründet werden kann (Übungsaufgabe).

Definition Eine Folge $(x_n)_{n\in\mathbb{N}} \subset X$ heißt Cauchy-Folge (bzgl. der Metrik d), falls zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ existiert, sodass

$$d(x_n, x_k) < \varepsilon \qquad \text{für alle} \quad n, k > N$$

gilt. Der metrische Raum (X, d) wird vollständig genannt, falls jede Cauchy-Folge konvergiert.

Bemerkungen

1. *Achtung*: Nicht jeder metrische oder normierte Raum ist vollständig (siehe die Beispiele weiter unten).

2. *Ausblick**: Jeder metrische Raum (X, d) besitzt eine *Vervollständigung*, d.h. es existiert ein anderer metrischer Raum $(\hat{X}, \hat{d})$ der zum einen vollständig ist und zum anderen X als dichte Teilmenge enthält. Insbesondere gilt $X \subseteq \hat{X}$ sowie $\hat{d}(x, y) = d(x, y)$ für alle $x, y \in X$. Zum Beispiel ist $\mathbb{R}$ die Vervollständigung von $\mathbb{Q}$.

3. Ein vollständiger normierter Raum wird Banach-Raum genannt. Wird darüber hinaus die Norm von einem Skalarprodukt erzeugt, so sprechen wir von einem Hilbert-Raum.

Kugeln und Sphären In jedem metrischen Raum ist

$$\overline{B}_\varrho(x_*) := \{x \in X \,:\, d(x, x_*) \le \varrho\} \qquad \text{bzw.} \qquad B_\varrho(x_*) := \{x \in X \,:\, d(x, x_*) < \varrho\}$$

die abgeschlossene bzw. offene Kugel mit Radius $\varrho > 0$ und Mittelpunkt x_*. Die Menge

$$S_\varrho(x_*) = \overline{B}_\varrho(x_*) \setminus B_\varrho(x_*) = \{x \in X \,:\, d(x, x_*) = \varrho\}$$

bezeichnet die entsprechende Sphäre.[6]

Bemerkung: Wenn wir mit mehreren Metriken arbeiten, so werden wir abgewandelte Bezeichnungen verwenden, in denen noch zusätzlich noch die Abhängigkeit von der Metrik d kenntlich gemacht wird.

Ausblick: Reihen in normierten Räumen Ausgehend vom Konvergenzbegriff können wir — ganz analog zu *Analysis 1* – auch Reihen betrachten, wobei die Glieder nun aus einem normierten Raum stammen dürfen. Die Grundidee kann wieder in der Formel

$$\sum_{k=1}^{\infty} x_k = \lim_{n\to\infty} \sum_{k=1}^{n} x_k$$

zusammengefasst werden und meint, dass eine unendliche Summe immer der Grenzwert einer Folge von endlichen Summen ist (eben der *Partialsummenfolge*). Wir werden im Fortgang der Vorlesung Beispiele für solche Reihen kennenlernen und schrittweise die Theorie aufbauen. Insbesondere wird es auch wieder einen absoluten Konvergenzbegriff geben.

1.2 Normen in kartesischen Räumen

Vorbemerkung Das für uns wichtigste Beispiel für einen normierten (und damit auch metrischen) Raum ist der $\mathbb{R}^m$. Darüber hinaus sind aber auch Funktionen- und Folgenräume sehr wichtig in der modernen Mathematik, wobei wir erste Beispiele im Anschluss diskutieren werden.

Wir werden im Folgenden sehen, dass es viele Möglichkeiten gibt, den $\mathbb{R}^m$ mit einer Norm auszustatten, wobei je nach Anwendung mal die eine, mal die andere besser geeignet sein wird.

Definition (**wichtige Normen**) Wir setzen

$$\|x\|_\infty := \max_{j\in\{1,\dots,m\}} |x_j|$$

sowie

$$\|x\|_p := \left(\sum_{j=1}^{m} |x_j|^p\right)^{1/p},$$

wobei der Parameter $p \in [1, \infty)$ oftmals Exponent genannt wird.[7]

Lemma (**nützliche Beobachtung**) Es gilt

$$|x_k| \le \|x\|_p$$

für jedes $x \in \mathbb{R}^m$ sowie alle $k \in \{1, \dots, m\}$ und alle $p \in [1, \infty]$.

[6] B und S beziehen sich auf '*ball*' und '*sphere*'. Deutschsprachige Mathematiker nennen daher Kugeln manchmal auch *Bälle*.

[7] Wir werden gleich zeigen, dass diese Definition für jedes p wirklich eine Norm liefert.

Beweis Für $p = \infty$ gilt

$$\big|x_k\big| \le \sup_{j=\{1,\,\ldots,\,m\}} \big|x_j\big| = \|x\|_\infty$$

und im Fall von $p < \infty$ folgt die Behauptung aus

$$\big|x_k\big|^p \le \sum_{j=1}^m \big|x_j\big|^p = \|x\|_p^p$$

nach Ziehen der p-ten Wurzel. □

Lemma (**fundamentale Ungleichungen**) Für alle $x, y \in \mathbb{R}^m$ und jedes $p \in [1,\,\infty]$ gilt die Minkowski-Ungleichung

$$\|x + y\|_p \le \|x\|_p + \|y\|_p$$

sowie die Hölder-Ungleichung

$$\sum_{j=1}^m \big|x_j\big|\big|y_j\big| \le \|x\|_p \, \|y\|_{p'} \,.$$

Hierbei ist p' der konjugierte Exponent zu p und kann aus der Formel

$$\frac{1}{p} + \frac{1}{p'} = 1 \qquad \text{bzw.} \qquad p' = \frac{1}{1 - \dfrac{1}{p}} = \frac{p}{p-1}$$

berechnet werden, wobei die Sonderregeln $1/\infty = 0$ und $1/0 = \infty$ vereinbart seien.[8]

Beweis *Vorbereitung*: Für $1 < p < \infty$ und zwei nichtnegative reelle Zahlen a und b gilt stets die Young-Ungleichung

$$a\,b \le \frac{a^p}{p} + \frac{b^{p'}}{p'} \,.$$

In der Tat, für $a = 0$ oder $b = 0$ ist diese Aussage trivial und für $a > 0$, $b > 0$ folgt wegen $1/p + 1/p' = 1$ die Abschätzung

$$\ln(a\,b) = \ln\Big(\big(a^p\big)^{1/p}\big(b^{p'}\big)^{1/p'}\Big) = \frac{1}{p}\ln\big(a^p\big) + \frac{1}{p'}\ln\big(b^{p'}\big) \le \ln\left(\frac{1}{p}\,a^p + \frac{1}{p'}\,b^{p'}\right)$$

aus den Rechenregeln sowie der Konkavität des Logarithmus und impliziert das Hilfsresultat nach Anwendung der monotonen Exponentialfunktion auf beiden Seiten.

Hölder-Ungleichung: Für $p = 1$ und $p' = \infty$ bzw. für $p = \infty$ und $p' = 1$ ergibt sich die gewünschte Ungleichung aus einfachen Argumenten (Übungsaufgabe). In allen anderen Fällen gilt $1 < p$, $p' < \infty$ und wir können O.b.d.A.[9] annehmen, dass $\|x\|_p > 0$ und $\|y\|_{p'} > 0$ gilt, da andernfalls nach der nützlichen Beobachtung alle x_j oder alle

[8] Die Sonderregeln implizieren insbesondere $1' = \infty$ und $\infty' = 1$.

[9] o.B.d.A. meint *ohne Beschränkung der Allgemeinheit*.

y_j verschwinden und die Behauptung wegen $0 \leq 0$ trivialerweise richtig ist. Aus der Young-Ungleichung — angewendet mit $a = |x_j| / \|x\|_p$ und $b = y_j / \|y\|_{p'}$ — ergibt sich

$$\frac{|x_j|}{\|x\|_p} \frac{|y_j|}{\|y\|_{p'}} \leq \frac{1}{p} \left(\frac{|x_j|}{\|x\|_p} \right)^p + \frac{1}{p'} \left(\frac{|y_j|}{\|y\|_{p'}} \right)^{p'}$$

für alle j und durch Summation erhalten wir via

$$\frac{1}{\|x\|_p \|y\|_{p'}} \sum_{j=1}^m |x_j| \, |y_j| \leq \frac{1}{p \|x\|_p^p} \sum_{j=1}^m |x_j|^p + \frac{1}{p' \|y\|_{p'}^{p'}} \sum_{j=1}^m |y_j|^{p'} = \frac{1}{p} + \frac{1}{p'} = 1$$

das gewünschte Ergebnis.

Minkowski-Ungleichung: Die Fälle $p = 1$ und $p = \infty$ sind eine Übungsaufgabe. Für $1 < p < \infty$ können wir O.B.d.A. $\|x + y\|_p > 0$ annehmen und bemerken, dass nach der reellen Dreiecksungleichung

$$\begin{aligned} \|x + y\|_p^p &= \sum_{j=1}^m |x_j + y_j|^p = \sum_{j=1}^m |x_j + y_j| \, |x_j + y_j|^{p-1} \\ &\leq \sum_{j=1}^m |x_j| \, |x_j + y_j|^{p-1} + \sum_{j=1}^m |y_j| \, |x_j + y_j|^{p-1} \end{aligned}$$

gilt. Die Hölder-Ungleichung garantiert

$$\sum_{j=1}^m |x_j| \, |x_j + y_j|^{p-1} \leq \left(\sum_{j=1}^m |x_j|^p \right)^{1/p} \left(\sum_{j=1}^m \left(|x_j + y_j|^{p-1} \right)^{p'} \right)^{1/p'} = \|x\|_p \, \|x + y\|_p^{p-1}$$

sowie analog

$$\sum_{j=1}^m |y_j| \, |x_j + y_j|^{p-1} \leq \|y\|_p \, \|x + y\|_p^{p-1} \, ,$$

wobei wir benutzt haben, dass $(p - 1) \, p' = p$ gilt. Insgesamt ergibt sich

$$\|x + y\|_p^p \leq \left(\|x\|_p + \|y\|_p \right) \|x + y\|^{p-1}$$

und das gewünschte Ergebnis folgt nach Division durch $\|x + y\|^{p-1}$. □

Bemerkungen

1. Für $1 < p < \infty$ gilt $1 < p' < \infty$ und wir können leicht zeigen, dass $p'' = p$ gilt. Beachte, dass das Zeichen $'$ in diesem Kontext keine Ableitung meint, sondern aus einem Exponenten einen anderen macht.

2. Die Hölder-Ungleichung und das Konzept des konjugierten Exponenten sind sehr wichtig und tauchen in vielen Bereichen der Mathematik auf.

3. Es gilt $2' = 2$, aber $p \neq p'$ für jeden anderen Exponenten. Beide Tatsachen spiegeln einige tiefe Erkenntnisse wider, die wir uns aber erst schrittweise klar

machen können. Wir wollen aber hier schon festhalten, dass nur für $p = 2$ die p-Norm durch ein Skalarprodukt erzeugt ist, nämlich durch

$$\langle x,\, y\rangle = \sum_{j=1}^{m} x_j\, y_j\,.$$

Insbesondere gilt die oben erwähnte Parallelogramm-Identität zwar für $p = 2$, aber nicht für $p \neq 2$.

4. Die Hölder-Ungleichung mit $p = 2$ ist gerade die Cauchy-Schwarz-Ungleichung.

Folgerung Für jeden Exponenten $p \in [1,\, \infty]$ definiert $\|\cdot\|_p$ eine Norm auf $\mathbb{R}^m$.

Beweis Alle drei Normeigenschaften der Funktion $\|\cdot\|_p : \mathbb{R}^m \to \mathbb{R}$ können leicht nachgerechnet werden (Übungsaufgabe), wobei die entsprechende Dreiecksungleichung gerade die Minkowski-Ungleichung ist.

Bemerkungen

1. Die p-Normen illustrieren, dass ein gegebener Vektorraum (hier $\mathbb{R}^m$) auf mehrere verschiedene Weisen zu einem normierten Raum gemacht werden kann.
2. Der Fall $p = 2$ ist besonders wichtig, wobei dann $\|x\|_2 = |x|$ gilt, d.h. die 2-Norm liefert gerade den *euklidischen* Abstand, den wir bereits aus der Schule kennen.
3. Die Normen für $p = 1$ bzw. $p = \infty$ werden auch Summennorm (oder Manhattan-Norm) bzw. Maximumsnorm genannt.
4. Im Fall $m = 1$ (eine Raumdimension) gilt $\|x\|_p = |x|$, d.h. im $\mathbb{R}^1$ sind alle p-Normen identisch.

Lemma (**Eigenschaften der p-Norm im $\mathbb{R}^m$**) Sei $x \in \mathbb{R}^m$ beliebig. Dann gilt

$$\|x\|_\infty \leq \|x\|_p \leq m^{1/p}\, \|x\|_\infty$$

für alle Exponenten $p \in [1,\, \infty]$ und damit auch

$$\lim_{p\to\infty} \|x\|_p = \|x\|_\infty$$

Hierbei ist x ein beliebiger Punkt im $\mathbb{R}^m$.

Beweis Doppelungleichung: Da die Behauptung für $p = \infty$ wegen $m^{1/\infty} = m^0 = 1$ trivialerweise richtig ist, können wir $1 \leq p < \infty$ annehmen. Für jedes $k \in \{1,\, \ldots,\, m\}$ gilt

$$\big|x_k\big| \leq \|x\|_p$$

nach der nützlichen Beobachtung und wenn wir auf der linken Seite das Maximum über k bilden, erhalten wir $\|x\|_\infty \leq \|x\|_p$. Andererseits gilt auch $\big|x_j\big| \leq \|x\|_\infty$ für alle $j \in \{1,\, \ldots,\, m\}$. Dies impliziert

$$\|x\|_p^p = \sum_{j=1}^{m} \big|x_j\big|^p \leq \sum_{j=1}^{m} \|x\|_\infty^p = m\, \|x\|_\infty^p$$

und damit den zweiten Teil der Doppelungleichung nach dem Ziehen der p-ten Wurzel.

Konvergenz: Die Behauptung ergibt sich mit dem Sandwichprinzip direkt aus der bereits bewiesenen Doppelungleichung. □

Bemerkungen

1. Aus dem Lemma folgt, dass für zwei Exponenten p, $q \in [1, \infty]$ immer

$$\|x\|_p \leq m^{1/p} \|x\|_q \,, \qquad \|x\|_q \leq m^{1/q} \|x\|_p$$

für alle $x \in \mathbb{R}^m$ gilt. Insbesondere impliziert dies, dass eine Folge aus $\mathbb{R}^m$ genau dann bzgl. der p-Norm konvergiert, wenn Sie bzgl. der q-Norm konvergiert und dass eine Teilmenge des $\mathbb{R}^m$ genau dann bzgl. der p-Norm abgeschlossen bzw. offen ist, wenn sie es bzgl. der q-Norm ist (siehe dazu weiter unten).

2. *Verallgemeinerung*: Sind $\|\cdot\|_\mathrm{a}$ und $\|\cdot\|_\mathrm{b}$ zwei verschiedene Normen auf einem *endlich-dimensionalen* Vektorraum X, so sind diese Normen äquivalent, d.h. es existieren positive Konstanten C_a, C_b, sodass

$$\|x\|_\mathrm{a} \leq C_\mathrm{a} \|x\|_\mathrm{b} \,, \qquad \|x\|_\mathrm{b} \leq C_\mathrm{b} \|x\|_\mathrm{a}$$

für alle $x \in X$ gilt.[10]

Achtung: In einem unendlich-dimensionalen Vektorraum gilt diese Aussage nicht, d.h. dort kann es sehr wohl nicht äquivalente Normen geben.[11]

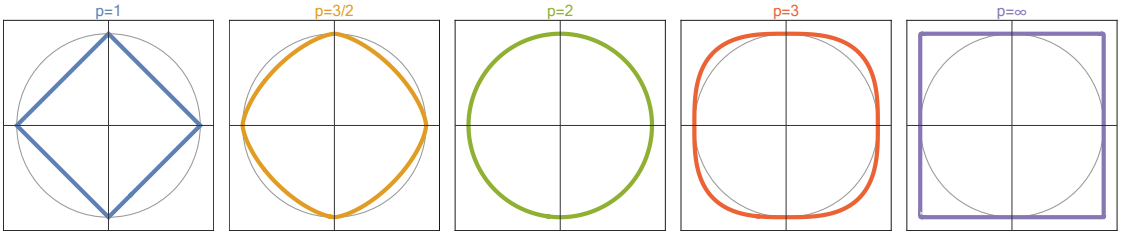

Abbildung Die Einheitssphären im $\mathbb{R}^2$, wobei jeweils eine andere Wahl des Normparameters p zugrunde liegt und der euklidische Standardkreis ($p = 2$) jeweils grau dargestellt ist.

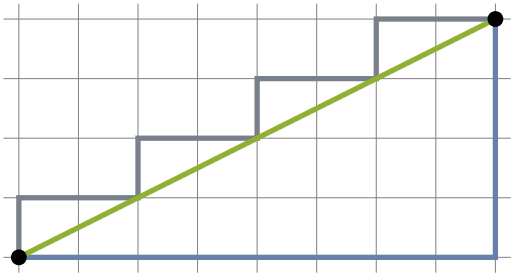

Abbildung In der euklidischen Metrik ($p = 2$) gibt es zu zwei beliebig gewählten Punkten (schwarz) immer genau einen kürzesten Verbindungsweg (grüne Linie), aber in der Manhattan-Metrik ($p = 1$) gibt es viele optimale Wege (zwei sind blau dargestellt). Das allgemeine Konzept *geodätischer Kurven* können wir in dieser Vorlesung aber nicht studieren.

Lemma (**über Konvergenz im $\mathbb{R}^m$**) Sei $p \in [1, \infty]$ ein beliebiger Exponent. Eine Folge $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^m$ konvergiert genau dann bzgl. der p-Norm gegen $x_\infty \in \mathbb{R}^m$, wenn sie komponentenweise konvergiert, d.h. wenn für jedes $j \in \{1, \ldots, m\}$ die Zahlenfolge $(x_{n,j})_{n\in\mathbb{N}}$ gegen $x_{\infty,j}$ konvergiert.

[10] Einen Beweis werden Sie zum Beispiel in der Einführungsvorlesung über Numerische Mathematik kennenlernen.

[11] Wir werden diesen Aspekt in der Vorlesung *Funktionalanalysis* genauer studieren und verstehen.

Beweis *Hinrichtung*: Es gelte $x_\infty = \lim_{n\to\infty} x_n$ im Sinne der Konvergenz bzgl. der p-Norm. Für jedes feste j gilt dann

$$\left|x_{n,\,j} - x_{\infty,\,j}\right| \leq \left\|x_n - x_\infty\right\|_p \quad \xrightarrow{n\to\infty} \quad 0$$

und wir schließen, dass $x_{\infty,\,j} = \lim_{n\to\infty} x_{n,\,j}$ im Sinne der Konvergenz reeller Zahlen gilt. *Rückrichtung*: Aus der Annahme $x_{\infty,\,j} = \lim_{n\to\infty} x_{n,\,j}$ für alle $j \in \{1,\,\ldots,\,m\}$ folgt

$$\left\|x_n - x_\infty\right\|_1 = \sum_{j=1}^{m} \left|x_{n,\,j} - x_{\infty,\,j}\right| \quad \xrightarrow{n\to\infty} \quad 0\,,$$

wobei wir benutzt haben, dass die Summe vom m reellen Nullfolgen wieder eine Nullfolge ist. Insbesondere konvergiert x_n für $n \to \infty$ gegen x_∞ bzgl. der 1-Norm. Da die Eigenschaften der p-Norm beide Teile der Doppelabschätzung

$$\left\|x_n - x_\infty\right\|_p \leq m^{1/p} \left\|x_n - x_\infty\right\|_\infty \leq m^{1/p} \left\|x_n - x_\infty\right\|_1$$

für alle $n \in \mathbb{N}$ implizieren, ergibt sich auch die Konvergenz bzgl. der p-Norm. □

Bemerkung Die Äquivalenzaussage des Theorems scheint auf den ersten Blick trivial zu sein und besitzt naheliegende Verallgemeinerungen in jedem anderen normierten Raum *endlicher* Dimension. In unendlich-dimensionalen Räumen, gibt es aber keine entsprechende Aussage. Siehe zum Beispiel die Eigenschaften der punktweisen Konvergenz in den weiter unten diskutierten Doppelfolgenräumen.

Beispiel Die vektorielle Konvergenz

$$x_n = \begin{pmatrix} \cos\big(\exp(-n)\big) \\ \dfrac{(1+3\,n)^2}{-1+2\,n^2} \end{pmatrix} \quad \xrightarrow{n\to\infty} \quad x_\infty = \begin{pmatrix} 1 \\ \dfrac{9}{2} \end{pmatrix}$$

gilt bzgl. jeder p-Norm im $\mathbb{R}^2$, eben weil sowohl die Folge ersten als auch die Folge der zweiten Komponenten konvergiert.

Theorem (**Vollständigkeit des** $\mathbb{R}^m$) Für jedes $p \in [0,\,\infty]$ ist der $\mathbb{R}^m$ ausgestattet mit der p-Norm ein vollständiger normierter Raum.

Beweis Sei $(x_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge im $\mathbb{R}^m$. Für jedes $j \in \{1,\,\ldots,\,m\}$ impliziert die nützliche Beobachtung via

$$\left|x_{n,\,j} - x_{k,\,j}\right| \leq \left\|x_n - x_k\right\|_p\,,$$

dass die Zahlenfolge $(x_{n,\,j})_{n\in\mathbb{N}}$ eine reelle Cauchy-Folge ist und daher — vgl. das Cauchy-Kriterium für Konvergenz aus *Analysis 1* — einen Grenzwert besitzt, den wir $x_{\infty,\,j}$ nennen wollen und als j-te Komponente von $x_\infty \in \mathbb{R}^m$ auffassen können. Nach Konstruktion konvergiert nun x_n für $n \to \infty$ komponentenweise gegen x_∞ und das vorherige Lemma garantiert die entsprechende Konvergenz bzgl. der p-Norm. □

Normen für Matrizen Die Menge aller reellen $k\times l$-Matrizen (mit k Zeilen und l Spalten) ist ein Vektorraum und in natürlicher Weise isomorph zum $\mathbb{R}^{k\,l}$.[12] Für eine gegebene Matrix $A = (A_{i,\,j})_{i=1\ldots k,\,j=1\ldots l}$ können wir daher durch

$$\|A\|_p = \left(\sum_{i=1}^{k}\sum_{j=1}^{l} \bigl|A_{i,\,j}\bigr|^p\right)^{1/p} \qquad \text{bzw.} \qquad \|A\|_\infty = \max_{i=1\ldots k\,,\,j=1\ldots l} \bigl|A_{i\,,j}\bigr|$$

die entsprechende p-Norm einführen. Es gibt außerdem noch andere wichtige Normen, wie zum Beispiel die *Frobenius-Norm* für quadratische Matrizen.

Ausblick: Matrixexponential* Ist A eine quadratische $m\times m$-Matrix, so wird durch

$$\exp(A) := \sum_{k=0}^{\infty} \frac{1}{k!} A^k = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!} A^k$$

das entsprechende *Matrixexponential* definiert, wobei die Partialsummenfolge bzgl. aller p-Normen konvergiert. Wir werden dieses Konzept im Kapitel über Differentialgleichungen genauer studieren.

1.3 Normen für stetige und beschränkte Funktionen

Vorbemerkung Neben dem $\mathbb{R}^m$ als Prototyp eines endlichdimensionalen normierten Raumes gibt es auch viele unendlichdimensionale Vektorräume mit sinnvollen Normen, zum Beispiel sogenannte *Funktionräume*, deren Elemente bzw. *Punkte* Funktionen mit speziellen Eigenschaften sind.

Im Folgenden betrachten wir ein festes Intervall $J \subset \mathbb{R}$ mit endlicher Länge $|J| < \infty$, wobei wir die Zahlen in J mit t bezeichnen und reellwertige Funktionen auf J als $x : J \to \mathbb{R}$ (oder $y : J \to \mathbb{R}$) schreiben.[13] Wir interessieren uns für den *Funktionenraum*[14]

$$\mathsf{BC}(J) := \bigl\{x : J \to \mathbb{R} \;:\; x \text{ ist beschränkt und stetig}\bigr\}\,,$$

der in natürlicher Weise ein reeller Vektorraum ist, wobei die Addition sowie die skalare Multiplikation durch

$$\bigl(x+y\bigr)(t) = x(t) + y(t)\,, \qquad \bigl(\lambda\, x\bigr)(t) = \lambda\, x(t)$$

gegeben sind.[15] Beachte, dass für ein kompaktes Intervall J jede stetige Funktion beschränkt sein muss, aber dass dies auf offenen oder halboffenen Intervallen nicht richtig ist.

[12] Siehe die Vorlesung *Lineare Algebra*.

[13] Wir hatten schon am Ende der Vorlesung *Analysis 1* die Konvergenz von Funktionenfolgen studiert, aber damals eine andere Notation verwendet, nämlich x statt t und f statt x. Der Wechsel in der Schreibweise macht deutlich, dass die Elemente eines abstrakten normierten Raumes auch Funktionen sein können, ändert aber nichts an der Gültigkeit der mathematischen Ergebnisse.

[14] B und C stehen für *bounded* and *continuous*.

[15] Der Buchstabe λ bezeichnet hier eine reelle Zahl. Wir können natürlich auch zwei Funktion via

$$\bigl(x\,y\bigr)(t) = x(t)\,y(t)$$

miteinander multiplizieren, d.h. $\mathsf{BC}(J)$ ist nicht nur ein reeller Vektorraum, sondern sogar eine *reelle Algebra*. Dieser Aspekt wird aber in diesem Abschnitt keine Rolle spielen.

p-Normen für Funktionen Für jede beschränkte und stetige Funktion auf J sind die Ausdrücke

$$\|x\|_\infty := \sup_{t\in J} \bigl|x(t)\bigr| \qquad \text{und} \qquad \|x\|_p := \left(\int_J \bigl|x(t)\bigr|^p \,\mathrm{d}t \right)^{1/p}$$

wohldefiniert, wobei in der zweiten Formel $1 \leq p < \infty$ vorausgesetzt wird und das Integral im Riemannschen Sinne zu verstehen ist.[16]

Lemma (**fundamentale Ungleichungen für Funktionen**) Die Minkowski- und die Hölderungleichung sind sinngemäß auch für Funktionen erfüllt. Außerdem gilt

$$\|x\|_p \leq |J|^{1/p-1/q} \, \|x\|_q$$

für jede Funktion $x \in \mathsf{BC}(J)$ sowie alle Exponenten p, $q \in [1, \infty]$ mit $p \leq q$.

Beweis *Teil 1*: Die beiden Ungleichungen können analog zu oben hergeleitet werden: Wir müssen nur die endlichen Summen durch Integrale ersetzen. *Teil 2*: Für jeden Exponenten $r \in [1, \infty]$ folgt

$$\begin{aligned} \int_J \bigl|x(t)\bigr|^p \,\mathrm{d}t = \int_J 1 \cdot \bigl|x(t)\bigr|^p \,\mathrm{d}t &\leq \left(\int_J 1^{r'} \,\mathrm{d}t \right)^{1/r'} \left(\int_J \Bigl(\bigl|x(t)\bigr|^p \Bigr)^r \,\mathrm{d}t \right)^{1/r} \\ &\leq |J|^{1/r'} \left(\int_J \bigl|x(t)\bigr|^{p\,r} \,\mathrm{d}t \right)^{1/r} \end{aligned}$$

aus der Hölder-Ungleichung, wobei r' der zu r konjugierte Exponent ist. Mit der speziellen Wahl

$$r := \frac{q}{p}, \qquad \frac{1}{r} = \frac{p}{q} = p\,\frac{1}{q}, \qquad \frac{1}{r'} = 1 - \frac{p}{q} = p\left(\frac{1}{p} - \frac{1}{q}\right)$$

können wir diese Abschätzung als

$$\|x\|_p^p \leq |J|^{p\,(1/p-1/q)} \, \|x\|_q^p$$

schreiben und erhalten die Behauptung nach Ziehen der p-ten Wurzel auf beiden Seiten. □

Folgerung Für jeden Exponenten $p \in [1, \infty]$ ist $\|\cdot\|_p$ eine Norm auf $\mathsf{BC}(J)$.

Beweis Die Minkowski-Ungleichung für Funktionen liefert die Dreiecksungleichung und die Homogenität kann einfach nachgerechnet werden. Außerdem gilt offensichtlich stets $\|x\|_p \geq 0$. Sei nun $x \in \mathsf{BC}(J)$ eine Funktion, die nicht der konstanten Nullfunktion entspricht, wobei letztere gerade der Nullvektor im Funktionenraum ist. Dann existiert

[16] Ist J ein kompaktes Intervall, so handelt es sich um ein eigentliches Riemann-Integral, andernfalls um ein uneigentliches. Siehe dazu jeweils *Analysis 1*.

mindestens ein $t_* \in J$ mit $x(t_*) \neq 0$ und aufgrund der Stetigkeit von x schließen wir, dass es eine reelle Zahl $\eta_* > 0$ sowie ein Teilintervall $J_* \subset J$ mit $|J_*| > 0$ gibt, sodass

$$\big|x(t)\big| \geq \eta_* \qquad \text{für alle} \quad t \in J_* \,.$$

Im Fall von $p = \infty$ folgt sofort $\|x\|_\infty \geq \eta_* > 0$ und für $1 \leq p < \infty$ ergibt sich

$$\|x\|_p \geq \left(\int\limits_{J_*} \big|x(t)\big|^p \,\mathrm{d}t \right)^{1/p} \geq \left(\int\limits_{J_*} \eta_*^p \,\mathrm{d}t \right)^{1/p} = |J_*|^{1/p}\, \eta_* > 0$$

aus den Eigenschaften der Integration. Insbesondere haben wir damit gezeigt, dass $\|x\|_p = 0$ nur für $x = 0$ gilt. □

Ausblick*

1. In der Literatur finden Sie oftmals den Funktionenraum

$$\mathsf{C}(J) := \big\{ x : J \to \mathbb{R} \;:\; x \text{ stetig} \big\} \,,$$

dessen Elemente zwar noch stetig, aber nicht unbedingt beschränkt und damit nicht unbedingt integrierbar sind. Für ein kompaktes Intervall gilt $\mathsf{BC}(J) = \mathsf{C}(J)$, andernfalls $\mathsf{BC}(J) \subsetneqq \mathsf{C}(J)$.

2. In der Mathematik und in den Anwendungswissenschaften spielen verschiedene Arten von normierten Funktionenräume eine herausragende Rolle. Beispiele sind die *Lebesgue-* und *Sobolev-Räume*.

verschiedene Konvergenzbegriffe Nach unseren Erkenntnissen aus *Analysis 1* und *Analysis 2* kann eine Funktionenfolge $(x_n)_{n\in\mathbb{N}} \subset \mathsf{BC}(J)$ auf verschiedene Weisen gegen eine Grenzfunktion $x_\infty \in \mathsf{BC}(J)$ konvergieren:

(K_∞) Die Konvergenz in der ∞-Norm meint

$$\|x_n - x_\infty\|_\infty = \sup_{t\in J} \big|x_n(t) - x_\infty(t)\big| \quad \xrightarrow{n\to\infty} \quad 0$$

und stimmt mit der gleichmäßigen Konvergenz aus *Analysis 1* überein.

($\mathrm{K_p}$) Für jedes $p \in [1, \infty)$ ist

$$\|x_n - x_\infty\|_p = \left(\int\limits_{J} \big|x_n(t) - x_\infty(t)\big|^p \,\mathrm{d}t \right)^{1/p} \quad \xrightarrow{n\to\infty} \quad 0$$

gleichbedeutend mit der Konvergenz in der p-Norm.

($\mathrm{K_{p.w.}}$) Bei punktweiser Konvergenz gilt

$$\big|x_n(t) - x_\infty(t)\big| \quad \xrightarrow{n\to\infty} \quad 0$$

für jedes feste Argument $t \in J$ im Sinne der Konvergenz reeller Zahlenfolgen. Diese Konvergenz wird jedoch nicht durch eine Norm auf $\mathsf{BC}(J)$ vermittelt.

Theorem (**Zusammenhang zwischen den Konvergenzbegriffen**) Für alle Exponenten p und q mit $1 \leq p < q < \infty$ gelten die logischen Implikationen

$$(\mathrm{K}_\infty) \quad \Longrightarrow \quad (\mathrm{K}_q) \quad \Longrightarrow \quad (\mathrm{K}_p) \quad \Longrightarrow \quad (\mathrm{K}_{\text{p.w.}})\,,$$

aber die Umkehrungen sind im Allgemeinen falsch.

Beweis Die Kette der ersten drei Implikationen folgt unmittelbar aus dem Lemma über die fundamentalen Ungleichungen für Funktionen. In *Analysis 1* hatten wir schon gesehen, dass die Konvergenz bzgl. der Supremumsnorm ($p = \infty$) die punktweise Konvergenz impliziert, aber die analoge Aussage für Integralnormen ($p < \infty$) werden wir erst später beweisen können. □

Theorem (**Vollständigkeit bzgl. der Supremumsnorm**) $\mathsf{BC}(J)$ ist bzgl. der ∞-Norm vollständig.

Beweis Sei $(x_n)_{n\in\mathbb{N}} \subset \mathsf{BC}(J)$ eine Cauchy-Folge bzgl. $\|\cdot\|_\infty$ und sei $\varepsilon > 0$ beliebig fixiert. Dann existiert ein Index $N \in \mathbb{N}$, sodass

$$\|x_n - x_k\|_\infty < \varepsilon \qquad \text{für alle} \quad n,\, k > N\,.$$

Für jedes $t \in J$ gilt damit

$$\big|x_n(t) - x_k(t)\big| < \varepsilon \qquad \text{für alle} \quad n,\, k > N\,,$$

d.h. $(x_n(t))_{n\in\mathbb{N}}$ ist eine Cauchy-Folge reeller Zahlen und die Vollständigkeit von $\mathbb{R}$ garantiert, dass diese für $n \to \infty$ gegen einen Grenzwert konvergiert, den wir $x_\infty(t)$ nennen wollen. Wir haben damit gezeigt, dass x_n für $n \to \infty$ punktweise gegen eine Grenzfunktion $x_\infty : J \to \mathbb{R}$ konvergiert. Wenn wir in der zweiten Formel k gegen ∞ laufen lassen, ergibt sich außerdem

$$\big|x_n(t) - x_\infty(t)\big| < \varepsilon \qquad \text{für alle} \quad n > N$$

und damit auch

$$\|x_n - x_\infty\|_\infty \leq \varepsilon \qquad \text{für alle} \quad n > N$$

nach Supremumsbildung bzgl. $t \in J$. Wir schließen nun (da ε beliebig fixiert war), dass x_n für $n \to \infty$ nicht nur punktweise, sondern sogar gleichmäßig gegen x_∞ konvergiert und das damit x_∞ selbst eine stetige Funktion auf J ist (siehe *Analysis 1*). Mit der Dreiecksungleichung kann auch gezeigt werden, dass x_∞ beschränkt ist. Es gilt also $x_\infty \in \mathsf{BC}(J)$. □

Achtung $\mathsf{BC}(J)$ ist für $p < \infty$ **nicht** vollständig bzgl. der p-Norm. Siehe dazu die Übungen.

Vorlesungswoche 02

1.4 normierte Doppelfolgenräume

Vorbemerkung Wir diskutieren nun weitere Beispiele für unendlich-dimensionale normierte Räume, wobei deren Elemente bzw. *Punkte* nun Zahlenfolgen über einer unendlichen Indexmenge sind. Solche Doppelfolgen können sowohl als Funktionen auf der Indexmenge, aber auch als Vektoren mit unendlich vielen Komponenten angesehen werden.

Doppelfolgen Eine reelle Zahlenfolge über der Indexmenge $\mathbb{Z}$ ist eine Abbildung $x : \mathbb{Z} \to \mathbb{R}$, wobei wir oftmals

$$(x_j)_{j\in\mathbb{Z}} = \big(\dots,\ x_{-3}\,,\ x_{-2}\,,\ x_{-1}\,,\ x_0\,,\ x_1\,,\ x_2\,,\ x_3\ ,\ \dots\big)$$

anstelle von x schreiben. Die Menge aller Doppelfolgen wird mit $\mathbb{R}^{\mathbb{Z}}$ bezeichnet und ist via

$$(x+y)_j = x_j + y_j\,, \qquad (\lambda\, x)_j = \lambda x_j$$

in natürlicher Weise ein reeller Vektorraum. Für jede Doppelfolge $x \in \mathbb{R}^{\mathbb{Z}}$ sind die Ausdrücke

$$\|x\|_\infty := \sup_{j\in\mathbb{Z}} \big|x_j\big|$$

und

$$\|x\|_p := \left(\sum_{j=-\infty}^{+\infty} |x_j|^p\right)^{1/p} = \left(\lim_{k\to\infty} \sum_{j=-k}^{k} |x_j|^p\right)^{1/p}$$

mit Parameter $p \in [1,\,\infty)$ wohldefiniert, sofern wir den Wert $+\infty$ zulassen.[17]

Beispiele

1. Für die konstante Doppelfolge $x_j = 1$ gilt offensichtlich

$$\|x\|_\infty = 1\,, \qquad \|x\|_p^p = \lim_{k\to\infty} \sum_{j=-k}^{k} 1 = \lim_{k\to\infty} (2\,k+1) = \infty\,,$$

   wobei die Konvergenz im Sinne der uneigentlichen Konvergenz aus *Analysis 1* zu verstehen ist.

[17] Die letzte Formel stellt klar, dass auch jede doppelt-unendliche Summe via

$$\sum_{j=-\infty}^{\infty} \xi_j = \lim_{k\to\infty} \sum_{j=-k}^{+k} \xi_j$$

als Grenzwert einer Folge endlicher Summen zu verstehen ist.

2. Mit $x_j = 1/(1+j^2)$ ergibt sich

$$\|x\|_\infty = x_0 = 1\,, \qquad \|x\|_p = \sum_{j=-\infty}^{\infty} \frac{1}{\big(1+j^2\big)^p} = \frac{\Gamma\big(p-\frac{1}{2}\big)}{\Gamma(p)} < \infty\,,$$

wobei die Berechnung des konkreten Reihenwerts nicht einfach ist.[18]

3. Für jeden reellen Parameter $\mu > 0$ betrachten wir die Doppelfolge mit

$$x_j = \exp\big(-\mu\,|j|\big)$$

und erhalten

$$\|x\|_\infty = x_0 = 1\,.$$

Da immer $x_j = x_{-j}$ gilt, erhalten wir

$$\begin{aligned}\|x\|_p^p &= \left(\sum_{j=-\infty}^{-1} x_j^p\right) + x_0^p + \left(\sum_{j=+1}^{+\infty} x_j^p\right) = x_0^p + 2\sum_{j=+1}^{+\infty} x_j^p \\ &= 1 + 2\sum_{j=+1}^{+\infty} \exp\left(-\mu\,j\,p\right) = 1 + 2\sum_{j=+1}^{+\infty} \big(\exp\left(-\mu\,p\right)\big)^j \\ &= 1 + 2\,\frac{\exp\left(-\mu\,p\right)}{1-\exp\left(-\mu\,p\right)} = \frac{1+\exp\left(-\mu\,p\right)}{1-\exp\left(-\mu\,p\right)}\end{aligned}$$

und damit

$$\|x\|_p = \left(\frac{1+\exp\left(-\mu\,p\right)}{1-\exp\left(-\mu\,p\right)}\right)^{1/p},$$

wobei wir auch die Summenformel für geometrische Reihen ausgewertet haben.

Lemma (**fundamentale Ungleichungen**) Für zwei beliebige Doppelfolgen x, y und jeden Exponenten $p \in [0,\,\infty]$ gilt die Minkowski-Ungleichung

$$\|x+y\|_p \le \|x\|_p + \|y\|_p$$

sowie die Hölder-Ungleichung

$$\sum_{j=-\infty}^{+\infty} |x_j|\;|y_j| \le \|x\|_p\;\|y\|_{p'}\,,$$

wobei p' wieder der konjugierte Exponent zu p ist und jede Norm auch den Wert $+\infty$ annehmen darf. Außerdem gilt die Einbettungsungleichung

$$\|x\|_q \le \|x\|_p$$

für jede Doppelfolge x und für beliebige Exponenten $1 \le p < q \le \infty$.

[18] Die Formeln können zum Beispiel mit MATHEMATICA bestimmt werden. Die Gamma-Funktion Γ hatten wir in *Analysis 1* als Verallgemeinerung der Fakultät eingeführt.

Beweis *Teile 1 und 2*: Die Hölder- und die Minkowski-Ungleichung können analog zu den entsprechenden Ungleichungen im $\mathbb{R}^n$ hergeleitet werden.

Teil 3a: Wir beweisen die Behauptung zunächst für $q = \infty$, wobei wir o.B.d.A. $p < \infty$ sowie $\|x\|_p < \infty$ annehmen dürfen (andernfalls ist die Behauptung trivialerweise richtig). Für jede Doppelfolge x und jedes $m \in \mathbb{Z}$ gilt

$$|x_m| = \left(|x_m|^p\right)^{1/p} \leq \left(\sum_{j=-\infty}^{+\infty} |x_j|^p\right)^{1/p} = \|x\|_p$$

und nach Supremumsbildung über $m \in \mathbb{Z}$ erhalten wir $\|x\|_\infty \leq \|x\|_p$.

Teil 3b: Für $q < \infty$ können wir o.B.d.A. $1 \leq p < q$ sowie $\|x\|_p < \infty$ voraussetzen. Dann gilt

$$|x_j|^q = |x_j|^{q-p}\, |x_j|^p \leq \|x\|_\infty^{q-p}\, |x_j|^p \leq \|x\|_p^{q-p}\, |x_j|^p$$

für alle $j \in \mathbb{Z}$, wobei wir $|x_j| \leq \|x\|_\infty$ sowie $\|x\|_\infty \leq \|x\|_p$ aus dem letzten Beweisschritt benutzt haben. Die Eigenschaften der Summen- und Reihenbildung implizieren die Abschätzung

$$\sum_{j=-\infty}^{+\infty} |x_j|^q \leq \|x\|_p^{q-p} \sum_{j=-\infty}^{+\infty} |x_j|^p = \|x\|_p^{q-p}\, \|x\|_p^p = \|x\|_p^q$$

und damit die Behauptung. □

Definition-Lemma (**wichtige Doppelfolgenräume**) Für jedes $p \in [1, \infty]$ ist die Menge

$$\ell^p(\mathbb{Z}) := \left\{x \in \mathbb{R}^{\mathbb{Z}} \,:\, \|x\|_p < \infty\right\},$$

ein normierter Raum bzgl. der p-Norm. Dabei gilt

$$\ell^p(\mathbb{Z}) \subset \ell^q(\mathbb{Z})$$

für alle $1 \leq p < q \leq \infty$.

Beweis Mit der Minkowski-Ungleichung können wir leicht nachrechnen, dass $\ell^p(\mathbb{Z})$ reeller Vektorraum — genauer gesagt, ein linearer Unterraum des Vektorraumes $\mathbb{R}^{\mathbb{Z}}$ — ist und dass $\|\cdot\|_p$ alle Norm-Eigenschaften erfüllt. Aus der Einbettungsungleichung ergibt sich außerdem die letzte Behauptung. □

Bemerkungen

1. Für $1 \leq p < q \leq \infty$ gilt zwar $\|x\|_q \leq \|x\|_p$ (siehe das Lemma oben), aber es gibt *keine* Konstante C, sodass $\|x\|_p \leq C\, \|x\|_q$ für alle Doppelfolgen x gelten würde. Insbesondere ist der Raum $\ell^p(\mathbb{Z})$ immer eine *echte Teilmenge* von $\ell^q(\mathbb{Z})$.

2. Mit etwas mehr Aufwand können wir

   $$\|x\|_\infty = \lim_{p\to\infty} \|x\|_p$$

   für jede Doppelfolge x zeigen, wobei die Konvergenz auf der rechten Seite sich wieder auf reelle Zahlen bezieht.

3. Für $p = 2$ (aber nicht für $p \neq 2$) gibt es ein Skalarprodukt auf $\ell^2(\mathbb{Z})$, nämlich

$$\langle x,\, y \rangle = \sum_{j=-\infty}^{\infty} x_j\, y_j\,,$$

wobei die Hölder-Ungleichung sicherstellt, dass die unendliche Summe auf der rechten Seite für $x,\, y \in \ell^2(\mathbb{Z})$ immer im Sinne einer absolut konvergenten Reihe definiert ist.

Folgen von Doppelfolgen Bei Folgen aus dem Doppelfolgenraum $\mathbb{R}^{\mathbb{Z}}$ müssen wir wieder zwei Indizes verwenden. Zum Beispiel wird durch

$$x_{n,\,j} := \exp\left(-\frac{n+2}{1+n\,j^2}\right)$$

eine Folge $(x_n)_{n\in\mathbb{N}}$ von Doppelfolgen definiert, wobei

$$x_n = (x_{n,\,j})_{j\in\mathbb{Z}}$$

gerade die n-te Doppelfolge ist, deren Glieder durch j indiziert sind.

Definition Eine Folge $(x_n)_{n\in\mathbb{N}}$ von Doppelfolgen konvergiert für $n \to \infty$ punktweise gegen die Doppelfolge x_∞, falls

$$x_{\infty,\,j} = \lim_{n\to\infty} x_{n,\,j}$$

für jeden Index $j \in \mathbb{Z}$ gilt.

Klarstellung Für Folgen in $\ell^p(\mathbb{Z})$ haben wir — analog zu dem oben diskutierten Raum der stetigen Funktionen — unterschiedliche Konvergenzbegriffe: Die soeben eingeführte punktweise Konvergenz sowie die weiter oben abstrakt definierte Konvergenz bzgl. einer p-Norm, wobei letztere gerade meint, dass $\|x_n - x_\infty\|_p$ für $n \to \infty$ gegen 0 konvergiert. Diese Konvergenzbegriffe sind verschieden und wir müssen bei jeder Grenzwertaussage immer deutlich machen, in welchem Sinn diese zu verstehen ist. Wir schreiben zum Beispiel oftmals

$$x_n \xrightarrow{\;n\to\infty\;} x_\infty \quad \text{punktweise} \qquad \text{bzw.} \qquad x_n \xrightarrow{\;n\to\infty\;} x_\infty \quad \text{in der } p\text{-Norm}\,,$$

aber es gibt auch andere Möglichkeiten, die Art der Konvergenz festzuhalten bzw. anzugeben.

Lemma (**wichtige Tatsache**) Für jedes $p \in [1,\, \infty]$ impliziert die Normkonvergenz in $\ell^p(\mathbb{Z})$ die punktweise Konvergenz.

Beweis Sei $(x_n)_{n\in\mathbb{N}}$ eine Folge von Doppelfolgen aus $\ell^p(\mathbb{Z})$, die bzgl. der p-Norm gegen $x_\infty \in \ell^p(\mathbb{Z})$ konvergiert und sei $k \in \mathbb{Z}$ ein beliebiger, aber fester Index. Im Fall von $1 \leq p < \infty$ erhalten wir via

$$\big|x_{n,\,k} - x_{\infty,\,k}\big|^p \leq \sum_{j=-\infty}^{+\infty} \big|x_{n,\,j} - x_{\infty,\,j}\big|^p = \|x_n - x_\infty\|_p^p$$

und damit

$$0 \leq \left|x_{n,k} - x_{\infty,k}\right| \leq \left\|x_n - x_\infty\right\|_p$$

nach Ziehen der p-ten Wurzel, wohingegen für $p = \infty$ die Abschätzung

$$0 \leq \left|x_{n,k} - x_{\infty,k}\right| \leq \sup_{j\in\mathbb{Z}} \left|x_{n,j} - x_{\infty,j}\right|_\infty = \left\|x_n - x_\infty\right\|_\infty$$

gilt. Nach Voraussetzung konvergiert in beiden Fällen $\left\|x_n - x_\infty\right\|_p$ für $n \to \infty$ gegen 0 und mit dem Vergleichsprinzip für reelle Zahlenfolgen schließen wir, dass

$$\left|x_{n,k} - x_{\infty,k}\right| \xrightarrow{n\to\infty} 0 \qquad \text{bzw.} \qquad x_{n,k} \xrightarrow{n\to\infty} x_{\infty,k}$$

gilt. Da k beliebig war, folgt die Behauptung.

Achtung Die umgekehrte Aussage ist im Allgemeinen falsch, d.h. aus punktweiser Konvergenz können wir nicht auf die Normkonvergenz schließen.

Beispiele

1. Die Funktionenfolge $(x_n)_{n\in\mathbb{N}}$ mit

$$x_{n,j} = \frac{1 + \cos(j/10)}{2 + \ln(n) + j^2}$$

konvergiert für $n \to \infty$ in jeder p-Norm gegen die triviale Doppelfolge x_∞ mit $x_{\infty,j} = 0$, denn wegen

$$\left|x_{n,j} - x_{\infty,j}\right| = \left|x_{n,j}\right| \leq \frac{2}{2 + \ln(n) + j^2}$$

ergibt sich

$$\left\|x_n - x_\infty\right\|_\infty \leq \frac{2}{2 + \ln(n)} \xrightarrow{n\to\infty} 0\,.$$

Für $p = 1$ gilt

$$\begin{aligned}\left\|x_n - x_\infty\right\|_1 &= \left|x_{n,0}\right| + 2\sum_{j=1}^{\infty} \left|x_{n,j}\right| \\ &\leq \frac{2}{\ln(n)} + 2\sum_{j=1}^{\infty} \frac{2}{\ln(n) + j^2} \leq \frac{2}{\ln(n)} + 4\int_0^\infty \frac{\mathrm{d}t}{\ln(n) + t^2} \\ &\leq \frac{2}{\ln(n)} + \frac{4\arctan\left(\sqrt{\ln(n)}\right)}{\sqrt{\ln(n)}} \xrightarrow{n\to\infty} 0\,,\end{aligned}$$

wobei wir das Integralkriterium für unendliche Summen (siehe *Analysis 1*) sowie die Eigenschaften des Arkustangens verwendet haben. Man kann auch

$$\left\|x_n - x_\infty\right\|_p \xrightarrow{n\to\infty} 0$$

für jeden Exponenten p mit $1 < p < \infty$ zeigen, aber die expliziten Abschätzungen sind aufwändiger.[19]

2. Die Formel

$$x_{n,\,j} = \exp\left(-\frac{|j|}{n}\right)$$

definiert eine Folge von Doppelfolgen, wobei x_n wegen

$$\|x_n\|_p = \left(\frac{1+\exp(-p/n)}{1-\exp(-p/n)}\right)^{1/p}, \qquad \|x_n\|_\infty = 1$$

für jedes $p \in [1,\,\infty]$ zu $\ell^p(\mathbb{Z})$ gehört (die Formeln können mit $\mu = 1/n$ aus einem der obigen Beispiele abgelesen werden). Die Folge $(x_n)_{n\in\mathbb{N}}$ konvergiert für $n \to \infty$ offensichtlich punktweise gegen die konstante Doppelfolge x_∞ mit

$$x_{\infty,\,j} = 1\,,$$

aber diese Konvergenz gilt nicht bzgl. der p-Norm. Für $p < \infty$ folgt dies zum Beispiel aus der Tatsache, dass x_∞ gar nicht zu $\ell^p(\mathbb{Z})$ gehört. Für ∞ können wir leicht zeigen, dass $\|x_n - x_\infty\|_\infty = 1$ gilt und daher für $n \to \infty$ nicht gegen 0 konvergiert.

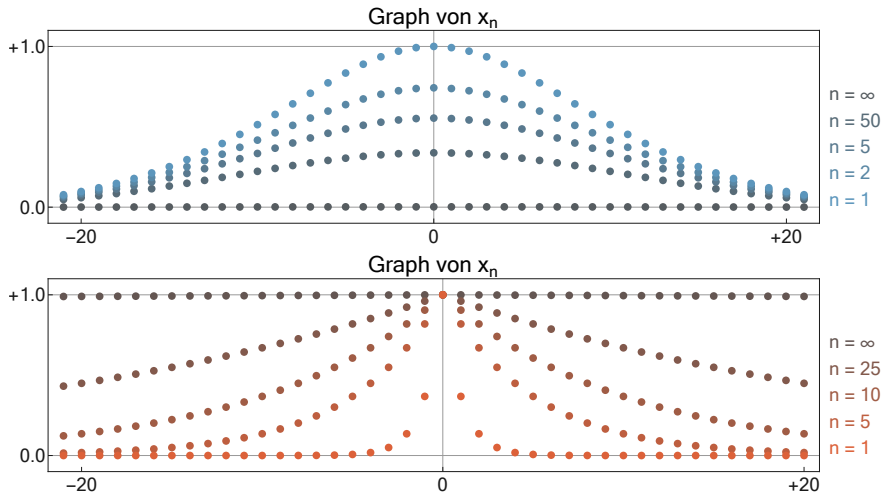

Abbildung Die Funktionenfolgen aus dem ersten (oben) und dem zweiten (unten) Beispiel.

Theorem (**Vollständigkeit**) Für jeden Exponenten $p \in [1,\,\infty]$ ist der normierte Raum $\left(\ell^p(\mathbb{Z}),\,\|\cdot\|_p\right)$ vollständig.

Beweis *Teil 1*: Wir betrachten eine Cauchy-Folge $(x_n)_{n\in\mathbb{N}}$ in $\ell^p(\mathbb{Z})$ und wollen zeigen, dass diese bzgl. der p-Norm konvergiert. Für jedes $j \in \mathbb{Z}$ gilt $|x_{n,\,j} - x_{k,\,j}| \le \|x_n - x_k\|_p$ und wir schließen, dass die Zahlenfolge $(x_{n,\,j})_{n\in\mathbb{N}}$ eine reelle Cauchy-Folge ist und deshalb nach *Analysis 1* gegen ein Grenzwert konvergiert, den wir $x_{\infty,\,j}$ nennen wollen.

[19] Alternativ können wir mit der Interpolationsungleichung

$$\|x\|_p \le \|x\|_1^{1/p}\,\|x\|_\infty^{1/p'}$$

argumentieren, die recht einfach bewiesen werden kann.

Wir haben damit gezeigt, dass x_n für $n \to \infty$ punktweise gegen eine Doppelfolge x_∞ konvergiert.

Teil 2: Wir wollen nun indirekt $\lim_{n\to\infty} \|x_n - x_\infty\|_p = 0$ zeigen und nehmen daher an, dies sei nicht der Fall. Dann existiert ein $\varepsilon > 0$ sowie eine Indexfolge $(n_l)_{l\in\mathbb{N}}$ mit $\lim_{l\to\infty} n_l = \infty$, sodass

$$\|x_{n_l} - x_\infty\|_p > \varepsilon$$

für alle $l \in \mathbb{N}$ gilt, und die Cauchy-Eigenschaft liefert einen Index $N \in \mathbb{N}$ mit

$$\|x_n - x_k\|_p < \varepsilon$$

für alle n, $k > N$. Im Fall von $p = \infty$ ergibt sich

$$|x_{n,\,j} - x_{k,\,j}| \le \|x_n - x_k\|_p < \varepsilon$$

für alle $j \in \mathbb{Z}$. Wenn wir auf der linken Seite den Grenzübergang $k \to \infty$ durchführen und anschließend das Supremum über j bilden, so erhalten wir

$$\|x_n - x_\infty\|_\infty \le \varepsilon$$

für alle $n > N$ und damit auch für $n = n_l$, sofern l hinreichend groß ist. Das ist aber ein Widerspruch und unsere Annahme muss falsch gewesen sein. Im Fall von $1 \le p < \infty$ bemerken wir, dass die Abschätzung

$$\sum_{j=-M}^{M} |x_{n,\,j} - x_{k,\,j}|^p \le \|x_n - x_k\|_p^p < \varepsilon^p$$

für alle n, $k > N$ und alle $M \in \mathbb{N}$ gilt. Wenn wir auf der linken Seite zunächst k and anschließend M nach ∞ laufen lassen, erhalten wir

$$\|x_n - x_\infty\|_p^p \le \varepsilon^p$$

für alle $n \in \mathbb{N}$ und nach dem Ziehen der p-ten Wurzel wieder einen Widerspruch.

Teil 3: Die Dreiecksungleichung impliziert $\|x_\infty\|_p \le \|x_\infty - x_n\|_p + \|x_n\|_p < \infty$ für jedes n und damit auch $x_\infty \in \ell^p(\mathbb{Z})$. Insgesamt haben wir damit gezeigt, dass die Cauchy-Folge $(x_n)_{n\in\infty}$ nicht nur punktweise, sondern auch bzgl. der p-Norm gegen x_∞ konvergiert. □

1.5 topologische Grundbegriffe

Vorbemerkung Im Folgenden ist (X, d) ein beliebiger metrischer Raum, aber bei der ersten Lektüre können Sie sich X als den $\mathbb{R}^2$ und d als den euklidischen Abstand (bzw. die 2-Norm) vorstellen.

Definition Eine Teilmenge $U \subseteq X$ heißt

1. offen in X, falls für jedes $x \in U$ ein Radius $\varepsilon > 0$ existiert, sodass die Kugel $B_\varepsilon(x)$ ganz in U liegt,
2. abgeschlossen in X, falls die Komplementmenge $X \setminus U$ offen ist.

Bemerkungen

1. Ist U offen und $x \in U$ ein Punkt in U, so nennen wir U auch eine offene Umgebung von x.
2. *Mengen sind keine Türen*: Viele Mengen sind weder abgeschlossen noch offen.
3. Die Konzepte *offene* und *abgeschlossene* Mengen sind von zentraler Bedeutung in der modernen Mathematik. Sie bilden die Grundlage der *Topologie*.
4. Die Hausdorffsche Trennungseigenschaft besagt, dass es zu je zwei verschiedenen Punkten x, $y \in X$ immer zwei offene Mengen U, $V \subset X$ gibt, sodass

$$x \in U\,, \qquad y \in V\,, \qquad U \cap V = \emptyset$$

gilt. In der Tat, wir können zum Beispiel immer $U = B_\varrho(x)$ und $V = B_\varrho(y)$ mit $\varrho = \frac{1}{3}\, d(x, y)$ wählen.

Beispiele

1. Jede offene Kugel $B_\varrho(x_*)$ ist offen.[20]

   Beweis: Sei $x \in B_\varrho(x_*)$ beliebig und sei ε ein Radius mit $0 < \varepsilon < \varrho - d(x, x_*)$. Für jedes $y \in B_\varepsilon(x)$ gilt dann $d(y, x_*) \le d(y, x) + d(x, x_*) < \varepsilon + d(x, x_*) < \varrho$ und damit $y \in B_\varrho(x_*)$. Damit haben wir $B_\varepsilon(x) \subset B_\varrho(x_*)$ gezeigt. □

   Bemerkung: Die Beweisidee ist relativ einfach und kann im $\mathbb{R}^2$ auch gut visuell verstanden werden (siehe das Bild).

2. Jede abgeschlossene Kugel $\overline{B}_\varrho(x_*)$ ist abgeschlossen.

   Beweis: Für beliebiges $x \in X \setminus \overline{B}_\varrho(x_*)$ wählen wir ε mit $0 < \varepsilon < d(x, x_*) - \varrho$. Jeder Punkt $y \in B_\varepsilon(x)$ erfüllt $d(y, x_*) > d(x, x_*) - d(x, y) > (\varepsilon + \varrho) - \varepsilon > \varrho$ und gehört damit auch zu $X \setminus \overline{B}_\varrho(x_*)$. Insbesondere ist das Komplement von $\overline{B}_\varrho(x_*)$ offen in X. □

3. Jede Sphäre $S_\varrho(x_*)$ ist abgeschlossen, aber nicht offen. Das Gleiche gilt für die Einpunktmenge $\{x_*\}$, die wir auch als (entartete) Sphäre vom Radius 0 ansehen können.

[20] Diese Aussage klingt tautologisch, aber wir müssen erst zeigen, dass die Formel für eine offene Kugel wirklich eine offene Menge im Sinne der Definition beschreibt bzw. dass unsere Bezeichnungen „offene Kugel“ sowie „offene Menge“ konsistent sind.

4. Die Mengen $\emptyset$ und X sind beide sowohl offen und als auch abgeschlossen.

   Bemerkung: Im $\mathbb{R}^m$ sind $\emptyset$ und X die einzigen Teilmengen, die sowohl offen als auch abgeschlossen sind.

5. Wir werden weiter unten sehen, dass offene bzw. abgeschlossene Mengen oftmals in natürlicher Weise durch strikte bzw. nicht-strikte Ungleichungen beschrieben werden können.

6. Für jede Teilmenge des U des $\mathbb{R}^m$ und zwei beliebige Exponenten p, $q \in [1, \infty]$ gilt: U ist genau dann offen bzw. abgeschlossen bzgl. der p-Norm, wenn U offen bzw. abgeschlossen bzgl. der q-Norm ist. [21]

Spezialfall Intervalle Sind a und b zwei reelle Zahlen mit $a < b$ so gilt:

$(-\infty, a)$, (a, b), $(b, +\infty)$	offen, aber nicht abgeschlossen in $\mathbb{R}$
$(-\infty, a]$, $[a, b]$, $[b, +\infty)$	nicht offen, aber abgeschlossen in $\mathbb{R}$
$(a, b]$, $[a, b)$	weder offen noch abgeschlossen in $\mathbb{R}$

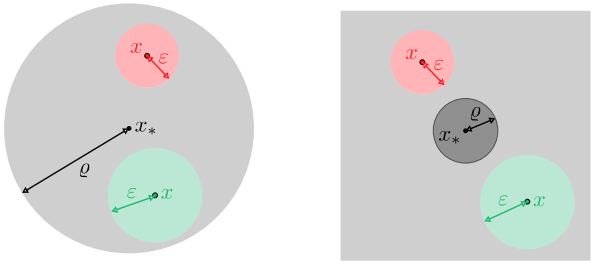

Abbildung *Links*: Die offene Kugel $B_\varrho(x_*)$ (grau, hier dargestellt im $\mathbb{R}^2$ und bzgl. der 2-Norm, d.h. des euklidischen Betrages) ist offen, denn mit jedem ihrer Punkte x (rot oder grün) enthält sie auch eine kleine Kugel um diesen Punkt. *Rechts*: Das Komplement (hellgrau) der abgeschlossenene Kugel $\overline{B}_\varrho(x_*)$ (dunkelgrau) ist auch offen.

Lemma (**Offenheit/Abgeschlossenheit und Mengenoperationen**) Es gelten die folgenden Aussagen:

1a. Die Vereinigung *beliebig* vieler offener Mengen ist offen.

1b. Der Durchschnitt *endlich* vieler offener Mengen ist offen.

2a. Die Vereinigung *endlich* vieler abgeschlossener Mengen ist abgeschlossen.

2b. Der Durchschnitt *beliebig* vieler abgeschlossener Mengen ist abgeschlossen.

Über den Durchschnitt unendlich vieler offener oder die Vereinigung unendlich vieler abgeschlossener Mengen kann im Allgemeinen keine Aussage getroffen werden.

Beweis *Teil 1a*: Seien L irgendeine Indexmenge[22] und U_l für jedes $l \in L$ eine offene Menge in X. Außerdem sei ein Punkt $x_* \in \bigcup_{l\in L} U_l$ beliebig fixiert. Dann existiert mindestens ein $l_* \in L$ mit $x \in U_{l_*}$ und — da U_{l_*} offen ist — ein Radius ε_*, sodass die Kugel $B_{\varepsilon_*}(x_*)$ zu U_{l_*} und damit auch zu $\bigcup_{l\in L} U_l$ gehört.

[21] Diese Aussage ergibt sich unmittelbar aus der Äquivalenz der Normen. Die analoge Aussage im Funktionenraum $\mathsf{BC}(J)$ ist jedoch falsch.

[22] L kann endlich viele, abzählbar unendlich viele oder sogar überabzählbar unendlich viele Elemente enthalten.

Teil 1b: Sei $L = \{1, \ldots, k\}$ eine endliche Indexmenge und seien $U_1, \ldots, U_k$ offene Mengen in X. Ist $x_* \in \bigcap_{l\in L} = U_1 \cap \ldots U_k$ beliebig fixiert, so existiert für jedes l ein Radius ε_l mit $B_{\varepsilon_l}(x_*) \subset U_l$. Für $\varepsilon_* := \min\{\varepsilon_1, \ldots, \varepsilon_k\} > 0$ liegt die Kugel $B_{\varepsilon_*}(x_*)$ in allen U_l und damit auch in $\bigcap_{l\in L} U_l$.[23]

Teil 2: Beide Behauptungen ergeben sich mit den mengentheoretischen Formeln

$$X \setminus \left(\bigcup_{l\in L} U_l\right) = \bigcap_{l\in L} \left(X \setminus U_l\right), \qquad X \setminus \left(\bigcap_{l\in L} U_l\right) = \bigcup_{l\in L} \left(X \setminus U_l\right)$$

direkt aus dem ersten Teil. □

Gegenbeispiel Die Formeln

$$\bigcap_{l\in\mathbb{N}} B_{1+1/l}(x) = \overline{B}_1(x), \qquad \bigcup_{l\in\mathbb{N}} \overline{B}_{1-1/l}(x) = B_1(x)$$

können leicht verifiziert werden und zeigen, dass der Durchschnitt unendlich vieler offener Mengen abgeschlossen sein kann und dass die Vereinigung unendlich vieler abgeschlossener Mengen offen sein kann.

Ausblick* Neben metrischen Räumen gibt es das deutlich allgemeinere Konzept eines *topologischen Raumes*, wobei ein solcher dadurch definiert wird, dass man eine Familie von Teilmengen angibt, die *offen* sein sollen. Dabei müssen natürlich einige Spielregeln erfüllt sein (im Wesentlichen sollen die im Lemma gelisteten Eigenschaften gelten), aber es gibt in einem allgemeinen topologischen Raum weder Abstände noch Kugeln. Die Topologie hat zu tiefen Einsichten geführt und spielt nicht nur in der Analysis, sondern auch in der Algebra und der Geometrie eine herausragende Rolle. Im Rahmen dieser Vorlesung werden wir aber keine allgemeinen topologischen Räume studieren, sondern nur metrische oder normierte.

Lemma (**äquivalente Charakterisierung von Konvergenz**) Seien $(x_n)_{n\in\mathbb{N}}$ bzw. x_* eine Folge bzw. ein Punkt in X. Dann gilt $x_* = \lim_{n\to\infty} x_n$ genau dann, wenn für jede offene Umgebung U von x_* höchstens endlich viele Folgenglieder x_n nicht zu U gehören, d.h. im Komplement $X \setminus U$ liegen.

Beweis *Hinrichtung*: Sei $x_* =: x_\infty$ Grenzwert der Folge und sei U eine beliebige offene Menge in X mit $x_\infty \in U$. Dann existiert ein $\varepsilon > 0$ mit $B_\varepsilon(x_\infty) \subseteq U$ sowie ein $N \in \mathbb{N}$ mit $d(x_n, x_\infty) < \varepsilon$ für alle $n > N$. Für jedes dieser n gilt dann $x \in B_\varepsilon(x_\infty)$ und damit $x_n \in U$, d.h. es können nur die Folgenglieder $x_1, \ldots, x_N$ im Komplement von U liegen.

Rückrichtung: Für jedes $\varepsilon > 0$ können höchstens endliche viele Folgenglieder x_n außerhalb der offenen Umgebung $U := B_\varepsilon(x_*)$ liegen. Insbesondere existiert $n \in N$, sodass $x_n \in U$ und damit $d(x_n, x_*) < \varepsilon$ für alle $n > N$ gilt. Also konvergiert x_n für $n \to \infty$ gegen x_*. □

Lemma (**äquivalente Charakterisierung für abgeschlossene Mengen**) Eine Menge $U \subseteq X$ ist genau dann abgeschlossen, wenn die folgende Eigenschaft erfüllt ist: Ist $(x_n)_{n\in\mathbb{N}}$ eine Folge, deren Glieder alle zu U gehören und die bzgl. der Metrik d gegen einen Grenzwert $x_\infty \in X$ konvergiert, so gehört dieser auch zu U.

[23] Beachte, dass das Minimum endlich vieler positiver Zahlen selbst positiv ist.

Beweis *Hinrichtung*: Ist U abgeschlossen und gilt $x_\infty = \lim_{n\to\infty} x_n$ für eine Folge $(x_n)_{n\in\mathbb{N}} \subset U$, so zeigen wir $x_\infty \in U$ durch folgenden indirekten Beweis: Angenommen, dies wäre nicht der Fall. Dann gehört x_∞ zur offenen Menge $X \setminus U$ und es existiert ein $\varepsilon > 0$, so dass $B_\varepsilon(x_\infty)$ ganz in $X \setminus U$ liegt. Andererseits muss es wegen der Konvergenz einen Index $N \in \mathbb{N}$ geben, sodass $d(x_n, x_\infty) < \varepsilon$ für alle $n > N$ gilt. Das bedeutet aber, dass jedes dieser x_n in $B_\varepsilon(x_\infty)$ und damit in $X \setminus U$ liegt und nicht zu U gehören kann. Das ist der gesuchte Widerspruch.

Rückrichtung: Wir nehmen an, U sei nicht abgeschlossen. Dann gibt es mindestens einen Punkt $x_* \in X \setminus U$ mit der Eigenschaft, dass keine Kugel um x_* ganz in $X \setminus U$ liegt, sondern immer mindestens einen Punkt aus U enthält. Insbesondere können wir für jedes $n \in \mathbb{N}$ einen Punkt $x_n \in X$ wählen, der sowohl in $B_{1/n}(x_*)$ als auch in U liegt. Dadurch erhalten wir eine Folge $(x_n)_{n\in\mathbb{N}}$ aus U, die wegen $d(x_n, x_*) < 1/n$ für $n \to \infty$ gegen den Grenzwert $x_\infty = x_*$ konvergiert. Die vorausgesetze Folgeneigenschaft impliziert $x_* \in U$ und damit einen Widerspruch zur Wahl von x_*. Also muss unsere Annahme falsch gewesen sein. □

Beispiel Wir betrachten $X = \mathbb{R}^1$ mit dem euklidischen Abstand sowie die durch

$$x_n := 1 - 1/n$$

definierte Folge, die offensichtlich gegen $x_\infty = 1$ konvergiert. Alle Folgenglieder sind Elemente der offenen Menge $B_1(0)$, aber der Grenzwert nicht. Andererseits liegen alle x_n auch in der abgeschlossenen Menge $\overline{B}_1(0)$ und daher muss dies auch für x_∞ gelten.

Rand einer Menge

Definition Sei $U \subseteq X$ eine beliebige Teilmenge von X. Ein Punkt $x \in X$ heißt innerer bzw. äußerer Punkt von U, falls es einen Radius $\varepsilon > 0$ gibt, sodass $B_\varepsilon(x)$ ganz in U bzw. ganz in $X \setminus U$ liegt. Andernfalls wird x Randpunkt von U genannt.

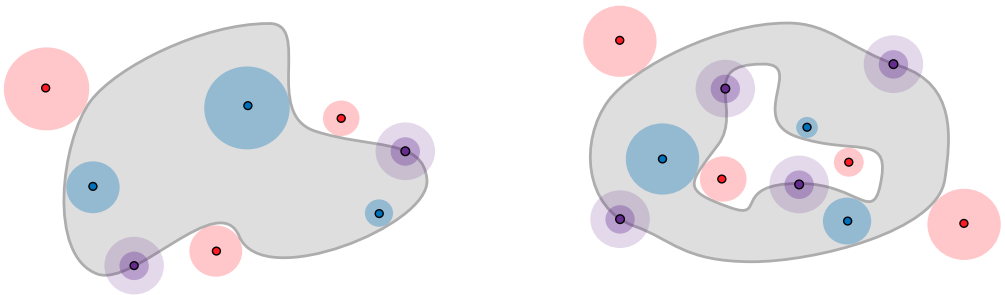

Abbildung Innere Punkte (blau) bzw. äußere Punkte (rot) einer gegebenen Menge U (grau; links ohne, rechts mit Loch) haben die Eigenschaft, dass jede hinreichend kleine Kugel um diesen Punkt auch zu U bzw. zur Komplementmenge $X \setminus U$ (weiß) gehört (im Bild ist immer der jeweils maximale Radius dargestellt). Randpunkte (lila) haben die Eigenschaft, dass jede noch so kleine Kugel um diesen Punkt sowohl Punkte aus U als auch Punkte aus $X \setminus U$ enthält.

Bemerkungen

1. Ist x ein innerer oder äußerer Punkt von U, so liegen alle Kugeln mit Mittelpunkt x und hinreichend kleinem Radius ganz in U bzw. in $X \setminus U$. Dies folgt, da mit $0 < \tilde{\varepsilon} < \varepsilon$ auch $B_{\tilde{\varepsilon}}(x) \subset B_\varepsilon(x)$ gilt. Analoges gilt für jeden äußeren Punkt.

2. Ist x ein Randpunkt von U, so enthält jede noch so kleine Kugel um x immer Punkte aus U und Punkte aus $X \setminus U$.

3. Innere bzw. äußere Punkte von U gehören zu U bzw. zu $X \setminus U$. Randpunkte von U können im Allgemeinen zu U oder zu $X \setminus U$ gehören.

4. Ein innerer bzw. äußerer Punkt von U ist äußerer bzw. innerer Punkt von $X \setminus U$. Jeder Randpunkt von U ist auch Randpunkt von $X \setminus U$.

Bezeichnungen Ausgehend von einer Teilmenge $U \subseteq X$ führen wir die folgenden anderen Teilmengen ein:[24]

$$\begin{aligned} \partial U &:= \operatorname{bnd}(U) := \{x \in X : x \text{ ist Randpunkt von } U\} && \underline{\text{Rand von } U} \\ U^\circ &:= \operatorname{int}(U) := \{x \in X : x \text{ ist innerer Punkt von } U\} && \underline{\text{Inneres von } U} \\ \overline{U} &:= \operatorname{cls}(U) := X \setminus \{x \in X : x \text{ ist äußerer Punkt von } U\} && \underline{\text{Abschluss von } U} \end{aligned}$$

Lemma (**wichtige Eigenschaften**) Für jede Teilmenge $U \subseteq X$ gilt

$$\operatorname{int}(U) = U \setminus \operatorname{bnd}(U), \qquad \operatorname{cls}(U) = U \cup \operatorname{bnd}(U).$$

Insbesondere ist U genau dann offen bzw. abgeschlossen, wenn der Rand $\operatorname{bnd}(U)$ zu $X \setminus U$ bzw. zu U gehört.

Beweis Übungsaufgabe. Siehe auch das Bild. □

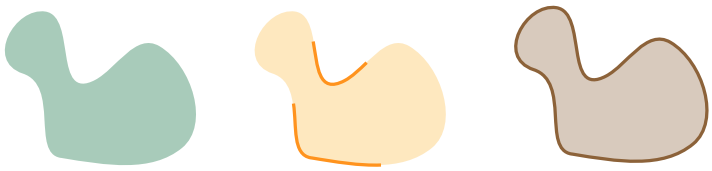

Abbildung Schematische Darstellung zum Rand von Teilmengen des $\mathbb{R}^2$: Bei einer offenen (grün) bzw. einer abgeschlossenen (braun) Menge U gehören alle Randpunkte zu $X \setminus U$ bzw. zu U. Bei allen anderen Mengen (gelb) gehören einige Randpunkte zu U, andere aber zu $X \setminus U$.

Bemerkungen

1. Weitere wichtige Eigenschaften von $\operatorname{int}(U)$, $\operatorname{bnd}(U)$ und $\operatorname{cls}(U)$ werden wir in den Übungen diskutieren.

2. Es gilt

$$\operatorname{bnd}\big(B_\varrho(x)\big) = S_\varrho(x), \qquad \operatorname{int}\big(B_\varrho(x)\big) = B_\varrho(x), \qquad \operatorname{cls}\big(B_\varrho(x)\big) = \overline{B}_\varrho(x)$$

   sowie

$$\operatorname{bnd}\big(\overline{B}_\varrho(x)\big) = S_\varrho(x), \qquad \operatorname{int}\big(\overline{B}_\varrho(x)\big) = B_\varrho(x), \qquad \operatorname{cls}\big(\overline{B}_\varrho(x)\big) = \overline{B}_\varrho(x).$$

   Für jede offene oder abgeschlossene Kugel gilt also: Der Rand / das Innere / der Abschluss ist immer die entsprechende Sphäre / offene Kugel / abgeschlossene Kugel.

3. Für jede Teilmenge U des $\mathbb{R}^m$ gilt: Die Mengen $\operatorname{int}(U)$, $\operatorname{bnd}(U)$ und $\operatorname{cls}(U)$ hängen nicht von der Wahl der konkreten Norm ab.[25]

[24] Im Englischen spricht man von '*boundary*', '*interior*' und '*closure*'.
[25] Dies ergibt sich wieder aus der Äquivalenz aller Normen.

4. Viele praktisch relevante Teilmengen des $\mathbb{R}^2$ bzw. $\mathbb{R}^3$ haben die Eigenschaft, dass ihr Rand eine (im Allgemeinen gekrümmte) Kurve bzw. Fläche ist und direkt mithilfe der geometrischen Anschauung bestimmt werden kann.

   Achtung: Es gibt auch Teilmengen des $\mathbb{R}^2$, deren Rand nicht *eindimensional*, sondern zweidimensional oder gar gebrochen-dimensional ist. Siehe das folgende Beispiel bzw. die *Kochsche Schneeflocke* für eine Menge mit *fraktalem* Rand.

5. Für eine Teilmenge eines Funktionen- oder Folgenraumes ist die Bestimmung des Randes im Allgemeinen nicht einfach.

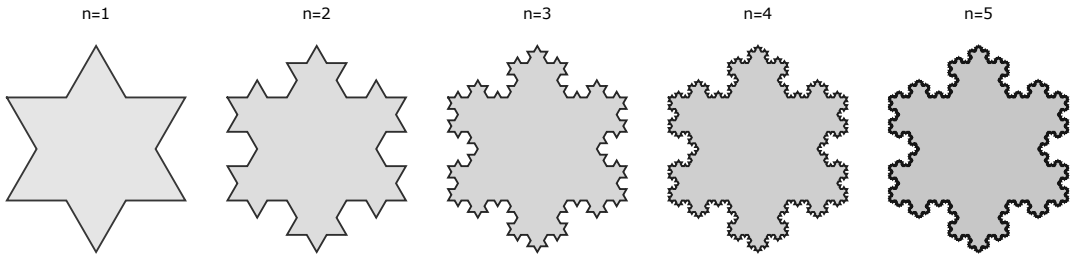

Abbildung Die Kochsche Schneeflocke entsteht durch einen Grenzprozess $n \to \infty$ aus einer rekursiv definierten Folge von polygonal berandeten Teilmengen des $\mathbb{R}^2$ (dargestellt sind nur die ersten fünf Folgenglieder). Der Rand der Schneeflocke ist eine fraktale Kurve und besitzt die Dimension $\ln(4)/\ln(3) \approx 1.26$.

Beispiel Wir betrachten $X = \mathbb{R}^2$ ausgestattet mit einer beliebigen p-Norm sowie die Teilmenge

$$U = \mathbb{Q}^2 = \left\{(x_1, x_2) \in \mathbb{R}^2 \,:\, x_1, x_2 \in \mathbb{Q}\right\},$$

die aus allen Punkten der Zahlenebene besteht, die zwei rationale Koordinaten besitzen. Da jede reelle Zahl beliebig genau durch rationale Zahlen approximiert werden kann, ergibt sich

$$\operatorname{bnd}(U) = \mathbb{R}^2, \qquad \operatorname{int}(U) = \emptyset, \qquad \operatorname{cls}(U) = \mathbb{R}^2.$$

Insbesondere besitzt diese Menge überhaupt keine inneren oder äußeren Punkte, sondern jeder Punkt der Ebene ist Randpunkt von U.

Lemma (**äquivalente Charakterisierung des Randes**) Ein Punkt $x_* \in X$ ist genau dann ein Randpunkt von $U \subseteq X$, wenn er als Grenzwert einer Folge aus U *und* als Grenzwert einer Folge aus $X \setminus U$ dargestellt werden kann.

Beweis *Hinrichtung*: Sei $x_* \in \operatorname{bnd}(U)$ ein beliebiger Randpunkt von U. Für jedes $n \in \mathbb{N}$ liegt die Kugel $B_1/n(x_*)$ weder in $X \setminus U$ noch in U und daher können wir $x_n \in B_1/n(x_*) \cap U$ und $y_n \in B_1/n(x_*) \cap (X \setminus U)$ wählen. Insgesamt ergibt sich eine Folge $(x_n)_{n\in\mathbb{N}}$ und U sowie eine Folge $(y_n)_{n\in\mathbb{N}}$ und $X \setminus U$, die wegen $d(x_n, x_*) < 1/n$ und $d(y_n, x_*) < 1/n$ beide gegen x_* konvergieren.

Rückrichtung: Seien nun $(x_n)_{n\in\mathbb{N}} \subseteq U$ und $(y_n)_{n\in\mathbb{N}} \subseteq X \setminus U$ zwei Folgen, die gegen denselben Grenzwert $x_* \in X$ konvergieren. Für jedes $\varepsilon > 0$ enhält $B_\varepsilon(x_*)$ sowohl Punkte aus U als auch Punkte aus $X \setminus U$ (siehe die äquivalente Charakterisierung von Konvergenz) und kann daher weder in $X \setminus U$ noch in U liegen. Also ist x_* weder äußerer noch innerer Punkt von U, sondern muss im Rand von U liegen. □

Bemerkung

1. Ein innerer Punkt von U kann niemals Grenzwert einer Folge aus $X \setminus U$ sein, aber kann auf viele Arten als Grenzwert eine Folge von U dargestellt werden. Analoge Aussagen gelten für jede äußeren Punkt von U.

2. Ist U eine endliche Punktmenge, so ist jeder dieser Punkte auch Randpunkt von U. Das Lemma gilt auch in diesem Fall, allerdings ist jede konvergente Folge in U fast konstant.

Stetigkeit von Abbildungen zwischen metrischen Räumen

Erinnerung Sei $f : X \to \tilde{X}$ eine Abbildung. Dann bezeichnen

$$f^{-1}\big(\tilde{x}\big) := \big\{x \in X \,:\, f(x) = \tilde{x}\big\} \qquad \text{bzw.} \qquad f^{-1}\big(\tilde{U}\big) := \big\{x \in X \,:\, f(x) \in \tilde{U}\big\}$$

die Urbildmenge des Punktes $\tilde{x} \in \tilde{X}$ bzw. der Menge $\tilde{U} \subseteq \tilde{X}$ unter der Abbildung f. Insbesondere sind diese Mengen selbst dann wohldefiniert, wenn f keine Umkehrabbildung besitzt. [26]

Definition Eine Abbildung $f : X \to \tilde{X}$ zwischen zwei metrischen Räumen wird stetig im Punkt $x_* \in X$ genannt, falls für jedes $\tilde{\varepsilon} > 0$ ein $\varepsilon > 0$ existiert, sodass die Implikation

$$d(x,\, x_*) < \varepsilon \quad \Longrightarrow \quad \tilde{d}\big(f(x),\, f(x_*)\big) < \tilde{\varepsilon}$$

für alle $x \in X$ gilt.

Bemerkungen

1. Die Definition verallgemeinert den punktweisen Stetigkeitsbegriff aus *Analysis 1* in natürlicher Weise. Sie kann auch dann verwendet werden, wenn f nur auf einer Teilmenge $D \subset X$ definiert ist, aber dies wird im Moment keine Rolle spielen.

2. Wir nennen f stetig, falls f in jedem Punkt $x_* \in X$ stetig ist.

3. Statt *Abbildung* können wir *Funktion* sagen. In dieser Vorlesung sprechen wir von einer *skalaren Funktion*, wenn $\tilde{X} = \mathbb{R}$ gilt, d.h. wenn f jeden Punkt aus X auf eine reelle Zahl abbildet, wobei wir dann immer stillschweigend voraussetzen, dass $\tilde{d}$ der euklidische Betrag ist.

4. Wir nennen f Lipschitz-stetig, falls es eine Konstante $L > 0$ gibt, sodass

   $$\tilde{d}\big(f(x),\, f(y)\big) \le L\, d(x,\, y)$$

   für alle $x,\, y \in X$ gilt. Wie schon in *Analysis 1* können wir leicht zeigen, dass jede Lipschitz-stetige Funktion stetig sein muss (man wähle $\varepsilon = \tilde{\varepsilon}/L$ in der obigen Definition), aber dass die Umkehrung im Allgemeinen nicht richtig ist.

[26] Beachte, dass f genau dann invertierbar ist, wenn die Urbildmenge $f^{-1}(\tilde{x})$ für jedes $\tilde{x} \in \tilde{X}$ aus genau einem Element besteht.

Theorem (**äquivalente Charakterisierung von Stetigkeit**) Die folgenden vier Aussagen sind paarweise äquivalent:

1. Die Abbildung $f : X \to \tilde{X}$ ist stetig.
2. Jede konvergente Folge aus X wird unter f auf eine konvergente Folge in $\tilde{X}$ abgebildet.
3. Das Urbild jeder offenen Teilmenge von $\tilde{X}$ ist offen in X.
4. Das Urbild jeder abgeschlossenen Teilmenge von $\tilde{X}$ ist abgeschlossen in X.

Beweis $\underline{1 \Rightarrow 2}$: Sei $(x_n)_{n\in\mathbb{N}}$ eine konvergente Folge aus X mit Grenzwert x_∞ und sei $\tilde{\varepsilon} > 0$ beliebig. Wir wählen $\varepsilon > 0$ wie in der Definition von Stetigkeit sowie $N \in \mathbb{N}$, sodass $d(x_n, x_\infty) < \varepsilon$ und damit auch $d(f(x_n), f(x_\infty)) < \tilde{\varepsilon}$ für alle $n > N$ gilt. Da $\tilde{\varepsilon}$ beliebig war, konvergiert $f(x_n)$ für $n \to \infty$ gegen $f(x_\infty)$ bzgl. der Metrik $\tilde{d}$ in $\tilde{X}$.

$\underline{2 \Rightarrow 1}$: Wir wählen $x_* \in X$ beliebig und nehmen an, es gäbe ein $\tilde{\varepsilon} > 0$, für dass in der obigen Definition kein entsprechendes $\varepsilon > 0$ existiert. Dann finden wir für jedes $n \in \mathbb{N}$ ein $x_n \in X$, sodass $d(x_n, x_*) < 1/n$ und $\tilde{d}(f(x_n), f(x_*)) \geq \tilde{\varepsilon}$ gilt (denn sonst wäre $\varepsilon = 1/n$ ja eine zulässige Wahl). Nach Konstruktion konvergiert x_n für $n \to \infty$ gegen $x_\infty = x_*$ und die Voraussetzung garantiert $f(x_*) = \lim_{n\to\infty} f(x_n)$ und damit $\lim_{n\to\infty} \tilde{d}(f(x_n), f(x_*)) = 0$. Dies ist ein Widerspruch zu $\tilde{d}(f(x_n), f(x_*)) \geq \tilde{\varepsilon} > 0$ für alle n, d.h. f ist in x_* stetig.

$\underline{1 \Rightarrow 3}$: Sei $\tilde{U} \subseteq \tilde{X}$ offen und sei $x_* \in U := f^{-1}\big(\tilde{U}\big)$ beliebig fixiert. Wir wählen zunächst $\tilde{\varepsilon} > 0$, sodass $B_{\tilde{\varepsilon}}(f(x_*))$ ganz in $\tilde{U}$ liegt, und anschließend $\varepsilon > 0$ wie in der Definition. Für alle Punkte $x \in B_\varepsilon(x_*)$ ergibt sich $\tilde{d}(f(x_n), f(x_*)) < \tilde{\varepsilon}$ und damit $f(x_n) \in B_{\tilde{\varepsilon}}(f(x_*)) \subset \tilde{U}$ aus der Stetigkeit von f in x_*. Insbesondere liegt die Kugel $B_\varepsilon(x_*)$ ganz in U, dem Urbild von $\tilde{U}$ unter f. Da x_* beliebig war, folgt insgesamt die Offenheit von U.

$\underline{3 \Rightarrow 1}$: Seien $x_* \in X$ und $\tilde{\varepsilon} > 0$ beliebig. Nach Voraussetzung ist das Urbild von $B_{\tilde{\varepsilon}}(f(x_*))$ offen in X und muss daher eine Kugel von Radius ε um x_* enthalten. Für jedes $x \in B_\varepsilon(x_*)$ gilt dann $f(x) \in B_{\tilde{\varepsilon}}(f(x_*))$ nach Konstruktion und wir schließen, dass f in x_* stetig ist.

$\underline{3 \Leftrightarrow 4}$: Beide Implikationen ergeben sich unmittelbar aus der Formel

$$f^{-1}\big(\tilde{X} \setminus \tilde{U}\big) = X \setminus f^{-1}\big(\tilde{U}\big)\,,$$

die ganz allgemein für die Urbilder jeder Abbildung gilt und mit elementarer Aussagenlogik hergeleitet werden kann (Übungsaufgabe). □

Bemerkungen

1. Das Theorem ist sowohl aus praktischer als auch aus theoretischer Sicht sehr nützlich. Insbesondere können wir analog zu den Argumenten aus *Analysis 1* wieder Rechenregeln für stetige Funktionen herleiten. Zum Beispiel ist für zwei stetige Funktionen $f : X \to \tilde{X}$ und $\tilde{f} : \tilde{X} \to \hat{X}$ ihre Komposition $\tilde{f} \circ f : X \to \hat{X}$ selbst stetig. Ist $\tilde{X}$ sogar ein normierter Raum, so ist die Summe zweier stetiger Funktionen f, $g : X \to \tilde{X}$ auch stetig usw. [27]

[27] Beachte, dass die Summe zweier Funktionen nur dann wohldefiniert ist, wenn es eine Addition im Bildraum gibt.

2. *Achtung*: Im Allgemeinen hat eine stetige Abbildung **nicht** die Eigenschaft, dass das Bild einer offenen Menge offen ist. Zum Beispiel bildet die reelle Sinusfunktion das offene Intervall $(0,\, \pi)$ auf das nicht-offene Intervall $(0,\, 1]$ ab. Wir werden aber unten sehen, dass das Bild einer kompakten Menge unter einer stetigen Abbildung immer kompakt ist.

Folgerung Ist $(X,\, \|\cdot\|)$ ein normierter Raum, so ist die Norm als skalare Funktion von X nach $\mathbb{R}$ stetig.

*Zusatz**: In jedem metrischen Raum ist die Metrik d stetig als skalare Funktion von $X \times X$ nach $\mathbb{R}$, sofern $X \times X$ mit einer natürlichen Produktmetrik ausgestattet ist.

Beweis Für jede konvergente Folge $(x_n)_{n\in\mathbb{N}}$ mit Grenzwert x_∞ gilt nach Dreiecksungleichung

$$\big|\|x_n\| - \|x_\infty\|\big| \le \big\|x_n - x_\infty\big\| \quad \xrightarrow{n\to\infty} \quad 0$$

und die zweite Behauptung folgt aus dem Theorem.

*Zusatz**: Die Definition einer Produktmetrik (siehe Übungen) impliziert, dass eine Folge $\big((x_n,\, y_n)\big)_{n\in\mathbb{N}}$ in $X \times X$ genau dann gegen einen Grenzwert $(x_\infty,\, y_\infty)$ konvergiert, wenn sie „komponentenweise“ konvergiert, d.h. wenn

$$x_\infty = \lim_{n\to\infty} x_n \qquad \text{sowie} \qquad y_\infty = \lim_{n\to\infty} y_n$$

gilt. Insbesondere ergibt sich

$$d(x_n,\, x_\infty) + d(y_n,\, y_\infty) \quad \xrightarrow{n\to\infty} \quad 0\,, \qquad \max\big\{d(x_n,\, x_\infty),\, d(y_n,\, y_\infty)\big\} \quad \xrightarrow{n\to\infty} \quad 0$$

und damit die Behauptung nach elementaren Umformungen. □

Lemma (**abstrakte Beispiele für offene und abgeschlossene Mengen**) Sei $f : X \to \mathbb{R}$ eine stetige skalare Funktion und sei $c \in \mathbb{R}$ eine beliebige reelle Zahl. Dann sind die Mengen

$$\big\{x \in X \,:\, f(x) < c\big\}\,, \qquad \big\{x \in X \,:\, f(x) > c\big\}$$

beide offen, wohingegen jede der drei Mengen

$$\big\{x \in X \,:\, f(x) \le c\big\}\,, \qquad \big\{x \in X \,:\, f(x) = c\big\}\,, \qquad \big\{x \in X \,:\, f(x) \ge c\big\}$$

abgeschlossen ist.

Beweis Bei den Mengen handelt es sich um die Urbilder der in $\mathbb{R}$ offenen Mengen $(-\infty,\, c)$, $(c,\, +\infty)$ bzw. der in $\mathbb{R}$ abgeschlossenen Mengen $(-\infty,\, c]$, $[c,\, c] = \{c\}$, $[c,\, +\infty)$. □

Vorlesungswoche 03

ergänzende Betrachtungen*

Definition Wir nennen eine offene (bzw. abgeschlossene) Teilmenge $U \subseteq X$ zusammenhängend, falls sie nicht als disjunkte Vereinigung zweier oder mehrerer offener (bzw. abgeschlossener), jeweils nichtleerer Mengen dargestellt werden kann.

Bemerkungen

1. Salopp gesprochen gilt: Eine Menge $U \subseteq X$ ist genau dann zusammenhängend, wenn sie nicht in zwei oder mehrere, voneinander getrennte Teile zerfällt (siehe die Bilder).
2. Man kann *Zusammenhang* auch für Mengen definieren, die weder offen noch abgeschlossen sind, aber diese spielen in dieser Vorlesung keine Rolle.
3. Wir werden später die verwandten Konzepte *wegzusammenhängend* und *einfach zusammenhängend* einführen.

Abbildung Zwei Beispiele (türkis und grün) sowie zwei Gegenbeispiele (orange und rosa) für eine zusammenhängende Menge im $\mathbb{R}^2$. Alle Menge sind hier offen, d.h. der Rand gehört jeweils nicht dazu, aber es gibt ganz analoge Beispiele und Gegenbeispiel mit abgeschlossenen Mengen.

Definition Eine Abbildung $f : X \to \tilde{X}$ heißt Homöomorphismus, sofern sie stetig und bijektiv ist und außerdem auch die Umkehrabbildung $f^{-1} : \tilde{X} \to X$ stetig ist.

Beispiel Die Mengen

$$P := (0,\, \infty) \times (-\pi,\, +\pi) = \big\{(r,\, \theta) \,:\, r > 0\,,\ -\pi < \theta < \pi\big\}\,,$$

und

$$G := \big\{(x_1,\, x_2) \in \mathbb{R}^2 \,:\, x_1 > 0\big\}$$

sind Teilmengen des $\mathbb{R}^2$ und können daher jeweils als ein metrischer Raum bzgl. des euklidischen Abstandes angesehen werden. Die ebenen Polarkoordinaten

$$(r,\, \theta) \in P \quad \mapsto \quad (x_1,\, x_2) = \big(r\cos(\theta),\, r\sin(\theta)\big) \in G$$

definieren einen Homöomorphismus von P nach G, wobei die Umkehrabbildung durch

$$(x_1,\, x_2) \in G \quad \mapsto \quad (r,\, \theta) = \left(\sqrt{x_1^2 + x_2^2},\, \operatorname{sgn}(x_2) \cos\left(x_1/\sqrt{x_1^2 + x_2^2}\right)\right) \in P$$

gegeben ist und einen Homöomorphismus von G nach P darstellt.

Bemerkung: Die ebenen Polarkoordinaten können auch als stetige Abbildung von $[0,\, \infty) \times \mathbb{R} \to \mathbb{R}^2$ angesehen werden, die dann allerdings nicht mehr invertierbar ist.

Achtung Aus der Invertierbarkeit und der Stetigkeit einer Abbildung folgt im Allgemeinen nicht die Stetigkeit der Umkehrabbildung.

Beispiel: Wir betrachten das halboffene Intervall $(-\pi,\, +\pi]$ bzw. die Einheitssphäre im $\mathbb{R}^2$ als metrische Räume, wobei die jeweilige Metrik durch den eindimensionalen bzw. zweidimensionalen euklidischen Abstand erzeugt sei. Dann ist die Abbildung f mit

$$\theta \in (-\pi,\, +\pi] \quad \mapsto \quad f(\theta) = \big(\cos(\theta),\, \sin(\theta)\big) \in \big\{(x_1,\, x_2) \,:\, x_1^2 + x_2^2 = 1\big\}$$

stetig und auch invertierbar, aber die Umkehrabbildung ist nicht stetig. In der Tat, mit $\theta_n = -\pi + 1/n$ konvergiert $f(\theta_n)$ für $n \to \infty$ gegen $f(+\pi)$ im Bildraum von f (und damit im Urbildraum von f^{-1}), aber $1/n$ konvergiert im Urbildraum von f (das ist der Bildraum von f^{-1}) nicht gegen $+\pi$, sondern gegen $-\pi$.[28]

metrische Produkträume Seien $(X_1,\, d_1)$ und $(X_2,\, d_2)$ zwei metrische Räume und sei $p \in [1,\, \infty)$ bzw. $p = \infty$ ein gegebener Exponent. Dann wird durch

$$d\big((x_1,\, x_2),\, (y_1,\, y_2)\big) := \Big(d_1(x_1,\, y_1)^p + d_2(x_2,\, y_2)^p\Big)^{1/p}$$

bzw.

$$d\big((x_1,\, x_2),\, (y_1,\, y_2)\big) := \max\big\{d_1(x_1,\, y_1)\,,\; d_2(x_2,\, y_2)\big\}$$

jeweils in natürlicher Weise ein Metrik auf dem Produktraum

$$X_1 \times X_2 := \big\{(x_1,\, x_2) \,:\, x_1 \in X_1\,,\; x_2 \in X_2\big\}$$

definiert (siehe die Übungen).[29] Mithilfe unserer Kenntnisse über metrische Räume können wir eine Vielzahl von Aussagen herleiten, zum Beispiel:

1. Sind $O_1 \subseteq X_1$ offen und $O_2 \subseteq X_2$ offen, so ist $O_1 \times O_2 \subseteq X_1 \times X_2$ offen bzgl. d. Analoge Aussagen gelten für abgeschlossene Mengen.

2. Eine Folge $(x_{n,1},\, x_{n,2})_{n\in\mathbb{N}}$ aus $X_1 \times X_2$ konvergiert bzgl. d genau dann gegen $(x_{\infty,1},\, x_{\infty,2})$, wenn die beiden Komponentenfolgen $(x_{n,1})_{n\in\mathbb{N}}$ bzw. $(x_{n,2})_{n\in\mathbb{N}}$ bzgl. d_1 bzw. d_2 gegen $x_{\infty,1}$ bzw. $x_{\infty,2}$ konvergieren.

Wir können die obige Konstruktion mühelos auf jedes *endliche* Produkt $X_1 \times \ldots \times X_m$ verallgemeinern und bemerken, dass auf diese Weise gerade die p-Normen auf dem $\mathbb{R}^m$ aus dem euklidischen Betrag auf $\mathbb{R}$ gewonnen werden können.

Merkregel Die Produkträume metrischer Räume sind wieder metrisch und der $\mathbb{R}^m$ kann in diesem Sinne als das m-fache Produkt von $\mathbb{R}$ mit sich selbst betrachtet werden.

[28] Dieses Beispiel hat wieder etwas mit ebenen Polarkoordinaten zu tun und die Unstetigkeit der Umkehrabbildung kann sehr gut mit einer Skizze visualisiert werden.

[29] Wir könnten natürlich andere Notationen verwenden und zum Beispiel die Abhängigkeit von d_1, d_2 und p durch eine geeignete Indexschreibweise verdeutlichen.

Teilmengen metrischer Räume als metrische Räume Jede Teilmenge $\hat{X} \subset X$ kann in natürlicher Weise als metrischer Raum $(\hat{X},\, \hat{d})$ betrachtet werden, wobei die Metrik $\hat{d} : \hat{X} \times \hat{X} \to \mathbb{R}$ durch die Metrik d induziert wird, d.h. es gilt $\hat{d}(x,\, y) = d(x,\, y)$ für alle $x,\, y \in \hat{X}$.[30]

Für jeden Punkt $x_* \in X$ und jeden Radius gibt es dann die offene Kugel $B_\varrho(x_*)$ sowie die offene Kugel $\hat{B}_\varrho(x_*)$, wobei sich die erste auf $(X,\, d)$ und die zweite auf $(\hat{X},\, \hat{d})$ bezieht. Wir rechnen leicht nach, dass

$$\hat{B}_\varrho(x_*) = B_\varrho(x_*) \cap \hat{X}$$

gilt, und eine analoge Formel ergibt sich für die abgeschlossenen Kugeln.

Achtung Beim Übergang von X zu $\hat{X}$ kann sich die Offenheit/Abgeschlossenheit einer Menge $U \subset \hat{X} \subset X$ ändern (analoges wird auch für die Kompaktheit gelten).

Beispiel: Seien $X = \mathbb{R}$ und $\hat{X} = (-1,\, +1)$, wobei der Abstand in beiden Mengen durch den euklidischen Betrag gegeben sei, d.h. es gilt $d(x,\, y) = |x - y|$. Die Menge $U = [0,\, 1)$ ist im metrischen Raum $(X,\, d)$ nicht abgeschlossen, aber im Raum $(\hat{X},\, \hat{d})$ ist sie abgeschlossen (Übungsaufgabe). [31]

Merkregel Jede Teilmenge eines metrischen Raumes ist zwar auch ein metrischer Raum, aber die Beziehung zwischen beiden ist subtiler als es auf den ersten Blick scheint.[32]

Lineare Abbildungen zwischen zwei normierten Räumen Sind $(X,\, \|\cdot\|)$ und $(Y,\, |||\cdot|||)$ gegebene normierte Räume, so können wir mit unserem Wissen über Normen und lineare Vektorräume zeigen, dass eine *lineare* Abbildung $f : X \to Y$ genau dann stetig ist, wenn die sogenannte *Operatornorm*

$$\|f\|_{\text{op}} := \sup\big\{|||f(x)||| \,:\, \|x\| = 1\big\}$$

im eigentlichen Sinne wohldefiniert ist, d.h. wenn das Supremum auf der rechten Seite eine nichtnegative reelle Zahl ist. Insbesondere ist die Menge aller linearen und stetigen Abbildungen von X nach Y in natürlicher Weise selbst ein normierter Raum und spielt in der *Funktionalanalysis* eine wesentliche Rolle. Wenn die Dimension des Vektorraumes X *endlich* ist, so ist jede lineare Abbildung auf X automatisch stetig,[33] aber auf unendlichen dimensionalen Räumen gibt es viele lineare Abbildungen, die unstetig sind und daher verblüffende Eigenschaften aufweisen.

[30] Streng genommen sind $\hat{d}$ und d verschiedene Abbildungen, da sie auf unterschiedlichen Mengen definiert sind (nämlich auf $\hat{X} \times \hat{X}$ und $X \times X$). Sie beschreiben aber natürlich denselben Abstandsbegriff. Man sagt auch, $\hat{s}$ ist die Einschränkung von d auf $\hat{X}$.

[31] Den wesentlichen Unterschied zwischen $(X,\, d)$ und $(\hat{X},\, \hat{d})$ erkennen wir an der Folge $(x_n)_{n\in\mathbb{N}}$ mit $x_n = 1 - 1/n$: Sie liegt in U und konvergiert in X gegen den Grenzwert $x_\infty = 1$, der aber nicht zu U gehört. Im Raum $(\hat{X},\, \hat{d})$ konvergiert diese Folge jedoch nicht, eben weil 1 gar nicht zu $\hat{X}$ gehört.

[32] Es gibt natürlich eine entsprechende Theorie, aber diese werden wir in dieser Vorlesung nicht entwickeln. Beachte auch, dass jede Teilmenge eines normierten Raumes zwar ein metrischer, aber im Allgemeinen kein normierter Raum sein wird. Eine Kugel im $\mathbb{R}^m$ ist zum Beispiel kein Vektorraum.

[33] Der Grund ist, dass jede lineare Abbildung nach Wahl einer Basis mit einer Matrix identifiziert werden kann. Siehe die Vorlesung *Lineare Algebra*.

1.6 Kompaktheit

Vorbemerkung Ein wichtiges und immer wiederkehrendes Konzept in der gesamten Mathematik ist die Kompaktheit von Teilmengen eines metrischen Raumes (X, d), wobei es zwei verschiedene Möglichkeiten gibt, diesen Begriff einzuführen.

Definition Wir nennen eine Teilmenge $K \subseteq X$

1. überdeckungskompakt, falls jede offene Überdeckung von K in X eine endliche Teilüberdeckung besitzt, und

2. folgenkompakt, wenn jede Folge in K mindestens einen Häufungspunkt *in* K besitzt.

Bemerkungen

1. Der erste Teil der Definition ist wie folgt zu verstehen: Ist I eine beliebige Indexmenge und $(O_i)_{i\in I}$ eine Familie[34] offener Mengen $O_i \subseteq X$ mit $K \subseteq \bigcup_{i\in I} O_i$, dann existieren *endlich* viele Indizes $i_1, \ldots, i_N$, sodass schon $K \subseteq O_{i_1} \cup \ldots \cup O_{i_N}$ gilt. Im Fall einer endlichen Menge I ist diese Aussage trivialerweise richtig, aber die endliche Teilüberdeckung muss bei einer kompakten Menge K auch dann existieren, wenn die Indexmenge I abzählbar oder überabzählbar unendlich viele Elemente enthält.

2. Ein Häufungspunkt ist — ganz analog zur Begriffsbildung in *Analysis 1* — immer Grenzwert einer Teilfolge. Der zweite Teil der Definition meint also zum einen, dass für jede Folge $(x_n)_{n\in\mathbb{N}}$, deren Glieder x_n alle in K liegen, mindestens eine strikt monotone Indexfolge $(n_k)_{k\in\mathbb{N}} \subset \mathbb{N}$ mit $\lim_{k\to\infty} n_k = \infty$ sowie ein $x_* \in X$ existieren, sodass $d(x_{n_k}, x_*)$ für $k \to \infty$ gegen 0 konvergiert. Der zweite Teil der Definition fordert außerdem, dass der Häufungspunkt x_* auch in der Menge K (und nicht im Komplement $X \setminus K$) liegt.

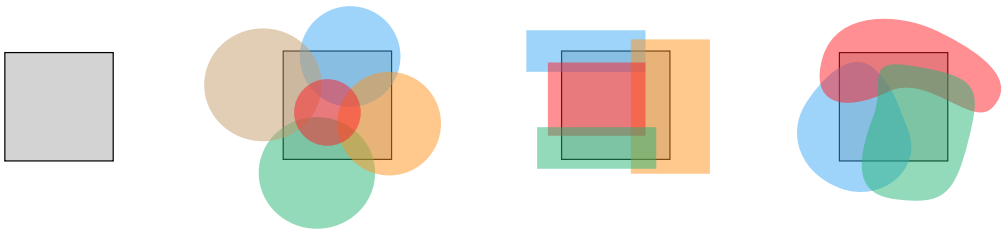

Abbildung Drei verschiedene offene Überdeckungen eines abgeschlossenen Quadrats in $\mathbb{R}^2$ (grau) mit endlich vielen offenen Mengen (farbig). Bei einer kompakten Menge kann aus einer Überdeckung mit unendlich vielen offenen Mengen immer eine endliche Teilüberdeckung ausgewählt werden.

Beispiel Ein kompaktes Intervall $[a, b]$ ist als Teilmenge des $\mathbb{R}$ folgenkompakt (siehe den Satz von Bolzano-Weierstraß in *Analysis 1*). Die Menge $[0, 1] \cup [2, 3]$ ist zwar kein Intervall, ist aber auch folgenkompakt.

[34] *Familie* ist ein anderes Wort für *Menge*.

Gegenbeispiel Das offene Intervall $(-2, +2)$ ist in $\mathbb{R}$ weder überdeckungs- noch folgenkompakt. In der Tat, die durch $O_n := (-2+1/n, +2-1/n)$ definierte abzählbare offene Überdeckung $(O_n)_{n\in\mathbb{N}}$ besitzt offensichtlich keine endliche Teilüberdeckung und die Folge $(x_n)_{n\in\mathbb{N}}$ mit $x_n = 2 - 1/n$ liegt zwar im Intervall und konvergiert, aber der Grenzwert $x_\infty = 2$ liegt außerhalb.

Abbildung *Links*: Konstruktionsidee für eine offene Überdeckung einer offenen Menge, die keine endliche Teilüberdeckung besitzt: Für *jeden* Punkt y in einer gegebenen offenen Kreisscheibe B (grau) wählen wir eine offene Kreisscheibe O_y um y, wobei der Radius gerade der halbe Abstand von y zum Rand ∂B ist (hier dargestellt für 5 farbige Punkte). Die Familie $(O_y)_{y\in U}$ stellt eine überabzählbare offene Überdeckung von B dar, besitzt aber keine endliche Teilüberdeckung. In der Tat, für jede Wahl $y_1, \ldots, y_N$ kann B nicht vollständig in $O_{y_1} \cup \ldots \cup O_{y_N}$ enthalten sein, da eine solche Menge einen positiven Abstand zum Rand ∂B aufweist. *Rechts*: Bei einer analogen Konstruktion für die abgeschlossene Kreisscheibe $\overline{B}$ muss auch für jeden Randpunkt $y \in \partial B$ (lila) eine offene Menge O_y gewählt werden (zum Beispiel irgendeine offene Kreisscheibe um y). Dies ändert die Diskussion dramatisch, denn nun wird $(O_y)_{y\in\overline{B}}$ immer eine endliche Teilüberdeckung von $\overline{B}$ besitzen ($\overline{B}$ ist ja kompakt).

Theorem (**Äquivalenz der beiden Kompaktheitsdefinitionen**) Eine Menge $K \subseteq X$ ist genau dann überdeckungskompakt, wenn sie folgenkompakt ist.

Hinweis: Der nun folgende Beweis ist sehr trickreich und alles andere als einfach zu verstehen. Sie sollten ihn trotzdem durcharbeiten, da Sie auf diese Weise viel über mathematische Argumentationstechniken lernen können. Für die Abschlußprüfung ist dieser Beweis aber nicht relevant.

Beweis* *Teil 1, Vorbereitung*: Wir nehmen an, dass K folgenkompakt ist und betrachten eine beliebige Familie $(O_i)_{i\in I}$ offener Mengen mit $K \subseteq \bigcup_{i\in I} O_i$. Wir definieren für jedes $n \in \mathbb{N}$ die Menge

$$K_n := \{x \in K \ : \ \text{es existiert ein } i \in I \text{ mit } B_{1/n}(x) \subseteq O_i\}$$

und bemerken, dass

$$K_1 \subseteq K_2 \subseteq K_3 \subseteq \ldots \qquad \text{sowie} \qquad K_1 \cup K_2 \cup K_3 \cup \ldots = K\,,$$

gilt, d.h. die Mengen K_n werden mit wachsendem n immer größer und schöpfen zusammen ganz K aus. Wir zeigen nun indirekt, dass $K_{n_\#} = K$ für ein geeignetes $n_\# \in \mathbb{N}$ gilt. Angenommen, dies sei nicht der Fall. Dann können wir sukzessive eine Folge $(x_n)_{n\in\mathbb{N}}$ in K definieren, indem wir x_1 beliebig und anschließend immer $x_{n+1} \in K$ als Element von $K \setminus K_n$ wählen. Nach Voraussetzung an K existiert eine Indexfolge $(n_k)_{k\in\mathbb{N}}$ sowie ein Häufungspunkt $x_* \in K$ mit $x_* = \lim_{k\to\infty} x_{n_k}$. Der Punkt x_* muss zu einer der Mengen K_n gehören, d.h. wir finden ein $n_* \in \mathbb{N}$ sowie ein $i_* \in I$ mit $x_* \in K_{n_*}$ und $B_{1/n_*}(x_*) \subset O_{i_*}$. Durch eine einfache Nebenrechnung mit der Dreiecksungleichung verifizieren wir $B_{1/(2\,n_*)}(x) \subset B_{1/n_*}(x_*) \subseteq O_{i_*}$ für jedes $x \in B_{1/(2\,n_*)}(x_*)$ und schließen,

dass die Kugel $B_{1/(2\,n_*)}(x_*)$ eine Teilmenge von $K_{2\,n_*}$ ist. Insbesondere gehören fast alle Glieder der konvergenten Teilfolge $(x_{n_k})_{k\in\mathbb{N}}$ zur Menge $K_{2\,n_*}$. Andererseits gilt für fast alle $k \in \mathbb{N}$ auch $n_k > 2\,n_*$ und damit $x_{n_k} \notin K_{n_k} \supseteq K_{2\,n_*}$, d.h. x_{n_k} liegt außerhalb von $K_{2\,n_*}$. Wir haben damit den gewünschten Widerspruch konstruiert und schließen, dass unsere Annahme falsch war bzw. dass $K = K_{n_\#}$ für ein $n_\# \in \mathbb{N}$ gilt. Insbesondere hat $\varepsilon_\# = 1/n_\#$ die Eigenschaft, dass für jedes $x \in K$ ein $i \in I$ existiert, sodass $B_{\varepsilon_\#}(x) \subset O_i$. Mit diesem wichtigen Hilfsresultat können wir nun die Existenz einer endlichen Teilüberdeckung zeigen.

Teil 1, Hauptargument: Wir nehmen als Antithese an, $(O_i)_{i\in I}$ besitzt keine endliche Teilüberdeckung von K, und definieren rekursiv zwei Folgen $(x_n)_{n\in\mathbb{N}}$ in X sowie $(i_n)_{n\in\mathbb{N}}$ in I: Als Startwerte wählen wir x_1 beliebig und i_1 mit $B_{\varepsilon_\#}(x_1) \subseteq O_{i_1}$. Ausgehend von den jeweils ersten n Folgengliedern wählen wir dann im n+1-ten Schritt x_{n+1} und i_{n+1} derart, dass sie den Bedingungen

$$x_{n+1} \in K \setminus (O_{i_1} \cup \ldots \cup O_{i_n})\,, \qquad B_{\varepsilon_\#}(x_{n+1}) \subseteq O_{i_{n+1}}$$

genügen, wobei die Antithese sowie das Hilfsresultat sicherstellen, dass eine solche Wahl immer möglich ist. Unsere Definition impliziert $d(x_n,\, x_l) \geq \varepsilon_\#$ für alle $n,\, l \in \mathbb{N}$ und wir schließen, dass $(x_n)_{n\in\mathbb{N}}$ keine konvergente Teilfolge enthalten kann. Da dies aber der Folgenkompaktheit von K widerspricht, war die obige Antithese falsch und K ist überdeckungskompakt.

Teil 2: Sei nun K überdeckungskompakt und $(x_n)_{n\in\mathbb{N}}$ eine beliebige Folge aus K. Als Antithese nehmen wir diesmal an, dass sie keine konvergente Teilfolge besitzt. Dann gibt es für jedes $y \in K$ einen Radius $\varepsilon_y > 0$, sodass $O_y := B_{\varrho_y}(y)$ nur endlich viele Glieder der Folge $(x_n)_{n\in\mathbb{N}}$ enthält. In der Tat, wenn ϱ_y nicht existieren würde, könnten wir für jedes $k \in \mathbb{N}$ ein $x_{n_k} \in B_{1/k}(y)$ wählen und würden eine Teilfolge $(x_{n_k})_{k\in\mathbb{N}}$ erhalten, die gegen y konvergiert. Die Familie $(O_y)_{y\in K}$ ist offensichtlich eine offene Überdeckung von K und wir können nach Voraussetzung eine endliche Teilüberdeckung von K finden, die zu N Punkten $y_1, \ldots, y_N$ gehört. Insbesondere liegen alle Glieder von $(x_n)_{n\in\mathbb{N}}$ in der Menge $O_{y_1} \cup \ldots \cup O_{y_N}$, aber diese Menge enthält nach Konstruktion nur endlich viele Folgenglieder. Da dies ein Widerspruch ist, muss K folgenkompakt sein. □

Bemerkungen

1. Im Folgenden sprechen wir einfach von *kompakten* Mengen.

2. Eine einfache, aber wichtige Folgerung ist, dass für jede konvergente Folge $(x_n)_{n\in\mathbb{N}}$ mit Grenzwert x_∞ die Menge

$$K := \{x_1,\ x_2,\ x_3\,,\ \ldots\} \cup \{x_\infty\}$$

   auch überdeckungskompakt ist.

3. *Ausblick**: In jedem topologischen Raum folgt aus der Überdeckungskompaktheit die Folgenkompaktheit, aber die umgekehrte Aussage ist nicht mehr unbedingt richtig.

Lemma (**Eigenschaften kompakter Mengen**) Jede kompakte Menge $K \subseteq X$ ist abgeschlossen und beschränkt. Letzteres meint, dass $K \subset \overline{B}_\varrho(x)$ für ein $x \in X$ und einen Radius $0 < \varrho < \infty$ gilt.

Beweis *Teil 1*: Angenommen, K sei nicht abgeschlossen. Dann gibt es mindestens einen Randpunkt $x_* \in \operatorname{bnd}(K)$ der nicht zu K gehört sowie (siehe die äquivalente Charakterisierung von Randpunkten) eine Folge $(x_n)_{n\in\mathbb{N}}$ aus K mit $x_* = \lim_{n\to\infty} x_n$. Jede Teilfolge konvergiert aber auch gegen x_*, d.h. diese Folge besitzt keinen Häufungspunkt in K. Dies ist ein Widerspruch und unsere Annahme war falsch.

Teil 2: Wir nehmen an, K sei nicht beschränkt und wählen $x_\# \in X$ beliebig. Da für jedes $n \in \mathbb{R}$ die Menge K nicht in $\overline{B}_n(x_\#)$ enthalten ist, können wir einen Punkt $x_n \in K$ so wählen, dass $d(x_n, x_\#) > n$ gilt. Wir erhalten insgesamt eine Folge $(x_n)_{n\in\mathbb{N}}$, die aufgrund der Kompaktheit von K eine konvergente Teilfolge besitzt. Insbesondere existiert eine Indexfolge $(n_k)_{k\in\mathbb{N}}$ sowie ein Häufungspunkt x_*, sodass $\lim_{k\to\infty} n_k = \infty$ sowie $\lim_{k\to\infty} d(x_{n_k}, x_*) = 0$ für alle $k \in \mathbb{N}$ gilt. Die Dreiecksungleichung impliziert

$$d(x_*, x_\#) \geq d(x_{n_k}, x_\#) - d(x_{n_k}, x_*) \geq n_k - d(x_{n_k}, x_*)$$

für alle k und der Limes $k \to \infty$ liefert einen Widerspruch, da die rechte Seite gegen $\infty - 0 = \infty$ konvergiert. Also ist K beschränkt. □

Theorem (**Satz von Heine-Borel**) Im $\mathbb{R}^m$ ist eine Menge genau dann kompakt, wenn sie beschränkt und abgeschlossen ist.

Beweis *Vorüberlegung*: Die Hinrichtung ergibt sich aus dem vorherigen Lemma. Für die Rückrichtung betrachten wir eine abgeschlossene und beschränkte Menge $K \subset \mathbb{R}^m$ sowie eine beliebige Folge $(x_n)_{n\in\mathbb{N}} \subseteq K$. Für jedes $j \in \{1, \ldots, m\}$ ist die entsprechende Komponentenfolge, also die reelle Zahlenfolge $(x_{n,j})_{n\in\mathbb{N}}$, beschränkt in $\mathbb{R}$ (dies ergibt sich zum Beispiel aus der nützlichen Beobachtung) und besitzt nach dem Satz von Bolzano-Weierstraß aus *Analysis 1* eine konvergente Teilfolge. Da aber verschiedene Werte von j zu unterschiedlichen Teilfolgen gehören können, müssen wir etwas subtiler argumentieren. Insbesondere werden wir die gegebene Folge in m aufeinanderfolgenden Schritten sukzessive „ausdünnen“, um schließlich eine Teilfolge mit guten Eigenschaften zu erhalten.

Wahl der Teilfolge: Wir wählen zunächst eine Teilfolge, sodass die entsprechenden ersten Komponenten gegen einen Grenzwert $x_{*,1} \in \mathbb{R}$ konvergieren. Abweichend von der üblichen Notation bezeichnen wir diese Teilfolge mit $(x_n^{(1)})_{n\in\mathbb{N}}$, sodass

$$x_{n,1}^{(1)} \quad \xrightarrow{n\to\infty} \quad x_{*,1}$$

gilt. Im zweiten Schritt wählen wir eine Teilfolge der im ersten Schritt gewählten Teilfolge, sodass die zweiten Komponenten gegen eine reelle Zahl $x_{*,2}$ konvergieren, wobei die Konvergenz der ersten Komponenten durch den Übergang zur Teilfolge nicht verändert wird. Auf diese Weise erhalten wir eine Folge $(x_n^{(2)})_{n\in\mathbb{N}}$ mit

$$x_{n,1}^{(2)} \quad \xrightarrow{n\to\infty} \quad x_{*,1}\,, \qquad x_{n,2}^{(2)} \quad \xrightarrow{n\to\infty} \quad x_{*,2}\,,$$

wobei diese Folge auch Teilfolge der Ausgangsfolge ist. In einem dritten Schritt können wir dann wieder eine Teilfolge wählen, um zusätzlich auch die Konvergenz der dritten Komponenten sicherzustellen. Nach genau m Schritten ergibt sich schließlich eine Folge $(x_n^{(m)})_{n\in\mathbb{N}}$ als Teilfolge von $(x_n)_{n\in\mathbb{N}}$, sodass

$$x_{n,j}^{(m)} \quad \xrightarrow{n\to\infty} \quad x_{*,j}$$

für alle $j = 1, \ldots, m$ gilt. Insbesondere konvergiert diese Folge komponentenweise gegen den Grenzwert $x_\infty = (x_{\infty,1}, \ldots, x_{\infty,m}) \in \mathbb{R}^m$ und damit auch bzgl. jeder Norm in $\mathbb{R}^m$. Da die Menge K abgeschlossen ist, muss auch $x_\infty \in K$ gelten (siehe die äquivalente Charakterisierung abgeschlossener Mengen). □

Bemerkungen

1. Wir haben im Beweis des Theorems die Folgenkompaktheit jeder abgeschlossenen und beschränkten Teilmenge des $\mathbb{R}^m$ gezeigt. Es gibt auch alternative Beweise, die auf dem Konzept der Überdeckungskompaktheit beruhen, siehe zum Beispiel [For, Abschnitt I.3 in Band 2].

2. Die Aussage gilt für jede Norm im $\mathbb{R}^m$ und ganz allgemein in jedem *vollständigen* normierten Raum *endlicher* Dimension.

3. Im Allgemeinen ist eine beschränkte und abgeschlossene Menge nicht kompakt. Dies gilt insbesondere in einem endlich-dimensionalen normierten Raum, der nicht vollständig ist, sowie in jedem unendlich-dimensionalen normierten Raum (egal ob er vollständig ist oder nicht).

4. *Gegenbeispiel**: Die Menge $X = (-1, 0) \cup (0, +1)$ besteht aus zwei disjunkten offenen Intervallen und ist bzgl. des euklidischen Abstands ein metrischer Raum. Das Intervall $U = (0, 1)$ ist als Teilmenge dieses Raumes nicht kompakt, obwohl sie dort beschränkt und abgeschlossen ist (Übungsaufgabe).

   *Bemerkung**: Als Teilmenge des metrischen Raumes $\mathbb{R}$ ist U zwar auch nicht kompakt, aber dort ist sie auch nicht abgeschlossen. Dieses Beispiel zeigt noch einmal: Wenn wir Teilmengen von metrischen Räumen als metrische Räume betrachten (bzgl. der induzierten Metrik), können manchmal unerwartete Dinge passieren.

Kompaktheit der abgeschlossenen Einheitskugel

1. Die abgeschlossene Einheitskugel im $\mathbb{R}^m$ ist kompakt bzgl. jeder p-Norm. Das garantiert gerade der Satz von Heine-Borel.

2. Der Raum $\mathbb{Q}^m$ ist — ausgestattet mit dem Abstand bzgl. einer beliebigen p-Norm — auch ein metrischer Raum, allerdings kein vollständiger. Insbesondere ist die entsprechende abgeschlossene Einheitskugel weder folgen- noch überdeckungskompakt.

   dreifacher Beweis für $m = 1$*: Wir wählen x_* als *irrationale* Zahl zwischen 0 und 1 und betrachten im Folgenden die abgeschlossene Einheitskugel in $\mathbb{Q}$, d.h. die Menge aller rationalen Zahlen $q \in \mathbb{Q}$ mit $-1 \leq q \leq +1$, die wir mit E bezeichnen wollen.

   (a) Wir können eine Approximationsfolge $(q_n)_{n\in\mathbb{N}}$ wählen, die nur aus rationalen Zahlen $q_n \in \mathbb{Q}$ mit $0 < q_n < 1$ besteht und gegen die irrationale Zahl x_* konvergiert. Eine solche Folge besitzt offensichtlich keinen Häufungspunkt innerhalb von E.

   (b) Für jedes $p \in \mathbb{Q}$ setzen wir

   $$O_p := \left\{ q \in \mathbb{Q} : |q - x_*| > \tfrac{1}{2} |p - x_*| \right\}$$

   und bemerken, dass O_p eine offene Teilmenge von $\mathbb{Q}$ ist, die den Punkt p enthält. Insbesondere ist $(O_p)_{p\in\mathbb{Q}}$ eine offene Überdeckung von ganz $\mathbb{Q}$ und damit auch E. Für endliche viele rationale Zahlen $p_1, \ldots, p_N$ ist

   $$\varepsilon := \tfrac{1}{2} \min \left\{ |p_1 - x_*|, \ldots, |p_N - x_*| \right\}$$

immer eine positive Zahl und wir schließen, dass $O_{p_1} \cup \ldots \cup O_{p_N}$ keinen Punkt aus dem Intervall $[x_* - \varepsilon,\, x_* + \varepsilon]$ enthält und daher keine Überdeckung von E sein kann.

(c) Wir können sogar eine disjunkte und abzählbar unendliche Überdeckung von E abgeben. Dazu wählen wir eine strikt monoton wachsende Folge $(x_n)_{\in\mathbb{N}}$ *irrationaler* Zahlen x_n, die für $n \to \infty$ von unten gegen x_* konvergieren (zum Beispiel $x_* = 1/\sqrt{2}$ und $x_n = 1/\sqrt{2} - 1/n$) und definieren

$$O_{\mathrm{L}} := \big\{ q \in \mathbb{Q} \,:\, -1 < q < x_1 \big\}\,, \qquad O_{\mathrm{R}} := \big\{ q \in \mathbb{Q} \,:\, x_* < q < 1 \big\}$$

sowie $O_n = \big\{ q \in \mathbb{Q} \,:\, x_n < q < x_{n+1} \big\}$ für alle $n \in \mathbb{N}$. Die Familie $\big\{ O_{\mathrm{L}}\,,\ O_{\mathrm{R}}\,,\ O_1\,,\ O_2\,,\ O_3\,,\ \ldots \big\}$ besteht aus paarweise disjunkten Mengen, die in Q jeweils offen sind und deren Vereinigung gerade E ist. Insbesondere kann es keine endliche Teilüberdeckung geben.

3. Die Einheitskugel in $(\mathsf{BC}(J),\, \|\cdot\|_\infty)$ besteht aus allen beschränkten und stetigen Funktionen $x : J \to \mathbb{R}$ mit $\|x\|_\infty = \sup_{t\in J} \big|x(t)\big| \leq 1$, aber diese Menge ist nicht kompakt. Für $J = [-1,\, +1]$ besitzen zum Beispiel die durch

$$x_n(t) = t^n \qquad \text{oder} \qquad x_n(t) = \cos\big(n\,\pi\,t\big)$$

definierten Funktionenfolgen jeweils keinen Häufungspunkt bzgl. der ∞-Norm. Im ersten Beispiel sieht man das daran, dass x_n für $n \to \infty$ punktweise gegen eine unstetige Funktion konvergiert (siehe *Analysis 1*). Für das zweite Beispiel ist die Nichtexistenz konvergenter Teilfolgen schwieriger zu zeigen, kann aber mit Plots leicht plausibilisiert werden.

4. *Ausblick**: In einem Banach-Raum — also in einem vollständigen normierten Raum — sind abgeschlossene Kugeln genau dann kompakt, wenn die Dimension des Raumes endlich ist.

Kompaktheit und Stetigkeit

Lemma (**Bild kompakter Mengen**) Seien $f : X \to \tilde{X}$ stetig und $K \subseteq X$ kompakt. Dann ist die Bildmenge

$$f(K) := \big\{ f(x) \,:\, x \in K \big\}$$

kompakt in $\tilde{X}$.

Beweis Sei $(\tilde{O}_i)_{i\in I}$ eine offene Überdeckung von $\tilde{K} := f(K)$ und sei $O_i := f^{-1}(\tilde{O}_i)$ für jedes $i \in I$ die Urbildmenge von $\tilde{O}_i$ unter f. Die Stetigkeit von f impliziert, dass jede Menge O_i offen in X ist und mit elementarer Aussagenlogik zeigen wir, dass $K \subseteq \bigcup_{i\in I} O_i$ gilt. Insbesondere ist $(O_i)_{i\in I}$ eine offene Überdeckung von K und besitzt nach Voraussetzung eine endliche Teilüberdeckung, d.h. wir können endlich viele Indizes $i_1,\, \ldots,\, i_N$ wählen, sodass $K \subseteq O_{i_1} \cup \ldots \cup O_{i_N}$. Mit aussagenlogischen Nebenrechnungen verifizieren wir $f\big(O_i\big) = \tilde{O}_i$ für alle $i \in I$ sowie die Formel

$$\tilde{K} = f(K) \subseteq f\big(O_{i_1}\big) \cup \ldots \cup f\big(O_{i_N}\big) = \tilde{O}_{i_1} \cup \ldots \cup \tilde{O}_{i_N}\,,$$

und haben damit eine endliche Teilüberdeckung von $\tilde{K}$ gefunden. □

Theorem (**Satz von Minimum und Maximum**) Seien $f : X \to \mathbb{R}$ stetig und $K \subseteq X$ kompakt. Dann nimmt f auf K sein Minimum und sein Maximum an, d.h. es existieren Punkte $\underline{x}$, $\overline{x} \in K$, sodass

$$f(\underline{x}) \leq f(x) \leq f(\overline{x})$$

für alle $x \in K$ gilt.[35]

Beweis Wir beweisen nur die Existenz eines Maximierers in K, denn die Existenz des Minimierers ergibt sich aus analogen Argumenten. Nach dem vorherigen Lemma ist die Bildmenge von f kompakt in $\mathbb{R}$ und damit auch beschränkt, d.h.

$$\overline{m} := \sup\{f(x) \,:\, x \in K\}$$

ist eine reelle Zahl. Für jedes $n \in \mathbb{N}$ wählen wir nun einen Punkt $x_n \in K$ mit

$$\overline{m} - 1/n \leq f(x_n) \leq \overline{m}$$

und erhalten eine Folge $(x_n)_{n\in\mathbb{N}}$ in K mit $\overline{m} = \lim_{n\to\infty} f(x_n)$. Die Folgenkompaktheit von K garantiert die Existenz einer Teilfolge $(x_{n_k})_{k\in\mathbb{N}}$ mit $\lim_{k\to\infty} x_{n_k} = \overline{x}$, wobei der Grenzwert $\overline{x}$ zu K gehört (siehe die äquivalente Charakterisierung von abgeschlossenen Mengen und beachte, dass K als kompakte Menge abgeschlossen ist). Die Stetigkeit von f sowie die Wahl der Folge $(x_n)_{n\in\mathbb{N}}$ garantiert

$$f(\overline{x}) = \lim_{k\to\infty} f(x_{n_k}) = \lim_{n\to\infty} f(x_n) = \overline{m}\,,$$

d.h. $\overline{x}$ ist wirklich Maximierer von f in K. □

Bemerkungen

1. Die gleichen Grundideen — also die Kompaktheit einer maximierenden Folge sowie die Stetigkeit von f — hatten wir schon in *Analysis 1* beim Beweis des entsprechenden Satzes verwendet.

2. Wie schon in *Analysis 1* nennen wir $\overline{x}$ bzw. $\underline{x}$ einen Maximierer bzw. einen Minimierer von f in K, wohingegen $f(\overline{x})$ bzw. $f(\underline{x})$ das entsprechende Maximum bzw. Minimum ist.

3. Das Lemma bezieht sich auf Extremstellen von f in K und macht keine Aussagen über das Verhalten von f auf der Komplementmenge $X \setminus K$.

   Beispiel: Für $X = \mathbb{R}$ betrachten wir $K = [-1, +1]$ sowie die stetige Funktion $f : \mathbb{R} \to \mathbb{R}$ mit $f(x) = \exp(x)$. Das Minimum bzw. das Maximum von K wird in $\underline{x} = -1$ bzw. in $\overline{x} = +1$ angenommen, denn es gilt $\mathsf{e}^{-1} \leq f(x) \leq \mathsf{e}^{+1}$ für alle $x \in K$. Außerhalb von K nimmt f aber sowohl kleinere als auch größere Werte an.

4. Beachte, dass es im Allgemeinen mehrere Maximierer von f in K — also verschiedene Kandidaten für $\overline{x}$ — geben kann und dass das Maximum $f(\overline{x})$ das *globale* Maximum von f in K ist. Die Frage, ob es auch *lokale Maxima* von f in

[35] Der Oberstrich bezieht sich in diesem Theorem nicht auf den Abschluss einer Menge, sondern $\overline{x}$ bezeichnet einen Punkt in X, in dem f sein Maximum annimmt.

K gibt, werden wir erst im Rahmen der Differentialrechnung studieren. Analoge Aussagen betreffen die Minima.

Beispiel: Auf dem kompakten Intervall $[-2,\, +2]$ nimmt die Funktion $f : \mathbb{R} \to \mathbb{R}$ mit $f(x) = (1 - x^2)^2 = x^4 - 2\,x^2 + 1$ ihr Maximum in jedem Punkt $\overline{x} \in \{-2,\, +2\}$ an, wohingegen das Minimum in den Punkten $\underline{x} \in \{-1,\, +1\}$ realisiert wird. Darüber hinaus gibt es den lokalen Maximierer $x = 0$.

5. Im Theorem ist es wichtig, dass f den Raum X nach $\mathbb{R}$ abbildet. In einem anderen metrischen Bildraum (zum Beispiel $\tilde{X} = \mathbb{R}^m$) gibt es nämlich in der Regel keine Ordnungsrelation und die Begriffe *Maximum* und *Minimum* können dann nicht sinnvoll definiert werden.

Lemma (**gleichmäßige Stetigkeit**) Eine stetige Abbildung $f : X \to \tilde{X}$ ist auf jeder kompakten Teilmenge $K \subseteq X$ gleichmäßig stetig, d.h. für jedes $\tilde{\varepsilon} > 0$ existiert ein $\varepsilon > 0$, sodass die Implikation

$$d(x,\, x_*) < \varepsilon \quad \Longrightarrow \quad \tilde{d}(x,\, x_*) < \tilde{\varepsilon}$$

für alle $x,\, x_* \in K$ gilt.[36]

Beweis Wir hatten in *Analysis 1* die analoge Aussage für Funktionen $f : \mathbb{R} \to \mathbb{R}$ und kompakte Intervalle $K = [a,\, b]$ hergeleitet. Alle Beweisideen können problemlos übertragen werden; wir müssen nur in allen Formeln den reellen Abstand im Urbild- bzw. Bildbereich durch d bzw. $\tilde{d}$ ersetzen. □

1.7 Banachscher Fixpunktsatz

Vorbemerkung In diesem Abschnitt beweisen wir den wichtigsten Fixpunktsatz der Mathematik, der sehr viele Anwendungen besitzt. Zum Beispiel werden wir am Ende dieses Semesters mit seiner Hilfe den Hauptsatz über gewöhnliche Differentialgleichungen, d.h. den Satz von Picard-Lindelöf, herleiten. Obwohl der Banachsche Fixpunktsatz sehr abstrakt ist, sind die zugrunde liegenden Beweisideen verblüffend einfach und ausgesprochen elegant.

Hinweis: Der Banachsche Fixpunktsatz und sein Beweis gehören an allen Universitäten zum Prüfungskanon der mathematischen Analysis.

Theorem (**Fixpunktsatz von Banach**) Seien $(X,\, d)$ ein vollständiger metrischer Raum, $A \subseteq X$ eine abgeschlossene Teilmenge und $f : X \to X$ eine stetige Abbildung. Außerdem sei f eine kontraktive Selbstabbildung von A, d.h.

1. es gilt $f(x) \in A$ für jedes $x \in A$,

2. es existiert eine reelle Zahl κ mit $0 \leq \kappa < 1$, sodass

$$d\big(f(x),\, f(y)\big) \leq \kappa\, d(x,\, y)$$

für alle $x,\, y \in A$ gilt.

[36] Beachte, dass hier ε von $\tilde{\varepsilon}$, aber nicht von x und x_* abhängen darf. Insofern ist die gleichmäßige Stetigkeit mehr als Stetigkeit in jedem Punkt $x_* \in K$, da dann ε auch von x_* abhängen darf.

Dann besitzt f genau einen Fixpunkt in A, d.h. es existiert ein eindeutiges $x_* \in A$ mit

$$f(x_*) = x_* \,.$$

Darüber hinaus konvergiert für *jeden* Startpunkt $x_0 \in A$ die durch die Rekursionsvorschrift

$$x_{n+1} := f(x_n)$$

definierte Folge gegen diesen Fixpunkt x_*, wobei die beiden Fehlerabschätzungen

$$d(x_n,\, x_*) \le \frac{\kappa^n}{1-\kappa}\, d(x_0,\, x_1)\,, \qquad d(x_n,\, x_*) \le \frac{\kappa}{1-\kappa}\, d(x_{n-1},\, x_n)$$

für jedes $n \in \mathbb{N}$ gelten.

Beweis *Eindeutigkeit*: Wir nehmen an, dass x_* und y_* zwei verschiedene Fixpunkte von f in A sind, d.h. dass

$$f(x_*) = x_*\,, \qquad f(y_*) = y_*\,, \qquad d(x_*,\, y_*) > 0$$

gilt. Aus der Kontraktionseigenschaft ergibt sich dann direkt via

$$d(x_*,\, y_*) = d\big(f(x_*),\, f(y_*)\big) \le \kappa\, d(x_*,\, y_*) < d(x_*,\, y_*)$$

ein Widerspruch. Es kann also höchstens einen Fixpunkt von f in A geben.

Existenz und Konvergenz: Sei $(x_n)_{n\in\mathbb{N}}$ eine beliebige Folge aus A mit $x_{n+1} = f(x_n)$ für alle $n \in \mathbb{N}$. Dann gilt

$$d(x_{n+1},\, x_{n+2}) = d\big(f(x_n),\, f(x_{n+1})\big) \le \kappa\, d(x_{n+1},\, x_n)$$

für alle $n \in \mathbb{N}$ und durch vollständige Induktion zeigen wir leicht, dass damit auch

$$d(x_n,\, x_{n+1}) \le \kappa^n\, d(x_0,\, x_1)$$

für jedes $n \in \mathbb{N}$ erfüllt ist. Für beliebige Indizes n, k mit $n < k$ ergibt sich

$$\begin{aligned} d(x_n,\, x_k) &= d(x_n,\, x_{n+1}) + d(x_{n+1},\, x_{n+2}) + \ldots + d(x_{k-1},\, x_k) \\ &\le \big(\kappa^n + \kappa^{n+1} + \ldots + \kappa^{k-1}\big)\, d(x_0,\, x_1) \\ &= \frac{\kappa^n - \kappa^k}{1-\kappa}\, d(x_0,\, x_1) \\ &\le \kappa^n\, \frac{d(x_0,\, x_1)}{1-\kappa}\,. \end{aligned}$$

Die rechte Seite in dieser Abschätzungskette hängt nicht mehr von k ab und konvergiert für $n \to \infty$ im Sinne der reellen Zahlen gegen 0. Insbesondere ist damit gezeigt, dass $(x_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge bzgl. der Metrik d ist. Aufgrund der Vollständigkeit von $(X,\, d)$ existiert ein Grenzwert $x_\infty \in X$ und die Abgeschlossenheit von A garantiert, dass x_∞ in A liegt. Nach Grenzübergang in der Rekursionsvorschrift erhalten wir

$$x_\infty = f(x_\infty)\,.$$

Insbesondere ist $x_* := x_\infty$ ein Fixpunkt von f und nach dem ersten Beweisteil auch der einzige in A.

Approximationsfehler: Wir hatten bereits gezeigt, dass für $n < k$ die Ungleichung

$$d(x_n,\, x_k) \leq \frac{\kappa^n - \kappa^k}{1-\kappa}\, d(x_0,\, x_1)$$

gilt, und wenn wir k bei festem n gegen ∞ laufen lassen, erhalten wir die erste der gewünschten Fehlerabschätzungen. Die Dreiecksungleichung, die Rekursionsvorschrift sowie die Kontraktionseigenschaft implizieren außerdem

$$\begin{aligned} d(x_n,\, x_*) &\leq d(x_n,\, x_{n+1}) + d(x_{n+1},\, x_*) \\ &= d\big(f(x_{n-1}),\, f(x_n)\big) + d\big(f(x_n),\, f(x_*)\big) \\ &\leq \kappa\, d(x_{n-1},\, x_n) + \kappa\, d(x_n,\, x_*) \end{aligned}$$

und die zweite Fehlerbschätzung ergibt sich nach elementaren Umformungen. □

Bemerkungen

1. Man nennt f im Theorem oftmals auch Kontraktion von A. Beachte auch, dass κ die Lipschitz-Konstante von f auf A ist und dass die Forderung $\kappa < 1$ ganz wesentlich ist (siehe den Beweis und die Beispiele weiter unten). Die Stetigkeit von f außerhalb von A, d.h. auf der Menge $X \setminus A$, ist hingegen nicht wichtig und muss eigentlich auch nicht gefordert werden.

2. In der Literatur wird oftmals nur der Spezialfall $X = A$ formuliert und bewiesen. Dies ist zwar keine wesentliche Einschränkung, aber die Unterscheidung von X und A ist für viele Zwecke sehr nützlich. Beachte insbesondere, dass bei uns die Kontraktionsungleichung nur auf A, und nicht auf ganz X gelten muss.

3. Die Stärke des Theorems liegt unter anderem darin, dass nicht nur die Existenz und Eindeutigkeit eines Fixpunktes garantiert wird, sondern dass zusätzlich auch ein sehr robustes Approximationsverfahren mitgeliefert wird. Insbesondere ist es nicht notwendig, dass der Startwert x_0 hinreichend nah an x_* liegt.

4. Die erste Fehlerabschätzung ist ein Beispiel für eine a-priori-Schranke, denn wir können sie allein aus κ, x_0 und $x_1 = f(x_0)$ berechnen und müssen x_n überhaupt nicht kennen. Sie garantiert insbesondere, dass der Abstand von x_n zu x_* in der Metrik d exponentiell abklingt.[37]

5. Die zweite Fehlerabschätzung ist eine a-posteriori-Schranke, die wir erst dann auswerten können, wenn wir x_n durch n Interationsschritte berechnet haben (was für große n oder komplizierte f sehr mühsam bzw. aufwendig sein kann). Dafür wird sie aber in der Regel deutlich besser als die erste Fehlerschranke sein.[38]

6. Die Vollständigkeit von X und die Abgeschlossenheit von A sind im Beweis sehr wichtig, denn sie garantieren, dass die Cauchy-Folge $(x_n)_{n\in\mathbb{N}}$ einen Grenzwert in X besitzt und dass dieser in A liegt. Ohne Vollständigkeit von X bzw. ohne die Abgeschlossenheit von A ist nicht nur der Beweis nicht mehr richtig, sondern die Aussage des Theorems ist im Allgemeinen sogar falsch (siehe die Beispiele).

[37] Beachte, dass $\kappa^n = \exp(-\mu\, n)$ gilt, wobei $\mu = |\ln(\kappa)| > 0$ die *Abklingrate* ist.

[38] Die Vor- und Nachteile der verschiedenen Arten von Fehlerabschätzungen werden Sie in den Vorlesungen zur *Numerischen Mathematik* noch genauer diskutieren.

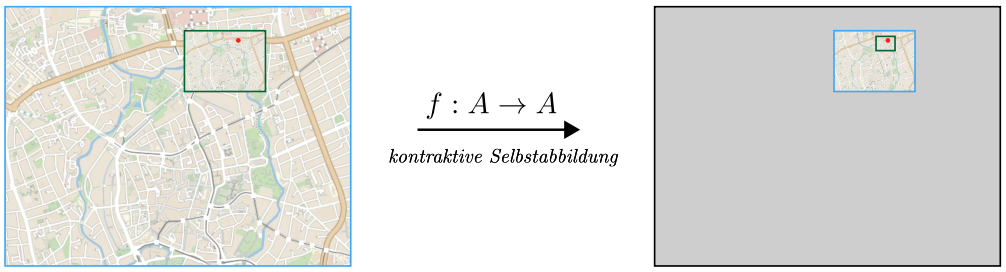

Abbildung Wenn Sie einen Stadtplan von Braunschweig innerhalb des dargestellten Kartenbereichs A betrachten, so wird es immer genau einen Ort geben, dessen Positionen auf der Karte und in der Realität zusammenfallen.

eindimensionaler Spezialfall Seien $X = \mathbb{R}$, A ein abgeschlossenes Intervall und $f : \mathbb{R} \to \mathbb{R}$ eine stetig differenzierbare Funktion, die A in sich abbildet. Dann gilt die Lipschitz-Abschätzung

$$\big|f(x) - f(y)\big| = \big|f'(\xi)\,(x-y)\big| \le \kappa\big|x-y\big| \qquad \text{mit} \qquad \kappa := \sup_{x\in A}\big|f'(x)\big|$$

für alle $x, y \in A$ nach dem Mittelwertsatz der Differentialrechnung.[39] Insbesondere besitzt im Fall $\kappa < 1$ die Funktion f im Intervall A genau einen Fixpunkt x_* mit $x_* = f(x_*)$.

Beispiele

1. Die reelle Kosinus-Funktion bildet offensichtlich das Intervall $A = [-1, +1]$ und sich ab, wobei $\kappa = \sup_{-1\le x\le+1}\big|\sin(x)\big| = \sin(1) < 1$ gilt. Also gibt es zwischen -1 und $+1$ genau eine Lösung x_* der Gleichung $x = \cos(x)$.[40]

   Achtung: Die reelle Kosinusfunktion bildet den metrischen Raum $\mathbb{R}$ in sich ab und ist dort Lipschitz-stetig mit Lipschitz-Konstante $\sup_{x\in\mathbb{R}}\big|\sin(x)\big| = 1$, aber eben nicht kontraktiv. Sie besitzt in $\mathbb{R}$ zwar genau einen Fixpunkt, aber diese Eindeutigkeit kann nicht mit dem Banachschen Fixpunktsatz begründet werden.

2. Wir betrachten die reelle Funktion $f(x) = x - \frac{1}{4}x^2 + \frac{1}{2}$ und zeigen mit kleinen Nebenrechnungen, dass diese das Intervall $A = [1, 2]$ in sich abbildet und außerdem eine Kontraktion ist. Der eindeutige Fixpunkt ist die irrationale Zahl $\sqrt{2}$.

   Achtung: Die Funktion f bildet auch die Menge $A \cap \mathbb{Q} = \{q \in \mathbb{Q} : 1 \le q \le 2\}$ kontraktiv in sich ab, aber enthält keinen Fixpunkt von f. Dies widerspricht nicht dem Banachschen Fixpunktsatz, denn $\mathbb{Q}$ ist zwar ein metrischer Raum (bzgl. des euklidischen Abstands), aber eben kein vollständiger.

3. Die Funktion $f(x) = 1 + \frac{1}{2}\sin(x)$ ist eine Kontraktion auf ganz $\mathbb{R}$ (die Lipschitz-Konstante ist $1/2$ wegen $f'(x) = \frac{1}{2}\cos(x)$) und besitzt dort nach dem Satz von Banach (ausgewertet mit $X = A = \mathbb{R}$) genau einen Fixpunkt $x_* \approx 1.4987$.

   Achtung: Für $X = \mathbb{R}$ und $A = [0, 1]$ kann das Theorem jedoch nicht verwendet werden, da dann f das Intervall A *nicht* in sich abbildet.

[39] Siehe *Analysis 1* und beachte, dass die Zwischenstelle ξ zwischen den Punkten x, y und damit auch im Intervall A liegt.

[40] Diese kann sehr einfach numerisch berechnet werden: Schalten Sie Ihren Taschenrechner ein und drücken Sie immer wieder auf die Kosinus-Taste. Das entspricht gerade der Banach-Iteration aus dem Theorem (mit Startwert $x_0 = 0$) und liefert die Approximation $x_* \approx 0.739085$ nach relativ wenigen Schritten.

4. Ein weiteres illustratives Gegenbeispiel ist die Funktion $f(x) = 2\sin x$. Sie besitzt in $\mathbb{R}$ genau drei Fixpunkte, nämlich $x = 0$ und $x \approx \pm 1.89549$, aber sie ist nicht kontraktiv. Der Banachsche Fixpunktsatz liefert daher keine hilfreichen Aussagen.

Merkregel Etwas vereinfacht kann man sagen: Die wesentlichen Voraussetzungen im Banachschen Fixpunktsatz sind *Vollständigkeit*, *Selbstabbildung* und *Kontraktivität*.

Anwendung: Integralgleichungen* Sei $J = [0,\, 1]$ das kompakte Einheitsintervall und sei $A = X = \mathsf{BC}(J)$, wobei wir diesen Funktionenraum mit der ∞-Norm ausstatten wollen, um die Vollständigkeit sicherzustellen. Außerdem sei $\beta : J \times J \to \mathbb{R}$ eine stetige Funktion mit

$$\kappa := \sup\Big\{ \big|\beta(t,\, s)\big| \,:\, t,\, s \in [0,\, 1] \Big\} < 1\,.$$

Dann existiert für jedes $\xi \in X$ genau ein $x_* \in X$, sodass

$$x_*(t) = \int_0^1 \beta(t,\, s)\, x_*(s)\, \mathrm{d}s + \xi(t)$$

für alle t mit $0 \le t \le 1$ gilt. In der Tat, die Funktion x_* ist der eindeutige Fixpunkt der Kontraktion $f : X \to X$ mit[41]

$$f\big(x\big)(t) = \xi(t) + \int_0^1 \beta(t,\, s)\, x(s)\, \mathrm{d}s\,,$$

wobei die oben eingeführte Zahl κ auch die Kontraktionskonstante von f ist. Für beliebige $x,\, y \in X$ gilt nämlich

$$\begin{aligned}
\Big|f\big(x\big)(t) - f\big(y\big)(t)\Big| &= \left| \int_0^1 \beta(t,\, s)\, \big(x(s) - y(s)\big)\, \mathrm{d}s \right| \\
&\le \int_0^1 \big|\beta(t,\, s)\big|\, \big|x(s) - y(s)\big|\, \mathrm{d}s \\
&\le \int_0^1 \kappa\, \big\|x - y\big\|_\infty\, \mathrm{d}s = \kappa\, \big\|x - y\big\|_\infty
\end{aligned}$$

und nach Supremumsbildung über $t \in [0,\, 1]$ ergibt sich $\big\|f(x) - f(y)\big\|_\infty \le \kappa\, \|x - y\|_\infty$.

Bemerkung: Die obige Formel ist eine spezielle *Integralgleichung* und x_* wird *Lösung* genannt. Integralgleichungen — sowie die eng verwandten Differentialgleichungen — spielen eine herausragende Rolle in der mathematischen Physik und es ist sehr wichtig zu verstehen, unter welchen Voraussetzungen Lösungen existieren, wann diese eindeutig oder mehrdeutig sind und wie sie approximativ berechnet werden können. In vielen Fällen (aber leider nicht in allen) liefert der Banachsche Fixpunktsatz die entsprechenden Antworten.

[41] Beachte, dass x und $f(x)$ hier zum Funktionenraum $\mathsf{BC}(J)$ gehören und damit jeweils eine skalare Funktion in der Variablen $t \in J$ darstellen. Insbesondere ist $f\big(x\big)(t)$ der Wert, den die skalare Funktion $f(x)$ im Punkt t annimmt.

Kapitel 2

Differentialrechnung

Vorlesungswoche 04

Vorbemerkung In diesem Kapitel entwickeln wir die Differentialrechnung für Funktionen, die eine Teilmenge des $\mathbb{R}^n$ in den $\mathbb{R}^m$ abbilden. Dabei werden wir standardmäßig den $\mathbb{R}^m$ (und analog den $\mathbb{R}^n$) immer mit dem euklidischen Betrag (bzw. der 2-Norm) ausstatten und schreiben

$$x = (x_1, \ldots, x_m) \qquad \text{bzw.} \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

sowie

$$|x| = \sqrt{x_1^2 + \ldots + x_m^2}\,, \qquad x \cdot y = x_1\, y_1 + \ldots + x_m\, y_m\,,$$

wobei $\cdot$ das euklidische Skalarprodukt ist. Insbesondere gilt immer die Formel $|x|^2 = x \cdot x$ sowie die Cauchy-Schwarz-Ungleichung $|x \cdot y| \leq |x|\ |y|$.

2.1 Kurven

Ziel In diesem Abschnitt behandeln wir die Grundlagen der mathematischen Theorie von *Kurven* (oder *Wegen*) und stellen wichtige Resultate bereit, die in der Analysis eine wichtige Rolle spielen werden. In der Differentialrechnung entspricht dies dem Sonderfall $n = 1$, das heißt der Definitionsbereich der betrachteten Abbildungen — die wir in diesem Abschnitt mit γ bezeichnen — ist eine Teilmenge des $\mathbb{R}^1$, der Bildbereich aber eine Teilmenge des $\mathbb{R}^m$.

Definition Eine parametrisierte Kurve (im Raum $\mathbb{R}^m$) ist eine stetige Abbildung $\gamma : I \to \mathbb{R}^m$, die auf einem Intervall $I \subseteq \mathbb{R}$ definiert ist.[1]

[1] Der Fall $m = 1$ ist in den folgenden Betrachtungen zugelassen, aber bei der ersten Lektüre sollten Sie sich immer vorstellen, dass $m = 2$ oder $m = 3$ gilt. Der Fall $m \geq 4$ ist auch sehr wichtig, aber dann wird unsere Anschauung in aller Regel versagen.

Bemerkungen

1. Die Punktmenge

$$\operatorname{im}(\gamma) := \big\{\gamma(t) \,:\, t \in I\big\} \subset \mathbb{R}^m$$

wird auch Bild (oder '*Image*') der parametrisierten Kurve genannt und ist das geometrische Objekt, das man landläufig *Kurve* nennt. Wir werden diese Menge oftmals mit Γ bezeichnen und sagen dann, die Abbildung γ parametrisiert die Menge Γ.

2. Die Koordinate im Urbildbereich bezeichnen wir meist mit t, denn sie kann oftmals in natürlicher Weise als Zeit interpretiert werden. In diesem Sinne ist $\gamma(t)$ die momentane Position eines gedachten Teilchens, das sich entlang von Γ bewegt.

3. In der Mathematik lässt man häufig das Attribut „parametrisiert" weg und nennt sowohl die Abbildung γ als auch die Punktmenge $\Gamma = \operatorname{im}(\gamma)$ schlicht *Kurve*. Das kann gerade am Anfang einige Verwirrung stiften, aber in aller Regel wird durch den Kontext klar, ob wir Kurve in dem einen oder dem anderen Sinne meinen.[2]

4. Eine parametrisierte Kurve im $\mathbb{R}^m$ können wir auch immer *komponentenweise* betrachten. Insbesondere existieren Funktionen $\gamma_1, \ \ldots, \gamma_m : I \to \mathbb{R}$, sodass

$$\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \vdots \\ \gamma_m(t) \end{pmatrix}$$

für alle $t \in I$ gilt.

Bemerkung: Wir haben hier $\gamma(t)$ als Spaltenvektor geschrieben, werden manchmal aber auch die Tupel-Notation verwenden.

5. Wir haben zunächst keine Annahmen an das Intervall I (bzw. den *Zeitbereich*) gemacht. Es kann endlich oder unendlich, offen oder abgeschlossen sein.

6. Unsere Intuition besagt, dass eine Kurve eine „eindimensionale Menge " ist[3] und für hinreichend gute Funktionen γ wird das auch so sein. Es gibt aber auch seltsame Kurven, wie zum Beispiel fraktale Kurven (etwa den Rand der Kochschen Schneeflocke, siehe oben) oder flächen- bzw. raumfüllende Kurven (siehe das Bild weiter unten).

7. Ist $I = [a, b]$ ein abgeschlossenes Intervall, so werden $\gamma(a)$ bzw. $\gamma(b)$ der Anfangs- bzw. der Endpunkt von γ genannt. Gilt darüber hinaus $\gamma(a) = \gamma(b)$, so ist γ eine geschlossene Kurve.

8. Wenn zwei verschiedene Zeiten $t_1, t_2 \in I$ mit

$$\gamma(t_1) = \gamma(t_2) \neq \gamma(t)$$

für alle $t \in I \setminus \{t_1, t_2\}$ existieren, so sprechen wir von einem Doppelpunkt, es sei denn es handelt sich um den Anfangs- und Endpunkt einer geschlossenen Kurve. Analog werden Dreifach- und Vierfachpunkte definiert.

[2] Eine ähnliche Zweideutigkeit werden wir beim Studium der Flächen antreffen.

[3] Wir werden erst später definieren, was die Dimension einer sogenannten *Mannigfaltigkeit* eigentlich ist. Im Moment appellieren wir an Ihre geometrische Anschauung.

9. Anstelle von $\mathbb{R}^m$ können wir im Prinzip auch andere Bildmengen betrachten und zum Beispiel parametrisierte Kurven in Funktionenräumen studieren. Diese werden aber in dieser Vorlesung noch keine Rolle spielen.

Beispiele

1. Die geschlossene Kurve $\gamma : [0,\, 2\,\pi] \to \mathbb{R}^2$ mit

$$\gamma(t) = \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}$$

besitzt die Bildmenge

$$\Gamma = \big\{(x_1,\, x_2) \,:\, x_1^2 + x_2^2 = 1\big\} = S_1(0) \subset \mathbb{R}^2$$

und wird die *Standardparametrisierung* der Einheitskreislinie genannt.

2. Die Abbildung

$$\gamma(t) = \begin{pmatrix} \cos(\omega\, t) \\ \sin(\omega\, t) \\ \eta\, t \end{pmatrix}, \qquad t \in I = \mathbb{R}$$

parametrisiert eine dreidimensionale und unendlich ausgedehnte Schraubenlinie, wobei die Parameter ω bzw. η die *Winkelgeschwindigkeit* bzw. die *Ganghöhe* festlegen.

3. Die ebene Gerade, die durch zwei gegebene Punkte μ, $\eta \in \mathbb{R}^2$ läuft, kann durch

$$\gamma(t) = (1-t)\,\mu + t\,\eta = (1-t)\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + t\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$$

mit $t \in I = \mathbb{R}$ parametrisiert werden. Mit $I = [0,\, 1]$ beschreibt die Formel gerade die Verbindungsstrecke zwischen μ und η.

4. Die parametrisierte Kurve

$$\gamma(t) = \begin{pmatrix} 2\cos(t) - \cos(2\,t) \\ \sin(2\,t) \end{pmatrix}, \quad t \in [0,\, 2\,\pi]$$

ist unter dem Namen Torpedo-Kurve bekannt.

5. Die geschlossene Kurve

$$\gamma(t) = \big(1 + \cos(t)\big)\begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}, \qquad t \in [0,\, 2\pi]$$

wird Kardioide genannt.

6. Die Formel

$$\gamma(t) = \sin(2\,t)\begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}, \qquad t \in [0,\, 2\pi]$$

beschreibt ein vierblättriges Kleeblatt (bzw. das Quadrifolium).

7. Die nicht-geschlossene Kurve

$$\gamma(t) = \begin{pmatrix} t - \tanh(t) \\ \dfrac{10}{\cosh(t)} \end{pmatrix}, \qquad t \in \mathbb{R}$$

ist eine Traktrix bzw. Schleppkurve.

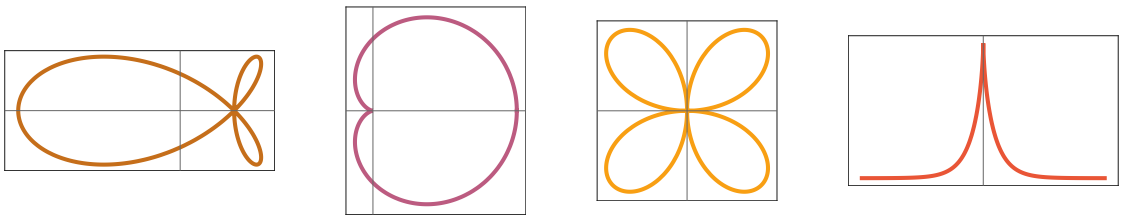

Abbildung Die Torpedo-Kurve (ein Dreifachpunkt), die Kardioide (ohne Mehrfachpunkte) sowie das Kleeblatt (ein Vierfachpunkt) sind spezielle geschlossene Kurven in der Ebene. Die Traktrix ist jedoch nicht geschlossen. Beachte, dass immer $\Gamma = \operatorname{im}(\gamma)$, also das Bild der parametrisierten Kurve dargestellt ist.

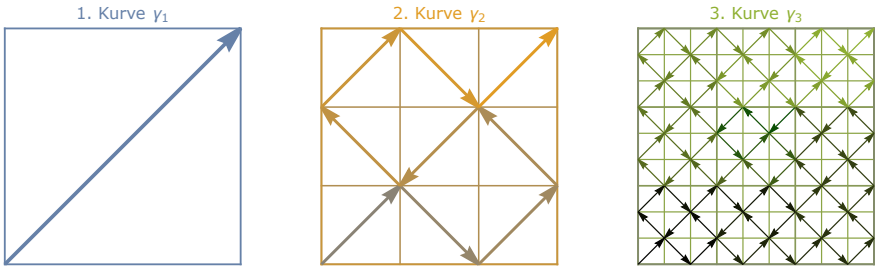

Abbildung Beispiel für die rekursive, selbstähnliche Konstruktion einer *flächenfüllenden* Kurve als gleichmäßiger Grenzwert einer Folge stückweise affiner Kurven $\gamma_n : [0, 1] \to \mathbb{R}^2$. Die stetige Limes-Abbildung $\gamma_\infty : [0, 1] \to \mathbb{R}^2$ ist eine Kurve im Sinne der obigen Definition, aber ihr Bild füllt das ganze Quadrat aus und ist damit „zweidimensional“. Diese verblüffende Eigenschaft hat damit zu tun, dass die Kurve γ_∞ weder differenzierbar noch stückweise differenzierbar ist (siehe dazu weiter unten). Das Bild einer stückweise differenzierbaren Kurve ist nämlich immer „eindimensional“.

Implizite Darstellung von Kurven Die Lösungsmenge von $m - 1$ skalaren Gleichungen für die Variablen $x_1, \ldots, x_m$ (bzw. für den Vektor $x \in \mathbb{R}^m$) kann oftmals als die Bildmenge Γ einer parametrisierten Kurve γ angesehen werden. Zum Beispiel beschreibt die Gleichung

$$x_1^2 + x_2^2 = \varrho^2 \qquad \text{bzw.} \qquad \left(x_1^2 + x_2^2\right)^2 - 2\,x_1\left(x_1^2 + x_2^2\right) - x_2^2 = 0$$

eine Kreislinie vom Radius ϱ bzw. die Kardioide und alle Lösungen $x \in \mathbb{R}^3$ des Gleichungssystems

$$x_1^2 + 2\,x_2^2 + (x_1 - x_3)^2 = 1\,, \qquad x_1 + x_2 = x_3$$

formen eine Ellipse, die schief im $\mathbb{R}^3$ liegt. Es ist im Allgemeinen aber nicht möglich, aus den Gleichungen direkt eine entsprechende Parametrisierung abzulesen oder umgekehrt. Wir werden auf dieses Problem später noch einmal zurückkommen.

Definition Eine parametrisierte Kurve $\tilde{\gamma} : \tilde{I} \to \mathbb{R}^m$ wird Reparametrisierung von $\gamma : I \to \mathbb{R}^m$ genannt, wenn es eine strikt monotone und bijektive Abbildung $h : \tilde{I} \to I$ (der sogenannte Parameterwechsel) gibt, sodass

$$\tilde{\gamma}(\tilde{t}) = \gamma\big(h(\tilde{t})\big)$$

für alle $\tilde{t} \in \tilde{I}$ gilt.

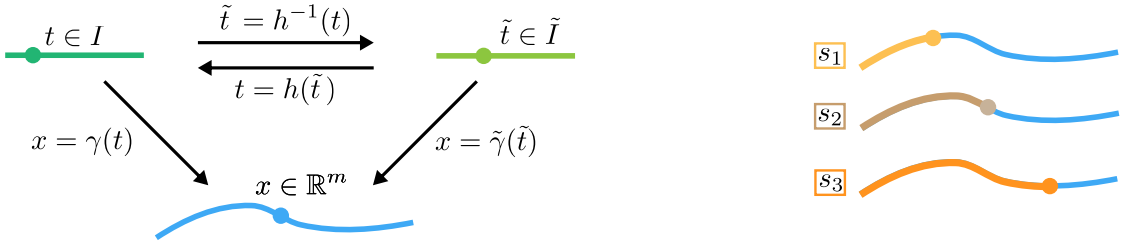

Abbildung *Links*: Illustration der Konzepte *Parametrisierung* und *Reparametrisierung* von Kurven. *Rechts*: Zum Bogenlängenparameter (siehe weiter unten), der oftmals mit s und nicht mit $\tilde{t}$ bezeichnet wird: Jeder Kurvenpunkt kann eindeutig durch eine entsprechende Länge charakterisiert werden.

Bemerkungen

1. Insbesondere gilt

   $$\tilde{\Gamma} = \operatorname{im}(\tilde{\gamma}) = \operatorname{im}(\gamma) = \Gamma\,,$$

   d.h. die Abbildungen $\tilde{\gamma}$ bzw. γ beschreiben dasselbe geometrische Objekt $\tilde{\Gamma} = \Gamma$, aber mittels verschiedener Parameter, nämlich mit $\tilde{t}$ bzw. $t = h(\tilde{t})$.

2. Eine *geometrische* Eigenschaft von $\Gamma = \tilde{\Gamma}$ (zum Beispiel die Länge) wird nicht davon abhängen, welche Parametrisierung gewählt wird, d.h. ob die Rechnungen mit γ oder mit $\tilde{\gamma}$ durchgeführt werden. Die Kunst besteht oftmals darin, eine gute Parametrisierung zu finden, mit der die Rechnungen möglichst einfach werden.

3. Eine alternative Schreibweise für die reparametrisierte Kurve ist $\tilde{\gamma} = \gamma \circ h$.

4. Der Parameterwechsel h ist immer invertierbar, d.h. es existiert die Umkehrabbildung $\tilde{h} := h^{-1} : I \to \tilde{I}$. Insbesondere gilt: Wenn $\tilde{\gamma}$ eine Reparametrisierung von γ ist, so ist auch γ eine Reparametrisierung von $\tilde{\gamma}$.

5. Wenn h wachsend bzw. fallend ist, so sagen wir, die Reparametrisierung erhält oder ändert den Durchlaufsinn.

6. Eine besonders wichtiger Parameterwechsel betrifft die Bogenlänge, die wir weiter untern studieren werden.

Beispiel Die Abbildung $\tilde{\gamma} : [0,\, 2\,\pi/\omega] \to \mathbb{R}^2$ mit

$$\tilde{\gamma}(\tilde{t}) = \begin{pmatrix} \cos(\omega\,\tilde{t}) \\ \sin(\omega\,\tilde{t}) \end{pmatrix}$$

ist eine Reparametrisierung der oben angegebenen Standardparametrisierung der Einheitskreislinie Γ, die nun mit konstanter *Winkelgeschwindigkeit* $\omega > 0$ durchlaufen wird. Der entsprechende Parameterwechsel ist durch $t = h(\tilde{t}) = \omega\,\tilde{t}$ gegeben. Die Formel

$$\tilde{\gamma}(\tilde{t}) = \begin{pmatrix} \cos\big(\tilde{t} + \mu\sin(\tilde{t})\big) \\ \sin\big(\tilde{t} + \mu\sin(\tilde{t})\big) \end{pmatrix}, \qquad \tilde{t} \in [0,\, 2\,\pi]$$

mit Konstante $-1 < \mu < +1$ liefert auch eine Parametrisierung der Kreislinie Γ bzw. eine Reparametrisierung von γ, wobei in diesem Fall die Winkelgeschwindigkeit nicht mehr konstant ist, sondern selbst variiert.

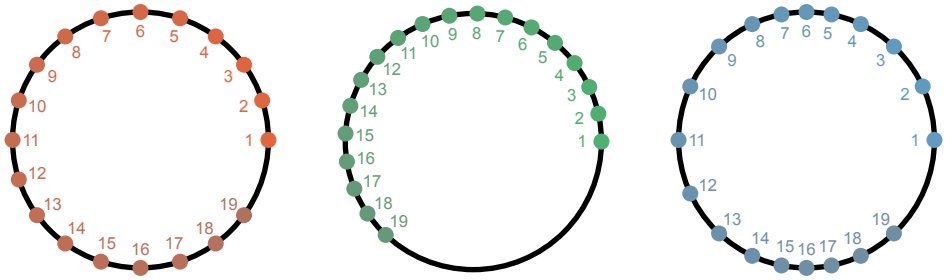

Abbildung Drei verschiedene Parametrisierungen einer Kreislinie im $\mathbb{R}^2$, die alle unterschiedlichen Durchlaufgeschwindigkeiten entsprechen. Die Punkte markieren jeweils den Kurvenpunkt $\gamma(t_k)$ zu äquidistant gewählten Zeiten $t_k = k\,\Delta t$.

Merkregel Verschiedene parametrisierte Kurven können zu demselben geometrischen Objekt gehören.

Ableitungen von Kurven

Definition Eine parametrisierte Kurve heißt differenzierbar, wenn die Ableitung

$$\frac{\mathrm{d}\gamma}{\mathrm{d}t}(t) := \dot\gamma(t) := \lim_{t\to t_*} \frac{\gamma(t) - \gamma(t_*)}{t - t_*}$$

in jedem $t_* \in I$ wohldefiniert ist.

Bemerkungen

1. Bei Kurven benutzen wir in dieser Vorlesung in aller Regel die Punkt- statt die Strichnotation, d.h. wir schreiben $\dot\gamma(t_*)$ statt $\gamma'(t_*)$.[4]

2. Die Definition von Differenzierbarkeit ist ganz analog zu *Analysis 1*, aber diesmal ist $\dot\gamma(t)$ ein Element des $\mathbb{R}^m$, wobei

$$\dot\gamma(t_*) := \begin{pmatrix} \dot\gamma_1(t_*) \\ \vdots \\ \dot\gamma_m(t_*) \end{pmatrix}$$

   für alle $t_* \in I$ gilt. Der Vektor $\dot\gamma(t)$ wird in der Geometrie als Tangentialvektor und in der Physik als Geschwindigkeitsvektor bezeichnet.

3. In der Formel für $\dot\gamma(t)$ wird immer stillschweigend $t_* \in I$, $t \in I$ sowie $t \neq t_*$ vorausgesetzt.[5] Ist t_* ein Randpunkt von I, so ist $\dot\gamma(t_*)$ im Sinne einer einseitigen Ableitung zu verstehen.

[4] Sie dürfen in den Hausaufgaben aber gerne die Strichnotation verwenden.

[5] Beachte auch, dass der Differenzenquotient wohldefiniert ist, da im Zähler zwar ein Vektor, im Nenner aber immer eine reelle Zahl steht. Der Quotient aus zwei Vektoren ist nicht definiert.

4. Ganz analog zu der obigen Definition werden die Konzepte stetig differenzierbar und k-mal (stetig) differenzierbar eingeführt. Bei einer zweimal differenzierbaren Funktion existiert in jedem $t_* \in I$ der Vektor

$$\ddot{\gamma}(t_*) = \begin{pmatrix} \ddot{\gamma}_1(t_*) \\ \vdots \\ \ddot{\gamma}_m(t_*) \end{pmatrix} = \frac{\mathrm{d}\dot{\gamma}}{\mathrm{d}t}(t) = \lim_{t \to t_*} \frac{\dot{\gamma}(t) - \dot{\gamma}(t_*)}{t - t_*},$$

der in der Physik Beschleunigungsvektor genannt wird.

5. Eine differenzierbare Kurve wird regulär genannt, wenn $\dot{\gamma}(t) \neq 0$ für alle $t \in I$ gilt. Geometrisch bedeutet dies, dass die Durchlaufgeschwindigkeit eines gedachten Punktes niemals verschwindet.

6. Viele praktisch relevante Kurven sind nur stückweise stetig differenzierbar, das heißt es existieren endliche viele Zeiten $t_k \in I$, sodass γ in t_k nicht differenzierbar ist. In jedem dieser Punkte müssen dann aber die einseitigen Ableitungen

$$\dot{\gamma}(t_k + 0) := \lim_{t \searrow t_k} \frac{\gamma(t) - \gamma(t_k)}{t - t_k}, \qquad \dot{\gamma}(t_k - 0) := \lim_{t \nearrow t_k} \frac{\gamma(t) - \gamma(t_k)}{t - t_k}$$

wohldefiniert sein. Geometrisch entspricht $\gamma(t_k)$ einem *Knickpunkt*. Fast alle der im Folgenden abgeleiteten Resultate gelten sinngemäß auch für Kurven, die nur stückweise stetig differenzierbar sind.

Schnittwinkel Sind $\gamma : [a, b] \to \mathbb{R}^m$ und $\hat{\gamma} : [\hat{a}, \hat{b}] \to \mathbb{R}^m$ zwei differenzierbare und reguläre Kurven, die sich im Punkt $\gamma(t_*) = \hat{\gamma}(\hat{t}_*)$ schneiden, so kann der entsprechende Schnittwinkel durch

$$\theta = \arccos\left(\frac{\dot{\gamma}(t_*) \cdot \dot{\hat{\gamma}}(\hat{t}_*)}{|\dot{\gamma}(t_*)|\,|\dot{\hat{\gamma}}(\hat{t}_*)|} \right)$$

berechnet werden, wobei θ immer Werte in $[0, \pi]$ annimmt.[6]

Merkregel: Winkel werden in der höheren Mathematik immer mit Tangentialvektoren berechnet.

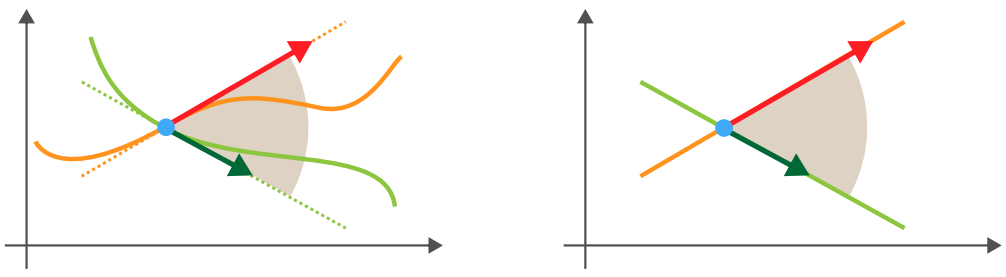

Abbildung Der Schnittwinkel (braun) zwischen zwei sich schneidenden Kurven (orange und grün) wird immer durch die Tangentialvektoren (rot und dunkelgrün) im Schnittpunkt (blau) festgelegt. *Links*: Der allgemeine Fall mit zwei gekrümmten Kurven. *Rechts*: Der aus der Schule bekannte Spezialfall mit zwei Geraden.

[6] In zwei Raumdimensionen ($m = 2$) werden wir später das etwas allgemeinere Konzept eines *orientierten Winkel* einführen, bei dem auch negative Winkel zugelassen sind.

Ausblick: Geometrie planarer Kurven* Für eine zweimal stetig differenzierbare Kurve $\gamma : I \to \mathbb{R}^2$ kann ihre Krümmung durch

$$\kappa(t) := \frac{\dot\gamma_1(t)\,\ddot\gamma_2(t) - \ddot\gamma_1(t)\,\dot\gamma_2(t)}{|\dot\gamma(t)|^3}$$

berechnet werden, wobei $|\dot\gamma(t)| = \sqrt{\big(\dot\gamma_1(t)\big)^2 + \big(\dot\gamma_2(t)\big)^2}$ gilt. Darüber hinaus beschreibt

$$\beta_1(t) = \frac{1}{|\dot\gamma(t)|}\begin{pmatrix}+\dot\gamma_1(t)\\ +\dot\gamma_2(t)\end{pmatrix} \qquad \text{bzw.} \qquad \beta_2(t) = \frac{1}{|\dot\gamma(t)|}\begin{pmatrix}-\dot\gamma_2(t)\\ +\dot\gamma_1(t)\end{pmatrix}$$

den normalisierten Tangentialvektor bzw. den normalisierten Normalenvektor und beide zusammen bilden das Frenetsche Zweibein.

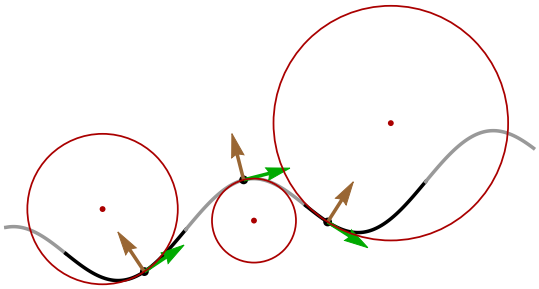

Abbildung Beispiel für eine stetig differenzierbare Kurve, wobei die Bereiche mit negativer bzw. positiver Krümmung in hell- bzw. dunkelgrau gezeichnet wurden. Die Vektoren repräsentieren das mitbewegte Frenetsche Zweibein in drei ausgewählten Kurvenpunkten, für die zusätzlich auch der Krümmungskreis (rot) dargestellt ist. Der Betrag der Krümmung ist dabei gerade der Kehrwert des Radius.

Bemerkung Die Geometrie von Kurven mit $m = 3$ (sogenannte *Raumkurven*) oder $m \geq 4$ ist komplizierter. Insbesondere gibt es immer $m-1$ skalare Krümmungen sowie ein Frenetsches m-Bein.

Alternative Notationen In der Physik und der Geometrie bezeichnet man die Abbildung γ oftmals nicht explizit, sondern benutzt dieselben Buchstaben wie für Vektoren in $\mathbb{R}^m$. Im Fall $m = 3$ schreibt man also

$$x(t) = \begin{pmatrix}x_1(t)\\ x_2(t)\\ x_3(t)\end{pmatrix} \qquad \text{und} \qquad \dot x(t) = \begin{pmatrix}\dot x_1(t)\\ \dot x_2(t)\\ \dot x_3(t)\end{pmatrix}$$

um deutlich zu machen, dass die räumlichen Koordinaten x_j sich entlang der Kurve mit dem Parameter t ändern. Diese Notation ist sehr intuitiv, benutzt aber die x_j in zweifacher Bedeutung, nämlich einmal als Koordinate in $\mathbb{R}^m$ und einmal als Komponenten einer Funktion $I \to \mathbb{R}^m$). Eine Mischform der Notation ist $x_j = \gamma_j(t)$. Wir können und werden mit beiden Notationen arbeiten, wobei je nach Kontext und Vorliebe mal die eine und mal die andere „besser geeignet“ ist.

Länge einer Kurve

Vorbemerkung Im Folgenden setzen wir voraus, dass $I = [a, b]$ ein abgeschlossenes Intervall ist und dass die parametrisierte Kurve $\gamma : I \to \mathbb{R}^m$ stetig differenzierbar ist. Alle Konzepte können aber analog für stückweise differenzierbare Kurven eingeführt werden.

Definition Für eine stetig differenzierbare Kurve $\gamma : [a, b] \to \mathbb{R}^m$ wird

$$\operatorname{len}(\gamma) := \int_a^b \left|\dot\gamma(t)\right| \mathrm{d}t = \int_a^b \left|\frac{\mathrm{d}\gamma}{\mathrm{d}t}(t)\right| \mathrm{d}t$$

die Länge von γ genannt.

Bemerkungen

1. Es gilt $\left|\dot\gamma(t)\right| = \sqrt{\left(\dot\gamma_1(t)\right)^2 + \ldots + \left(\dot\gamma_m(t)\right)^2}$.
2. Die Formel werden wir weiter unten geometrisch motivieren.
3. Für viele praktisch relevante Kurven (zum Beispiel Ellipsen) können wir das Integral der Länge nicht exakt, sondern nur approximativ berechnen.

Beispiele

1. Ist γ die oben angegeben Parametrisierung der Verbindungstrecke zwischen zwei Punkten $\mu, \eta \in \mathbb{R}^m$, so gilt $\dot\gamma(t) = \eta - \mu$ für alle t und wir erhalten mit

   $$\operatorname{len}(\gamma) = \int_0^1 |\eta - \mu| \, \mathrm{d}t = |\eta - \mu| \int_0^1 \mathrm{d}t = |\eta - \mu|$$

   das erwartetet Resultat.

2. Der Rand der zweidimensionalen Kreisscheibe $\overline{B}_\varrho(0)$ mit Radius ϱ wird durch

   $$\gamma(t) = \begin{pmatrix} \varrho \cos(t) \\ \varrho \sin(t) \end{pmatrix}, \qquad t \in I = [0, 2\pi]$$

   parametrisiert, wobei dann

   $$\dot\gamma(t) = \begin{pmatrix} -\varrho \sin(t) \\ +\varrho \cos(t) \end{pmatrix}, \qquad \left|\dot\gamma(t)\right| = \varrho$$

   gilt und wir mit

   $$\operatorname{len}(\gamma) = \int_0^{2\pi} \varrho \, \mathrm{d}t = 2\pi\varrho$$

   die bekannte Formel für den Kreisumfang wiederentdecken. Beachte, dass für $I = [0, \pi]$ das Längenintegral den Wert $\varrho\pi$ als Länge des Halbkreises liefert. Mit der Wahl $I = [0, 4\pi]$ erhalten wir jedoch $\operatorname{len}(\gamma) = 4\pi\varrho$, da nun der Kreis zweimal durchlaufen wird.

   Achtung: Bei der Berechnung geometrischer Längen muss die Parametrisierung γ so gewählt werden, dass die Punkte in Γ nur einmal durchlaufen werden, wobei endlich viele Ausnahmen (zum Beispiel in Anfangs- und Endpunkten) zulässig bzw. unproblematisch sind.

3. Die Formel

$$\gamma(t) = \begin{pmatrix} \varrho_1 \cos(t) \\ \varrho_2 \sin(t) \end{pmatrix}, \qquad t \in I = [0,\, 2\,\pi]$$

parametrisiert eine achsenparallele Ellipse.[7] Deren Länge ist durch das *elliptische* Integral

$$\operatorname{len}(\gamma) = \int_0^{2\,\pi} \sqrt{\varrho_1^2 \cos^2(t) + \varrho_2^2 \sin^2(t)}\, \mathrm{d}t$$

gegeben, für das es aber keine explizite Formel gibt (es sei denn, es gilt $\varrho_1 = \varrho_2$).

4. Für die Kardioide erhalten wir

$$\dot\gamma(t) = -\sin(t) \begin{pmatrix} +\cos(t) \\ +\sin(t) \end{pmatrix} + \big(1 + \cos(t)\big) \begin{pmatrix} -\sin(t) \\ +\cos(t) \end{pmatrix},$$

wobei auf der rechten Seite zwei zueinander senkrechte Vektoren stehen. Aus direkten Rechnungen sowie dem Additionstheorem $\cos(t) + 1 = 2\,\cos^2\big(\frac{1}{2}\,t\big)$ folgt

$$\big|\dot\gamma(t)\big| = \sqrt{\sin^2(t) + \big(1 + \cos(t)\big)^2} = \sqrt{2\,(1 + \cos(t))} = 2\,\big|\cos\big(\tfrac{1}{2}\,t\big)\big|,$$

und wir berechnen

$$\begin{aligned} \operatorname{len}(\gamma) &= 2 \int_0^{2\,\pi} \big|\cos\big(\tfrac{1}{2}\,t\big)\big|\, \mathrm{d}t = 2 \int_0^{\pi} \cos\big(\tfrac{1}{2}\,t\big)\, \mathrm{d}t - 2 \int_\pi^{2\,\pi} \cos\big(\tfrac{1}{2}\,t\big)\, \mathrm{d}t \\ &= 4\big[\sin\big(\tfrac{1}{2}\,t\big)\big]_{t=0}^{t=\pi} - 4\big[\sin\big(\tfrac{1}{2}\,t\big)\big]_{t=\pi}^{t=2\,\pi} = 8\,\sin\big(\tfrac{1}{2}\,\pi\big) = 8 \end{aligned}$$

mithilfe des Hauptsatzes der Differential- und Integralrechnung.

Lemma (**Länge als geometrische Eigenschaft**) Ist $h : [\tilde a,\, \tilde b] \to [a,\, b]$ ein stetig differenzierbarer Parameterwechsel, so gilt $\operatorname{len}(\gamma) = \operatorname{len}(\tilde\gamma)$ für $\tilde\gamma = \gamma \circ h$.

Beweis Wir betrachten zunächst den Fall, dass h monoton wachsend ist, d.h. dass $h'\big(\tilde t\big) \geq 0$ für alle $\tilde t \in [\tilde a,\, \tilde b]$ gilt. Die Kettenregel impliziert

$$\frac{\mathrm{d}\tilde\gamma_j}{\mathrm{d}\tilde t}(\tilde t) = \frac{\mathrm{d}\gamma_j}{\mathrm{d}t}\big(h(\tilde t)\big)\, \frac{\mathrm{d}h}{\mathrm{d}\tilde t}(\tilde t) \qquad \text{und damit} \qquad \left|\frac{\mathrm{d}\tilde\gamma}{\mathrm{d}\tilde t}(\tilde t)\right| = \left|\frac{\mathrm{d}\gamma}{\mathrm{d}t}\big(h(\tilde t)\big)\right| \frac{\mathrm{d}h}{\mathrm{d}\tilde t}(\tilde t)$$

und mit der Substitutionsregel für Integrale[8] erhalten wir

$$\operatorname{len}(\tilde\gamma) = \int_{\tilde a}^{\tilde b} \left|\frac{\mathrm{d}\tilde\gamma}{\mathrm{d}\tilde t}(\tilde t)\right| \mathrm{d}\tilde t = \int_a^b \left|\frac{\mathrm{d}\gamma_j}{\mathrm{d}t}(t)\right| \mathrm{d}t = \operatorname{len}(\gamma)\,.$$

Der Beweis für monoton fallende Reparametrisierungen erfordert einige Vorzeichenwechsel, ist aber sonst ganz analog. □

[7] Die Ellipse Γ enthält die Punkte $(\pm\varrho_1,\, 0)$ und $(0,\, \pm\varrho_2)$, wobei dies den Zeiten $t \in \{0,\, \pi,\, \frac{1}{2}\,\pi,\, \frac{3}{2}\,\pi\}$ entspricht. Außerdem gilt stets

$$\frac{x_1^2}{\varrho_1^2} + \frac{x_2^2}{\varrho_2^2} = 1$$

mit $x_1 = \gamma_1(t)$ und $x_2 = \gamma_1(t)$.

[8] Beachte, dass symbolisch $\mathrm{d}t = \frac{\mathrm{d}h}{\mathrm{d}\tilde t}(\tilde t)\, \mathrm{d}\tilde t$ gilt.

Bemerkung Das Lemma zeigt, dass die Länge invariant unter Reparametrisierung und damit eine *geometrische Größe* ist, d.h. eine Eigenschaft der Bildmenge $\Gamma = \operatorname{im}(\gamma)$. Wir brauchen aber eine Parametrisierung γ, um die Länge des geometrischen Objektes Γ überhaupt ausrechnen zu können.

Lemma (**untere Schranke für die Länge**) Es gilt

$$\operatorname{len}(\gamma) \geq \big|\gamma(b) - \gamma(a)\big|\,,$$

das heißt die Länge einer Kurve ist niemals kleiner als der euklidische Abstand ihrer Endpunkte.[9]

Beweis Wir definieren $d \in \mathbb{R}^m$ komponentenweise durch

$$d_j := \int_a^b \dot\gamma_j(t)\,\mathrm{d}t = \gamma_j(b) - \gamma_j(a)\,,$$

wobei das zweite Gleichheitszeichen den Fundamentalsatz der Analysis widerspiegelt. Andererseits gilt

$$\big|d\big|^2 = d \cdot d = \sum_{j=1}^m d_j\, d_j = \sum_{j=1}^m d_j \int_a^b \dot\gamma_j(t)\,\mathrm{d}t = \int_a^b \left(\sum_{j=1}^m d_j\, \dot\gamma_j(t)\right) \mathrm{d}t = \int_a^b d \cdot \dot\gamma(t)\,\mathrm{d}t$$

und die Cauchy-Schwarz-Ungleichung liefert $d \cdot \dot\gamma(t) \leq \big|d\big|\,\big|\dot\gamma(t)\big|$. Wir erhalten

$$\big|d\big|^2 \leq \int_a^b \big|d\big|\,\big|\dot\gamma(t)\big|\,\mathrm{d}t = \big|d\big| \int_a^b \big|\dot\gamma(t)\big|\,\mathrm{d}t = \big|d\big|\operatorname{len}(\gamma)$$

und damit die Behauptung. □

Parametrisierung durch Bogenlänge

Bogenlänge Ist $\gamma : [a,\, b] \to \mathbb{R}^m$ eine stetig differenzierbare und reguläre Kurve, so wird durch

$$l(t) := \int_a^t \big|\dot\gamma(\tau)\big|\,\mathrm{d}\tau$$

eine strikt monoton wachsende und stetig differenzierbare Funktion l definiert. Diese bildet das Intervall $[a,\, b]$ bijektiv auf das Intervall $[0,\, \operatorname{len}(\gamma)]$ ab und wird die Bogenlängenfunktion von γ genannt. Sie ist insbesondere invertierbar und ihre Ableitung

$$\dot l(t) = \frac{\mathrm{d}l}{\mathrm{d}t}(t) = \big|\dot\gamma(t)\big| > 0$$

heißt infinitisimales Längenelement von γ. Gilt $\dot l(t) = 1$ für alle $t \in [a,\, b]$, so sagt man, γ ist nach Bogenlänge parametrisiert. Parametrisierte Kurven mit dieser Eigenschaft sind besonders wichtig und nützlich.

[9] Mit anderen Worten: Der kürzeste Weg zwischen zwei gegebenen Punkten ist die entsprechende Verbindungsstrecke.

Lemma (**Reparametrisierung nach Bogenlänge**) Jede stetig differenzierbare und reguläre Kurve kann nach ihrer Bogenlänge parametrisiert werden.

Beweis Wir setzen $\tilde{I} := [0, \operatorname{len}(\gamma)]$ sowie $h(\tilde{t}) := l^{-1}(\tilde{t})$, wobei l^{-1} gerade die Umkehrfunktion der Bogenlängenfunktion l ist. Mit der Kettenregel und unter Verwendung der Ableitungsregel für Umkehrfunktionen verifizieren wir

$$\left|\frac{\mathrm{d}\tilde{\gamma}}{\mathrm{d}\tilde{t}}(\tilde{t})\right| = \left|\frac{\mathrm{d}\gamma}{\mathrm{d}t}\big(h(\tilde{t})\big)\right| \frac{\mathrm{d}h}{\mathrm{d}\tilde{t}}(\tilde{t}), \qquad \frac{\mathrm{d}h}{\mathrm{d}\tilde{t}}(\tilde{t}) = \frac{1}{\frac{\mathrm{d}l}{\mathrm{d}t}\big(h(\tilde{t})\big)} = \frac{1}{\left|\frac{\mathrm{d}\gamma}{\mathrm{d}t}\big(h(\tilde{t})\big)\right|}$$

und schließen, dass

$$\frac{\mathrm{d}\tilde{l}}{\mathrm{d}\tilde{t}}(\tilde{t}) = \left|\frac{\mathrm{d}\tilde{\gamma}}{\mathrm{d}\tilde{t}}(\tilde{t})\right| = 1$$

für alle $t \in \tilde{I}$ gilt. Insbesondere ist die Kurve $\tilde{\gamma} = \gamma \circ l^{-1}$ nach Bogenlänge parametrisiert. □

Bemerkung

1. Für viele Kurven ist die explizite Berechnung der Bogenlängen-Parametrisierung nicht möglich, da entweder die Integrale der Bogenlängenfunktion nicht exakt ausgewertet werden können oder weil keine geschlossene Formel für ihre Umkehrfunktion verfügbar ist. Die Existenz der entsprechenden Reparametrisierung ist aber immer gesichert und spielt in vielen Beweisen eine wesentliche Rolle.

2. Der Bogenlängenparameter wird in der Geometrie und der Physik meist mit s (und nicht wie im Beweis mit $\tilde{t}$) bezeichnet.

3. In Physikernotation stellt die Formel $\mathrm{d}s = |\dot{\gamma}(t)|\,\mathrm{d}t = |\dot{x}|\,\mathrm{d}t$ das Gesetz für die zeitliche Änderung des Bogenlängenparameters s dar und taucht auch in der symbolischen Substitutionsformel des ersten Kurvenintegrals auf.

Beispiele

1. Die Bogenlängenparametrisierung einer Kreislinie vom Radius ϱ um den Ursprung ist durch

   $$\tilde{\gamma}(\tilde{t}) = \begin{pmatrix} \varrho \cos\left(\varrho^{-1}\,\tilde{t}\right) \\ \varrho \sin\left(\varrho^{-1}\,\tilde{t}\right) \end{pmatrix}, \qquad \tilde{t} \in \tilde{I} = [0,\, 2\,\pi\,\varrho]$$

   gegeben, denn es gilt offensichtlich

   $$\left|\frac{\mathrm{d}\tilde{\gamma}}{\mathrm{d}\tilde{t}}(\tilde{t})\right| = 1$$

   für alle $\tilde{t}$.

2. Um die Bogenlängenparametrisierung der Kardioide zu berechnen, bestimmen wir zunächst die Bogenlängenfunktion der oben angegebenen Parametrisierung. Dies liefert

   $$l(t) = \int_0^t \left|\dot{\gamma}(\tau)\right| \mathrm{d}\tau = 2 \int_0^t \left|\cos\left(\tfrac{1}{2}\,t\right)\right| \mathrm{d}t = \begin{cases} 4 \sin\left(\frac{1}{2}\,t\right) & \text{für } 0 \le t \le \pi, \\ 8 - 4 \sin\left(\frac{1}{2}\,t\right) & \text{für } \pi \le t \le 2\,\pi, \end{cases}$$

wobei die Funktion l das Intervall $I = [0, 2\pi]$ bijektiv und strikt monoton wachsend auf das Intervall $\tilde{I} = [0, 8]$ abbildet. Die Umkehrfunktion von l ist gerade der gesuchte Parameterwechsel h und wenn wir die Formel $\tilde{t} = l(t)$ nach t auflösen, erhalten wir

$$h(\tilde{t}) = \begin{cases} 2\arcsin(\frac{1}{4}\tilde{t}) & \text{für } 0 \le \tilde{t} \le 4, \\ 2\pi + 2\arcsin(\frac{1}{4}\tilde{t} - 2) & \text{für } 4 \le \tilde{t} \le 8. \end{cases}$$

Die gesuchte Parametrisierung der Kardioide ist durch $\tilde{\gamma}(\tilde{t}) = \gamma\big(h(\tilde{t})\big)$ gegeben, wobei wir im konkreten Fall und mittels trigonometrischer Identitäten[10] die vereinfachten Formeln

$$\tilde{\gamma}(\tilde{t}) = \big(2 - \tfrac{1}{8}\tilde{t}^2\big) \begin{pmatrix} 1 - \frac{1}{8}\tilde{t}^2 \\ \frac{1}{8}\tilde{t}\sqrt{16 - \tilde{t}^2} \end{pmatrix} \qquad \text{für} \quad \tilde{t} \in [0, 4]$$

und

$$\tilde{\gamma}(\tilde{t}) = \big(-6 + 2\tilde{t} - \tfrac{1}{8}\tilde{t}^2\big) \begin{pmatrix} -7 + 2\tilde{t} - \frac{1}{8}\tilde{t}^2 \\ \frac{1}{8}(\tilde{t} - 8)\sqrt{-48 + 16\tilde{t} - \tilde{t}^2} \end{pmatrix} \qquad \text{für} \quad \tilde{t} \in [4, 8]$$

herleiten können.

Bemerkung: Eine (allerdings recht aufwendige) Probe zeigt, dass in der Tat $\big|\frac{\mathrm{d}\tilde{\gamma}}{\mathrm{d}\tilde{t}}(\tilde{t})\big| = 1$ für alle $\tilde{t} \in \tilde{I}$ gilt.

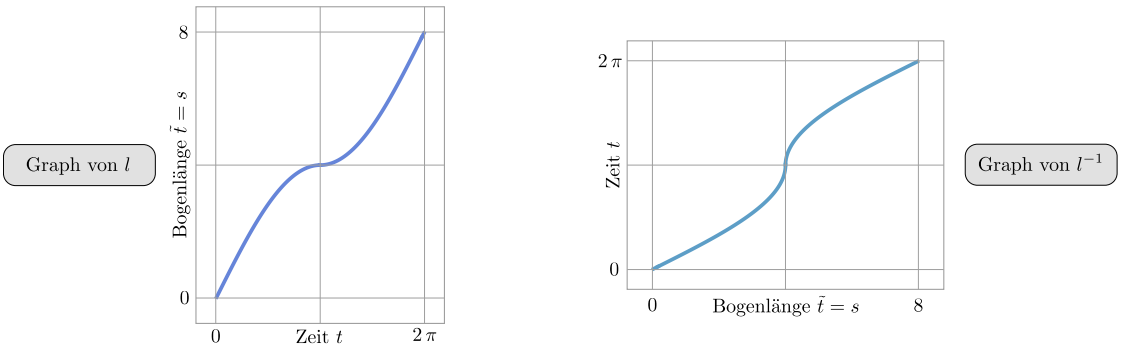

Abbildung Bei der Bogenlängen-Parametrisierung der Kardioide muss die strikt monoton wachsende Bogenlängenfunktion berechnet und anschließend invertiert werden. Insbesondere gilt $\tilde{h} = h^{-1} = l$ und $h = \tilde{h}^{-1} = l^{-1}$ und die Formeln $\tilde{t} = \tilde{h}(t)$ bzw. $t = h(\tilde{t})$ beschreiben den Übergang von t zu $\tilde{t}$ bzw. von $\tilde{t}$ zu t.

Kurvenintegrale

Definition Ist $f : \mathbb{R}^m \to \mathbb{R}$ eine stetige Funktion, so wird

$$\int_\gamma f(x)\,\mathrm{d}s := \int_a^b f\big(\gamma(t)\big)\big|\dot{\gamma}(t)\big|\,\mathrm{d}t = \int_a^b f\big(\gamma(t)\big)\left|\frac{\mathrm{d}\gamma}{\mathrm{d}t}(t)\right|\mathrm{d}t$$

als das Kurvenintegral 1. Art von f bezeichnet. Für eine stetige Funktion $f : \mathbb{R}^m \to \mathbb{R}^m$ nennen wir

$$\int_\gamma f(x) \cdot \mathrm{d}x := \int_a^b f\big(\gamma(t)\big) \cdot \dot{\gamma}(t)\,\mathrm{d}t = \int_a^b f\big(\gamma(t)\big) \cdot \frac{\mathrm{d}\gamma}{\mathrm{d}t}(t)\,\mathrm{d}t$$

das Kurvenintegral 2. Art.

[10] Zum Beispiel gilt $\cos(\arcsin(\xi)) = \sqrt{1 - \xi^2}$ für alle $\xi \in [-1, 1]$.

Bemerkungen

1. Der wesentliche Unterschied zwischen den beiden Arten ist der folgende: Ein Kurvenintegral der 1. Art bzw. 2. Art ist für eine *skalare Funktion* bzw. ein *Vektorfeld* definiert, also wenn $\mathbb{R}$ bzw. $\mathbb{R}^m$ der Bildbereich von f ist. Beachte auch, dass beide Arten von Kurvenintegralen immer eine *reelle Zahl* liefern.

2. Analoge Formeln werden verwendet, wenn f nur auf einer Teilmenge $D \subset \mathbb{R}^m$ definiert ist und die Kurve γ in D verläuft, d.h. wenn $\gamma(t) \in D$ für alle $t \in [a,\, b]$ gilt.

3. Die Länge einer Kurve ist das Kurvenintegral 1. Art für die konstante Funktion mit $f(x) = 1$ für alle $x \in \mathbb{R}^m$.

4. Für Kurvenintegrale kann analog zum obigen Lemma die Invarianz unter Reparametrisierung bewiesen werden, wobei sich bei Integralen der 2. Art das Vorzeichen unter orientierungsändernden Reparametrisierungen umkehrt (wegen des geänderten Durchlaufsinns).

5. *Achtung*: In der Literatur existieren viele verschiedene Notationen für Kurvenintegrale und Sie müssen sich immer klar machen, für welche Objekte (Zahlen, Vektoren, Funktionen usw.) die einzelnen Bausteine in den Formeln stehen.

Beispiele

1. Für die Parametrisierung der Kreisscheibe vom Radius ϱ und die skalare Funktion mit $f(x_1,\, x_2) = x_1^2\, x_2^2$ ergibt sich

$$f\big(\gamma(t)\big) = \varrho^4\, \cos^2(t)\, \sin^2(t) = \varrho^4 \left(\cos^2(t) - \cos^4(t)\right), \qquad \big|\dot\gamma(t)\big| = \varrho$$

   und damit

$$\int\limits_\gamma f(x)\, \mathrm{d}s = \varrho^5 \int\limits_0^{2\,\pi} \left(\cos^2(t) - \cos^4(t)\right) \mathrm{d}t = \tfrac{1}{4}\, \pi\, \varrho^5$$

   als das Kurvenintegral 1. Art, wobei das Integral zum Beispiel mit mehrfacher partieller Integration oder durch Verwendung der Euler-Formel berechnet werden kann.

   Bemerkung: Das bei Kurvenintegralen der 1. Art zu berechnende reelle Riemann-Integral kann formal wie folgt ermittelt werden: Wir substituieren $x_j = \gamma_j(t)$ im Argument von f und ersetzen $\mathrm{d}s$ durch den Ausdruck $\big|\dot\gamma(t)\big|\, \mathrm{d}t$. Außerdem liefern die Randpunkte von I gerade die Integrationsgrenzen.

2. Für die soeben betrachtete Kreislinie sowie das Vektorfeld

$$f(x_1,\, x_2) = \begin{pmatrix} x_1 - x_2 \\ 2\, x_1 \end{pmatrix}$$

   ergibt sich

$$\begin{aligned} f(x) \cdot \dot\gamma(t) &= \begin{pmatrix} \varrho\, \cos(t) - \varrho\, \sin(t) \\ 2\, \varrho\, \cos(t) \end{pmatrix} \cdot \begin{pmatrix} -\varrho\, \sin(t) \\ +\varrho\, \cos(t) \end{pmatrix} \\ &= \varrho^2 \left(1 + \cos^2(t) - \cos(t)\, \sin(t)\right) \end{aligned}$$

und damit

$$\int_\gamma f(x) \cdot \mathrm{d}x = \varrho^2 \int_0^{2\,\pi} \bigl(1 + \cos^2(t) - \cos(t)\,\sin(t)\bigr)\,\mathrm{d}t = 3\,\pi\,\varrho^2$$

nach kleineren Nebenrechnungen.

Bemerkung: Bei Kurvenintegralen der 2. Art substituieren wir $x = \gamma(t)$ sowie $\mathrm{d}x = \dot\gamma(t)\,\mathrm{d}t$.

Kurvenintegrale 2. Art in der Physik Beschreibt γ die Bahn eines geladenen Teilchens mit Ladung q und ist f die elektrische Feldstärke, so ist

$$q \int_\gamma f(x) \cdot \mathrm{d}x$$

gerade die vom elektrischen Feld am Teilchen verrichtete Arbeit. Bei einer analogen mechanischen Interpretation ist q die Masse und f ein Beschleunigungsfeld.

Lemma (**Standardabschätzung für Kurvenintegrale**) Mit den Notationen von oben gilt

$$\Bigl|\int_\gamma f(x)\,\mathrm{d}s\Bigr| \le M\ \mathrm{len}(\gamma)\,, \qquad \Bigl|\int_\gamma f(x)\cdot\,\mathrm{d}x\Bigr| \le M\ \mathrm{len}(\gamma)\,,$$

wobei $M := \max_{t\in[a,\,b]} \bigl|f(\gamma(t))\bigr|$ das Maximum von $|f|$ entlang der Kurve γ ist.

Beweis Für eine skalare Funktion $f : \mathbb{R}^m \to \mathbb{R}$ ergibt sich die erste Abschätzung via

$$\left|\int_a^b f(\gamma(t))\,\bigl|\dot\gamma(t)\bigr|\,\mathrm{d}t\right| \le \int_a^b \bigl|f(\gamma(t))\bigr|\,\bigl|\dot\gamma(t)\bigr|\,\mathrm{d}t \le M \int_a^b \bigl|\dot\gamma(t)\bigr|\,\mathrm{d}t$$

unmittelbar aus der Definition eines Kurvenintegrals 1. Art sowie den Eigenschaften von Riemann-Integralen. Für ein Integral 2. Art mit einem Vektorfeld $f : \mathbb{R}^m \to \mathbb{R}^m$ argumentieren wir analog, benutzen aber zusätzlich, dass

$$\bigl|f\bigl(\gamma(t)\bigr)\cdot\dot\gamma(t)\bigr| \le \bigl|f\bigl(\gamma(t)\bigr)\bigr|\,\bigl|\dot\gamma(t)\bigr| \le M\bigl|\dot\gamma(t)\bigr|$$

aus der Cauchy-Schwarz-Ungleichung folgt. □

Ausblick: Kurvenintegrale und zweidimensionale Flächeninhalte In zwei Raumdimensionen (also für $m = 2$) ist es oftmals möglich, den Rand $\Gamma = \partial K$ einer gegebenen kompakten Menge $K \subset \mathbb{R}^2$ durch eine stückweise stetig differenzierbare Kurve $\gamma : [a,\,b] \to \mathbb{R}^2$ zu parametrisieren, wobei dann $\gamma(a) = \gamma(b)$ gilt. Unter der Zusatzvoraussetzung, dass der Rand ∂K im mathematisch positiven Sinne — das heißt *gegen den Uhrzeigersinn* — durchlaufen wird, kann der Flächeninhalt von K durch die Formel

$$\mathrm{area}(K) = \int_a^b \gamma_1(t)\,\dot\gamma_2(t)\,\mathrm{d}t = -\int_a^b \gamma_2(t)\,\dot\gamma_1(t)\,\mathrm{d}t$$

berechnet werden, wobei sich die beiden bestimmten Integrale in der Formel nur im Vorzeichen unterscheiden.[11]

Bemerkung

1. Wir werden dieses Resultat erst am Ende von *Analysis 3* beweisen können, wollen aber schon festhalten, dass dieses oftmals als *Greensches Theorem* bezeichnet wird und ein Spezialfall des sehr viel allgemeineren *Satzes von Stokes* ist. In der angegebenen Form gilt es aber nur im $\mathbb{R}^2$.[12]

2. Wird ∂K im Uhrzeigersinn (also im mathematisch negativen Sinne) durchlaufen, so liefert die obige Formel immer noch den Flächeninhalt, aber diesmal mit einem falschen Vorzeichen.

3. Die obige Formel wird oftmals als

$$\operatorname{area}(K) = \int\limits_{\gamma} x_1 \,\mathrm{d}x_2 = -\int\limits_{\gamma} x_2 \,\mathrm{d}x_1$$

geschrieben. Beide Formulieren sind natürlich äquivalent, wobei die Substitution $x_j = \gamma_j(t)$ zugrunde liegt, die insbesondere $\mathrm{d}x_j = \dot{\gamma}_j(t)\,\mathrm{d}t$ impliziert.

Beispiele

1. Wir betrachten noch einmal die Parametrisierung einer achsenparallelen Ellipse (siehe die Beispiele zur Längenberechnung) und erhalten via

$$\operatorname{area}(K) = \int\limits_0^{2\pi} \gamma_1(t)\,\dot{\gamma}_2(t)\,\mathrm{d}t = \varrho_1\,\varrho_2 \int\limits_0^{2\pi} \cos^2(t)\,\mathrm{d}t = \pi\,\varrho_1\,\varrho_2$$

das aus der Schule bekannte Resultat. Für $\varrho_1 = \varrho_2$ ergibt sich der Flächeninhalt einer Kreisscheibe. Beachte, dass die verwendete Parametrisierung die Ellipse wirklich entgegen dem Uhrzeigersinn durchläuft.

2. Die Kardioide beschreibt den Rand einer kompakten Menge K und via

$$\gamma_1(t)\,\dot{\gamma}_2(t) = \big(\cos(t) - \cos^2(t)\big)\,\big(\sin^2(t) + \cos(t) - \cos^2(t)\big)$$

sowie elementaren Integrationstechniken[13] berechnen wir

$$\operatorname{area}(K) = \tfrac{3}{2}\,\pi$$

für die von γ eingeschlossene Fläche K.

[11] Die letzte Aussage ergibt sich via

$$\int\limits_a^b \gamma_1(t)\,\dot{\gamma}_2(t)\,\mathrm{d}t + \int\limits_a^b \gamma_2(t)\,\dot{\gamma}_1(t)\,\mathrm{d}t = \int\limits_a^b \Big(\dot{\gamma}_1(t)\,\gamma_2(t) + \gamma_1(t)\,\dot{\gamma}_2(t)\Big)\,\mathrm{d}t$$
$$= \int\limits_a^b \frac{\mathrm{d}}{\mathrm{d}t}\Big(\gamma_1(t)\,\gamma_2(t)\Big)\;\mathrm{d}t = \Big[\gamma_1(t)\,\gamma_2(t)\Big]_{t=a}^{t=b} = 0$$

aus dem Hauptsatz der Differential- und Integralrechung, da die Geschlossenheit der Kurve die Formeln $\gamma_1(a) = \gamma_1(b)$ und $\gamma_2(a) = \gamma_2(b)$ sicherstellt.

[12] Beachte, dass im $\mathbb{R}^3$ der Rand einer hinreichend guten Menge K eine zweidimensionale Fläche und keine eindimensionale Kurve sein wird.

[13] Mit der Euler-Formel können wir zum Beispiel eine alternative Darstellungsformel herleiten, in der nur Terme der Bauart $\cos(k\,t)$ und $\sin(k\,t)$ auftauchen. Diese können wir dann sehr leicht integrieren.

Kurven und Polygonzüge*

Polygonzüge als stückweise stetig differenzierbare Funktionen Sei $I = [a, b]$ ein abgeschlossenes Intervall, sei $T = \{t_0, t_1, \ldots, t_N\}$ eine entsprechende Zerlegung mit

$$a = t_0 < t_1 < \ldots < t_{N-1} < t_N = b$$

und seien $\xi_0, \ldots, \xi_N \in \mathbb{R}^m$ paarweise verschiedene Punkte. Die parametrisierte Kurve $\gamma_{\text{poly}} : [a, b] \to \mathbb{R}^m$ mit

$$\gamma_{\text{poly}}(t) := \frac{t - t_{n-1}}{t_n - t_{n-1}}\,\xi_{n-1} + \frac{t_n - t}{t_n - t_{n-1}}\,\xi_n \qquad \text{für} \quad t \in [t_{n-1}, t_n]$$

und alle $n \in \{1, \ldots, N\}$ nennen wir den Polygonzug durch die gegebenen Punkte ξ_n. Diese Kurve ist immer stetig und stückweise stetig differenzierbar, wobei für jedes n die Formeln

$$\gamma_{\text{poly}}(t_n) = \xi_n\,, \qquad \dot\gamma_{\text{poly}}(t) = \frac{\xi_n - \xi_{n-1}}{t_n - t_{n-1}} \qquad \text{für} \quad t_{n-1} < t < t_n$$

erfüllt sind. Insbesondere ist der Tangentialvektor $\dot\gamma_{\text{poly}}(t)$ auf jedem Zeitintervall (t_{n-1}, t_n) konstant und bis auf den skalaren Normierungsfaktor $1/(t_n - t_{n-1})$ durch den Differenzenvektor $\xi_n - \xi_{n-1}$ gegeben.

Die Länge eines Polygonzuges berechnet sich mit der Definition von oben zu

$$\begin{aligned}\text{len}\,(\gamma_{\text{poly}}) &= \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \left|\frac{\mathrm{d}\gamma_{\text{poly}}}{\mathrm{d}t}(t)\right| \mathrm{d}t = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \frac{\left|\gamma_{\text{poly}}(t_n) - \gamma_{\text{poly}}(t_{n-1})\right|}{t_n - t_{n-1}}\,\mathrm{d}t \\ &= \sum_{n=1}^{N} \left|\gamma_{\text{poly}}(t_n) - \gamma_{\text{poly}}(t_{n-1})\right| = \sum_{n=1}^{N} \left|\xi_n - \xi_{n-1}\right|,\end{aligned}$$

das heißt die Integralformel für die Länge liefert gerade die Summe der Längen der Verbindungstrecken.

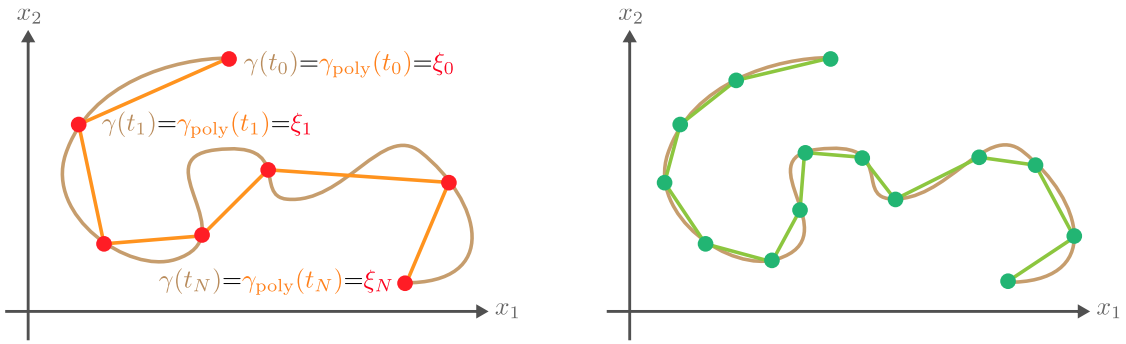

Abbildung Eine hinreichend gute Kurve (braun) kann beliebig gut durch Polygonzüge approximiert werden. Mit dieser Idee kann jedes Kurvenintegral näherungsweise in Form einer endlichen Summe berechnet werden.

Approximation durch Polygonzüge Seien $\gamma : I \to \mathbb{R}^m$ eine mindestens zweimal stetig differenzierbare Kurve und T eine Zerlegung des Intervalles $I = [a, b]$. Mit der speziellen Wahl

$$\xi_n := \gamma(t_n)$$

erhalten wir analog zu oben einen Polygonzug $\gamma_{\text{poly}}(t_n)$, der zu den Zeiten t_n durch dieselben Punkte wie γ läuft (siehe dazu das Bild). Mit überschaubarem Aufwand — wir wenden kompontenweise die Ergebnisse der Differential- und Integralrechung aus *Analysis 1* an — können wir für die Approximationsfehler die Abschätzungen

$$\|\gamma - \gamma_{\text{poly}}\|_\infty \leq C \, \|\ddot{\gamma}\|_\infty \, |T|^2 \,, \qquad \|\dot{\gamma} - \dot{\gamma}_{\text{poly}}\|_\infty \leq C \, \|\ddot{\gamma}\|_\infty \, |T|$$

etablieren, wobei die Konstante C nicht von γ oder T abhängt,

$$|T| = \max_{n \in \{1,\dots,N\}} \left|t_n - t_{n-1}\right|$$

die Feinheit der Zerlegung quantifiziert, und die Normterme durch

$$\|\gamma - \gamma_{\text{poly}}\|_\infty = \max_{t \in I} \left|\gamma(t) - \gamma_{\text{poly}}(t)\right| , \qquad \|\ddot{\gamma}\|_\infty = \max_{t \in I} \left|\ddot{\gamma}(t)\right|$$

definiert sind. Insgesamt schließen wir, dass jede feste Kurve γ bei immer feiner werdender Zerlegung von $[a, b]$ — also im Limes $|T| \to 0$ — immer besser durch den entsprechenden Polygonzug approximiert wird.

Insbesondere gilt für alle hinreichend kleine $|T|$ die Näherungsformel

$$\operatorname{len}(\gamma) \approx \operatorname{len}(\gamma_{\text{poly}}) = \sum_{n=1}^{N} \left|\xi_n - \xi_{n-1}\right| ,$$

die wir auch als heuristische Herleitung der der Integralformel für die Länge einer nicht-polygonalen Kurve interpretieren können. Etwas allgemeiner können wir

$$\int\limits_\gamma f(x) \, \mathrm{d}s \approx \int\limits_{\gamma_{\text{poly}}} f(x) \, \mathrm{d}s \approx \sum_{n=1}^{N} f\big(\xi_n\big) \big|\xi_n - \xi_{n-1}\big|$$

und

$$\int\limits_\gamma f(x) \cdot \mathrm{d}x \approx \int\limits_{\gamma_{\text{poly}}} f(x) \cdot \mathrm{d}x \approx \sum_{n=1}^{N} f\big(\xi_n\big) \cdot \big(\xi_n - \xi_{n-1}\big)$$

zeigen, wobei die entsprechenden Fehlerterme auch noch von der Güte von f abhängen. Beachte auch, dass die rechte Seite in jeder der beiden Formeln als verallgemeinerte Riemann-Summe für das jeweilige Integral auf der linken Seite interpretiert werden kann.

Vorlesungswoche 05

2.2 partielle Differenzierbarkeit

Vorbemerkung In der Mathematik gibt es zwei unterschiedliche Zugänge, um die Differentialrechnung für Funktionen in mehreren Variablen einzuführen bzw. aufzubauen, nämlich den *partiellen* und den *totalen* Ableitungsbegriff. Beide Konzepte meinen am Ende etwas leicht anderes, wobei nur das zweite aus theoretischer Sicht zufriedenstellend ist. Für praktische Zwecke ist aber das erste meist besser geeignet.

In diesem Abschnitt betrachten wir Abbildungen (oder Funktionen) $f : U \to \mathbb{R}^m$, die auf einer *offenen* Teilmenge $U \subseteq \mathbb{R}^n$ definiert sind.[14] Insbesondere enthält die Menge U keinen ihrer Randpunkte und zu jedem Punkt $x_* \in U$ existiert ein (vielleicht sehr kleines) $\varepsilon_* > 0$, sodass $B_{\varepsilon_*}(x_*)$ noch ganz in U liegt. Heuristisch meint dies, dass wir ausgehend von x_* immer wenigstens ein bisschen in *jede* Richtung laufen können, ohne den Definitionsbereich von f zu verlassen. Diese Eigenschaft wird sich als sehr nützlich und wichtig erweisen.

Erinnerung

1. Die äquivalente Charakterisierung von Stetigkeit aus dem vorherigen Kapitel impliziert, dass f genau dann stetig im Punkt $x_* \in U$ ist, falls für jede Folge $(x_k)_{k\in\mathbb{N}}$ aus U die Implikation

$$x_k \xrightarrow{k\to\infty} x_* \qquad \Longrightarrow \qquad f(x_k) \xrightarrow{k\to\infty} f(x_*)$$

   gilt, wobei links bzw. rechts die Konvergenz im $\mathbb{R}^n$ bzw. im $\mathbb{R}^m$ gemeint ist.

2. Wir nennen die Abbildung f stetig, falls sie in jedem Punkt $x_* \in U$ stetig ist.

3. Wir hatten im letzten Kapitel auch gezeigt, dass eine Folge endlich-dimensionaler Vektoren genau dann konvergiert, wenn sie komponentenweise konvergiert. Insbesondere gilt die Äquivalenz

$$x_k \xrightarrow{k\to\infty} x_* \qquad \Longleftrightarrow \qquad x_{k,j} \xrightarrow{k\to\infty} x_{*,j} \quad \text{für alle} \quad j \in \{1, \ldots, n\}\,,$$

   wobei links bzw. rechts die Konvergenz in $\mathbb{R}^n$ bzw. in $\mathbb{R}$ gemeint ist. Eine analoge Aussage charakterisiert die Konvergenz von $f(x_k)$.

partielle Ableitungen skalarer Funktionen

Vorbemerkung Wir betrachten zunächst den Fall $m = 1$ — d.h. skalare Funktionen auf der Menge $U \subseteq \mathbb{R}^n$ — und bezeichnen mit

$$f(x) \qquad \text{oder} \qquad f(x_1, \ldots, x_n)$$

den reellen Funktionswert von f im Punkt $x = (x_1, \ldots, x_n) \in U$. Dabei schreiben wir die Komponenten des Arguments x meist als n-Tupel, da die entsprechende Notation als Spaltenvektor sehr unübersichtlich wäre.

[14] Die Frage, ob bzw. wie Ableitungen in Randpunkten definiert werden können, werden wir später diskutieren.

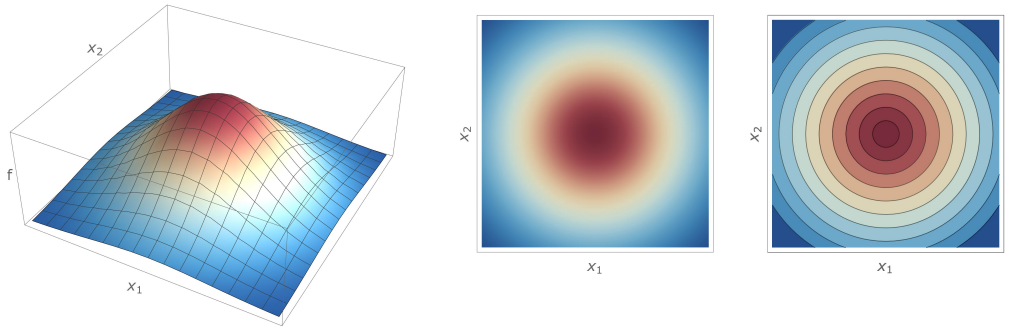

Abbildung Skalare Funktionen in zwei Variablen ($m = 1$ und $n = 2$) können auf verschiedene Weisen visualisiert werden: Als Flächenplot (links), als Dichteplot (Mitte) oder als Konturplot (rechts). Hier dargestellt für die Gaußsche Glockenfunktion $f(x_1, x_2) = \exp\left(-x_1^2 - x_2^2\right)$.

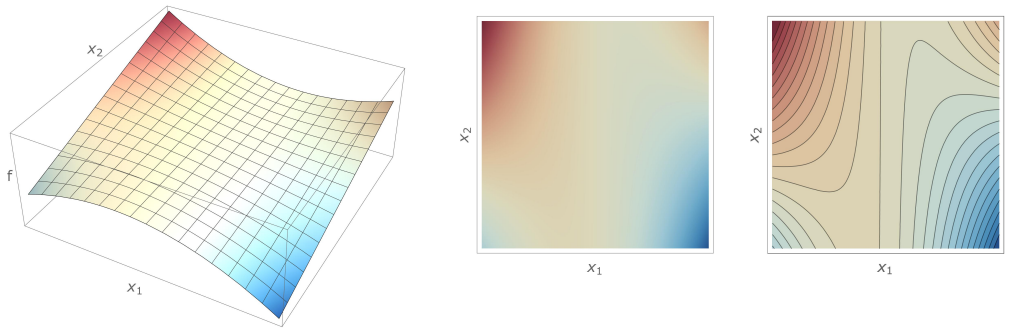

Abbildung Drei verschiedene Darstellungen der Funktion $f(x_1, x_2) = x_1^2\, x_2 - 2\, x_1$.

Erinnerung Für eine Funktion $f : \mathbb{R} \to \mathbb{R}$ in der Variablen $x \in \mathbb{R}$ ist die Ableitung in $x_* \in \mathbb{R}$ durch

$$f'(x_*) := \lim_{x \to x_*} \frac{f(x) - f(x_*)}{x - x_*} = \lim_{h \to 0} \frac{f(x_* + h) - f(x_*)}{h}$$

definiert, wobei die beiden Grenzwertformeln äquivalent sind (via $h = x - x_*$ bzw. $x = x_* + h$). Die zweite dieser Formeln werden wir nun verallgemeinern.

Definition Sei $x_* = (x_{*,1}, \ldots, x_{*,n})$ ein gegebener Punkt in der Menge U. Wir sagen, f besitzt im Punkt x_* die partielle Ableitung nach x_j, falls der Grenzwert

$$\partial_{x_j} f(x_*) := \lim_{h \to 0} \frac{f(x_* + h\, e_j) - f(x_*)}{h}$$

wohldefiniert ist, wobei $e_j \in \mathbb{R}^n$ den j-ten, n-dimensionalen kartesischen Einheitsvektor bezeichnet und h für eine reelle Zahl steht.[15] Existiert $\partial_{x_j} f(x_*)$ für jedes $j = \{1, \ldots, n\}$, so nennen wir f partiell differenzierbar im Punkt x_*.

[15] Für $n = 2$ gilt also

$$e_1 = \begin{pmatrix} e_{1,1} \\ e_{1,2} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad e_2 = \begin{pmatrix} e_{2,1} \\ e_{2,2} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

wobei wir oftmals auch die Tupel-Notation $e_1 = (1, 0)$, $e_2 = (0, 1)$ verwenden. Im Fall $n = 3$ existieren die drei Einheitsvektoren $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$.

Bemerkungen

1. Partielle Differenzierbarkeit ist zunächst wieder eine punktweise Eigenschaft. Ist diese in jedem Punkt $x_* \in U$ erfüllt, so nennen wir die Funktion f partiell differenzierbar (auf der Menge U). In diesem Fall sind $\partial_{x_1} f, \dots, \partial_{x_n} f$ selbst Funktionen mit Definitionsbereich U und Wertebereich $\mathbb{R}$.

2. Wir setzen bei allen Betrachtungen zur partiellen Differenzierbarkeit immer stillschweigend $h \neq 0$ voraus. Beachte auch, dass wir im Differenzenquotienten in der obigen Definition durch die reelle Zahl h und nicht durch die Differenz von zwei Vektoren dividieren.

3. Die Offenheit von U garantiert, dass für jedes $x_* \in U$ und alle $h \in \mathbb{R}$ mit hinreichend kleinem Betrag der Punkt $x_* + h\,e_j$ ebenfalls in U liegt und dass der Differenzenquotient daher wohldefiniert ist.[16]

4. In der Literatur gibt es weitere Notationen für partielle Ableitungen, zum Beispiel
$$\partial_{x_j} f(x_*) = \frac{\partial f}{\partial x_j}(x_*) = f_{;x_j}(x_*)\,.$$
Die dritte Schreibweise werden wir in dieser Vorlesung nicht verwenden, da mit ihr leicht Missverständnisse entstehen können.

5. Für $n = 1$ ist partielle Differenzierbarkeit die bekannte Differenzierbarkeit aus *Analysis 1* und es gilt
$$\partial_x f(x) = f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x)\,.$$
Hängt f jedoch nicht nur von einer Variablen $x \in \mathbb{R}$, sondern von n skalaren Variablen $x_1, \dots, x_n$ (bzw. einer vektorwertigen Variable $x \in \mathbb{R}^n$) ab, so schreiben wir weder f' noch $\mathrm{d}f/\mathrm{d}x$, sondern immer $\partial_{x_j} f$ oder $\partial f/\partial x_j$.

6. Wir können partielle Ableitungen auch wie folgt verstehen: f hängt zwar von den n unabhängigen Variablen $x_1, \dots, x_n$ ab, aber bei der Berechnung der partiellen Ableitung $\partial_{x_1} f(x_*)$ fixieren wir die Werte von $x_2, \dots, x_n$ (via $x_j = x_{*,j}$) und untersuchen die Differenzierbarkeit der eindimensionalen Funktion
$$\mathbb{R} \ni x_1 \quad \mapsto \quad f(x_1,\, x_{*,2},\, \dots,\, x_{*,n}) \in \mathbb{R}$$
im Punkt $x_{*,1}$. Analoges gilt für $\partial_{x_j} f(x_*)$ mit $j \in \{2, \dots, n\}$.

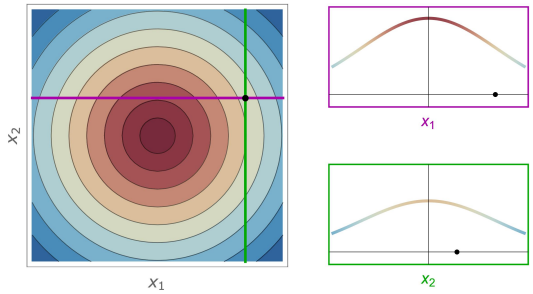

Abbildung Partielle Ableitungen für $n = 2$: Bei der Berechnung von $\partial_{x_1} f(x_*)$ bzw. $\partial_{x_2} f(x_*)$ betrachten wir x_2 bzw. x_1 als konstant und studieren f als Funktion von x_1 bzw. x_2. Der schwarze Punkt repräsentiert $x_* \in \mathbb{R}^2$.

[16] Genauer gesagt: Es muss $|h| < \operatorname{dist}(x, \partial U)$ gelten, wobei der Abstand von Punkten und Mengen bereits in den Übungen diskutiert wurde.

Beispiele

1. Wir betrachten die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ mit

$$f(x_1,\, x_2) = x_1\, {x_2}^2\,,$$

und fixieren zunächst $x_* = (x_{*,1},\, x_{*,2})$ beliebig. Für $h \neq 0$ berechnen wir

$$\begin{aligned}\frac{f(x_* + h\,e_1) - f(x_*)}{h} &= \frac{f(x_{*,1} + h,\, x_{*,2}) - f(x_{*,1},\, x_{*,2})}{h} \\ &= \frac{(x_{*,1} + h)\, x_{*,2}^2 - x_{*,1}\, x_{*,2}^2}{h} = \frac{h\, x_{*,2}^2}{h} \\ &= x_{*,2}^2 \quad \xrightarrow{h\to 0} \quad x_{*,2}^2\end{aligned}$$

sowie

$$\begin{aligned}\frac{f(x_* + h\,e_2) - f(x_*)}{h} &= \frac{f(x_{*,1},\, x_{*,2} + h) - f(x_{*,1},\, x_{*,2})}{h} \\ &= \frac{x_{*,1}\,(x_{*,2} + h)^2 - x_{*,1}\, x_{*,2}^2}{h} = \frac{2\, x_{*,1}\, x_{*,2}\, h + x_{*,1} h^2}{h} \\ &= 2\, x_{*,1}\, x_{*,2} + x_{*,1}\, h \quad \xrightarrow{h\to 0} \quad 2\, x_{*,1}\, x_{*,2}\end{aligned}$$

und haben damit die Existenz der partiellen Ableitungen

$$\partial_{x_1} f(x_{*,1},\, x_{*,2}) = x_{*,2}^2\,, \qquad \partial_{x_2} f(x_{*,1},\, x_{*,2}) = 2\, x_{*,1}\, x_{*,2}$$

gezeigt. Da unsere Argumente für alle $x_* \in \mathbb{R}^2$ gelten, können wir in den finalen Formeln auch x_1 statt $x_{*,1}$ und x_2 statt $x_{*,2}$ schreiben.

Bemerkung zum formalen Rechnen: Unser Ergebnis kann auch durch

$$\partial_{x_1}\big(x_1\, x_2^2\big) = x_2^2\,, \qquad \partial_{x_2}\big(x_1\, x_2^2\big) = 2\, x_1\, x_2$$

abgeleitet werden, wobei wir zuerst bei festgehaltenem x_2 nach x_1 und danach bei festgehaltenem x_1 nach x_2 differenziert haben. In der Praxis werden wir oftmals so rechnen, d.h. wir werden den Punkt x_* meist nicht explizit einführen. Es ist aber wichtig zu verstehen, dass partielle Differenzierbarkeit eine punktweise Eigenschaft ist, die in einzelnen Punkten erfüllt, in anderen Punkten aber verletzt sein kann.

2. Der Ausdruck

$$f(x_1,\, x_2,\, x_3) = x_1^3\, \cos(x_2) \;+\; (x_1 + x_2)\, x_3\, \sin(x_3)$$

definiert eine Funktion $f : \mathbb{R}^3 \to \mathbb{R}$, für die in jedem Punkt ihres Definitionsbereichs $U = \mathbb{R}^3$ alle drei partiellen Ableitungen existieren. Insbesondere gilt

$$\partial_{x_1} f(x_1,\, x_2,\, x_3) = 3\, x_1^2\, \cos(x_2) + x_3\, \sin(x_3)\,,$$

wobei wir den Ausdruck für f analog zu oben bei festgehaltenem x_2 und x_3 nach x_1 abgeleitet haben. Analog erhalten wir

$$\partial_{x_2} f(x_1,\, x_2,\, x_3) = -x_1^3\, \sin(x_2) + x_3\, \sin(x_3)$$

sowie

$$\partial_{x_3} f(x_1,\, x_2,\, x_3) = (x_1 + x_2)\, \sin(x_3) + (x_1 + x_2)\, x_3\, \cos(x_3)$$

durch partielle Differentiation nach x_2 bzw. x_3.

3. Bei allen theoretischen Betrachtungen werden wir die unabhängigen Variablen standardmäßig mit $x_1, \ldots, x_n$ bezeichnen. In der Praxis werden aber natürlich auch andere Schreibweisen verwendet, zum Beispiel

$$f(x, y) = x\,y^2\,, \qquad \partial_x f(x, y) = y^2\,, \qquad \partial_y f(x, y) = 2\,x\,y\,.$$

Wichtig ist, dass Sie sich in jedem Kontext immer klar machen, was die unabhängigen Variablen sind und wie viele es gibt. In der Physik werden die unabhängigen Variablen manchmal überhaupt nicht explizit geschrieben. Aus

$$E = m\,c^2$$

folgt zum Beispiel

$$\partial_m E = \frac{\partial E}{\partial m} = c^2\,, \qquad \partial_c E = \frac{\partial E}{\partial c} = 2\,m\,c\,,$$

denn E kann ja als Funktion in den unabhängigen Variablen m und c betrachtet werden.

4. Für $f : \mathbb{R}^3 \to \mathbb{R}$ mit

$$f(x, y, z) = x^2 + \big|y + z\big|$$

existieren nicht alle partiellen Ableitungen überall. Genauer gesagt: Es gilt

$$\partial_x f(x, y, z) = 2$$

für alle (x, y, z) und

$$\partial_y f(x, y, z) = \partial_z f(x, y, z) = \operatorname{sgn}(y + z)$$

in allen Punkten (x, y, z) mit $y + z \neq 0$, wobei

$$\operatorname{sgn}(s) = \begin{cases} -1 & \text{für } s < 0 \\ 0 & \text{für } s = 0 \\ +1 & \text{für } s > 0 \end{cases}$$

die Signums- bzw. Vorzeichenfunktion ist. Für $y + z = 0$ existieren aber die partiellen Ableitungen nach y und z jeweils nicht.

Rechenregeln Mit den Resultaten aus *Analysis 1* ergeben sich für zwei gegebene skalare Funktionen f und g die folgenden Gesetze:

1. *Linearität*: Es gilt

$$\partial_{x_j}\big(\alpha\,f(x) + \beta\,g(x)\big) = \alpha\,\partial_{x_j} f(x) + \beta\,\partial_{x_j} g(x)$$

sofern α und β reelle Zahlen sind (oder Ausdrücke, die nicht von x_j abhängen).

2. *Produkt- und Quotientenregel*: Es gilt

$$\partial_{x_j}\big(f(x)\,g(x)\big) = \big(\partial_{x_j} f(x)\big)\,g(x) + f(x)\,\big(\partial_{x_j} g(x)\big)$$

und

$$\partial_{x_j}\left(\frac{f(x)}{g(x)}\right) = \frac{\big(\partial_{x_j} f(x)\big)\,g(x) - f(x)\,\big(\partial_{x_j} g(x)\big)}{\big(g(x)\big)^2}\,,$$

wobei die letzte Regel nur in den Punkten mit $g(x) \neq 0$ gilt.

3. *Kettenregel*: Es gilt

$$\partial_{x_j}\phi\big(f(x)\big) = \phi'\big(f(x)\big)\,\partial_{x_j} f(x)$$

für jede stetig differenzierbare Funktion $\phi : \mathbb{R} \to \mathbb{R}$.

Bemerkung: Wir schon in *Analysis 1* kodieren die Rechenregeln Differenzierbarkeitsaussagen. Wenn zum Beispiel $f,\, g : U \to \mathbb{R}$ beide differenzierbar sind, so ist auch die Funktion $\alpha f + \beta g$ differenzierbar, wobei ihre Ableitung in jedem Punkt durch die angegebene Summenformel berechnet werden kann. Beachte aber, dass aus der Differenzierbarkeit von $\alpha\, f + \beta\, g$ nicht die Differenzierbarkeit von f und g folgt.[17]

Beispiele

1. Es gilt

$$\partial_{x_1}\left(\frac{x_1^2 + x_2}{x_2 + 2}\right) = \frac{2\,x_1}{x_2 + 2}\,, \quad \partial_{x_2}\left(\frac{x_1^2 + x_2}{x_2 + 2}\right) = \frac{(x_2 + 2) - (x_1^2 + x_2)}{(x_2 + 2)^2} = \frac{2 - x_1^2}{(x_2 + 2)^2}$$

für alle $x_1 \in \mathbb{R}$ und alle $x_2 \neq -2$.

2. Mit $\phi(s) = s^3$ liefert die Kettenregel

$$\partial_{x_j}\big(f(x_1,\, x_2,\, x_3)\big)^3 = 3\,\big(f(x_1,\, x_2,\, x_3)\big)^2\,\partial_{x_j} f(x_1,\, x_2,\, x_3)\,.$$

Analog folgt

$$\partial_{x_1}\big(\sin\big(x_1\,x_2^2\big)\big) = \cos\big(x_1\,x_2^2\big)\,x_2^2\,, \qquad \partial_{x_2}\big(\sin\big(x_1\,x_2^2\big)\big) = \cos\big(x_1\,x_2^2\big)\,2\,x_1\,x_2$$

mit $\phi(s) = \sin(s)$ und $f(x_1,\, x_2) = x_1\,x_2^2$.

partielle Ableitungen und Stetigkeit Die Existenz aller partiellen Ableitungen garantiert in höheren Dimensionen (d.h. für $n > 1$) noch nicht die Stetigkeit. Ein Standardgegenbeispiel ist die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ mit

$$f(x_1,\, x_2) = \begin{cases} \dfrac{x_1\,x_2}{(x_1^2 + x_2^2)^2} & \text{für } (x_1,\, x_2) \neq (0,\, 0)\,, \\ 0 & \text{für } x_1 = x_2 = 0\,, \end{cases}$$

wobei die Probleme hier im Koordinatenursprung $x_* = (0,\, 0)$ (und auch nur dort) auftauchen. Zum einen gilt

$$\partial_{x_1} f(0,\, 0) = \lim_{h\to 0} \frac{f(h,\, 0) - f(0,\, 0)}{h} = \lim_{h\to 0} \frac{0 - 0}{h} = 0$$

sowie

$$\partial_{x_2} f(0,\, 0) = \lim_{h\to 0} \frac{f(0,\, h) - f(0,\, 0)}{h} = \lim_{h\to 0} \frac{0 - 0}{h} = 0\,,$$

[17] Ein Standardgegenbeispiel mit $n = 1$ und $\alpha = \beta = 1$ ist $f(x) = \operatorname{sgn}(x)$ und $g(x) = -\operatorname{sgn}(x)$. Beide Funktionen sind in $x = 0$ nicht differenzierbar, aber $f + g$ schon (da die Sprünge von f und g sich aufheben).

d.h. beide partiellen Ableitungen sind in x_* definiert und nehmen jeweils den Wert 0 an. Andererseits ist f in x_* nicht stetig, denn für die Folge

$$x_k = (x_{k,1},\, x_{k,2}) = \left(\tfrac{1}{k},\, \tfrac{1}{k}\right)$$

ergibt sich

$$f\big(x_k\big) = \frac{\left(\frac{1}{k}\right)^2}{\left(\left(\frac{1}{k}\right)^2 + \left(\frac{1}{k}\right)^2\right)^2} = \frac{k^2}{4}\,,$$

d.h. x_k konvergiert für $k \to \infty$ zwar gegen x_*, aber $f(x_k)$ konvergiert nicht gegen $f(x_*) = 0$ (sondern im uneigentlichen Sinne gegen $+\infty$).

Merkregel Der Zusammenhang zwischen Stetigkeit und partieller Differenzierbarkeit ist für $n > 1$ deutlich subtiler als für $n = 1$. Das ist einer der mehreren Gründe, warum wir weiter unten das Konzept der *totalen Differenzierbarkeit* einführen werden.

Gradient einer skalaren Funktion Die partiellen Ableitungen von f können (sofern sie alle existieren) in jedem $x \in U$ zu einem n-dimensionalen Vektor vereinigt werden. Den entsprechenden Spaltenvektor

$$\operatorname{grad} f(x) = \begin{pmatrix} \partial_{x_1} f(x) \\ \vdots \\ \partial_{x_n} f(x) \end{pmatrix} = \begin{pmatrix} \partial_{x_1} f(x) & \dots & \partial_{x_n} f(x) \end{pmatrix}^T$$

nennen wir den Gradienten von f im Punkt x.

Achtung: Den entsprechenden Zeilenvektor werden wir unten als die Jacobi-Matrix von f bezeichnen.

*alternative Notation**: In der Physik und in den Ingenieurwissenschaften schreibt man auch gerne $\nabla f(x)$ statt $\operatorname{grad} f(x)$, wobei

$$\nabla = \begin{pmatrix} \partial_{x_1} \\ \vdots \\ \partial_{x_n} \end{pmatrix}$$

ein sogenannter Differentialoperator ist und als Nabla-Operator bezeichnet wird.[18]

Geometrische Interpretation des Gradienten Für jede Funktion $f : U \to \mathbb{R}$ und jeden Wert $c \in \mathbb{R}$ wird

$$N_f(c) = \{x \in U \;:\; f(x) = c\}$$

die entsprechende Niveaumenge (bzw. Kontur) genannt. Wir werden später sehen, dass $N_f(c)$ für $n = 2$ bzw. $n = 3$ oftmals als Kurve bzw. Fläche (und ganz allgemein als $n-1$-dimensionale Hyperfläche des $\mathbb{R}^n$) betrachtet werden kann.

Die wesentliche Beobachtung ist nun, dass für jeden Punkt $x \in N_f(c)$ der Vektor $\operatorname{grad} f(x)$ senkrecht auf $N_f(c)$ steht. Insbesondere liefert $\operatorname{grad} f(x)$ immer die Richtung des steilsten Anstiegs (und $-\operatorname{grad} f(x)$ damit die Richtung des steilsten Abstiegs). Wir werden dies weiter unten im Abschnitt über Richtungsableitungen beweisen.

[18] Der Nabla-Kalkül spielt in dieser Vorlesung keine Rolle.

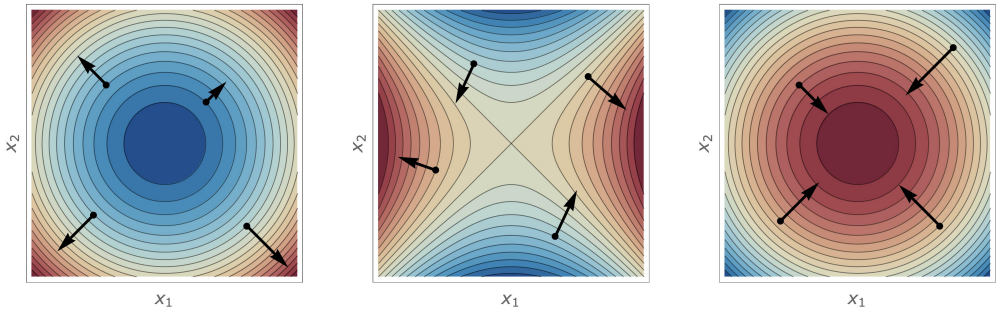

Abbildung Konturplots der skalaren Funktionen $f(x_1, x_2) = x_1^2 + x_2^2$ (links), $f(x_1, x_2) = x_1^2 - x_2^2$ (Mitte) und $f(x_1, x_2) = -x_1^2 - x_2^2$ (rechts), wobei Rot und Blau für große und kleine Werte stehen. Der Gradient $\operatorname{grad} f(x)$ wurde jeweils in vier ausgewählten Punkten x als Pfeil abgetragen. Er zeigt immer in Richtung des steilsten Anstiegs und steht senkrecht auf der entsprechenden Konturlinie. Beachte die Analogie zum Bergwandern, wobei die Konturen gerade die *Höhenlinien* sind.

höhere partielle Ableitungen und der Satz von Schwarz

Notation Wir schreiben

$$\partial_{x_i}\partial_{x_j} f(x) = \partial_{x_i}\big(\partial_{x_j} f(x)\big) \qquad \text{und} \qquad \partial_{x_i}^2 f(x) = \partial_{x_i}\big(\partial_{x_i} f(x)\big)$$

für zweifache partielle Ableitungen, d.h. für partielle Ableitungen von partiellen Ableitungen (sofern diese existieren). Alternative Schreibweisen sind

$$\partial_{x_i}\partial_{x_j} f(x) = \frac{\partial^2}{\partial x_i \partial x_j} f(x) \qquad \text{und} \qquad \partial_{x_i}^2 f(x) = \frac{\partial^2}{\partial {x_i}^2} f(x)$$

und analog werden dreifache, vierfache usw. Ableitungen eingeführt.

Bemerkung: Wir nennen eine k-fache partielle Ableitung auch partielle Ableitung der Ordnung k genannt. Für $n = 2$ sind also $\partial_{x_1}^2 f(x)$, $\partial_{x_1}\partial_{x_2} f(x)$, $\partial_{x_2}\partial_{x_1} f(x)$ und $\partial_{x_2}^2 f(x)$ die vier partiellen Ableitungen zweiter Ordnung im Punkt x und es gibt insgesamt 8 verschiedene Ableitungen dritter Ordnung. Für $n = 3$ gibt es 3 bzw. 9 bzw. 27 partielle Ableitungen erster bzw. zweiter bzw. dritter Ordnung.[19]

Definition Die Funktion $f : U \to \mathbb{R}$ heißt k-mal stetig partiell differenzierbar, wenn alle k-fachen partiellen Ableitungen in jedem Punkt aus U existieren und darüber hinaus auch stetig sind.

Bemerkungen

1. Da partielle Differenzierbarkeit und Stetigkeit punktweise Eigenschaften sind, gibt es auch ein punktweises Analogon zu dieser Definition.

2. *Vorwegnahme*: Wir werden weiter unten sehen, dass die Existenz *und* Stetigkeit aller partiellen Ableitungen (bis zur Ordnung k) schon die Existenz *und* Stetigkeit der entsprechenden totalen Ableitungen (bis zur Ordnung k) impliziert, wobei der Fall $k = 1$ der wichtigste ist. Die analoge Aussage ohne Stetigkeit gilt aber nicht.

[19] Sie haben es sicher schon ausgerechnet: Die allgemeine Formel ist n^k.

Beispiele

1. Für die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ mit

$$f(x_1,\, x_2) = \sin\left(x_1 + x_2^2\right)$$

berechnen wir zunächst

$$\partial_{x_1} f(x_1,\, x_2) = \cos\left(x_1 + x_2^2\right), \qquad \partial_{x_2} f(x_1,\, x_2) = 2\, x_2\, \cos\left(x_1 + x_2^2\right)$$

und anschließend

$$\begin{aligned} \partial_{x_1}^2 f(x_1,\, x_2) &= \partial_{x_1}\big(\cos\left(x_1 + x_2^2\right)\big) = -\sin\left(x_1 + x_2^2\right), \\ \partial_{x_2}\partial_{x_1} f(x_1,\, x_2) &= \partial_{x_2}\big(\cos\left(x_1 + x_2^2\right)\big) = -2\, x_2\, \sin\left(x_1 + x_2^2\right) \end{aligned}$$

sowie

$$\begin{aligned} \partial_{x_1}\partial_{x_2} f(x_1,\, x_2) &= \partial_{x_1}\big(2\, x_2\, \cos\left(x_1 + x_2^2\right)\big) = -2\, x_2\, \sin\left(x_1 + x_2^2\right), \\ \partial_{x_2}^2 f(x_1,\, x_2) &= \partial_{x_2}\big(2\, x_2\, \cos\left(x_1 + x_2^2\right)\big) = 2\, \cos\left(x_1 + x_2^2\right) - 4\, x_2^2\, \sin\left(x_1 + x_2^2\right). \end{aligned}$$

Insbesondere gilt in diesem Beispiel $\partial_{x_1}\partial_{x_2} f(x_1,\, x_2) = \partial_{x_2}\partial_{x_1} f(x_1,\, x_2)$ und damit der Satz von Schwarz (siehe unten) in jedem Punkt $(x_1,\, x_2)$.

2. In Physikernotation ergeben sich aus

$$f = \frac{x + y^2}{z}$$

die Formeln

$$\partial_x f = \frac{1}{z}, \qquad \partial_y f = \frac{2\, y}{z}, \qquad \partial_z f = -\frac{x + y^2}{z^2}$$

sowie

$$\begin{aligned} \partial_x^2 f &= 0, & \partial_y \partial_x f &= 0, & \partial_z \partial_x f &= -\frac{1}{z^2}, \\ \partial_x \partial_y f &= 0, & \partial_y^2 f &= \frac{2}{z}, & \partial_z \partial_y f &= -\frac{2\, y}{z^2}, \\ \partial_x \partial_z f &= -\frac{1}{z^2}, & \partial_y \partial_z f &= -\frac{2\, y}{z^2}, & \partial_z^2 f &= \frac{2\, x + 2\, y^2}{z^3}, \end{aligned}$$

wobei wir immer stillschweigend $z \neq 0$ vorausgesetzt haben. Auch hier gilt mit

$$\partial_x \partial_y f = \partial_y \partial_x f, \qquad \partial_y \partial_z f = \partial_z \partial_y f, \qquad \partial_x \partial_z f = \partial_z \partial_x f$$

der Satz von Schwarz in jedem zulässigen Punkt $(x,\, y,\, z)$.

3. Ein negatives Standardbeispiel ist

$$f(x_1,\, x_2) = \begin{cases} x_1\, x_2\, \dfrac{x_1^2 - x_2^2}{x_1^2 + x_2^2} & \text{für } (x_1,\, x_2) \neq (0,\, 0), \\ 0 & \text{für } (x_1,\, x_2) = (0,\, 0). \end{cases}$$

Für diese Funktion erhalten wir die ersten Ableitungen

$$\partial_{x_1} f(x_1,\, x_2) = \begin{cases} \dfrac{x_1^4\, x_2 + 4\, x_1^2\, x_2^3 - x_2^5}{\left(x_1^2 + x_2^2\right)^2} & \text{für } (x_1,\, x_2) \neq (0,\, 0)\,, \\ 0 & \text{für } (x_1,\, x_2) = (0,\, 0)\,, \end{cases}$$

$$\partial_{x_2} f(x_1,\, x_2) = \begin{cases} \dfrac{x_1^5 - 4\, x_1^3\, x_2^2 - x_1\, x_2^4}{\left(x_1^2 + x_2^2\right)^2} & \text{für } (x_1,\, x_2) \neq (0,\, 0)\,, \\ 0 & \text{für } (x_1,\, x_2) = (0,\, 0)\,, \end{cases}$$

wobei die Berechnung dieser Ableitungen in jedem Punkt $(x_1,\, x_2) \neq (0,\, 0)$ mit der Produkt- und der Quotientenregel gelingt, aber in $(0,\, 0)$ das Bestimmen von Grenzwerten erfordert. Die gemischten zweiten Abeitungen im Ursprung ergeben sich zu

$$\partial_{x_1}\partial_{x_2} f(0,\, 0) = \lim_{h\to 0} \frac{1}{h}\Big(\partial_{x_2} f(h,\, 0) - \partial_{x_2} f(0,\, 0)\Big) = \lim_{h\to 0} \frac{1}{h}\left(\frac{+h^5}{h^4} - 0\right) = +1\,,$$

$$\partial_{x_2}\partial_{x_1} f(0,\, 0) = \lim_{h\to 0} \frac{1}{h}\Big(\partial_{x_1} f(0,\, h) - \partial_{x_1} f(0,\, 0)\Big) = \lim_{h\to 0} \frac{1}{h}\left(\frac{-h^5}{h^4} - 0\right) = -1$$

und wir schließen, dass der Satz von Schwarz im Punkt $(0,\, 0)$ verletzt ist (in allen anderen Punkten gilt er aber). Das Problem ist, dass zwar die ersten partiellen Ableitungen in $(0,\, 0)$ stetig sind (Übungsaufgabe), aber die zweiten nicht mehr.

Theorem (**Satz von Schwarz**) Für eine zweimal stetig differenzierbare Funktion gilt

$$\partial_{x_i}\partial_{x_j} f(x) = \partial_{x_j}\partial_{x_i} f(x)$$

in jedem Punkt $x \in U$ und alle $i,\, j = \{1,\, \ldots,\, n\}$ mit $i \neq j$.

Beweis* *Vorbereitung*: Wir beweisen die Behauptung nur für $n = 2$, $i = 1$ und $j = 2$, aber alle Argumente können mühelos auf den allgemeinen Fall übertragen werden. Wir fixieren einen beliebigen Punkt $x_* = (x_{*,1},\, x_{*,2}) \in U$ und betrachten den Ausdruck

$$T(x_1,\, x_2) = \frac{f(x_1,\, x_2) + f(x_{*,1},\, x_{*,2}) - f(x_1,\, x_{*,2}) - f(x_{*,1},\, x_2)}{(x_1 - x_{*,1})\,(x_2 - x_{*,2})}\,,$$

wobei wir im Folgenden immer $x_1 \neq x_{*,1}$, $x_2 \neq x_{*,2}$ voraussetzen und außerdem annehmen, dass $|x_1 - x_{*,1}|$ und $|x_2 - x_{*,2}|$ immer so klein sind, dass $(x_1,\, x_2)$ ganz in U liegt (alternativ können Sie bei der ersten Lektüre $U = \mathbb{R}^2$ setzen).

Erstes Konvergenzresultat: Durch einfache Termumformungen erhalten wir die Darstellungsformel

$$T(x_1,\, x_2) = \frac{\dfrac{f(x_1,\, x_2) - f(x_1,\, x_{*,2})}{x_2 - x_{*,2}} - \dfrac{f(x_{*,1},\, x_2) - f(x_{*,1},\, x_{*,2})}{x_2 - x_{*,2}}}{x_1 - x_{*,1}}\,,$$

und haben damit $T(x_1,\, x_2)$ als Differenzenquotient (bzgl. der 1. Variable) eines anderen Differenzenquotienten (bzgl. der 2. Variablen) ausgedrückt. Für festgehaltenes x_2 ist die Hilfsfunktion ϕ mit

$$\phi(x_1) := \frac{f(x_1,\, x_2) - f(x_1,\, x_{*,2})}{x_2 - x_{*,2}}$$

differenzierbar in ihrer Variablen x_1 und besitzt die Ableitung

$$\phi'(x_1) = \frac{\partial_{x_1} f(x_1,\, x_2) - \partial_{x_1} f(x_1,\, x_{*,2})}{x_2 - x_{*,2}}\,.$$

Nach dem Mittelwertsatz der Differentialrechnung aus *Analysis 1* existiert daher ein Zwischenwert $\hat{x}_1$ mit

$$T(x_1,\, x_2) = \frac{\phi(x_1) - \phi(x_{*,1})}{x_1 - x_{*,1}} = \phi'(\hat{x}_1)\,,$$

wobei $\hat{x}_1$ zwischen $x_{*,1}$ und x_1 liegt und von $x_{*,1}$, x_1 sowie x_2 abhängt (was wir aber nicht explizit schreiben). Wir erhalten das Zwischenergebnis

$$T(x_1,\, x_2) = \frac{\partial_{x_1} f(\hat{x}_1,\, x_2) - \partial_{x_1} f(\hat{x}_1,\, x_{*,2})}{x_2 - x_{*,2}}\,,$$

wobei die rechte Seite ein Differenzenquotient bzgl. x_2 ist, sofern $\hat{x}_1$ als Parameter betrachtet wird. Für die entsprechende Hilfsfunktion ψ in der Variablen x_2 gilt

$$\psi(x_2) := \partial_{x_1} f(\hat{x}_1,\, x_2)\,, \qquad \psi'(x_2) = \partial_{x_2}\partial_{x_1} f(\hat{x}_1,\, x_2)$$

und eine erneute Anwendung des Mittelwertsatzes (diesmal auf die Funktion ψ) liefert

$$T(x_1,\, x_2) = \partial_{x_2}\partial_{x_1} f(\hat{x}_1,\, \hat{x}_2)\,.$$

Die reellen Zahlen $\hat{x}_1$ bzw. $\hat{x}_2$ hängen beide zwar in komplizierter Weise von x_1, x_2 und $x_{*,1}$, $x_{*,2}$ ab, liegen aber immer zwischen $x_{*,1}$ und x_1 bzw. zwischen $x_{*,2}$ und x_2. Deshalb folgt die Konvergenzaussage

$$T(x_1,\, x_2) \quad \xrightarrow{(x_1,\, x_2) \to (x_{*,1},\, x_{*,2})} \quad \partial_{x_2}\partial_{x_1} f(x_{*,1},\, x_{*,2})$$

aus der vorausgesetzten Stetigkeit der Funktion $\partial_{x_2}\partial_{x_1} f$ im Punkt $(x_{*,1},\, x_{*,2})$.

Zweites Konvergenzresultat: Es gilt auch die alternative Darstellungsformel

$$T(x_1,\, x_2) = \frac{\dfrac{f(x_1,\, x_2) - f(x_{*,1},\, x_2)}{x_1 - x_{*,1}} - \dfrac{f(x_1,\, x_{*,2}) - f(x_{*,1},\, x_{*,2})}{x_1 - x_{*,1}}}{x_2 - x_{*,2}}\,,$$

wobei die rechte Seite jetzt der x_2-Differenzenquotient eines x_1-Differenzenquotienten ist. Wir können alle Argumente von oben analog wiederholen und erhalten

$$T(x_1,\, x_2) \quad \xrightarrow{(x_1,\, x_2) \to (x_{*,1},\, x_{*,2})} \quad \partial_{x_1}\partial_{x_2} f(x_{*,1},\, x_{*,2})\,.$$

Die Behauptung folgt nach Vergleich der beiden Grenzwertformeln und weil $x_* \in U$ beliebig war. □

Bemerkung* Die Beweisidee kann informell wie folgt erklärt werden. Die partielle Ableitung von f nach x_1 kann bei festgehaltenem x_* durch den Differenzenquotienten

$$\mathsf{DQ}_{x_1} f(x_1,\, x_2) = \frac{f(x_1,\, x_{*,2}) - f(x_{*,1},\, x_{*,2})}{x_1 - x_{*,1}} \approx \partial_{x_1} f(x_*)$$

approximiert werden und analog wird $\mathsf{DQ}_{x_2} f(x_1,\, x_2)$ eingeführt. Die entscheidende Beobachtung ist nun, dass wir für jede Funktion f die Formel

$$\mathsf{DQ}_{x_1}\mathsf{DQ}_{x_2} f(x_1,\, x_2) = \mathsf{DQ}_{x_2}\mathsf{DQ}_{x_1} f(x_1,\, x_2)$$

nachrechnen können (siehe die beiden Darstellungsformeln für $T(x_1,\, x_2)$ im Beweis). Es ist also egal, in welcher Reihenfolge wir zwei verschiedene Differenzenquotienten bilden. Die technische Schwierigkeit besteht darin, den Grenübergang $x \to x_*$ durchzuführen und zu beweisen, dass $\mathsf{DQ}_{x_i}\mathsf{DQ}_{x_j} f(x_1,\, x_2)$ gegen $\partial_{x_i}\partial_{x_j} f(x_{*,1},\, x_{*,2})$ konvergiert.

Klarstellung In praktisch relevanten Fällen wird der Satz von Schwarz in aller Regel gelten, d.h. zwei verschiedene partielle Ableitungen dürfen miteinander vertauscht werden. Es gibt aber auch die entarteten Funktionen, bei denen dies nicht möglich ist (siehe das dritte Beispiel oben).

Verallgemeinerung Ist f sogar dreimal stetig partiell differenzierbar, so gilt

$$\partial_{x_i}\partial_{x_j}\partial_{x_k} f(x) = \partial_{x_i}\partial_{x_k}\partial_{x_j} f(x)\,, \qquad \partial_{x_i}\partial_{x_j}^2 f(x) = \partial_{x_j}\partial_{x_i}\partial_{x_j} f(x) = \partial_{x_j}^2\partial_{x_i} f(x)\,,$$

für alle paarweise verschiedenen Indizes i, j, $k = \{1, \ldots, n\}$, d.h. wir können sogar drei partielle Ableitungen beliebig vertauschen. Analoge Aussagen gelten für Funktionen, die vier- und fünfmal stetig partiell differenzierbar sind.

Hesse-Matrix einer skalaren Funktion Existieren sämtliche zweiten partiellen Ableitungen von f, so wird

$$\operatorname{Hess} f(x) = \begin{pmatrix} \partial_{x_1}^2 f(x) & \ldots & \partial_{x_1}\partial_{x_n} f(x) \\ \vdots & & \vdots \\ \partial_{x_n}\partial_{x_1} f(x) & \ldots & \partial_{x_n}^2 f(x) \end{pmatrix}$$

als die Hesse-Matrix von f im Punkt x bezeichnet. Wenn der Satz von Schwarz gilt, ist diese reelle und quadratische $n\times n$-Matrix symmetrisch und besitzt daher nur reelle Eigenwerte (siehe *Lineare Algebra 2*). Diese Eigenwerte werden eine prominente Rolle bei der Untersuchung lokaler Extremstellen spielen.

*alternative Notation**: In den Anwendungswissenschaften wird die Hesse-Matrix auch als $\nabla \otimes \nabla f(x)$ geschrieben.

*Ausblick**: Der Laplace-Operator $\Delta = \nabla \cdot \nabla$ ist der vielleicht wichtigste Differentialoperator. Er ist durch

$$\Delta f(x) = \partial_{x_1}^2 f(x) + \ldots + \partial_{x_n}^2 f(x)$$

definiert und entspricht gerade der *Spur* der Hesse-Matrix, also der Summe ihrer Diagonaleinträge.

partielle Ableitungen vektorwertiger Funktionen

Vorbemerkung Wir betrachten nun vektorwertige Funktionen $f : U \to \mathbb{R}^m$, die wieder auf einer offenen Menge $U \subseteq \mathbb{R}^n$ definiert sind, aber diesmal Werte im $\mathbb{R}^m$ annehmen (wobei m und n verschieden sein können). Insbesondere besitzt eine solche Funktion via

$$f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}$$

genau m skalare Komponentenfunktionen $f_i : U \to \mathbb{R}$. Wir werden $f(x)$ standardmäßig als Spaltenvektor mit m Zeilen schreiben, können im Prinzip aber auch die Tupel-Notation $f(x) = (f_1(x), \ldots, f_m(x))$ verwenden.

Alle Konzepte des vorherigen Abschnitts (partielle Differenzierbarkeit, stetige partielle Differenzierbarkeit, Satz von Schwarz) können komponentenweise übertragen bzw. ausgewertet werden. So ist die Funktion f zum Beispiel genau dann stetig partiell differenzierbar, wenn alle ihre Komponenten f_i diese Eigenschaft besitzen.

Jacobi-Matrix Die partiellen Ableitungen der Komponentenfunktionen f_i können (sofern sie alle existieren) in eine $m\times n$-Matrix einsortiert werden, die die Jacobi-Matrix von f genannt wird:

$$\mathrm{Jac} f(x) := \begin{pmatrix} \partial_{x_1} f_1(x) & \dots & \partial_{x_n} f_1(x) \\ \vdots & & \vdots \\ \partial_{x_1} f_m(x) & \dots & \partial_{x_n} f_m(x) \end{pmatrix}$$

Merkregel: Die partiellen Ableitungen der i-ten Komponentenfunktion f_i stehen in der i-ten Zeile der Jacobi-Matrix, wohingegen alle partiellen Ableitungen der Variablen x_j zusammen ihre j-te Spalte bilden.

Achtung Es ist sehr wichtig, dass die Jacobi-Matrix in der angegebenen Weise aus den partiellen Ableitungen zusammengesetzt wird. Werden zum Beispiel Zeilen und Spalten vertauscht, so gelten viele der weiter unten hergeleiteten Formeln nicht mehr. Zum Beispiel die allgemeine Form der Kettenregel.

Bemerkung Die Berechnung der Jacobi-Matrix ist eine lineare Operation, d.h. es gilt

$$\mathrm{Jac}\,\big(\alpha\, f + \beta\, g\big)(x) = \alpha\ \mathrm{Jac}\, g(x) + \beta\ \mathrm{Jac}\, g(x)$$

für je zwei partiell differenzierbare Funktionen f, g und beliebige reelle Zahlen α, β.

Alternative Notationen

1. Innerhalb der Mathematik gibt es leider keine einheitliche Schreibweise für Jacobi-Matrizen und in der Literatur wird anstelle von $\mathrm{Jac} f(x)$ auch

   $$\big(\mathrm{Jac}\, f\big)(x)\,, \qquad \mathrm{Jac}\big(f\big)(x)\,, \qquad Jf(x)\,, \qquad J_f(x)\,, \qquad \dots$$

   verwendet. Auch $\nabla f(x)$ ist nicht ungewöhnlich.

2. Physiker und Ingenieure setzen oftmals $y = f(x)$ und schreiben die Jacobi-Matrix symbolisch als

   $$\mathrm{Jac}\, f = \frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix} = \left(\frac{\partial y_i}{\partial x_j}\right)_{i=1\dots m,\, j=1\dots n},$$

   wobei die Notation auf der rechten Seite andeuten soll, dass y_i bzw. x_j in der i-ten Zeile bzw. j-ten Spalte auftaucht. Diese Notation hat viele Vorteile, aber auch einige Nachteile, und es wird niemanden überraschen, dass die mathematische Variante exakter, die andere aber oftmals praktischer ist.

Beispiele

1. Für $m = 2$, $n = 3$, $U = \mathbb{R}^3$ sei die stetig differenzierbare Funktion $f : \mathbb{R}^3 \to \mathbb{R}^2$ gegeben durch

$$f_1(x_1, x_2, x_3) = x_1 x_2 + x_3^2, \qquad f_2(x_1, x_2, x_3) = x_1^2 (x_2 + x_3).$$

Die Jacobi-Matrix berechnet sich zu

$$\operatorname{Jac} f(x) = \begin{pmatrix} x_2 & x_1 & 2x_3 \\ 2x_1(x_2 + x_3) & x_1^2 & x_1^2 \end{pmatrix},$$

wobei die erste bzw. zweite Zeile aus den partiellen Ableitungen von f_1 bzw. f_2 zusammengesetzt ist. Die erste bzw. zweite bzw. dritte Spalte entspricht dabei den Ableitungen nach x_1 bzw. x_2 bzw. x_3 (siehe auch die Merkregel).

2. Die Formeln

$$f_1(x_1, x_2) = x_1 + x_2, \qquad f_2(x_1, x_2) = x_1^2 x_2, \qquad f_3(x_1, x_2) = x_1 x_2^3$$

definieren eine Abbildung $f : \mathbb{R}^2 \to \mathbb{R}^3$ und mit einfachen Rechnungen erhalten wir die 3×2-Matrix

$$\operatorname{Jac} f(x) = \begin{pmatrix} 1 & 1 \\ 2x_1 x_2 & x_1^2 \\ x_2^3 & 3x_1 x_2^2 \end{pmatrix}$$

mit 3 Zeilen und 2 Spalten.

3. Im Fall von $m = 1$ gilt

$$\operatorname{Jac} f(x) = f(x) = \begin{pmatrix} \partial_{x_1} f_1(x) & \dots & \partial_{x_n} f_1(x) \end{pmatrix} = \big(\operatorname{grad} f(x)\big)^T,$$

d.h. die Jacobi-Matrix einer skalaren Funktion ist gerade der Zeilenvektor aller partiellen Ableitungen und damit der zum Gradienten transponierte Vektor.

4. Für $n = 1$ ergibt sich

$$\operatorname{Jac} f(x) = \begin{pmatrix} \partial_x f_1(x) \\ \vdots \\ \partial_x f_m(x) \end{pmatrix} = \begin{pmatrix} f_1'(x) \\ \vdots \\ f_m'(x) \end{pmatrix}$$

als Jacobi-Matrix einer vektorwertigen Funktion mit nur einer Variablen $x \in \mathbb{R}$. In diesem Fall kann f auch als parametrisierte Kurve interpretiert werden und $\operatorname{Jac} f(x)$ ist gerade der Tangentialvektor im Punkt x.[20]

5. Die Kreisfrequenz ω eines idealisierten Pendels kann mittels

$$\omega = \sqrt{\frac{g}{l}}$$

aus der Erdbeschleunigung g und der Fadenlänge l berechnet werden. In Physikernotation kann die entsprechende Jacobi-Matrix als

$$\frac{\partial \omega}{\partial(g, l)} = \begin{pmatrix} \partial_g \omega & \partial_l \omega \end{pmatrix} = \begin{pmatrix} \dfrac{1}{2\sqrt{g\,l}} & -\dfrac{\sqrt{g}}{2\sqrt{l^3}} \end{pmatrix}$$

geschrieben werden. Wichtig ist wieder, dass es sich um eine 1×2-Matrix handelt (und die Einträge richtig berechnet wurden).

[20] Beachte, dass wir im ersten Abschnitt dieses Kapitels eine andere Notation verwendet haben: t statt x, γ statt f und Punkt statt Strich für Ableitungen. Dies entspricht der ewigen Grundregel: *Mathematische Notationen sind immer kontextabhängig!*

Elemente der Vektoranalysis

Vektorfelder Besonders wichtig ist der Fall $m = n$. In diesem Fall nennen wir eine Abbildung $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^n$ ein (n-dimensionales) Vektorfeld auf U. Insbesondere stimmt die Anzahl der Komponenten f_i mit der Anzahl der Variablen x_j überein.

Beispiele Die Formeln

$$f(x) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad f(x) = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix}, \qquad f(x) = \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix}, \qquad f(x) = \begin{pmatrix} x_1 \\ x_1 \end{pmatrix}$$

mit $x = (x_1, x_2) \in U = \mathbb{R}^2$ definieren vier sehr einfache 2-dimensionale Vektorfelder (siehe die Bilder).

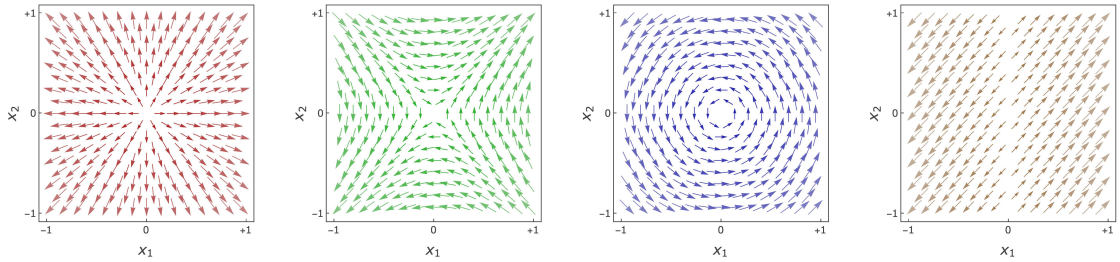

Abbildung Vektorfelder in zwei Dimensionen ($n = m = 2$) können sehr gut visualisiert werden: In ausgewählten Punkten $x \in U \subseteq \mathbb{R}^2$ wird $f(x) \in \mathbb{R}^2$ abgetragen. Hier dargestellt für die vier Beispiele.

Ausblick Eine differenzierbare Kurve $\gamma : I \to U$ wird Integralkurve an das Vektorfeld $f : U \to \mathbb{R}^n$ genannt, wenn

$$\dot\gamma(t) = f\big(\gamma(t)\big)$$

für alle $t \in I$ gilt. Wir werden am Ende von *Analysis 2* die Theorie solcher *autonomer Differentialgleichungen* genauer studieren.

spezielle Differentialoperatoren Für ein n-dimensionales partiell differenzierbares Vektorfeld $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^n$ wird durch

$$\operatorname{div} f(x) := \partial_{x_1} f_1(x) + \ldots + \partial_{x_n} f_n(x)$$

eine skalare Funktion auf U definiert, die wir die Divergenz von f nennen.

Im Fall von $n = 2$ bzw. $n = 3$ wird die Rotation von f durch

$$\operatorname{rot} f(x) := \partial_{x_1} f_2(x) - \partial_{x_2} f_1(x) \qquad \text{bzw.} \qquad \operatorname{rot} f(x) := \begin{pmatrix} \partial_{x_2} f_3(x) - \partial_{x_3} f_2(x) \\ \partial_{x_3} f_1(x) - \partial_{x_1} f_3(x) \\ \partial_{x_1} f_2(x) - \partial_{x_2} f_1(x) \end{pmatrix}$$

definiert und stellt eine skalare Funktion bzw. ein Vektorfeld dar.

Ausblick*

1. Die mathematische Bedeutung der Divergenz und der Rotation eines Vektorfeldes werden wir erst in *Analysis 3* richtig verstehen, wenn wir die *Integralsätze von Gauß und Stokes* formulieren und beweisen.

2. Aus physikalischer Sicht quantifiziert $\operatorname{div} f(x)$ in jedem Punkt $x \in U$ die *Quellen* und *Senken* des Vektorfeldes f. Zum Beispiel ist das Geschwindigkeitsfeld einer idealen Flüssigkeit immer divergenzfrei, da dort Masse nur transportiert, aber weder erzeugt noch vernichtet wird.

3. Die Rotation eines Vektorfeldes beschreibt, wie „verwirbelt“ das Vektorfeld f in der Nähe des Punktes x ist. Sie taucht zum Beispiel bei den Maxwellschen Differentialgleichungen für elektromagnetische Felder auf. Es existiert auch eine Rotation für n-dimensionale Vektorfelder, aber diese ist weder eine skalare Funktion noch ein Vektorfeld, sondern besitzt $\frac{1}{2}\,n\,(n-1)$ viele Komponenten der Bauart $\partial_{x_j} f_i(x) - \partial_{x_i} f_j(x)$.

Gradientenfelder Eine wichtige Rolle spielen auch Vektorfelder der Bauart

$$f(x) = \operatorname{grad} \Phi(x)\,,$$

wobei die skalare Funktion $\Phi : U \to \mathbb{R}$ ein Potential zu f genannt wird.

Ausblick*

1. Beispiele für Gradientenfelder in der Physik sind das elektrische Feld bzw. das Gravitationsfeld, dass von einer idealen Punktladung bzw. Punktmasse erzeugt wird, sowie das Geschwindigkeitsfeld einer sogenannten Potentialströmung.

2. Ein n-dimensionales und stetig partiell differenzierbares Vektorfeld $f : U \to \mathbb{R}^n$ kann nur dann ein Gradientenfeld sein, wenn

   $$\partial_{x_j} f_i(x) = \partial_{x_i} f_j(x)$$

   für alle $x \in U$ und alle Indizes $i,\, j \in \{1, \ldots, n\}$ mit $i \neq j$ gilt. Diese *notwendigen Bedingungen* ergeben sich unmittelbar aus dem Satz von Schwarz (angewendet auf Φ) und werden auch Integrabilitätsbedingungen genannt.[21] Mit ihrer Hilfe können wir beweisen, dass ein gegebenes Vektorfeld *kein* Gradientenfeld ist.

3. Die Frage, unter welchen Voraussetzungen die notwendigen Bedingungen auch *hinreichend* für die Existenz von Φ sind, wird in der *Potentialtheorie* studiert.

   *Ausblick**: Im Fall $n = 2$ wird sich zeigen, dass die notwendigen Bedingungen genau dann hinreichend sind, wenn die Menge U keine *Löcher* besitzt. Zum Beispiel besitzt eine Kreisscheibe kein Loch, ein Kreisring aber schon. Für $n \geq 3$ ist die Antwort jedoch subtiler.

4. Die Divergenz eines Gradientenfeldes ist gerade der Laplace-Operator, d.h. für jede skalare Funktion $\Phi : U \to \mathbb{R}$ gilt

   $$\Delta\Phi(x) = \operatorname{div} \operatorname{grad} \Phi(x) = \sum_{j=1}^{n} \partial_{x_j}^2 \Phi(x)\,.$$

   Wenn $\Delta\Phi(x) = 0$ für alle $x \in U$ gilt, so wird Φ *harmonisch* genannt.

[21] Alternativ können wir die Integrabilitätsbedingungen für Vektorfeld $f : U \to \mathbb{R}^n$ auch als $\operatorname{rot} f(x) = 0$ für alle $x \in U$ schreiben.

Vorlesungswoche 06

2.3 totale Differenzierbarkeit

Vorbemerkung Wir führen in diesem Abschnitt den Begriff der totalen Ableitung für eine Funktion $f : U \to \mathbb{R}^m$ ein und diskutieren insbesondere die Beziehung zu den partiellen Ableitungen. Dabei ist $U \subseteq \mathbb{R}^n$ offen und wir interpretieren wieder $x \in \mathbb{R}^n$ und $f(x) \in \mathbb{R}^m$ standardmäßig als Spaltenvektoren, d.h. wir schreiben

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \qquad f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}.$$

Bei den Argumenten von f verwenden wir jedoch wie bisher die Tupelnotation, da andernfalls die Formeln sehr unübersichtlich werden. Die einzigen Zeilenvektoren sind die Jacobi-Matrizen skalarer Funktionen.

zur Konvergenz bei Funktionen in höheren Dimensionen Seien $x_* \in U$ und $y_* \in \mathbb{R}^m$ gegeben. Analog zu *Analysis 1* schreiben wir

$$\lim_{x \to x_*} f(x) = y_*\,,$$

sofern

$$y_* = \lim_{k \to \infty} f(x_k)$$

für *jede* Folge $(x_k)_{k\to\infty}$ gilt, die den folgenden Eigenschaften genügt:

1. Jedes Folgenglied liegt in $U \setminus \{x_*\}$, d.h. $x_k \in U$ und $x_k \neq x_*$ für jedes $k \in \mathbb{N}$.
2. Es gilt $x_* = \lim_{k\to\infty} x_k$.

Bemerkungen:

1. Wie schon in *Analysis 1* benutzen wir die folgenden alternativen Notationen: Statt $\lim_{x\to x_*} f(x) = y_*$ schreiben wir auch
$$f(x) \quad \xrightarrow{x \to x_*} \quad y_*$$
oder sagen, $f(x)$ konvergiert für $x \to x_*$ gegen y_*.
2. Die äquivalente Charakterisierung von Stetigkeit kann auch wie folgt formuliert werden: f ist genau dann stetig in x_*, wenn $f(x_*) = \lim_{x\to x_*} f(x)$ gilt.
3. Die Definition von Konvergenz in normierten Räumen garantiert die logischen Äquivalenzen
$$\lim_{x \to x_*} f(x) = y_* \quad \Leftrightarrow \quad \lim_{x \to x_*} \big(f(x) - y_*\big) = 0 \quad \Leftrightarrow \quad \lim_{x \to x_*} \big|f(x) - y_*\big| = 0$$
und dass wir statt $x \to x_*$ auch $(x - x_*) \to 0$ oder $|x - x_*| \to 0$ schreiben könnten. Beachte aber, dass $\big|f(x)\big| \to \big|y_*\big|$ und $|x| \to |x_*|$ jeweils etwas anderes meinen, nämlich nicht die Konvergenz von Vektoren, sondern die Konvergenz ihrer Beträge.
4. Wir werden diese Konzepte nicht nur für f, sondern analog auch für andere Funktionen verwenden.

lineare Abbildungen und ihre Norm Eine Abbildung $L : \mathbb{R}^n \to \mathbb{R}^m$ wird bekanntlich linear genannt, falls

$$L\big(\lambda\, x + \tilde{\lambda}\, \tilde{x}\big) = \lambda\, L(x) + \tilde{\lambda}\, L(\tilde{x})$$

für alle x, $\tilde{x} \in \mathbb{R}^n$ und λ, $\tilde{\lambda} \in \mathbb{R}$ erfüllt ist und die Menge aller linearen Abbildungen wird mit

$$\mathsf{Lin}(\mathbb{R}^n, \mathbb{R}^m)$$

bezeichnet. Für jede lineare Abbildung L und jedes $x \in \mathbb{R}^n$ gilt

$$L(x) = x_1 L(e_1) + \ldots + x_n\, L(e_n)\,,$$

wobei $e_1, \ldots, e_n$ wieder die kartesischen Einheitsvektoren im $\mathbb{R}^n$ bezeichnen, und mit dieser Darstellungsformel sowie der Abschätzung $|x_j| \le |x|$ können wir leicht zeigen (Übungsaufgabe), dass die *Operatornorm*

$$\|L\|_{\mathrm{op}} := \sup\big\{\big|L(x)\big| \;:\; |x| = 1\big\}$$

für jedes $L \in \mathsf{Lin}(\mathbb{R}^n, \mathbb{R}^m)$ als reelle Zahl wohldefiniert ist[22] und dass die Abschätzung

$$\big|L(x)\big| \le \|L\|_{\mathrm{op}}\, |x|$$

für alle $x \in \mathbb{R}^n$ gilt, wobei die Betragszeichen auf der linken bzw. auf der rechten Seite sich auf die euklidische Norm im $\mathbb{R}^m$ bzw. $\mathbb{R}^n$ beziehen. Insbesondere ist jede lineare Abbildung aus $\mathsf{Lin}(\mathbb{R}^n, \mathbb{R}^m)$ Lipschitz-stetig (und damit auch stetig), wobei $\|L\|_{\mathrm{op}}$ gerade die optimale Lipschitz-Konstante ist.[23]

totale Ableitung als lineare Abbildung

Definition Die Abbildung $f : U \to \mathbb{R}^m$ heißt total differenzierbar im Punkt $x_* \in U$, falls eine lineare Abbildung $L_* : \mathbb{R}^n \to \mathbb{R}^m$ existiert, sodass

$$\frac{\big|f(x) - f(x_*) - L_*(x - x_*)\big|}{|x - x_*|} \quad \xrightarrow{|x-x_*|\to 0} \quad 0$$

im Sinne der Konvergenz reeller Zahlen gilt. Hierbei steht $|\cdot|$ im Zähler bzw. Nenner der linken Seite für die euklidische Norm im $\mathbb{R}^m$ bzw. $\mathbb{R}^n$.

Bemerkungen

1. Totale Differenzierbarkeit ist wieder eine punktweise Eigenschaft. Wenn diese für alle $x_* \in U$ gegeben ist, so nennen wir f total differenzierbar.

2. Auf den ersten Blick haben partielle und totale Differenzierbarkeit nicht viel miteinander zu tun. Wir werden aber sehen, das dem nicht so ist, sondern dass beide Konzepte sehr eng verwandt sind. Sie sind aber nicht identisch!

[22] Insbesondere gilt $\|L\|_{\mathrm{op}} \le \sum_{j=1}^n \big|L(e_j)\big| < \infty$.

[23] *Ausblick**: Lineare Abbildungen zwischen unendlich-dimensionalen normierten Räumen sind nicht unbedingt stetig, da ihre Operatornorm den Wert ∞ annehmen kann.

3. Die Konvergenzbedingung in der Definition kann mit der Substitution $x = x_* + h$ bzw. $h = x - x_*$ auch als

$$\frac{\big|f(x_* + h) - f(x_*) - L_*(h)\big|}{|h|} \quad \xrightarrow{|h|\to 0} \quad 0$$

geschrieben werden, wobei $h \in \mathbb{R}^n$ ein Vektor ist und immer stillschweigend $h \neq 0$ vorausgesetzt wird.

4. Wir schreiben im Folgenden meist

$$\mathrm{D}f(x_*) \quad \text{statt} \quad L_*$$

und nennen diese lineare Abbildung die totale Ableitung oder das Differential von f in x_*. In der mathematischen Literatur findet sich oftmals auch die Bezeichnung Fréchet-Ableitung.

Merkregel: Die totale Ableitung einer (meist nichtlinearen) Funktion f von $\mathbb{R}^n$ nach $\mathbb{R}^m$ ist (so sie existiert) in *jedem* Punkt eine *lineare* Abbildung zwischen diesen Räumen.

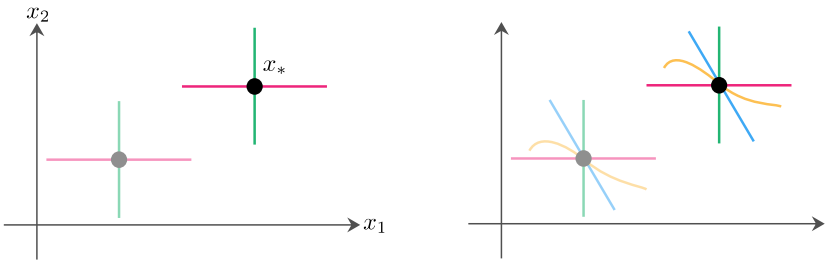

Abbildung *Links*: Die partielle Differenzierbarkeit einer Funktion f im Punkt x_* garantiert, dass die Funktion in allen Einheitsrichtungen differenzierbar ist. *Rechts*: Totale Differenzierbarkeit stellt sicher, dass die entsprechende Eigenschaft auch auf allen gedrehten oder gar gekrümmten Kurven gilt, die durch x_* laufen. Siehe dazu auch die Kettenregel und das Konzept der Richtungsableitung weiter unten.

Beispiele

1. Wir betrachten die skalare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ mit

$$f(x) = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{j,k}\, x_j\, x_k + \sum_{j=1}^{n} b_j\, x_j + c\,,$$

mit reellen Koeffizienten $a_{j,k}$, b_j, und c. Für einen gegebenen Punkt $x_* \in \mathbb{R}^n$ betrachten wir außerdem die durch

$$L_*(h) := \sum_{j=1}^{n} \sum_{k=1}^{n} (a_{j,k}\, x_{*,j}\, h_k + a_{j,k}\, x_{*,k}\, h_j) + \sum_{j=1}^{n} b_j\, h_j$$

definierte lineare Abbildung $L_* : \mathbb{R}^n \to \mathbb{R}$ und berechnen

$$f(x_* + h) - f(x_*) - L_*(h) = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{j,k}\, h_j\, h_k\,.$$

Wegen $|h_j| \leq |h|$ erhalten wir

$$\big|f(x_* + h) - f(x_*) - L_*(h)\big| \leq C\,|h|^2\,, \qquad C := \sum_{j=1}^{n} \sum_{k=1}^{n} \big|a_{j,k}\big|$$

und schließen, dass L_* die totale Ableitung von f in x_* ist.

2. Die Formel

$$f(x) = \begin{pmatrix} f_1(x_1,\, x_2) \\ f_2(x_1,\, x_2) \end{pmatrix} = \begin{pmatrix} x_2^2 \\ x_1\, x_2 \end{pmatrix}$$

beschreibt ein Vektorfeld $f : \mathbb{R}^2 \to \mathbb{R}^2$ und mit direkten Rechnungen verifizieren wir die Formel

$$f(x_* + h) - f(x_*) = \begin{pmatrix} 2\, x_{*,\,2}\, h_2 \\ x_{*,\,1}\, h_2 + x_{*,\,2}\, h_1 \end{pmatrix} + \begin{pmatrix} h_2^2 \\ h_1\, h_2 \end{pmatrix},$$

wobei wir auf der rechten Seite schon die linearen Terme bzgl. h von den höheren Ordnungstermen (quadratisch, kubisch usw.) getrennt haben. Insbesondere sehen wir, dass

$$\big(\mathrm{D}f(x_*)\big)(h) = \begin{pmatrix} 2\, x_{*,\,2}\, h_2 \\ x_{*,\,1}\, h_2 + x_{*,\,2}\, h_1 \end{pmatrix} = \begin{pmatrix} 0 & 2\, x_{*,\,2} \\ x_{*,\,2} & x_{*,\,1} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

für jedes $x_* \in \mathbb{R}^2$ im Sinne einer totalen Ableitung gilt, denn die entsprechenden Fehler- oder Restterme (siehe dazu auch weiter unten) können durch

$$\left| \begin{pmatrix} h_2^2 \\ h_1\, h_2 \end{pmatrix} \right| = \sqrt{h_2^4 + h_1^2\, h_2^2} \leq \sqrt{|h|^4 + |h|^2\, |h|^2} = \sqrt{2}\, |h|^2$$

abgeschätzt werden und konvergieren auch nach Division durch $|h|$ immer noch für $|h| \to 0$ gegen 0.

Bemerkung: Da x_* zwar einen festen, aber letztlich beliebigen Punkt in $\mathbb{R}^2$ beschreibt, können wir alternativ auch x statt x_* schreiben, wobei dann h und x als unabhängige Vektoren zu betrachten sind.

3. Jede lineare Abbildung $L \in L(\mathbb{R}^n,\, \mathbb{R}^m)$ ist total differenzierbar, wobei

$$L(x) - L(x_*) = L\big(x - x_*\big) \qquad \text{und damit} \qquad \mathrm{D}_L(x_*) = L$$

gilt. Insbesondere ist L in jedem Punkt ihre eigene Ableitung.

Ausblick: höhere totale Ableitungen Man kann auch zweite und dritte totale Ableitungen definieren, allerdings wird es dann noch abstrakter. In jedem Punkt $x_* \in U$ gilt zum Beispiel

$$\mathrm{DD}f(x_*) \in \mathsf{Lin}\big(\mathbb{R}^n,\, \mathsf{Lin}(\mathbb{R}^n,\, \mathbb{R}^m)\big)\,,$$

d.h. die totale Ableitung der totalen Ableitung von f ist (so sie existiert) eine lineare Abbildung von $\mathbb{R}^n$ in den Raum $\mathsf{Lin}(\mathbb{R}^n,\, \mathbb{R}^m)$, der ausgestattet mit der Operatornorm ein vollständiger normierter Raum ist. Wir werden im Rahmen dieser Vorlesung aber keine höheren totalen Ableitungen brauchen.[24]

[24] Mit der Zeit und durch viel Übung wachsen Ihr Wissen sowie Ihre mathematische Intuition und irgendwann werden Ihnen Konstrukte wie $\mathrm{DD}f(x_*)$ oder $\mathrm{DDD}f(x_*)$ ganz vertraut vorkommen.

Zusammenhang zwischen totalen und partiellen Ableitungen

Darstellungsformel mit Restglied Ist $f : U \to \mathbb{R}^n$ in $x_* \in U$ total differenzierbar, so nennen wir

$$r(x,\, x_*) := f(x) - f(x_*) - \big(\mathrm{D}f(x_*)\big)(x - x_*)$$

das entsprechende Restglied und erhalten die Darstellungsformel

$$f(x) = f(x_*) + \big(\mathrm{D}f(x_*)\big)(x - x_*) + r(x,\, x_*)$$

nach einer einfachen Termumstellung. Die totale Differenzierbarkeit impliziert dabei

$$\frac{r(x,\, x_*)}{\big|x - x_*\big|} \quad \xrightarrow{x \to x_*} \quad 0 \qquad \text{bzw.} \qquad \frac{\big|r(x,\, x_*)\big|}{\big|x - x_*\big|} \quad \xrightarrow{x \to x_*} \quad 0$$

wobei dies eine Konvergenz vom m-dimensionalen Vektoren bzw. ihrer Beträge ist.[25]

Ausblick: Wir werden unten sehen, dass die Darstellungsformel ein Spezialfall des *Satzes von Taylor* in höheren Dimensionen ist.

Lemma (**totale Differenzierbarkeit impliziert Stetigkeit**) Ist $f : U \to \mathbb{R}^m$ total differenzierbar in x_*, so ist f in x_* auch stetig.

Beweis Mit der Darstellungsformel schließen wir

$$\big|f(x) - f(x_*)\big| = \big|\big(\mathrm{D}f(x_*)\big)(x - x_*) + r(x,\, x_*)\big| \le \|\mathrm{D}f(x_*)\|_{\mathrm{op}} \big|x - x_*\big| + \big|r(x,\, x_*)\big|$$

und erhalten

$$\big|f(x) - f(x_*)\big| \le \|\mathrm{D}f(x_*)\|_{\mathrm{op}} \big|x - x_*\big| + \frac{\big|r(x,\, x_*)\big|}{\big|x - x_*\big|}\, |x - x_*| \quad \xrightarrow{x \to x_*} \quad 0\,.$$

Die Behauptung folgt nun aus der äquivalenten Charakterisierung von Stetigkeit. □

Lemma (**totale impliziert partielle Differenzierbarkeit**) Ist $f : U \to \mathbb{R}^m$ total differenzierbar in $x_* \in U$, so ist f in x_* auch partiell differenzierbar und es gilt

$$\big(\mathrm{D}f(x_*)\big)(h) = \mathrm{Jac}f(x_*)\, h\,,$$

für alle $h \in \mathbb{R}^n$.[26] Insbesondere ist die Jacobi-Matrix von f in x_* wohldefiniert und stellt die Matrixdarstellung der linearen Abbildung $\mathrm{D}f(x_*)$ bzgl. der kanonischen Basen in $\mathbb{R}^n$ und $\mathbb{R}^m$ dar.

[25] Beachte, dass der Ausdruck $r(x,\, x_*)/(x - x_*)$ für $m \neq 1$ als Quotient von zwei Vektoren nicht definiert ist. Vektoren dürfen nur durch Zahlen geteilt werden!

[26] Auf der linken Seite der Formel wird $\mathrm{D}f(x_*)$ als lineare Abbildung von $\mathbb{R}^n$ nach $\mathbb{R}^m$ auf den Vektor $h \in \mathbb{R}^n$ angewendet und liefert einen Vektor im $\mathbb{R}^m$. Die rechte Seite ist das Produkt der $m{\times}n$-Matrix $\mathrm{Jac}f(x_*)$ und des Spaltenvektors h (das ist eine $n{\times}1$-Matrix), was insgesamt eine $m{\times}1$-Matrix, also einen m-dimensionalen Spaltenvektor ergibt.

Beweis Nach Voraussetzung gilt

$$r(x,\, x_*) = \frac{f(x) - f(x_*) - \big(\mathrm{D}f(x_*)\big)\,(x - x_*)}{|x - x_*|} \quad \xrightarrow{x \to x_*} \quad 0$$

im Sinne der vektoriellen Konvergenz in $\mathbb{R}^m$. Wir werten die i-te Komponente dieses Resultats im Spezialfall $x = x_* + h\,e_j$ aus, wobei e_j den j-ten kartesischen Einheitsvektor im $\mathbb{R}^n$ bezeichnet und $h \neq 0$ eine reelle Zahl ist. Nach Einsetzen und einfachen Umformungen — die jeweils leicht anders für $h < 0$ bzw. $h > 0$ sind — erhalten wir

$$\frac{f_i(x_* + h\,e_j) - f_i(x_*)}{h} \quad \xrightarrow{h \to 0} \quad e_i \cdot \big(\mathrm{D}f(x_*)\big)(e_j)\,,$$

wobei $\cdot$ das Skalarprodukt im $\mathbb{R}^m$ meint und wir $e_i \cdot f(x) = f_i(x)$ benutzt haben. Dieses skalare Konvergenzresultat zeigt, dass die partielle Ableitung von f_i nach x_j in x_* existiert und durch

$$\partial_{x_j} f_i(x_*) = e_i \cdot \big(\mathrm{D}f(x_*)\big)(e_j)$$

gegeben ist. Wir haben damit die partielle Differenzierbarkeit von f im Punkt x_* gezeigt. Die letzte Formel impliziert außerdem (siehe die Vorlesung *Lineare Algebra*), dass in der kanonischen Matrixdarstellung der linearen Abbildung $\mathrm{D}f(x_*)$ der Eintrag in der i-ten Zeile und j-ten Spalte gerade die partielle Ableitung von f_i nach x_j im Punkt x_* ist.[27] □

Bemerkungen

1. *Klarstellung*: Der Zusammenhang zwischen linearen Abbildungen und ihren Matrixdarstellungen wird in der Vorlesung *Lineare Algebra* genauer diskutiert. Da wir in dieser Vorlesung meist nur mit den kanonischen Basen arbeiten, fällt der konzeptionelle Unterschied zwischen $\mathrm{D}f(x_*)$ und $\mathrm{Jac}f(x_*)$ nicht sonderlich ins Gewicht. Später wird es aber sehr wichtig sein, dass das Differential einer Abbildung im Gegensatz zur Matrix der partiellen Ableitungen nicht von der Wahl einer Basis bzw. von der Wahl der Koordinaten abhängt.

2. *Achtung*: Die Existenz aller partiellen Ableitungen impliziert alleine noch nicht die Existenz der totalen Ableitung. Siehe dazu das Gegenbeispiel im Abschnitt über *Richtungsableitungen.* Wir werden im Anschluss aber beweisen, dass die Existenz *und* Stetigkeit aller partiellen Ableitungen schon die Existenz der totalen Ableitung garantieren.

3. Die totale Differenzierbarkeit stellt die Stetigkeit von f in x_* sicher, aber wir hatten bereits oben gesehen, dass dies für die partielle Differenzierbarkeit *nicht* gilt.

[27] Insbesondere gilt

$$\mathrm{Jac}\, f(x_*) = \begin{pmatrix} | & & | \\ \big(\mathrm{D}f(x_*)\big)(e_1) & \dots & \big(\mathrm{D}f(x_*)\big)(e_n) \\ | & & | \end{pmatrix},$$

d.h. die j-Spalte der Jacobi-Matrix von f in x_* ist gerade das Bild von e_j unter der linearen Abbildung $\mathrm{D}f(x_*)$.

Theorem (**stetige partielle impliziert totale Differenzierbarkeit**) Ist f stetig partiell differenzierbar auf U, so ist f in jedem Punkt x_* total differenzierbar, wobei $\mathrm{D}f(x_*)$ die von der Jacobi-Matrix $\mathrm{Jac}f(x_*)$ vermittelte lineare Abbildung ist.

Beweis* *Spezialfall*: Unter der Zusatzannahme $m = 1$ zeigen wir, dass die — mittels der partiellen Ableitungen von f definierte — lineare Abbildung

$$h \in \mathbb{R}^n \quad \mapsto \quad \mathrm{Jac}f(x_*)h = \sum_{j=1}^{n} \partial_{x_j} f(x_*)\, h_j \in \mathbb{R}\,,$$

wirklich die totale Ableitung von f in x_* darstellt bzw. dass das skalare Restglied

$$r(x,\, x_*) = f(x) - f(x_*) - \sum_{j=1}^{n} \partial_{x_j} f(x_*)(x_j - x_{*,\,j})$$

für $x \to x_*$ im Betrag schneller gegen 0 konvergiert als $|x - x_*|$. In jedem Punkt $x \neq x_*$ betrachten wir dazu die Hilfsvektoren

$$\xi_0 := x_*\,, \qquad \xi_1 := x_* + (x_1 - x_{*,\,1})\, e_1\,, \qquad \xi_2 := x_* + (x_1 - x_{*,\,1})\, e_1 + (x_2 - x_{*,\,2})\, e_2\,,$$

usw. bis

$$\xi_n := x_* + (x_1 - x_{*,\,1})\, e_1 + (x_2 - x_{*,\,2})\, e_2 + \cdots + (x_n - x_{*,\,n})\, e_n\,.$$

Jeder der Vektoren ξ_j ist also so definiert, dass seine ersten j Komponenten die von x, die restlichen aber die von x_* sind. Insbesondere unterscheiden sich ξ_j und ξ_{j-1} für jedes $j \in \{1,\, \ldots,\, n\}$ jeweils nur in der j-ten Komponente. Wir können daher den Mittelwertsatz der Differentialrechnung aus *Analysis 1* auf die differenzierbare Funktion

$$x_j \quad \mapsto \quad f(x_1,\, \ldots,\, x_{j-1},\, x_j,\, x_{*,\,j+1},\, \ldots,\, x_{*,\,n})$$

anwenden. Dies liefert

$$f(\xi_j) - f(\xi_{j-1}) = \partial_{x_j} f(\hat{\xi}_j)\, \big(x_j - x_{*,\,j}\big)\,,$$

mit $\hat{\xi}_j = (x_1,\, \ldots,\, x_{j-1},\, \hat{x}_j,\, x_{*,\,j+1},\, \ldots,\, x_{*,\,n})$, wobei die reelle Zahl $\hat{x}_j$ zwischen $x_{*,\,j}$ und x_j liegt. Da nach Konstruktion außerdem

$$f(x) - f(x_*) = \sum_{j=1}^{n} f(\xi_j) - f(\xi_{j-1})$$

gilt, erhalten wir die Formel

$$r(x,\, x_*) = \sum_{j=1}^{n} \big(\partial_{x_j} f(\hat{\xi}_j) - \partial_{x_j} f(x_*)\big)\, (x_j - x_{*,\,j})\,,$$

aus der sich wegen $|x_j - x_{*,\,j}| \leq |x - x_*|$ die Abschätzung

$$\frac{\big|r(x,\, x_*)\big|}{\big|x - x_*\big|} \leq \sum_{j=1}^{n} \Big|\partial_{x_j} f(\hat{\xi}_j) - \partial_{x_j} f(x_*)\Big|$$

ergibt. Die rechte Seite hängt nicht nur von x_*, sondern auf komplizierte Art — nämlich indirekt über die Zwischenstellen $\hat{x}_1, \ldots, \hat{x}_n$ — auch von x ab. Sie verschwindet aber im Limes $x \to x_*$, da dann jede der Zahlen $\hat{x}_j$ gegen $x_{*,j}$ und jeder der Vektoren $\hat{\xi}_j$ gegen x_* konvergiert und weil alle partiellen Ableitungen von f nach Voraussetzung stetig in x_* sind. Damit haben wir die Behauptung für skalare Funktionen gezeigt.

allgemeiner Fall: Wir können die Argumente des Spezialfalls für jede Komponente f_i von f wiederholen und erhalten

$$\frac{\big|r_i(x,\, x_*)\big|}{\big|x - x_*\big|} \quad \xrightarrow{x \to x_*} \quad 0\,,$$

für jedes $i \in \{1, \ldots, m\}$, wobei

$$r_i(x,\, x_*) := f_i(x) - f_i(x_*) - \sum_{j=1}^{n} \partial_{x_j} f_i(x_*)(x_j - x_{*,j})$$

gerade die i-te Komponente des vektoriellen Restgliedes

$$r(x,\, x_*) = f(x) - f(x_*) - \mathrm{Jac} f(x_*)\,\big(x - x_*\big)$$

ist. Hieraus ergibt sich die Behauptung auch für vektorwertige Funktionen f. □

Bemerkung Die Voraussetzung im Theorem impliziert außerdem

$$\big\|\mathrm{D}f(x) - \mathrm{D}f(x_*)\big\|_{\mathrm{op}} \quad \xrightarrow{x \to x_*} \quad 0$$

in jedem Punkt $x_* \in U$, d.h. f ist auch stetig total differenzierbar in U. Umgekehrt können wir leicht zeigen, dass die Stetigkeit der totalen Ableitung in allen Punkten die Stetigkeit aller partiellen Ableitungen nach sich zieht. Wir werden daher im Folgenden von *stetig differenzierbaren* Funktionen sprechen und den Zusatz *partiell* oder *total* weglassen.

Zusammenfassung Die partielle und die totale Differenzierbarkeit sind zwei verschiedene Konzepte, wobei aus theoretischer Sicht die totale besser als die partielle ist und auch leichter verallgemeinert werden kann (etwa auf Abbildungen zwischen gekrümmten Flächen). Die beiden Eigenschaften *stetig partiell differenzierbar* und *stetig total differenzierbar* sind aber äquivalent und bei den in Anwendungen auftretenden Funktionen in aller Regel erfüllt. Insbesondere beweisen wir die Existenz der totalen Ableitung von f meist dadurch, dass wir die Existenz *und* Stetigkeit aller partiellen Ableitungen zeigen. Dies gelingt vor allem dann, wenn wir explizite Formeln für alle Einträge in der Jacobi-Matrix für f angeben können.

2.4 allgemeine Form der Kettenregel

Theorem (**Kettenregel in höheren Dimensionen**) Seien $U \subseteq \mathbb{R}^n$ und $V \subseteq \mathbb{R}^m$ zwei offene Mengen und seien $f : U \to V$ bzw. $g : V \to \mathbb{R}^l$ zwei Funktionen, die in $x_* \in U$ bzw. $f(x_*)$ total differenzierbar sind. Dann ist $g \circ f : U \to \mathbb{R}^l$ in x_* total differenzierbar und es gilt

$$\mathrm{D}\big(g \circ f\big)(x_*) = \big(\mathrm{D}g\big(f(x_*)\big)\big) \circ \mathrm{D}f(x_*)$$

für die totalen Ableitungen sowie

$$\mathrm{Jac}\,\big(g \circ f\big)(x_*) = \mathrm{Jac}\,g\big(f(x_*)\big)\ \mathrm{Jac}f(x_*)$$

für die entsprechenden Jacobi-Matrizen.[28]

Beweis Teil 1: Nach Voraussetzung gelten die beiden Darstellungsformeln

$$f(x) = f(x_*) + \big(\mathrm{D}f(x_*)\big)(x - x_*) + r(x,\, x_*)$$

und

$$g(y) = g(y_*) + \big(\mathrm{D}g(y_*)\big)(y - y_*) + s(y,\, y_*)\,,$$

mit $y_* = f(x_*)$, wobei die Restglieder $r(x,\, x_*) \in \mathbb{R}^m$ und $s(y,\, y_*) \in \mathbb{R}^l$ wohldefiniert sind und den Bedingungen

$$\frac{\big|r(x,\, x_*)\big|}{\big|x - x_*\big|} \quad \xrightarrow{x \to x_*} \quad 0\,, \qquad\qquad \frac{\big|s(y,\, y_*)\big|}{\big|y - y_*\big|} \quad \xrightarrow{x \to x_*} \quad 0$$

genügen. Wir werten nun die zweite Darstellungsformel mit $y = f(x)$ aus und setzen dabei die erste Darstellungsformel im Argument der linearen Abbildung $\mathrm{D}g(y_*)$ ein. Dies liefert die dritte Darstellungsformel

$$\big(g \circ f\big)(x) = \big(g \circ f\big)(x_*) + \Big(\mathrm{D}g\big(f(x_*)\big) \circ \mathrm{D}f(x_*)\Big)(x - x_*) + u(x,\, x_*) + v(x,\, x_*)\,,$$

wobei

$$u(x,\, x_*) = \big(\mathrm{D}g\big(f(x_*)\big)\big)\big(r(x,\, x_*)\big)\,, \qquad v(x,\, x_*) = s\big(f(x),\, f(x_*)\big)$$

zusammen das Restglied für $g \circ f$ darstellen. Im zweiten Teil des Beweises zeigen wir, dass der euklidische Betrag beider Terme im Limes $x \to x_*$ schneller gegen 0 konvergiert als $|x - x_*|$. Wenn uns dies gelungen ist, folgt die totale Differenzierbarkeit von $g \circ f$ sowie die erste Formel in der Behauptung direkt aus der dritten Darstellungsformel. Die entsprechende Formel für die Jacobi-Matrizen ergibt sich anschließend aus einem bekannten Satz der *Linearen Algebra* über die Komposition linearer Abbildungen und die Mutliplikation ihrer Matrix-Darstellungen.

Teil 2: Die obigen Resultate implizieren für das erste der neuen Restglieder die gewünschte Abschätzung via

$$\frac{\big|u(x,\, x_*)\big|}{\big|x - x_*\big|} \leq \frac{\big\|\mathrm{D}g\big(f(x_*)\big)\big\|_{\mathrm{op}}\,\big|r(x,\, x_*)\big|}{\big|x - x_*\big|} \quad \xrightarrow{x \to x_*} \quad 0\,.$$

[28] Auf der rechten Seite dieser Formel steht das Produkt einer $l \times m$-Matrix mit einer $m \times n$-Matrix und damit wirklich eine $l \times n$-Matrix

Die erste Darstellungsformel garantiert aber auch

$$\big|f(x) - f(x_*)\big| = \big|\big(\mathrm{D}f(x_*)\big)\big(x - x_*\big) + r(x,\, x_*)\big| \leq \|\mathrm{D}f(x_*)\|_{\mathrm{op}}\,\big|x - x_*\big| + \big|r(x,\, x_*)\big|$$

und wir erhalten

$$\begin{aligned}\frac{\big|v(x,\, x_*)\big|}{\big|x - x_*\big|} &= \frac{\big|s\big(f(x),\, f(x_*)\big)\big|}{\big|f(x) - f(x_*)\big|}\,\frac{\big|f(x) - f(x_*)\big|}{\big|x - x_*\big|} \\ &\leq \frac{\big|s\big(f(x),\, f(x_*)\big)\big|}{\big|f(x) - f(x_*)\big|}\left(\big\|\mathrm{D}f(x_*)\big\|_{\mathrm{op}} + \frac{\big|r(x,\, x_*)\big|}{\big|x - x_*\big|}\right).\end{aligned}$$

Da mit $x \to x_*$ auch $f(x) \to f(x_*)$ gilt, ergibt sich die zweite gewünschte Konvergenzaussage aus den obigen Eigenschaften der Restglieder $r(x,\, x_*)$ und $s(y,\, y_*)$. □

Korollar (**Ableitung der Umkehrfunktion**) Seien U, V zwei offene Mengen im $\mathbb{R}^n$ und $f : U \to V$ eine bijektive Abbildung mit Umkehrabbildung $f^{-1} : V \to U$. Außerdem sei f in x_* und f^{-1} in $f(x_*)$ total differenzierbar. Dann gilt

$$\mathrm{D}f^{-1}\big(f(x_*)\big) = \big(\mathrm{D}f(x_*)\big)^{-1} \quad \text{sowie} \quad \mathrm{Jac}\, f^{-1}\big(f(x_*)\big) = \big(\,\mathrm{Jac}\, f(x_*)\big)^{-1},$$

d.h. $\mathrm{D}f^{-1}\big(f(x_*)\big)$ ist die Umkehrabbildung von $\mathrm{D}f(x_*)$ und $\mathrm{Jac}\, f^{-1}\big(f(x_*)\big)$ ist die inverse Matrix zu $\mathrm{Jac}\, f(x_*)$.

Beweis Alle Behauptungen ergeben sich mit $m = n$ und $g = f^{-1}$ direkt aus der Kettenregel. In der Tat, es gilt $\big(g \circ f\big)(x) = x$ für jedes $x \in U$ und in Kombination mit direkten Rechnungen zeigen wir, dass damit $\mathrm{Jac}\,\big(g \circ f\big)(x)$ in jedem Punkt aus U die n-dimensionale Einheitsmatrix ist. □

Bemerkungen

1. Die Kettenregel ist die mit Abstand wichtigste Formel der Differentialrechnung und besitzt mannigfaltige Anwendungen in allen Bereichen der Mathematik und den Naturwissenschaften. Sie gilt immer bei totaler Differenzierbarkeit, aber nicht unbedingt bei partieller.

2. Die Ableitungsregel für die Umkehrfunktion ist auch sehr wichtig. Wir werden sie zum Beispiel weiter unten bei der Diskussion ebener Polarkoordinaten verwenden. Im Korollar haben wir neben der Differenzierbarkeit von f in x_* auch explizit die Differenzierbarkeit von f^{-1} im Punkt $f(x_*)$ gefordert. Wir werden später im *Satz über die Umkehrfunktion* sehen, dass diese zweite Bedingung wirklich äquivalent zur Invertierbarkeit von $\mathrm{D}f(x_*)$ bzw. $\mathrm{Jac}f(x_*)$ ist.

3. Neben der Kettenregel gibt es auch eine Summenregel für totale Ableitungen. Wir haben diese hier nicht explizit formuliert und bewiesen, da sie sich direkt aus der Definition und analog zur Diskussion bei partiellen Ableitungen ergibt.

4. Es existieren auch Varianten der Produkt- und der Quotientenregel (siehe die Übungen), aber da $\mathbb{R}^m$ für $m \neq 1$ kein Körper ist, gibt es kein direktes Analogon zu den Resultaten aus *Analysis 1*.[29]

[29] Mittels komplexer Zahlen kann eine Multiplikation und Division in $\mathbb{R}^2$ definiert werden und dies liefert einen weiteren Ableitungsbegriff, nämlich die *komplexe Differenzierbarkeit*. Wir werden die entsprechende Theorie in der Vorlesung *Funktionentheorie* ausführlich studieren.

5. Es gilt auch die folgende Verschärfung des Theorems: Sind g und f sogar stetig differenzierbar auf U, so ist auch $g \circ f$ stetig differenzierbar auf U.

   Beweis: Die Kettenregel impliziert, dass jede partielle Ableitung von $g \circ f$ existiert und gemäß der Matrizenmultiplikation als Summe von Produkten geschrieben werden kann. Jeder der beteiligten Faktoren entspricht dabei einer stetigen Funktion auf U, denn er ist entweder der Bauart $\partial_{y_i} g \circ f$ oder der Bauart $\partial_{x_j} f$. Die Rechenregeln für stetige Funktionen garantieren, dass auch jede partielle Ableitung von $g \circ f$ stetig auf U ist. □

Alternative Notation In den Anwendungswissenschaften setzt man gerne $y = f(x)$ sowie $z = g(y)$ und formuliert die Kettenregel als

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x},$$

wobei rechts das Produkt zweier Jacobi-Matrizen steht. Diese Schreibweise ist sehr intuitiv und verallgemeinert die eindimensionale Kettenregel

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y} \frac{\mathrm{d}y}{\mathrm{d}x},$$

in natürlicher Weise auf höhere Dimensionen. Beachte aber, dass die Reihenfolge der Faktoren in höheren Dimensionen sehr wichtig ist, da andernfalls die Matrizenmultiplikation nicht definiert sein muss oder ein falsches Ergebnis liefern könnte. Die Kettenregel kann auch komponentenweise als

$$\frac{\partial z_k}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial z_k}{\partial y_i} \frac{\partial y_i}{\partial x_j}$$

geschrieben werden,[30] wobei die Faktoren auf der rechten Seite diesmal sogar vertauscht werden dürfen (denn die Multiplikation von Zahlen ist im Gegensatz zur Multiplikation von Matrizen kommutativ).

Beispiele

1. Bei der Anwendung der Kettenregel empfiehlt es sich auch in der Mathematik oftmals, die Variablen der Funktionen f bzw. g mit unterschiedlichen Buchstaben zu bezeichnen, zum Beispiel mit x bzw. y. Für $n = 3$, $m = 2$ und $l = 3$ betrachten wir exemplarisch die Funktionen $f : \mathbb{R}^3 \to \mathbb{R}^2$ und $g : \mathbb{R}^2 \to \mathbb{R}^3$ mit

   $$f_1(x_1, x_2, x_3) = x_1 + x_2 + x_3, \qquad f_2(x_1, x_2, x_3) = x_1 x_2 x_3$$

   und

   $$g_1(y_1, y_2) = y_1 + y_2, \qquad g_2(y_1, y_2) = y_1^2, \qquad g_3(y_1, y_2) = y_2^2.$$

   Wir können die verkettete Abbildung $g \circ f : \mathbb{R}^3 \to \mathbb{R}^3$ in diesem recht überschaubaren Beispiel natürlich mit der Substitution

   $$y_1 = f_1(x_1, x_2, x_3), \qquad y_2 = f_2(x_1, x_2, x_3)$$

[30] Unter der *Einsteinschen Summenkonvention* kann das Summenzeichen sogar weggelassen werden, da in diesem Kalkül immer über doppelt auftretende Indizes (hier i) automatisch summiert wird. Wir wollen das in dieser Vorlesung aber *niemals* tun.

direkt ermitteln und anschließend komponentenweise nach den Variablen x_j differenzieren. Dies liefert

$$\big(g\circ f\big)(x) = \begin{pmatrix} x_1+x_2+x_3+x_1\,x_2\,x_3 \\ (x_1+x_2+x_3)^2 \\ x_1^2\,x_2^2\,x_3^2 \end{pmatrix}$$

sowie

$$\operatorname{Jac}\big(g\circ f\big)(x) = \begin{pmatrix} 1+x_2\,x_3 & 1+x_1\,x_3 & 1+x_1\,x_2 \\ 2\,x_1+2\,x_2+2\,x_3 & 2\,x_1+2\,x_2+2\,x_3 & 2\,x_1+2\,x_2+2\,x_3 \\ 2\,x_1\,x_2^2\,x_3^2 & 2\,x_1^2\,x_2\,x_3^2 & 2\,x_1^2\,x_2^2\,x_3 \end{pmatrix}.$$

Alternativ können wir aber auch die Kettenregel mit

$$\operatorname{Jac} f(x) = \begin{pmatrix} 1 & 1 & 1 \\ x_2\,x_3 & x_1\,x_3 & x_1\,x_2 \end{pmatrix}, \qquad \operatorname{Jac} g(y) = \begin{pmatrix} 1 & 1 \\ 2\,y_1 & 0 \\ 0 & 2\,y_2 \end{pmatrix}$$

verwenden. Durch Substitution erhalten wir zunächst

$$\operatorname{Jac} g\big(f(x)\big) = \begin{pmatrix} 1 & 1 \\ 2\,x_1+2\,x_2+2\,x_3 & 0 \\ 0 & 2\,x_1\,x_2\,x_3 \end{pmatrix}$$

und anschließend via

$$\operatorname{Jac}\big(g\circ f\big)(x) = \begin{pmatrix} 1 & 1 \\ 2\,x_1+2\,x_2+2\,x_3 & 0 \\ 0 & 2\,x_1\,x_2\,x_3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ x_2\,x_3 & x_1\,x_3 & x_1\,x_2 \end{pmatrix}$$

dieselbe 3×3-Matrix wie vorher.

2. Wir wollen ein Beispiel mit $n=2$, $m=3$, $l=2$ in Physikernotation rechnen, d.h. es gilt $x\in\mathbb{R}^2$, $y\in\mathbb{R}^3$ und $z\in\mathbb{R}^2$. Anstelle expliziter Funktionen f bzw. g betrachten wir die Formelsätze

$$y_1 = x_1+x_2\,, \qquad y_2 = x_1-x_2\,, \qquad y_3 = x_1\,x_2$$

bzw.

$$z_1 = y_1+y_2-y_3\,, \qquad z_2 = -y_3^2\,,$$

die y durch x bzw. z durch y ausdrücken. Die Kettenregel kann nun als

$$\begin{aligned} \begin{pmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} \end{pmatrix} &= \begin{pmatrix} \frac{\partial z_1}{\partial y_1} & \frac{\partial z_1}{\partial y_2} & \frac{\partial z_1}{\partial y_3} \\ \frac{\partial z_2}{\partial y_1} & \frac{\partial z_2}{\partial y_2} & \frac{\partial z_2}{\partial y_3} \end{pmatrix} \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \\ \frac{\partial y_3}{\partial x_1} & \frac{\partial y_3}{\partial x_2} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 & -1 \\ 0 & 0 & -2\,y_3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ x_2 & x_1 \end{pmatrix} \\ &= \begin{pmatrix} 2-x_2 & -x_1 \\ -2\,x_2\,y_3 & -2\,x_1\,y_3 \end{pmatrix} = \begin{pmatrix} 2-x_2 & -x_1 \\ -2\,x_1\,x_2^2 & -2\,x_1^2\,x_2 \end{pmatrix} \end{aligned}$$

geschrieben werden. Alternativ (oder als Probe) können wir dieses Ergebnis auch direkt aus

$$z_1 = 2\,x_1 - x_1\,x_2\,, \qquad z_2 = -x_1^2\,x_2^2$$

durch Ableiten von z_k nach x_j gewinnen.

3. Mit der Kettenregel können auch skalare oder vektorwertige Funktionen entlang von Kurven differenziert werden. Zum Beispiel ändert sich das dreidimensionale Vektorfeld

$$f(x) = \begin{pmatrix} x_1 + x_3 \\ x_2 - x_3 \\ x_1^2 + x_2^2 \end{pmatrix}, \quad x \in \mathbb{R}^3$$

entlang der parametrisierten Kurve

$$\gamma(t) = \begin{pmatrix} \cos(t) \\ \sin(t) \\ t \end{pmatrix}, \qquad t \in \mathbb{R}$$

gemäß der Rechnung

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} f\big(\gamma(t)\big) &= \mathrm{Jac} f\big(\gamma(t)\big)\ \mathrm{Jac}\,\gamma(t) \\ &= \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 2\cos(t) & 2\sin(t) & 0 \end{pmatrix} \begin{pmatrix} -\sin(t) \\ +\cos(t) \\ 1 \end{pmatrix} = \begin{pmatrix} -\sin(t) + 1 \\ +\cos(t) - 1 \\ 0 \end{pmatrix}, \end{aligned}$$

wobei $\mathrm{Jac}\,\gamma(t) = \dot\gamma(t)$ gerade den Tangentialvektor im Punkt $\gamma(t)$ beschreibt. Alternativ können wir auch hier erst $x = \gamma(t)$ in der Formel für $f(x)$ substituieren und anschließend nach der skalaren Variablen t differenzieren. Diese Beobachtung wird bei den Richtungsableitungen wichtig werden.

Richtungsableitungen und Gradienten

Definition Sei $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ in $x_* \in U$ total differenzierbar und sei $v \in \mathbb{R}^n$ beliebig mit $v \neq 0$. Dann nennen wir

$$\mathrm{D}_v f(x_*) = \big(\mathrm{D} f(x_*)\big)(v) = \mathrm{Jac} f(x_*)\,v$$

die Richtungsableitung von f im Punkt x_* in Richtung v.

Lemma (**Bedeutung der Richtungsableitung**) Mit den obigen Notationen gilt

$$\mathrm{D}_v f(x_*) = \lim_{t \to t_*} \frac{f(\gamma(t)) - f(\gamma(t_*))}{t - t_*}$$

für jede stetig differenzierbare Kurve $\gamma : I \to U$, die via

$$\gamma(t_*) = x_*\,, \qquad \dot\gamma(t_*) = v$$

zur Zeit t_* in Richtung v durch den Punkt x_* läuft.

Beweis Die Abbildung $\eta = f \circ \gamma : I \to \mathbb{R}^m$ ist auch eine Kurve und die Kettenregel (siehe auch die Rechnungen im Beispiel oben) garantiert ihre Differenzierbarkeit in t_* sowie die Formel

$$\dot\eta(t_*) = \operatorname{Jac}\eta(t_*) = \operatorname{Jac} f\big(\gamma(t_*)\big)\ \operatorname{Jac}\gamma(t_*) = \operatorname{Jac} f\big(\gamma(t_*)\big)\,\dot\gamma(t_*)\,,$$

wobei die rechte Seite gerade $\mathrm{D}_v f(x_*)$ ist. Andererseits gilt

$$\lim_{t\to t_*} \frac{f(\gamma(t)) - f(\gamma(t_*))}{t - t_*} = \lim_{t\to t_*} \frac{\eta(t) - \eta(t_*)}{t - t_*} = \dot\eta(t_*)$$

und die Behauptung folgt unmittelbar. □

Bemerkungen

1. Für den j-ten Einheitsvektor e_j im $\mathbb{R}^n$ ergibt sich

$$\mathrm{D}_{e_j} f(x_*) = \begin{pmatrix} \partial_{x_j} f_1(x_*) \\ \vdots \\ \partial_{x_j} f_m(x_*) \end{pmatrix},$$

   d.h. die partiellen Ableitungen der Komponenten von f nach x_j liefern zusammen gerade die Richtungsableitung in Richtung $v = e_j$. Diese Formel kann dabei aus der Definition abgelesen oder alternativ aus dem Lemma mit

$$\gamma(t) = x_* + (t - t_*)\, e_j$$

   hergeleitet werden.

2. Für eine skalare Funktion ($m = 1$) ist $\mathrm{D}_v f(x_*)$ immer eine reelle Zahl, sonst immer ein Vektor aus dem $\mathbb{R}^m$. Beachte auch, dass wir den Nullvektor in der Definition und im Lemma ausgeschlossen haben. Alle Formeln gelten aber auch in diesem Entartungsfall mit $\mathrm{D}_0 f(x_*) = 0$.

3. Die Definition impliziert

$$\mathrm{D}_{\lambda v} f(x_*) = \lambda\, \mathrm{D}_v f(x_*)$$

   für jeden skalaren Faktor $\lambda \in \mathbb{R}$. Aus diesem Grund wird die Richtungsableitung oftmals nur für *normierte Vektoren* $\nu \in \mathbb{R}^n$ mit $|\nu| = 1$ ausgewertet bzw. betrachtet. Beachte, dass für jedes $v \neq 0$ der Vektor $\nu = v/|v|$ normiert ist.

4. Ist f nur partiell, aber nicht total differenzierbar, so ist $\operatorname{Jac} f(x_*)\, v$ zwar immer noch definiert, aber es ist nicht sichergestellt, dass diese Größe der Limes von entsprechenden Differenzenquotienten ist. Siehe dazu das Gegenbeispiel.

Gegenbeispiel Wir betrachten die skalare Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ mit

$$f(x_1, x_2) = \begin{cases} 0 & \text{für } x_1 = 0 = x_2 \\ \dfrac{x_1^3}{x_1^2 + x_2^2} & \text{sonst} \end{cases}$$

und mithilfe der Definition partieller Ableitungen können wir im Ursprung $x_* = (0,\, 0)$ die Formel

$$\mathrm{Jac} f(x_*) = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

zeigen (siehe die Übungen). Für gegebenes $v \in \mathbb{R}^2$ mit $v \neq 0$ betrachten wir die Kurve $\gamma : \mathbb{R} \to \mathbb{R}^2$ mit

$$\gamma(t) = x_* + t\, v = \begin{pmatrix} t\, v_1 \\ t\, v_2 \end{pmatrix}$$

und zur Zeit $t_* = 0$ erhalten wir

$$\lim_{t \to t_*} \frac{f\big(\gamma(t)\big) - f\big(\gamma(t_*)\big)}{t - t_*} = \lim_{t \to 0} \frac{\dfrac{t^3 v_1^3}{t^2\, v_1^2 + t^2 v_2^2}}{t} = \frac{v_1^3}{v_1^2 + v_2^2}\,,$$

was aber nur für $v_2 = 0$ mit

$$\mathrm{Jac} f(x_*) \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = v_1$$

übereinstimmt. Insbesondere gelten in diesem Beispiel die Aussagen des Lemmas nicht für $x_* = 0$.

Erklärung: Die Funktion f besitzt im Punkt x_* alle partiellen Ableitungen und wir können sogar alle Richtungsableitungen als Ableitungen entlang geeigneter Kurven berechnen. Die Funktion f ist aber in x_* nicht total differenzierbar. Dies sieht man zum Beispiel daran, dass die partiellen Ableitungen in x_* nicht stetig sind (siehe wieder die Übungen).

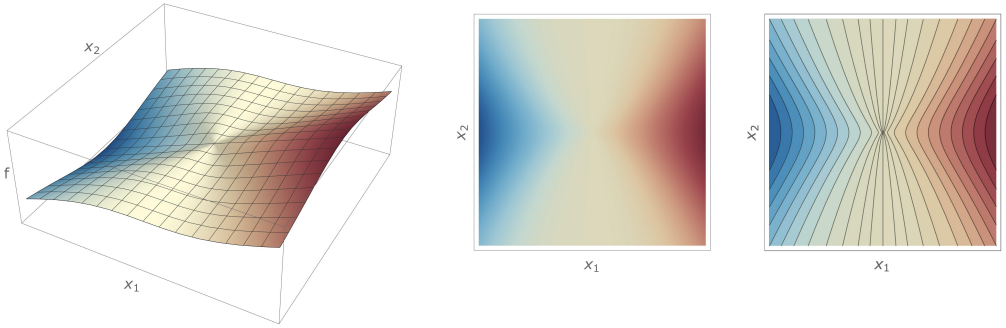

Abbildung Die skalare Funktion aus dem Gegenbeispiel ist im Ursprung — der in allen Plots dem Mittelpunkt entspricht — partiell, aber nicht total differenzierbar, wobei man dies im Flächenplot und im Konturplot mit etwas Erfahrung auch visuell erkennen kann. In jedem anderen Punkt liegt jedoch totale Differenzierbarkeit vor und das Lemma über die Richtungsableitungen kann gefahrlos angewendet werden.

Lemma (**Gradient als Richtung des steilsten Anstiegs**) Die skalare Funktion $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ sei im Punkt x_* total differenzierbar. Dann gilt

$$-\mathrm{D}_{\nu_*} f(x_*) \leq \mathrm{D}_\nu f(x_*) \leq +\mathrm{D}_{\nu_*} f(x_*)$$

für jeden normierten Vektor $\nu \in \mathbb{R}^n$ mit $|\nu| = 1$, wobei

$$\nu_* = \frac{\mathrm{grad}\, f(x_*)}{\big|\mathrm{grad}\, f(x_*)\big|}$$

der normierte Gradient von f in x_* ist.

Beweis Die Definitionen der Richtungsableitung und des Gradienten implizieren

$$\mathrm{D}_\nu f(x_*) = \mathrm{Jac} f(x_*)\nu = \operatorname{grad} f(x_*) \cdot \nu = v_* \cdot \nu = |v_*|\ \nu_* \cdot \nu\,,$$

sowie den Spezialfall

$$\mathrm{D}_{\nu_*} f(x_*) = |v_*|\ \nu_* \cdot \nu_* = |v_*|\ ,$$

wobei $\cdot$ das Skalarprodukt im $\mathbb{R}^n$ bezeichnet und wir die Abkürzung $v_* = \operatorname{grad} f(x_*)$ verwendet haben. Die Cauchy-Schwarz-Ungleichung angewendet auf ν_* und ν garantiert

$$-1 \leq \nu_* \cdot \nu \leq +1$$

und die Behauptung folgt nach Multiplikation mit $|v_*|$. □

Theorem (**Gradient und Niveaumengen**) Sei $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ im Punkt x_* total differenzierbar. Dann gilt

$$\operatorname{grad} f(\gamma(t_*)) \cdot \dot\gamma(t_*) = 0$$

für jede stetig differenzierbare Kurve $\gamma : I \to \mathbb{R}^n$, die zur Zeit $t_* \in I$ durch den Punkt x_* läuft und deren Bild vollständig in der Niveaumenge

$$N_f\big(f(x_*)\big) = \big\{x \in U\ : f(x) = f(x_*)\big\}$$

enthalten ist.

Beweis Wir können $v := \dot\gamma(t_*) \neq 0$ annehmen, da andernfalls nichts zu zeigen ist. Nach Voraussetzung gilt $x_* = \gamma(t_*)$ sowie $f\big(\gamma(t)\big) = f\big(\gamma(t_*)\big)$ für alle $t \in I$ und das Lemma zur Bedeutung der Richtungsableitung impliziert

$$\mathrm{D}_v f(x_*) = 0\,.$$

Andererseits folgt

$$\mathrm{D}_v f(x_*) = \mathrm{Jac}\, f(x_*)v = \operatorname{grad} f(x_*) \cdot v$$

aus den Definitionen von Richtungsableitung und Gradient. □

Bemerkungen

1. Das Lemma garantiert vereinfacht gesprochen, dass der Gradient von f in jedem Punkt x_* in die Richtung zeigt, in der die Funktion f *lokal* — das heißt aus Sicht einer kleinen Umgebung von x_* — am stärksten wächst. Das bedeutet aber nicht, dass der Gradient von f in jedem Punkt zu einem lokalen oder gar globalen Maximierer von f zeigt.[31] Analog zum Beweis des Lemmas können wir zeigen, dass der negative Gradient die Richtung des lokal stärksten Abfallens zeigt.

[31] Eine Analogie beim Bergwandern: Der Gradient ist sowas wie ein aktueller Wegweiser. Er zeigt immer bergauf, aber wenn Sie ein kleines Stück in die angezeigte Richtung gelaufen sind, gibt es einen neuen Wegweiser, der Sie nun in eine leicht andere Richtung schickt. Insgesamt gewinnen Sie zwar immer an Höhe, aber ihr Weg wird in aller Regel gekrümmt und verschlungen sein, da Sie ja die Marschrichtung ständig anpassen. Die eben beschriebene Wanderstrategie ist ein Beispiel für ein *Gradientenverfahren*. Mit ihr erreichen sie immer einen Gipfel und dort wird der Wegweiser den Nullvektor anzeigen. Es gibt aber leider keine Garantie, dass dieser der höchste aller Gipfel sein wird.

2. Salopp gesprochen besagt das Theorem, dass der Gradient von f im Punkt x_* immer senkrecht auf der Niveaumenge $N_f(f(x_*))$ bzw. auf dem entsprechenden *Tangentialraum* steht. Wir werden dies in *Analysis 3* genauer untersuchen, wollen aber schon vorwegnehmen, dass der Tangentialraum in jedem Punkt x_* ein linearer bzw. affiner Raum der Dimension $n-1$ ist.

3. Im Fall von $n = 2$ kann die Aussage des Theorems sehr gut visualisiert und damit auch intuitiv verstanden werden, da dann die Niveaumenge von f meist eine Niveaukurve ist, die in jedem Punkt eine Tangentialgerade besitzt (Ausnahmen bestätigen hier die Regel). Für $n = 3$ handelt es sich um eine gekrümmte Fläche im dreidimensionalen Raum bzw. um eine Tangentialebene, aber auch hier gibt es noch eine geometrische Anschauung.

4. Die von uns bewiesenen Eigenschaften des Gradienten einer skalaren Funktion werden Sie Ihr ganzes Studium begleiten, denn sie bilden den Ausgangspunkt für sehr viele Resultate und Approximationsverfahren in der *Optimierung* sowie der *Numerischen Mathematik*. Sie sollten daher unbedingt versuchen, diese Konzepte sowohl auf der heuristischen als auch auf der formalen Ebene zu verstehen.

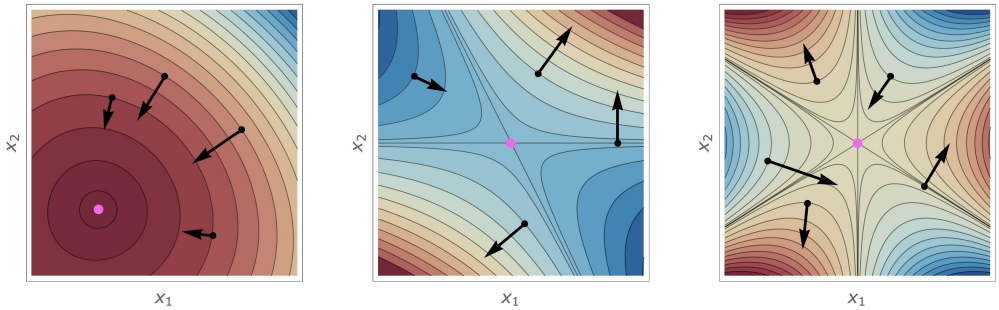

Abbildung Konturplots dreier Funktionen $f : \mathbb{R}^2 \to \mathbb{R}$ sowie der jeweilige Gradient $\operatorname{grad} f(x_*)$ in ausgewählten Punkten x_* (schwarz). Dieser Vektor charakterisiert das *lokale* Verhalten der Funktion f in der Nähe von x_*, jedoch nicht ihre globalen Eigenschaften. Jeder der drei lila Punkte ist *kritisch*, da dort der Gradient von f jeweils der Nullvektor ist. Es handelt sich um einen lokalen Minimierer, einen (Standard-)Sattelpunkt bzw. einen Affensattel.

Vorlesungswoche 07

Einschub: Matrizen und ihre euklidische Norm Durch die Formel

$$\big|A\big| := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} \qquad \text{für} \quad A = \begin{pmatrix} A_{11} & \dots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \dots & A_{mn} \end{pmatrix}$$

wird in natürlicher Weise ein Betrag (und damit eine Norm) auf dem Vektorraum aller reellen $m\times n$-Matrizen definiert.[32]

Lemma (Kompatibilitätssatz): Für jede $m\times n$-Matrix A und jede $l\times m$-Matrix B gilt

$$\big|B\,A\big| \le \big|B\big|\,\big|A\big|\,,$$

d.h. die Norm des Matrizenproduktes ist niemals größer als das Produkt der Normen.

Beweis: Die k-te Zeile von B definiert einen Vektor $b_k \in \mathbb{R}^m$ und analog sei $a_j \in \mathbb{R}^m$ die j-te Spalte von A. In der k-ten Zeile und j-ten Spalte der Produktmatrix $C = B\,A$ steht die reelle Zahl $C_{k\,j} = b_k \cdot a_j$, wobei $\cdot$ das Skalarprodukt im $\mathbb{R}^m$ ist. Die Cauchy-Schwarz-Ungleichung und sowie die unsere Definitionen implizieren

$$C_{kj}^2 = \big(b_k \cdot a_j\big)^2 \le \big|b_k\big|^2\,\big|a_j\big|^2\,, \qquad |b_k|^2 = \sum_{i=1}^{m} B_{ki}^2\,, \qquad |a_j|^2 = \sum_{i=1}^{m} A_{ij}^2$$

und wir erhalten mit

$$|C|^2 \le \sum_{k=1}^{l}\sum_{j=1}^{n} \big|b_k\big|^2\,\big|a_j\big|^2 = \left(\sum_{k=1}^{l}\big|b_k\big|^2\right)\left(\sum_{j=1}^{n}\big|a_j\big|^2\right) = \big|B\big|^2\,\big|A\big|^2$$

das gewünschte Ergebnis nach dem Ziehen der Wurzel auf beiden Seiten. □

Bemerkungen:

1. Nicht jede Norm für Matrizen ist mit der Multiplikation kompatibel. Es gilt zwar immer $\|BA\| \le c\,\|A\|\,\|B\|$ für eine Konstante c, aber diese wird in der Regel größer als 1 und damit *nicht kompatibel* sein.

2. Analog zum Beweis können wir zeigen, dass $|A\,x| \le |A|\,|x|$ für alle $x \in \mathbb{R}^n$ und $|B\,y| \le |B|\,|y|$ für alle $y \in \mathbb{R}^m$ gilt. Alternativ kann man diese Aussagen auch als Spezialfall des Lemmas ansehen (sofern die Notationen entsprechend angepasst werden).

Mittelwertsatz in höheren Dimensionen

Theorem **(zwei Versionen des Mittelwertsatzes)** Sei $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ stetig differenzierbar. Außerdem seien x, x_* zwei Punkte in U, sodass die Verbindungsstrecke zwischen x und x_*, d.h. das Bild der parametrisierten Kurve $\gamma : [0,\,1] \to \mathbb{R}^n$ mit

$$\gamma(t) = x_* + t\,(x - x_*) = (1-t)\,x_* + t\,x\,,$$

[32] Die Indizierung der Matrixelemente erfolgt nach der bekannten Eselsbrücke: *Zeilen zuerst, Spalten später*. Die euklidische Norm einer Matrix wird auch *Frobenius-Norm* genannt.

ganz in U liegt. Dann gilt

$$f(x) - f(x_*) = J\,\big(x - x_*\big) \qquad \text{mit} \qquad J_{ij} := \int_0^1 \partial_{x_j} f_i\big(\gamma(t)\big)\,\mathrm{d}t$$

sowie

$$\big|f(x) - f(x_*)\big| \le C\,\big|x - x_*\big| \qquad \text{mit} \qquad C := \max\big\{\big|\mathrm{Jac}\, f\big(\gamma(t)\big)\big| \,:\, 0 \le t \le 1\big\}\,.$$

Im Fall von $m = 1$ gilt außerdem

$$f(x) - f(x_*) = \mathrm{grad}\, f\big(\gamma(\tau)\big) \cdot (x - x_*)$$

für mindestens ein $\tau \in [0,\, 1]$.

Beweis Da U offen ist, existiert ein offenes Intervall I mit $[0,\, 1] \subset I$, sodass $g(t) \in U$ für alle $t \in I$ gilt.[33] Die parametrisierte Kurve $\eta := f \circ \gamma : I \to \mathbb{R}^m$ ist nach Kettenregel stetig differenzierbar und es gilt

$$\dot\eta(t) = \mathrm{Jac}\, f\big(\eta(t)\big)\,(x - x_*) \qquad \text{und damit} \qquad \dot\eta_i(t) = \sum_{j=1}^n \partial_{x_j} f_i\big(\gamma(t)\big)\,(x_j - x_{*,\,j})$$

für alle $i \in \{1,\, \ldots,\, m\}$. Der Hauptsatz der Differential- und Integralrechung sowie die Rechenregeln für eindimensionale Integrale — siehe jeweils *Analysis 1* — liefert

$$\eta_i(1) - \eta_i(0) = \int_0^1 \dot\eta_j(t)\,\mathrm{d}t = \sum_{j=1}^n \left(\int_0^1 \partial_{x_j} f_i\big(\gamma(t)\big)\,\mathrm{d}t\right)(x_j - x_{*,\,j}) = \sum_{i=1}^n J_{ij}\,(x_j - x_{*,\,j})\,.$$

Das ist aber gerade die i-te Komponente der ersten Behauptung, den es gilt $\eta(1) = f(x)$ und $\eta(0) = f(x_*)$. Das Lemma über die untere Schranke für die Länge einer Kurve sowie die Eigenschaften der euklidischen Matrixnorm implizieren außerdem

$$\big|\eta(1) - \eta(0)\big| \le \mathrm{len}\,(\eta) \le \int_0^1 \big|\dot\eta(t)\big|\,\mathrm{d}t \le \int_0^1 \big|\mathrm{Jac}\, f\big(\eta(t)\big)\big|\,\big|x - x_*\big|\,\mathrm{d}t \le \int_0^1 C\,\big|x - x_*\big|\,\mathrm{d}t$$

und damit die zweite Behauptung (auf der rechten Seite hängt der Integrand nicht von t ab). Sei nun $m = 1$. Dann ist η eine skalare Funktion und nach dem Mittelwertsatz der Integralrechnung aus *Analysis 1* existiert $\tau \in [0,\, 1]$ mit $\int_0^1 \dot\eta(t)\,\mathrm{d}t = \dot\eta(\tau)$. Weil auch $\dot\eta(\tau) = \mathrm{grad}\, f\big(\gamma(\tau)\big) \cdot (x - x_*)$ gilt, folgt die dritte Behauptung. □

Bemerkungen

1. Die $m \times n$-Matrix J entsteht durch komponentenweise Integration der Jacobi-Matrix von $f$entlang der Kurve γ.

2. Bei einer vektorwertigen Funktion f können wir den skalaren Mittelwertsatz (also die dritte Behauptung im Theorem) natürlich auf jede Komponente f_i anwenden. Wir werden dann aber im Allgemeinen m verschiedene Werte τ_i erhalten.

[33] Wir können zum Beispiel $I = (-\varepsilon,\, 1 + \varepsilon)$ setzen, wobei $\varepsilon > 0$ so gewählt wird, dass die offenen Kugeln $B_\varepsilon(x_*)$ und $B_\varepsilon(x)$ beide ganz in U liegen. Die Existenz von I stellt sicher, dass $\gamma(t)$ auch in $t = 0$ und $t = 1$ stetig differenzierbar ist und wir nicht mit einseitigen Ableitungen arbeiten müssen.

3. Ist $K \subset U$ kompakt und konvex, so impliziert der Mittelwertsatz, dass f auf K Lipschitz stetig mit Lipschitz-Konstante $L := \max\big\{\big|\operatorname{Jac} f(x)\big| : x \in K\big\}$ ist, wobei die Kompaktheit von K und die stetige Differenzierbarkeit von f sicherstellen, dass L als reelle Zahl wohldefiniert ist. Diese Aussage wird Schrankensatz genannt und recht häufig in Beweisen verwendet.

4. Eine Menge $K \subset \mathbb{R}^n$ heißt dabei konvex, falls sie mit je zwei Punkten x, x_* auch alle Punkte der jeweiligen Verbindungsstrecke — also alle Punkte der Bauart $(1-t)\,x_* + t\,x$ mit $t \in [0,\,1]$ — enthält. Kugeln und Quader sind zum Beispiel immer konvex (egal ob offen oder abgeschlossen), Kreisringe hingegen nicht.

5. Aus dem Schrankensatz folgt insbesondere: Wenn alle Ableitungen von f auf der Menge K verschwinden, so ist f auf K konstant. *Ausblick*: Diese Aussage gilt nicht nur für konvexe Mengen, sondern auf jeder zusammenhängenden Menge $K \subset U$.

nichtlineare Variablenwechsel und Transformationsformeln*

Vorbemerkung Wir betrachten zwei offene Mengen $U, V \subseteq \mathbb{R}^n$ sowie zwei stetig differenzierbare Abbildungen $f : U \to \mathbb{R}^m$ und $u : V \to U$, wobei wir annehmen, dass die Umkehrfunktion $v = u^{-1} : U \to V$ existiert und auch stetig differenzierbar ist. Insbesondere können wir u bzw. v als nichtlineare Variablen- oder Koordinatenwechsel interpretieren und die Kettenregel liefert mit

$$\operatorname{Jac} f(x) = \operatorname{Jac} g\big(v(x)\big)\,\operatorname{Jac} v(x)\,, \qquad \operatorname{Jac} g(y) = \operatorname{Jac} f\big(u(y)\big)\,\operatorname{Jac} u(y)$$

Beziehungen zwischen den Ableitungen von f und von $g = f \circ u$. Aus abstrakter Sicht ist damit alles gesagt, aber wir wollen nun mithilfe der Physikernotation diese Formeln auf eine andere Art herleiten bzw. verstehen. Dazu schreiben wir

$$x = u(y)\,, \qquad y = v(x)\,, \qquad z = g(y) = f(x)\,,$$

d.h. wir bezeichnen Punkte in U bzw. V bzw. $\mathbb{R}^m$ mit x bzw. y bzw. z.

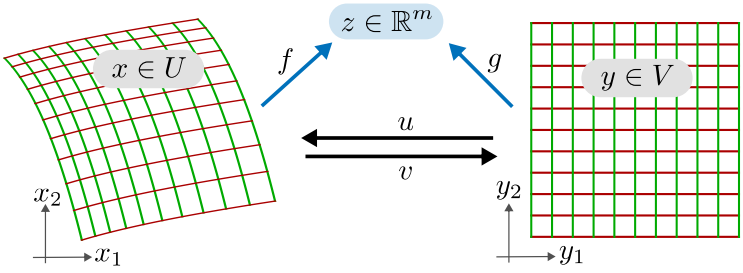

Abbildung Die Funktionen f und g beschreiben dieselbe Größe z, aber einmal durch die Variablen x und das andere Mal mittels der Variablen y. Die grünen bzw. roten Linien entsprechen auf beiden Seiten den Kurven $y_1 = \text{const}$ bzw. $y_2 = \text{const}$ (Niveau- oder Konturlinien der y_i).

Herleitung der allgemeinen Formeln Nach Konstruktion gilt

$$u\big(v(x)\big) = x \qquad \text{bzw.} \qquad v\big(u(y)\big) = y$$

und durch Differentiation nach x bzw. y erhalten wir mit der Kettenregel die Matrizengleichungen

$$\frac{\partial x}{\partial y}\,\frac{\partial y}{\partial x} = 1\,, \qquad \frac{\partial y}{\partial x}\,\frac{\partial x}{\partial y} = 1\,,$$

wobei $\mathbb{1}$ gerade die Einheitsmatrix (mit n Zeilen und n Spalten) ist. Oder anders gesagt: Die Jacobi-Matrix von u in y ist invers zur Jacobi-Matrix von v in x und diese Aussage kann komponentenweise als

$$\sum_{l=1}^{n} \frac{\partial x_k}{\partial y_l} \frac{\partial y_l}{\partial x_j} = \delta_j^k \,, \qquad \delta_j^k := \begin{cases} 1 & \text{für } j = k \\ 0 & \text{für } j \neq k \end{cases}$$

geschrieben werden, wobei δ_j^k das Kronecker-Delta ist. Die Kettenregel impliziert auch

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x} \,, \qquad \frac{\partial z}{\partial y} = \frac{\partial z}{\partial x} \frac{\partial x}{\partial y} \,,$$

wobei die linke Seite in der ersten bzw. zweiten Gleichung die Jacobi-Matrix von f bzw. g repräsentiert. Nach Übergang zur Komponentenschreibweise und nach Vertauschung von skalaren Faktoren erhalten wir

$$\frac{\partial z_i}{\partial x_j} = \sum_{l=1}^{n} \frac{\partial y_l}{\partial x_j} \frac{\partial z_i}{\partial y_l} \,, \qquad \frac{\partial z_i}{\partial y_j} = \sum_{l=1}^{n} \frac{\partial x_l}{\partial y_j} \frac{\partial z_i}{\partial x_l} \,.$$

Da diese Formeln für jedes z_i gelten, schreibt man oftmals auch

$$\frac{\partial}{\partial x_j} = \sum_{l=1}^{n} \frac{\partial y_l}{\partial x_j} \frac{\partial}{\partial y_l} \,, \qquad \frac{\partial}{\partial y_j} = \sum_{l=1}^{n} \frac{\partial x_l}{\partial y_j} \frac{\partial}{\partial x_l}$$

und erhält die Transformationsformel für die Differentialoperatoren erster Ordnung, mit der man bequem symbolisch rechnen kann.[34] Durch ähnliche Betrachtungen können wir auch zweite Ableitungen transformieren und erhalten

$$\frac{\partial^2}{\partial x_k \partial x_j} = \sum_{l=1}^{n} \frac{\partial}{\partial x_k} \left(\frac{\partial y_l}{\partial x_j} \frac{\partial}{\partial y_l} \right) = \sum_{l=1}^{n} \frac{\partial^2 y_l}{\partial x_k \partial x_j} \frac{\partial}{\partial y_l} + \sum_{l=1}^{n} \sum_{o=1}^{n} \frac{\partial y_l}{\partial x_j} \frac{\partial y_o}{\partial x_k} \frac{\partial^2}{\partial y_o \partial y_l}$$

sowie analog

$$\frac{\partial^2}{\partial y_k \partial y_j} = \sum_{l=1}^{n} \frac{\partial^2 x_l}{\partial y_k \partial y_j} \frac{\partial}{\partial x_l} + \sum_{l=1}^{n} \sum_{o=1}^{n} \frac{\partial x_l}{\partial y_j} \frac{\partial x_o}{\partial y_k} \frac{\partial^2}{\partial x_o \partial x_l}$$

durch Vertauschung der Buchstaben x und y.

Achtung: Wir müssen bei diesen Formeln sehr aufpassen, da sich leicht Fehler bei den Indizes einschleichen und Unsinn produzieren. Die Transformationsformeln für dritte oder vierte Ableitungen sehen natürlich noch komplizierter aus.

Polarkoordinaten in der Ebene Ein Beispiel mit $n = 2$ sind die zweidimensionalen Polarkoordinaten

$$x_1 = r \, \cos(\varphi) \,, \qquad x_2 = r \, \sin(\varphi) \,.$$

Die beiden kartesischen Variablen (oder Koordinaten) x_1 und x_2 werden dabei durch die Kombination aus Radius $r \geq 0$ und Winkel φ ersetzt, wobei diese an die Stelle von y_1 und y_2 treten und alle Formeln 2π-periodisch in φ sind.

[34] In die „Leerstelle“ bei $\partial \,/ \partial x_j$ usw. kann eine beliebige Größe eingesetzt werden, zum Beispiel z_i.

Bemerkung: Die Polarkoordinaten sind im kartesischen Ursprung $(x_1, x_2) = (0, 0)$ nicht definiert und in allen anderen Punkten (x_1, x_2) ist die Wahl des Winkels φ mehrdeutig. Die Polarkoordinaten können aber als Bijektion zwischen geeignet gewählten Mengen betrachtet werden, zum Beispiel zwischen den offenen Mengen

$$U := \big\{(x_1, x_2) : x_2 \neq 0 \text{ oder } x_1 > 0\big\}, \quad V := \big\{(r, \varphi) : r > 0 \text{ und } -\pi < \varphi < +\pi\big\},$$

wobei U gerade die in der negativ-reellen Achse geschlitzte Ebene ist.

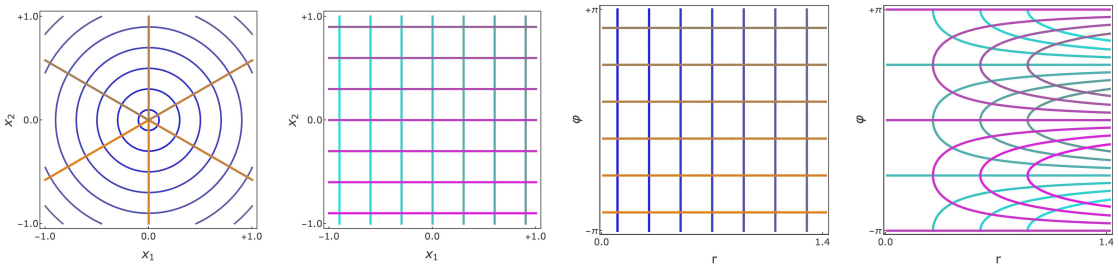

Abbildung Ausgewählte Niveaulinien der Polarkoordinaten in der (x_1, x_2)-Ebene (links) sowie der (r, φ)-Ebene (rechts). Blau bzw. Orange gehört zu $r = \text{const}$ bzw. $\varphi = \text{const}$, Türkis bzw. Lila zu $x_1 = \text{const}$ bzw. $x_2 = \text{const}$.

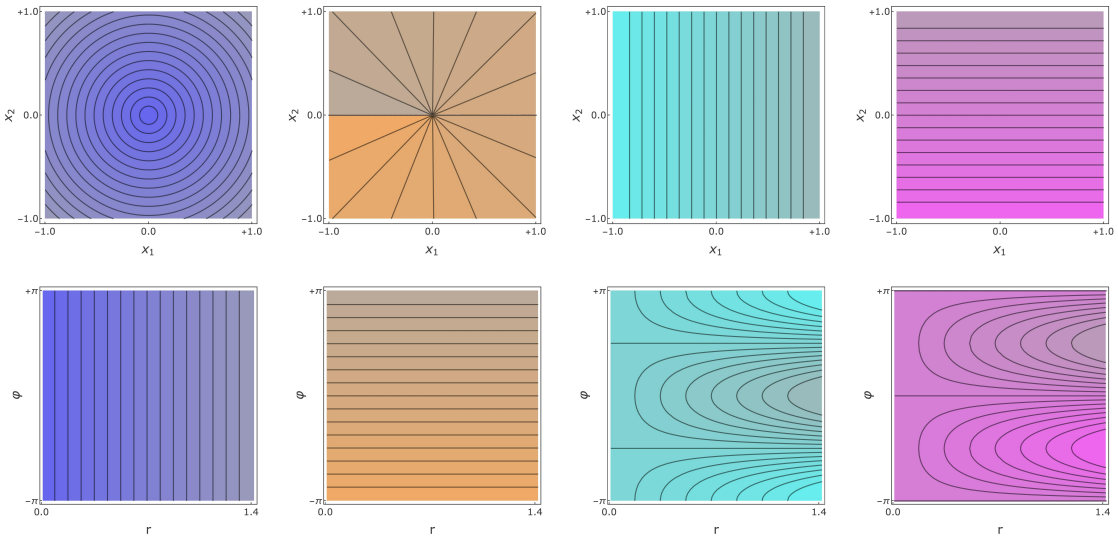

Abbildung Eine andere Visualisierung der Niveaumengen von r (Blau), φ (Orange), x_1 (Türkis) und x_2 (Lila). Oben in der (x_1, x_2)-Ebene, unten in der (r, φ)-Ebene. Hier benötigen wir jedoch acht Bilder, um dieselben Informationen darzustellen.

Transformationsformeln für Polarkoordinaten Durch direkte Differentiation erhalten wir

$$\frac{\partial(x_1, x_2)}{\partial(r, \varphi)} = \begin{pmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \varphi} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} \cos(\varphi) & -r\sin(\varphi) \\ \sin(\varphi) & r\cos(\varphi) \end{pmatrix}$$

und die Berechnung der inversen Matrix liefert[35]

$$\frac{\partial(r, \varphi)}{\partial(x_1, x_2)} = \begin{pmatrix} \frac{\partial r}{\partial x_1} & \frac{\partial r}{\partial x_2} \\ \frac{\partial \varphi}{\partial x_1} & \frac{\partial \varphi}{\partial x_2} \end{pmatrix} = \frac{1}{r}\begin{pmatrix} r\cos(\varphi) & r\sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{pmatrix}.$$

[35]Insbesondere können wir Ausdrücke für die partiellen Ableitungen von r und φ nach x_1 und x_2 angeben, ohne die komplizierten Formeln für r und φ als Funktionen in x_1 und x_2 explizit hinschreiben zu müssen.

Wir können nun die Transformationsformeln der ersten Ordnung ablesen und erhalten

$$\frac{\partial}{\partial x_1} = \frac{\partial r}{\partial x_1}\frac{\partial}{\partial r} + \frac{\partial \varphi}{\partial x_1}\frac{\partial}{\partial \varphi} = \cos(\varphi)\frac{\partial}{\partial r} - \frac{\sin(\varphi)}{r}\frac{\partial}{\partial \varphi}$$
$$\frac{\partial}{\partial x_2} = \frac{\partial r}{\partial x_2}\frac{\partial}{\partial r} + \frac{\partial \varphi}{\partial x_2}\frac{\partial}{\partial \varphi} = \sin(\varphi)\frac{\partial}{\partial r} + \frac{\cos(\varphi)}{r}\frac{\partial}{\partial \varphi}.$$

Für die zweiten Ableitungen verifizieren wir mit längeren Rechnungen

$$\frac{\partial^2}{\partial x_1^2} = \frac{\sin^2(\varphi)}{r}\frac{\partial}{\partial r} + \frac{\sin(2\varphi)}{r^2}\frac{\partial}{\partial \varphi} + \cos^2(\varphi)\frac{\partial^2}{\partial r^2} - \frac{\sin(2\varphi)}{r}\frac{\partial^2}{\partial r \partial \varphi} + \frac{\sin^2(\varphi)}{r^2}\frac{\partial^2}{\partial \varphi^2}$$
$$\frac{\partial^2}{\partial x_2^2} = \frac{\cos^2(\varphi)}{r}\frac{\partial}{\partial r} - \frac{\sin(2\varphi)}{r^2}\frac{\partial}{\partial \varphi} + \sin^2(\varphi)\frac{\partial^2}{\partial r^2} + \frac{\sin(2\varphi)}{r}\frac{\partial^2}{\partial r \partial \varphi} + \frac{\cos^2(\varphi)}{r^2}\frac{\partial^2}{\partial \varphi^2}$$

und erhalten

$$\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} = \frac{\partial^2}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2} + \frac{1}{r}\frac{\partial}{\partial r}$$

als Transformationsformel für den Laplace-Operator in zwei Dimensionen.[36] Beachte, dass in allen Formeln Singularitäten bei $r = 0$ auftreten (da dort der Variablenwechsel entartet).

räumliche Kugelkoordinaten Ein 3D-Analogon zu den Polarkoordinaten ist

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} r\cos(\varphi)\cos(\theta) \\ r\sin(\varphi)\cos(\theta) \\ r\sin(\theta) \end{pmatrix}$$

mit Radius r und Euler-Winkeln φ und θ, wobei man meist $r > 0$, $\varphi \in (-\pi, \pi)$ und $\theta \in (-\pi/2, \pi/2)$ annimmt (obwohl dann einige Punkte $(x_1, x_2, x_3) \in \mathbb{R}^3$ nicht dargestellt werden können). Analog zu oben ergeben sich die Transformationsformeln

$$\begin{aligned}
\frac{\partial}{\partial x_1} &= \cos(\varphi)\cos(\theta)\frac{\partial}{\partial r} - \frac{\sin(\varphi)}{r\cos(\theta)}\frac{\partial}{\partial \varphi} - \frac{\cos(\varphi)\sin(\theta)}{r}\frac{\partial}{\partial \theta} \\
\frac{\partial}{\partial x_2} &= \sin(\varphi)\cos(\theta)\frac{\partial}{\partial r} + \frac{\cos(\varphi)}{r\cos(\theta)}\frac{\partial}{\partial \varphi} - \frac{\sin(\varphi)\sin(\theta)}{r}\frac{\partial}{\partial \theta} \\
\frac{\partial}{\partial x_3} &= \sin(\theta)\frac{\partial}{\partial r} + \frac{\cos(\theta)}{r}\frac{\partial}{\partial \theta}
\end{aligned}$$

aus direkten Rechnungen mit partiellen Ableitungen.

[36] Diese Transformationsformel wird sehr oft in der Mathematik und der Physik verwendet.

2.5 Satz von Taylor

Vorbemerkung In diesem Abschnitt verallgemeinern wir den Satz von Taylor, den wir in *Analysis 1* für Funktionen in einer Variablen kennen gelernt hatten, auf skalare Funktionen mit mehreren Veränderlichen. Die resultierenden lokalen Approximationsformeln werden wir sehr oft verwenden, zum Beispiel beim Studium von lokalen Extremstellen oder um vereinfachte Formeln in den Natur- und Ingenieurwissenschaften abzuleiten.

Wir beschränken uns in der Darstellung auf eine skalare Funktion. Das ist aber keine wesentliche Einschränkung, denn bei einer vektorwertigen Abbildung kann die Taylor-Entwicklung komponentenweise angewendet werden. Wir werden in diesem Abschnitt meist die Tupel-Notation für Elemente des $\mathbb{R}^n$ verwenden.

Vereinbarung Im Folgenden ist U immer eine offene Teilmenge des $\mathbb{R}^n$, $x_* \in U$ ein beliebiger, aber fester Punkt und $f : U \to \mathbb{R}$ eine K-mal stetig differenzierbare Funktion. Insbesondere existieren alle partiellen Ableitungen von f bis zur Ordnung K als stetige Funktionen auf ganz U.[37]

Erinnerung Für $n = 1$ ist K-te Taylor-Polynom einer Funktion $f : I \to \mathbb{R}$ im Entwicklungspunkt $x_* \in I$ ist durch

$$\begin{aligned} T_{f,K,x_*}(x) &= \sum_{k=0}^{K} \frac{f^{(k)}(x_*)}{k!} (x - x_*)^K \\ &= f(x_*) + f'(x_*)(x - x_*) + \tfrac{1}{2} f''(x_*)(x - x_*)^2 + \ldots + \tfrac{1}{K!} f^{(K)}(x_*)(x - x_*)^K \end{aligned}$$

gegeben, wobei $I \subseteq \mathbb{R}$ ein offenes Intervall ist. Für das entsprechende Restglied gilt

$$\frac{|R_{f,K,x_*}(x)|}{|x - x_*|^K} \xrightarrow{x \to x_*} 0\,, \qquad R_{f,K,x_*}(x) := f(x) - T_{f,K,x_*}(x)\,,$$

d.h. sein Betrag konvergiert schneller gegen 0 als $|x - x_*|^K$. Insbesondere kann f in der Nähe von x_* durch T_{f,K,x_*} approximiert werden, wobei Koeffizienten dieses Polynoms mit Grad K vom Entwicklungspunkt x_* abhängen).

Beispiel: Die Taylor-Approximation dritter Ordnung der Sinusfunktion ist

$$T_{\sin,3,x_*}(x) = \sin(x_*) + \cos(x_*)(x - x_*) - \tfrac{1}{2}\sin(x_*)(x - x_*)^2 - \tfrac{1}{6}\cos(x_*)(x - x_*)^3$$

und mit $x_* = 0$ ergibt sich die besonders einfache Formel $T_{\sin,3,0}(x) = x - \frac{1}{6}x^3$.

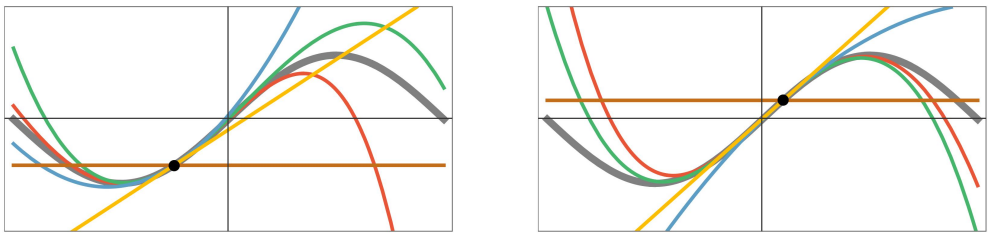

Abbildung Verschiedene Taylor-Approximationen der Sinusfunktion (grau) in zwei verschiedenen Entwicklungspunkten (links und rechts). Die Farben Braun/Gelb/Blau/Grün/Rot entsprechen den Ordnungen $K = 0/1/2/3/4$ und der schwarze Punkt repräsentiert $(x_*, \sin(x_*))$.

[37] Alternativ können wir fordern, dass die K-te totale Ableitung in jedem Punkt aus U existiert und außerdem stetig bzgl. der entsprechenden Operatornorm ist.

Formeln der Taylor-Polynome

Multi-Indizes Für ein n-Tupel $\kappa = (\kappa_1, \ldots, \kappa_n)$ bestehend aus natürlichen Zahlen (einschließlich der 0) und einen Vektor $x \in \mathbb{R}^n$ führen wir die folgenden Notationen ein:

$$\begin{array}{lcll} \kappa! & := & \kappa_1! \ldots \kappa_n! & \text{(die „Fakultät“)} \\ |\kappa| & := & \kappa_1 + \ldots + \kappa_n & \text{(die Ordnung)} \\ x^\kappa & := & x_1^{\kappa_1} \ldots x_n^{\kappa_n} & \text{(verallgemeinerte Potenz)} \\ \partial_x^\kappa & := & \partial_{x_1}^{\kappa_1} \ldots \partial_{x_n}^{\kappa_n} & \text{(Differentialoperator der Ordnung } |\kappa|) \end{array}$$

Dabei gilt wie immer $0! = 1! = 1$ sowie $x_j^0 = 1$. Außerdem sei $\partial_{x_j}^0 f(x) = f(x)$ vereinbart, das heißt die nullfache partielle Ableitung von f nach x_j ist f selbst.

Bemerkung Die Schreibweise mit Multi-Indizes ist sehr elegant, aber sicherlich gewöhnungsbedürftig. Sie sollten sich davon nicht abschrecken lassen und sich am Anfang auf die weiter unten angegebenen Taylor-Formeln für $n = 2$ bzw. $K = 1$ oder $K = 2$ konzentrieren, da man diese auch sehr gut ohne Multi-Index-Notation verstehen und memorieren kann. Später werden Sie dann mit allgemeineren Fällen konfrontiert sein und die Multi-Indizes schätzen lernen.

Beispiele

1. Für $n = 2$ erhalten wir

$$(0,\,0)! = 0!\,0! = 1 \quad (x_1,\,x_2)^{(0,\,0)} = 1 \quad \partial_{(x_1,\,x_2)}^{(0,\,0)} f(x) = f(x)$$

sowie

$$(1,\,0)! = 1!\,0! = 1 \quad (x_1,\,x_2)^{(1,\,0)} = x_1 \quad \partial_{(x_1,\,x_2)}^{(1,\,0)} f(x) = \partial_{x_1} f(x)$$

und

$$(0,\,1)! = 0!\,1! = 1 \quad (x_1,\,x_2)^{(0,\,1)} = x_2 \quad \partial_{(x_1,\,x_2)}^{(0,\,1)} f(x) = \partial_{x_2} f(x)\,.$$

Analog ergeben sich die Formeln

$$(2,\,1)! = 2!\,1! = 2\,, \quad (x_1,\,x_2)^{(2,\,1)} = x_1^2\,x_2\,, \quad \partial_{(x_1,\,x_2)}^{(2,\,1)} f(x) = \partial_{x_1}^2 \partial_{x_2} f(x)$$

und

$$(0,\,3)! = 0!\,3! = 6\,, \quad (x_1,\,x_2)^{(0,\,3)} = x_2^3\,, \quad \partial_{(x_1,\,x_2)}^{(0,\,3)} f(x) = \partial_{x_2}^3 f(x)\,.$$

2. Für $n = 3$ erhalten wir zum Beispiel

$$(1,\,3,\,2)! = 1!\,3!\,2! = 12\,, \qquad (4,\,2,\,0)! = 4!\,2!\,0! = 48$$

sowie

$$(x_1,\,x_2,\,x_3)^{(1,\,3,\,2)} = x_1\,x_2^3\,x_3^2\,, \qquad (x_1,\,x_2,\,x_3)^{(4,\,2,\,0)} = x_1^4\,x_2^2\,.$$

und

$$\partial_{(x_1,\,x_2,\,x_2)}^{(1,\,3,\,2)} f(x) = \partial_{x_1}^1 \partial_{x_2}^3 \partial_{x_3}^2 f(x)\,, \qquad \partial_{(x_1,\,x_2,\,x_2)}^{(4,\,2,\,0)} f(x) = \partial_{x_1}^4 \partial_{x_2}^2 f(x)\,.$$

Definition Das K-te Taylor-Polynom von f im Entwicklungspunkt x_* ist durch

$$T_{f,K,x_*}(x) := \sum_{|\kappa|\leq K} \frac{\partial_x^\kappa f(x_*)}{\kappa!}(x-x_*)^\kappa$$

definiert, wobei die Multi-Summe auf der rechten Seite der Formel für jeden Multi-Index $\kappa = (\kappa_1, \ldots, \kappa_n)$ mit Ordnung $\kappa_1 + \ldots + \kappa_n \leq K$ genau einen Summanden enthält. Den Ausdruck

$$R_{f,K,x_*}(x) := f(x) - T_{f,K,x_*}(x)$$

bezeichnen wir wieder als das entsprechende Restglied. [38]

Bemerkungen

1. Alternativ können wir das Taylor-Polynom auch als

$$T_{f,K,x_*}(x) = \sum_{k=0}^{K} \sum_{|\kappa|=k} \frac{\partial_x^\kappa f(x_*)}{\kappa!}(x-x_*)^\kappa\,,$$

schreiben, wobei die Multi-Summe nur anders gruppiert wurde: die erste Summe wird über den Ordnungsparameter k gebildet und die zweite Summe fasst die Beiträge aller Multi-Indizes mit Ordnung $\kappa_1 + \ldots + \kappa_n = k$ zusammen.

2. $T_{f,K,x_*}(x)$ kann sowohl als Polynom vom Grad K in den Variablen x_j, aber auch als Polynom in den Variablen $x_j - x_{*,j}$ betrachtet werden, wobei die jeweiligen Koeffizienten verschieden sind und von x_* abhängen. Dies ergibt sich unmittelbar aus den Rechenregeln der reellen Addition und Multiplikation, aber die konkreten Umrechnungformeln zwischen den Koeffizienten können unhandlich sein.

   Beispiel: Durch Ausmultiplizieren verifizieren wir die Formeln

$$a_* (x_j - x_{*,j})^2 = a_* x_j^2 - 2\, a_* x_{*,j}\, x_j^1 + a_* x_{*,j}^2\, x_j^0$$

   und

$$b_* x_j^2 = b_* (x_j - x_{*,j})^2 + 2\, b_* x_{*,j} (x_j - x_{*,j})^1 + b_* x_{*,j}^2 (x_j - x_{*,j})^0\,,$$

   wobei jeweils auf einer Seite ein Polynom in x_j, auf der anderen aber ein Polynom in $x_j - x_{*,j}$ steht.

3. Wenn die Wahl von f und x_* zweifelsfrei klar ist, schreiben wir oftmals einfach $T_K(x)$ statt $T_{f,K,x_*}(x)$.

4. Es gilt

$$T_K(x) - T_{K-1}(x) = \sum_{|\kappa|=K} \frac{\partial_x^\kappa f(x_*)}{\kappa!}(x-x_*)^\kappa\,,$$

   für die Differenz zweier aufeinanderfolgender Taylor-Polynome, wobei jeder Term auf der rechten Seite ein Monom in den Variablen x_1, ..., x_n vom Grad K ist. Die Abhängigkeit von den $x_{*,j}$ wird aber im Allgemeinen komplizierter sein.

[38] Beachte, dass $T_{f,K,x_*}(x)$ für alle $x \in \mathbb{R}^n$, $R_{f,K,x_*}(x)$ jedoch nur für $x \in U$ definiert ist.

5. Bei allen theoretischen Aussagen werden wir die Argumente von f wie immer mit $x \in \mathbb{R}^n$ bzw. mit $(x_1, \ldots, x_n)$ bezeichnen. In der Praxis treten natürlich auch andere Variablennamen auf. Im Fall von $n = 2$ bzw. $n = 3$ werden zum Beispiel gerne die alternativen Schreibweisen (x, y) bzw. (x, y, z) für zwei- bzw. dreidimensionale Punkte verwendet. Siehe das Beispiel weiter unten.

Spezialfall n=2 Für eine gegebene Funktion f in zwei Variablen $x = (x_1, x_2)$ und einen festen Punkt $x_* = (x_{*,1}, x_{*,2})$ erhalten wir

$$T_0(x) = f(x_*)$$

sowie

$$\begin{aligned} T_1(x) = &+ T_0(x) \\ &+ \partial_{x_1} f(x_*)\,(x_1 - x_{*,1}) \\ &+ \partial_{x_2} f(x_*)\,(x_2 - x_{*,2}) \end{aligned}$$

und

$$\begin{aligned} T_2(x) = &+ T_1(x) \\ &+ \tfrac{1}{2}\,\partial_{x_1}^2 f(x_*)\,(x_1 - x_{*,1})^2 \\ &+ 1\,\partial_{x_1}\partial_{x_2} f(x_*)\,(x_1 - x_{*,1})\,(x_2 - x_{*,2}) \\ &+ \tfrac{1}{2}\,\partial_{x_2}^2 f(x_*)\,(x_2 - x_{*,2})^2\,, \end{aligned}$$

wobei wir in jedem Schritt die jeweils neuen Terme angegeben haben. Insbesondere gilt: $T_0(x)$ hängt gar nicht von x ab, aber $T_1(x)$ bzw. $T_2(x)$ sind Polynome in x_1 und x_2 vom Grad 1 bzw. 2. Außerdem gilt

$$\begin{aligned} T_3(x) = &+ T_2(x) \\ &+ \tfrac{1}{6}\,\partial_{x_1}^3 f(x_*)\,(x_1 - x_{*,1})^3 \\ &+ \tfrac{1}{2}\,\partial_{x_1}^2\partial_{x_2} f(x_*)\,(x_1 - x_{*,1})^2\,(x_2 - x_{*,2}) \\ &+ \tfrac{1}{2}\,\partial_{x_1}\partial_{x_2}^2 f(x_*)\,(x_1 - x_{*,1})\,(x_2 - x_{*,2})^2 \\ &+ \tfrac{1}{6}\,\partial_{x_2}^3 f(x_*)\,(x_2 - x_{*,2})^3\,, \end{aligned}$$

wobei die neuen Terme vom Grad 3 den Multi-Indizes $(3, 0)$, $(2, 1)$, $(1, 2)$, $(0, 3)$ entsprechen, die alle die Ordnung 3 besitzen. Bei $T_4(x)$ kommen insgesamt 5 neue Terme hinzu, da es fünf zweidimensionale Multi-Indizes der Ordnung 4 gibt, nämlich $(4, 0)$, $(3, 1)$, $(2, 2)$, $(1, 3)$, $(0, 4)$.

Spezialfälle K=0, K=1, K=2 Ist f eine Funktion in n Variablen, so können wir für die ersten drei Taylor-Polynome die Multi-Index-Formeln in eine kompaktere Form bringen. Es gilt

$$T_0(x) = f(x_*)$$

und für $K = 1$ erhalten wir (Nachrechnen!)

$$\begin{aligned} T_1(x) - T_0(x) &= \sum_{j=1}^{n} \partial_{x_j} f(x_*)\,(x_j - x_{*,j}) \\ &= \operatorname{grad} f(x_*) \cdot (x - x_*) = \operatorname{Jac} f(x_*)\,(x - x_*)\,. \end{aligned}$$

Bei $K = 2$ können die Terme der Multi-Indizes zweiter Ordnung mittels der Hesse-Matrix ausgedrückt werden. Durch Nachrechnen verifizieren wir

$$\begin{aligned} T_2(x) - T_1(x) &= \sum_{j=1}^{n} \sum_{l=1}^{n} \tfrac{1}{2}\, \partial_{x_j} \partial_{x_l} f(x_*)\, (x_j - x_{*,j})\, (x_l - x_{*,l}) \\ &= \tfrac{1}{2}\, (x - x_*)^T \operatorname{Hess} f(x_*)\, (x - x_*)\,, \end{aligned}$$

wobei in der j-l-Doppelsumme die gemischten Terme zweimal auftreten und auf der rechten Seite die quadratische Matrix $\operatorname{Hess} f(x_*)$ von links bzw. rechts mit einem Zeilen- bzw. einem Spaltenvektor mutipliziert wird (sodass am Ende eine reelle Zahl entsteht). Zum Beispiel erhalten wir den Term

$$\tfrac{1}{2}\, \partial_{x_1} \partial_{x_2} f(x_*)\, (x_1 - x_{*,1})\, (x_2 - x_{*,2}) = \tfrac{1}{2}\, \partial_{x_2} \partial_{x_1} f(x_*)\, (x_2 - x_{*,2})\, (x_1 - x_{*,1})$$

zweimal, nämlich einmal mit $j = 1$ und $l = 2$ und einmal für $j = 2$ und $l = 1$. In der Multi-Index-Notation entsprechen beide aber dem n-dimensionalen Multi-Index $(1,\, 1,\, 0,\, \ldots,\, 0)$, dessen Fakultät gerade 1 und damit zweimal $\frac{1}{2}$ ist.

Spezialfall $n = K = 2$ In diesem Fall ergibt sich

$$\begin{aligned} &f(x_1,\, x_2) = \\ &\quad + f(x_*) \\ &\quad + \begin{pmatrix} \partial_{x_1} f(x_{*,1},\, x_{*,2}) & \partial_{x_2} f(x_{*,1},\, x_{*,2}) \end{pmatrix} \begin{pmatrix} x_1 - x_{*,1} \\ x_2 - x_{*,2} \end{pmatrix} \\ &\quad + \tfrac{1}{2} \begin{pmatrix} x_1 - x_{*,1} & x_2 - x_{*,2} \end{pmatrix} \begin{pmatrix} \partial_{x_1}^2 f(x_{*,1},\, x_{*,2}) & \partial_{x_1} \partial_{x_2} f(x_{*,1},\, x_{*,2}) \\ \partial_{x_2} \partial_{x_1} f(x_{*,1},\, x_{*,2}) & \partial_{x_2}^2 f(x_{*,1},\, x_{*,2}) \end{pmatrix} \begin{pmatrix} x_1 - x_{*,1} \\ x_2 - x_{*,2} \end{pmatrix}, \end{aligned}$$

wobei wir hier ausnahmsweise und zur besseren Übersicht den Entwicklungspunkt in Blau und die freien Variablen rot geschrieben haben. Insbesondere ist die rechte Seite ein quadratisches Polynom in x_1 und x_2, aber die Koeffizienten dieses Polynoms können in komplizierter Weise von $x_{*,1}$ und $x_{*,2}$ abhängen.

Beispiel Für die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ mit

$$f(x_1,\, x_2) := 2 + \cos(x_1)\, x_2$$

berechnen wir die ersten bzw. zweiten partiellen Ableitungen zu

$$\partial_{x_1} f(x_1,\, x_2) = -\sin(x_1)\, x_2\,, \qquad \partial_{x_2} f(x_1,\, x_2) = \cos(x_1)$$

bzw.

$$\partial_{x_1}^2 f(x_1,\, x_2) = -\cos(x_1)\, x_2\,, \quad \partial_{x_1}\, \partial_{x_2} f(x_1,\, x_2) = -\sin(x_1)\,, \quad \partial_{x_2}^2 f(x_1,\, x_2) = 0\,.$$

Im Entwicklungspunkt $(x_{*,1},\, x_{*,2}) = (0,\, 0)$ erhalten wir nach Einsetzen damit

$$T_0(x_1,\, x_2) = 2\,, \quad T_1(x_1,\, x_2) = 2 + x_2\,, \quad T_2(x_1,\, x_2) = 2 + x_2\,,$$

wobei hier T_1 und T_2 zufälligerweise zusammenfallen, da alle zweiten Ableitungen im Koordinatenursprung $(0,\, 0)$ verschwinden. Um auch noch T_3 anzugeben, berechnen wir

$$\partial_{x_1}^3 f(x_1,\, x_2) = \sin(x_1)\, x_2\,, \qquad \partial_{x_1}^2\, \partial_{x_2} f(x_1,\, x_2) = -\cos(x_1)$$

sowie

$$\partial_{x_1} \partial_{x_2}^2 f(x_1, x_2) = 0\,, \quad \partial_{x_2}^3 f(x_1, x_2) = 0\,.$$

Durch Auswertung im Entwicklungspunkt erhalten wir schließlich

$$T_3(x_1, x_2) = T_2(x_1, x_2) - \tfrac{1}{2}\, x_1^2\, x_2 = 2 + x_2 - \tfrac{1}{2}\, x_1^2\, x_2$$

als kubisches Polynom in x_1 und x_2, wobei die nichtverschwindenden Beiträge auf der rechten Seite den Multi-Indizes $(0, 0)$, $(0, 1)$ und $(2, 1)$ entsprechen. Natürlich ist dieses Beispiel sehr einfach, eben weil die meisten Ableitungen sich im gewählten Entwicklungspunkt zu Null ergeben.

Beispiel Wir betrachten noch einmal die Funktion aus dem letzten Beispiel, aber diesmal mit dem Entwicklungspunkt $(x_{*,1}, x_{*,2}) = (\pi/2, 1)$. Nach Einsetzen der entsprechenden Werte in die obigen Formeln für die ersten und zweiten partiellen Ableitungen erhalten wir diesmal die Taylor-Polynome

$$T_0(x_1, x_2) = \frac{4 + \sqrt{2}}{2}$$

sowie

$$T_1(x_1, x_2) = T_0(x_1, x_2) - \frac{\sqrt{2}}{2} \left(x_1 - \frac{\pi}{2}\right) + \frac{\sqrt{2}}{2}\,(x_2 - 1)\,,$$

und

$$T_2(x_1, x_2) = T_1(x_1, x_2) - \frac{\sqrt{2}}{4} \left(x_1 - \frac{\pi}{2}\right)^2 - \frac{\sqrt{2}}{2} \left(x_1 - \frac{\pi}{2}\right)(x_2 - 1)\,,$$

wobei wir $\cos(\pi/2) = \sin(\pi/2) = \sqrt{2}/2$ benutzt haben und T_3 diesmal nicht angeben wollen.

Beispiel Wir betrachten die (sehr einfache) polynomiale Funktion $f : \mathbb{R}^3 \to \mathbb{R}$ mit

$$f(x, y, z) = x\, y^2 + y\, z + z + 1\,,$$

wobei wir diesmal die Notation $(x, y, z) \in \mathbb{R}^3$ verwenden, und berechnen sowohl die Jacobi-Matrix (erste Ableitungen)

$$\begin{pmatrix} \partial_x f(x_*, y_*, z_*) & \partial_y f(x_*, y_*, z_*) & \partial_z f(x_*, y_*, z_*) \end{pmatrix} = \begin{pmatrix} y_*^2 & 2\, x_*\, y_* + z_* & y_* + 1 \end{pmatrix}$$

als auch die Hesse-Matrix (zweite Ableitungen)

$$\begin{pmatrix} \partial_x^2 f(x_*, y_*, z_*) & \partial_x \partial_y f(x_*, y_*, z_*) & \partial_x \partial_z f(x_*, y_*, z_*) \\ \partial_y \partial_x f(x_*, y_*, z_*) & \partial_y^2 f(x_*, y_*, z_*) & \partial_y \partial_z f(x_*, y_*, z_*) \\ \partial_z \partial_x f(x_*, y_*, z_*) & \partial_z \partial_y f(x_*, y_*, z_*) & \partial_z^2 f(x_*, y_*, z_*) \end{pmatrix} = \begin{pmatrix} 0 & 2\, y_* & 0 \\ 2\, y_* & 2\, x_* & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

wobei wir beide in einem beliebigen Entwicklungspunkt (x_*, y_*, z_*) ausgewertet haben. Wir erhalten nun schrittweise die Taylor-Polynome

$$\begin{aligned} T_0(x, y, z) &= +x_*\, y_*^2 + y_*\, z_* + z_* + 1\,, \\ T_1(x, y, z) &= T_0(x, y, z) + y_*^2\,(x - x_*) + (2\, x_*\, y_* + z_*)\,(y - y_*) + (y_* + 1)\,(z - z_*)\,, \\ T_2(x, y, z) &= T_1(x, y, z) + 2\, y_*\,(x - x_*)\,(y - y_*) + (y - y_*)\,(z - z_*) + x_*\,(y - y_*)^2\,. \end{aligned}$$

Diese Ausdrücke liefern in der Tat Polynome in den Variablen x, y, und z und besitzen den Grad 0, 1, bzw. 2. Alle dreifachen partiellen Ableitungen von f verschwinden bis auf den Term

$$\partial_x \, \partial_y^2 f(x_*, \, y_*, \, z_*) = 2 \, ,$$

der dem Multi-Index $(1, \, 2, \, 0)$ entspricht, und wir erhalten insgesamt das kubische Taylor-Polynom

$$T_3(x, \, y, \, z) = T_2(x, \, y, \, z) + \tfrac{1}{2} \, 2 \, (x - x_*) \, (y - y_*)^2 \, .$$

Ein Ausmultiplizieren aller Terme (Nachrechnen!) in T_3 liefert

$$T_3(x, \, y, \, z) = f(x, \, y, \, z) \, ,$$

d.h. in diesem Beispiel heben sich alle Beiträge vom Entwicklungspunkt bei T_3 (aber noch nicht bei T_1 und T_2) gegenseitig auf. Dies ist nicht überraschend, da f selbst ein kubisches Polynom ist. Siehe dazu unten die Folgerungen aus dem Satz von Taylor.

lokale Approximation durch Taylor-Polynome

Vorbemerkung Auch in höheren Dimensionen ($n > 1$) liefert das Taylor-Polynom $T_{f, \, K, \, x_*}(x)$ in der Nähe von x_* eine gute Näherung von f, wobei mit wachsendem K die Approximationsgüte immer besser wird, aber auch immer höhere Ableitungen von f benötigt werden. Wir werden nun die entsprechenden Restglied-Abschätzungen schrittweise herleiten.

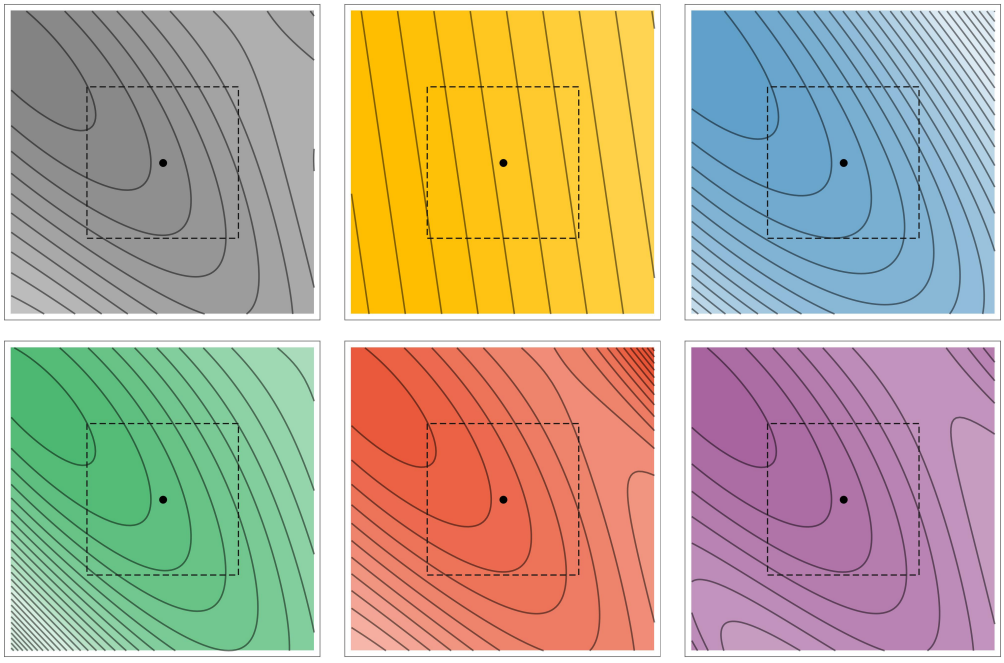

Abbildung Niveaulinien einer Funktion f (grau) sowie ihrer Taylor-Polynome T_1 (gelb), T_2 (blau), T_3 (grün) T_4 (rot), T_5 (lila) in einem fixierten Entwicklungspunkt x_* (schwarz). Beachte, dass die gelben/blauen/grünen/roten/lila Niveaulinien die Lösungen gewisser polynomieller Gleichungen vom Grad 1/2/3/4/5 darstellen und dass die Approximation in dem kleineren Fenster (gestrichelte Linien) besser ist.

Klarstellung In der Taylor-Theorie geht es zunächst um *lokale* Approximationen, d.h. $f(x) \approx T_{f,K,x_*}(x)$ gilt im Allgemeinen nur dann, wenn x hinreichend nahe bei x_* liegt bzw. wenn $|x - x_*|$ hinreichend klein ist. Die Frage, ob auch globale Näherungsformeln auf ganz U oder zumindest auf einer vorgegebenen Teilmenge von U existieren, ist natürlich auch wichtig, wird aber erst in einem zweiten Schritt oder manchmal mit ganz anderen Methoden untersucht.

Lemma (**Schmiege-Eigenschaft von Taylor-Polynomen**) Für jedem Multi-Index λ gilt

$$\partial_x^\lambda T_{f,K,x_*}(x_*) = \begin{cases} \partial_x^\lambda f(x_*) & \text{falls } |\lambda| \le K, \\ 0 & \text{sonst.} \end{cases}$$

Insbesondere stimmen im Entwicklungspunkt x_* (aber in der Regel auch nur dort) alle partiellen Ableitungen von f und von T_{f,K,x_*} bis zur Ordnung K überein.

Beweis Siehe die entsprechende Übungsaufgabe. □

Geometrische Interpretation Für $n = 2$ kann der Graph einer skalaren Funktion f als gekrümmte Fläche im $\mathbb{R}^3$ interpretiert werden, wohingegen der Graph des ersten Taylor-Polynoms T_1 immer eine Ebene beschreibt. Wir werden später sehen, dass der Graph von T_1 gerade die *Tangentialebene* an die Fläche im Punkt x_* ist. Die Graphen von T_2 und T_3 schmiegen sich auch an den Graphen von f an (sogar zu höherer Ordnung), aber es handelt sich nicht mehr um Ebenen, sondern um *algebraische Flächen*, die selbst gekrümmt sind. Bei T_2 spricht man auch von der *Schmiegquadrik*.

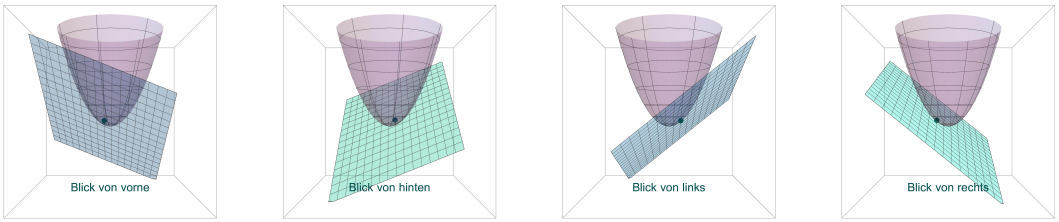

Abbildung Der Graph von T_1 ist eine Ebene, die sich im Entwicklungspunkt (schwarzer Punkt) an den Graphen von f (rosa, hier Rotations-Paraboloid) anschmiegt. Geometrisch gesehen handelt es sich dabei um eine Tangentialebene an eine gekrümmte Fläche.

Lemma (**spezielle Ableitungsformel**) Sei $v \in \mathbb{R}^n$ eine Richtung mit $|v| > 0$ und I ein offenes Intervall mit $0 \in I$, sodass das Bild der parametrisierten Kurve $\gamma : I \to \mathbb{R}^n$ mit

$$\gamma(t) = x_* + t\,v\,,$$

ganz in U liegt.[39] Dann gilt

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k} f\big(\gamma(t)\big) = \sum_{|\kappa|=k} \frac{k!}{\kappa!}\, \partial_x^\kappa f\big(\gamma(t)\big)\, v^\kappa$$

für jedes $k \le K$ und alle $t \in I$.[40]

[39] Die Bildmenge Γ ist ein Geradenstück mit Richtungsvektor v, dass den Punkt x_* enthält.

[40] Beachte, dass jeder Summand auf der rechten Seite eine reelle Zahl ist. Insbesondere ist $\partial_x^\kappa f\big(\gamma(t)\big)$ eine partielle Ableitung von f, die im Punkt $\gamma(t)$ ausgewertet wird, und v^κ steht für ein Produkt von Potenzen der Komponenten des Vektors v.

Beweis *analytischer Teil*: Wir betrachten die Funktion $g : I \to \mathbb{R}$ mit

$$g(t) := f\big(\gamma(t)\big)$$

und schließen mit (endlicher) Induktion über $k \in \{1, \ldots, K\}$, dass die Hilfsformel

$$g^{(k)}(t) = \sum_{j_1=1}^{n} \ldots \sum_{j_k=1}^{n} \partial_{x_{j_k}} \ldots \partial_{x_{j_1}} f\big(\gamma(t)\big)\, v_{j_1} \,\ldots\, v_{j_k}$$

erfüllt ist.[41] In der Tat, für $k = 1$ und wegen $\gamma(t) = (x_{*,1}+t\,v_1, \ldots, x_{*,n}+t\,v_n)$ impliziert die Kettenregel die Formel

$$g^{(1)}(t) = \frac{\mathrm{d}}{\mathrm{d}t} f\big(\gamma(t)\big) = \sum_{j_1=1}^{n} \partial_{x_{j_1}} f\big(\gamma(t)\big)\, v_{j_1}$$

und wir haben den Induktionsanfang etabliert.[42] Im Induktionsschritt $k \rightsquigarrow k+1$ folgt analog

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\partial_{x_{j_k}} \ldots \partial_{x_{j_1}} f\big(\gamma(t)\big)\, v_{j_1} \ldots v_{j_k}\Big) = \sum_{j_{k+1}=1}^{n} \partial_{x_{j_{k+1}}} \partial_{x_{j_k}} \ldots \partial_{x_{j_1}} f\big(\gamma(t)\big)\, v_{j_1} \ldots v_{j_k}\, v_{j_{k+1}}\,,$$

indem wir die Kettenregel diesmal auf die eindimensionale Hilfsfunktion

$$t \quad \mapsto \quad \partial_{x_{j_k}} \ldots \partial_{x_{j_1}} f\big(\gamma(t)\big)\, v_{j_1} \ldots v_{j_k}$$

anwenden. Nach Einsetzen in die Induktionsvoraussetzung erhalten wir schließlich

$$\begin{aligned} g^{(k+1)}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{j_1=1}^{n} \ldots \sum_{j_k=1}^{n} \partial_{x_{j_k}} \ldots \partial_{x_{j_1}} f\big(\gamma(t)\big)\, v_{j_1} \,\ldots\, v_{j_k}\right) \\ &= \sum_{j_1=1}^{n} \ldots \sum_{j_k=1}^{n} \frac{\mathrm{d}}{\mathrm{d}t}\Big(\partial_{x_{j_k}} \ldots \partial_{x_{j_1}} f\big(\gamma(t)\big)\, v_{j_1} \,\ldots\, v_{j_k}\Big) \\ &= \sum_{j_1=1}^{n} \ldots \sum_{j_k=1}^{n} \sum_{j_{k+1}=1}^{n} \partial_{x_{j_{k+1}}} \partial_{x_{j_k}} \ldots \partial_{x_{j_1}} f\big(\gamma(t)\big)\, v_{j_1} \,\ldots\, v_{j_k}\, v_{j_{k+1}} \end{aligned}$$

und damit die gewünschte Induktionsbehauptung.

kombinatorischer Teil: In der Hilfsformel von oben gibt es genau einen Summanden für jedes k-Tupel $(j_1, \ldots, j_k)$, das aus den Zahlen $\{1, \ldots, n\}$ gebildet werden kann, wobei jede dieser Zahlen einfach, mehrfach oder überhaupt nicht im Tupel auftreten kann. Sei nun $\kappa = (\kappa_1, \ldots, \kappa_k)$ ein fester Multi-Index der Ordnung $\kappa_1 + \ldots + \kappa_n = k$. Dann gibt es nach der Formel für Permutationen mit Wiederholung (siehe *Analysis 1*) insgesamt

$$\frac{k!}{\kappa_1! \ldots \kappa_n!} = \frac{k!}{\kappa!}$$

[41] Wie in *Analysis 1* meint $g^{(k)}$ die k-te Ableitung von g und die Symbole $j_1, \ldots, j_k$ stehen für k verschiedene Summationsindizes, wobei jeder von 1 bis n läuft.

[42] Wir können dieses Zwischenergebnis alternativ als

$$g'(t) = \mathrm{D}_v f\big(\gamma(t)\big) = \operatorname{grad} f\big(\gamma(t)\big) \cdot v$$

schreiben, aber für den Induktionsschritt ist die im Beweis verwendete Notation besser geeignet.

verschiedene Tupel $(i_1, \ldots, i_k)$, in denen der Index

$$1 \text{ genau } \kappa_1 \text{ mal}, \quad 2 \text{ genau } \kappa_2 \text{ mal}, \quad \ldots, \quad n \text{ genau } \kappa_n \text{ mal}$$

auftritt. Jedes dieser Tupel liefert aber denselben Beitrag zur Hilfsformel, denn nach dem Satz von Schwarz und aufgrund des Kommutativgesetzes der Multiplikation gilt nämlich

$$\partial_{x_{j_k}} \ldots \partial_{x_{j_1}} f\big(\gamma(t)\big) = \partial_{x_1}^{\kappa_1} \ldots \partial_{x_n}^{\kappa_n} f\big(\gamma(t)\big) = \partial_x^\kappa f\big(\gamma(t)\big)\,, \quad v_{j_1} \,\ldots\, v_{j_k} = v_1^{\kappa_1} \ldots v_n^{\kappa_n} = v^\kappa\,.$$

Insbesondere ergibt sich nun die Behauptung aus der Hilfsformel nach Umgruppierung und Zusammenfassung der Summanden. □

Theorem (**zwei Versionen des Satzes von Taylor**) Es gilt

$$\frac{\big|R_{f,\,K,\,x_*}(x)\big|}{\big|x - x_*\big|^K} \quad \xrightarrow{\;x \to x_*\;} \quad 0\,,$$

wobei auf der linken Seite immer stillschweigend $x \in U$ vorausgesetzt ist. Für jeden Radius $\varepsilon_* > 0$ mit $\overline{B}_{\varepsilon_*}(x_*) \subset U$ existiert außerdem eine Konstante C_*, sodass

$$\frac{\big|R_{f,\,K-1,\,x_*}(x)\big|}{\big|x - x_*\big|^K} \le C_*$$

für jedes x mit $|x - x_*| \le \varepsilon_*$ gilt. Beachte, dass die erste bzw. zweite Formel das Restglied der Ordnung K bzw. $K-1$ betrifft.

Beweis *Darstellung des Restglieds*: Wir fixieren $x \in U$ mit $0 < |x - x_*| \le \varepsilon_*$ und setzen $v := x - x_*$. Da x_* und $x_* + v$ innere Punkte von U sind, existiert ein offenes Intervall I mit $[0,\,1] \subset I$, sodass $\gamma(t) := x_* + t\,v$ für jedes $t \in I$ ein Punkt in U ist.[43] Wie schon im Beweis der speziellen Ableitungsformel betrachten wir die Hilfsfunktion $g : I \to \mathbb{R}$ mit $g(t) = f\big(\gamma(t)\big)$. Diese ist nach dem vorherigen Lemma K-mal stetig differenzierbar mit

$$g(1) = f(x)\,, \qquad \frac{g^{(k)}(0)}{k!} = \sum_{|\kappa|=k} \frac{\partial_x^\kappa f(x_*)}{\kappa!}\,(x - x_*)^\kappa$$

und die eindimensionale Version des Satzes von Taylor — genauer gesagt, die Langrange-Darstellung des entsprechenden Restgliedes — liefert

$$g(1) - \sum_{k=0}^{K-1} \frac{g^{(k)}(0)}{k!} = \frac{g^{(K)}(\tau)}{K!}\,,$$

wobei τ eine Zahl zwischen 0 und 1 ist, die in komplizierter Weise von x_* und x abhängen darf (und in der Regel wird). Diese Formel können wir als

$$R_{f,\,K-1,\,x_*}(x) = \sum_{|\kappa|=K} \frac{\partial_x^\kappa f\big(x_* + \tau\,(x - x_*)\big)}{\kappa!}\,(x - x_*)^\kappa$$

[43] Dieses technische Argument hatten wir auch im Beweis des Mittelwertsatzes verwendet.

schreiben und haben damit die Lagrange-Darstellung des Taylor-Restglieds in höheren Dimensionen identifiziert.

Abschätzung des Restglieds: Die reelle Zahl

$$M_* := \max\Big\{ \sum_{|\kappa|=K} \big|\partial_x^\kappa f(x)\big| \,:\, |x-x_*| \le \varepsilon_* \Big\}$$

ist nach dem Satz vom Maximum wohldefiniert, denn nach Voraussetzung sind die Beträge aller partiellen Ableitungen der Ordnung K von f stetig auf U und damit auch auf der kompakten Menge $\overline{B}_{\varepsilon_*}(x_*)$. Mit der Abschätzung

$$\big|R_{f,\,K-1,\,x_*}(x)\big| \le \sum_{|\kappa|=K} \frac{\big|\partial_x^\kappa f\big(x_* + \tau\,(x-x_*)\big)\big|}{\kappa!} \big|(x-x_*)^\kappa\big| \le C_*\,|x-x_*|^K$$

ergibt sich nun die zweite Version des Satzes von Taylor mit $C_* := M_* \sum_{|\kappa|=K} 1/\kappa!$, wobei wir benutzt haben, dass $x_* + \tau\,(x-x_*)$ auch in $\overline{B}_{\varepsilon_*}(x_*)$ liegt und dass

$$\big|(x-x_*)^\kappa\big| = |x_1-x_{*,1}|^{\kappa_1} \,\ldots\, |x_n-x_{*,n}|^{\kappa_n} \le |x-x_*|^{\kappa_1+\ldots+\kappa_n} = |x-x_*|^K$$

gilt. Aus der Definition der Taylor-Polynome sowie der Darstellungsformel ergibt sich

$$\begin{aligned} R_{f,\,K,\,x_*}(x) &= R_{f,\,K-1,\,x_*}(x) - \sum_{|\kappa|=K} \frac{\partial_x^\kappa f\big(x_*\big)}{\kappa!}\,(x-x_*)^\kappa \\ &= \sum_{|\kappa|=K} \frac{\partial_x^\kappa f\big(x_* + \tau\,(x-x_*)\big) - \partial_x^\kappa f\big(x_*\big)}{\kappa!}\,(x-x_*)^\kappa \end{aligned}$$

und wir erhalten

$$\frac{\big|R_{f,\,K,\,x_*}(x)\big|}{|x-x_*|^K} \le \sum_{|\kappa|=K} \frac{\big|\partial_x^\kappa f\big(x_* + \tau\,(x-x_*)\big) - \partial_x^\kappa f\big(x_*\big)\big|}{\kappa!}\,.$$

Die Größe τ hängt zwar von x ab, liegt aber immer zwischen 0 und 1. Insbesondere gilt $x_* + \tau\,(x-x_*) \to x_*$ für $x \to x_*$ und die erste Version des Satzes von Taylor ergibt sich aus der Stetigkeit der partiellen Ableitungen von f. □

Bemerkungen

1. Stark vereinfacht können wir den Beweis des Theorems wie folgt zusammenfassen: Die höherdimensionale Taylor-Formel ergibt sich aus der eindimensionalen Variante sowie der allgemeinen Kettenregel.

2. Die Formeln im Theorem werden oftmals als

$$R_{f,\,K-1,\,x_*}(x) = O\big(\,|x-x_*|^K\,\big)\,, \qquad R_{f,\,K,\,x_*}(x) = o\big(\,|x-x_*|^K\,\big)$$

   geschrieben, wobei O und o die Landau-Symbole sind (siehe unten).

3. Wie schon in *Analysis 1* gibt es auch für $n > 1$ neben der Lagrange-Darstellung aus dem Beweis weitere Darstellungsformeln für das Restglied. Es sei auf die Literatur sowie WIKIPEDIA verwiesen.

4. Das Theorem impliziert insbesondere die folgende Aussage: Wenn alle partiellen Ableitungen bis zur Ordnung K existieren und stetig sind, so können wir f in der Nähe von x_* durch T_{f,K,x_*} approximieren und die erste Version des Satzes von Taylor garantiert, dass der Betrag des entsprechenden Fehlerterms $|R_{f,K,x_*}|$ für $x \to x_*$ schneller gegen 0 konvergiert als $|x - x_*|^K$. Ist aber f sogar $K+1$-mal stetig differenzierbar (also etwas besser als notwendig), so garantiert die zweite Version (ausgewertet mit $K+1$ anstelle von K), dass dieser Fehlerterm sogar wie $|x - x_*|^{K+1}$ abklingt.

5. Ist f unendlich oft differenzierbar, so können wir die Taylor-Reihe

$$T_{f,\infty,x_*}(x) = \sum_{k=0}^{\infty} \sum_{|\kappa|=k} \frac{\partial_x^\kappa f(x_*)}{\kappa!} (x - x_*)^\kappa$$

betrachten, aber es gilt nicht unbedingt $f(x) = T_{f,\infty,x_*}(x)$ (sondern nur dann, wenn f eine *analytische Funktion* ist).

6. *Ergänzung**: In unserer Fassung der zweiten Version haben wir die Restglied- bzw. Fehlerabschätzung auf abgeschlossenen Kugeln um x_* etabliert. Analoge Resultate können auch für allgemeinere Mengen B (n-dimensionale Ellipsoiden, Quader, Polyeder, ...) hergeleitet werden. Wichtig ist nur, dass eine solche Menge B kompakt ist, ganz in U liegt und — siehe den Beweis des Lemmas — mit jedem ihrer Punkte x auch immer die gesamte Verbindungsstrecke zwischen x_* und x enthält. Die letzte Eigenschaft meint gerade, dass B sternförmig bzgl. x_* ist.

Hinweis Der Satz von Taylor ist ausgesprochen bedeutsam und besitzt mannigfaltige Anwendungen innerhalb und außerhalb der Mathematik. Er ist daher integraler Bestandteil des Prüfungskanons in der mathematischen Analysis und Sie müssen ihn unbedingt fehlerfrei formulieren und anwenden können. Der sehr technische Beweis des Satzes wird jedoch eher selten in Prüfungen abgefragt.

Korollar (**Taylor-Polynome von Polynomen**) Ist f selbst ein Polynom vom Grad N, so gilt

$$T_{f,K,x_*}(x) = f(x)$$

für alle x und alle $K \geq N$.

Beweis Als Polynom ist f unendlich oft stetig differenzierbar und durch direkte Rechnungen mit Monomen zeigen wir (siehe dazu die Übungsaufgabe zur Schmiege-Eigenschaft), dass

$$\partial_x^\kappa f(x) = 0$$

für jeden Multi-Index κ der Ordnung $|\kappa| \geq N+1$ gilt. Die Formel der Lagrange-Darstellung (siehe den Beweis des Satzes von Taylor) impliziert daher

$$R_{f,K,x_*}(x) = 0$$

für jedes $K \geq N$. □

Vorlesungswoche 08

Landau-Symbole* In *Analysis 1* hatten wir eine spezielle Notation für kleine skalare Fehlerterme kennengelernt, die wie folgt verallgemeinert werden kann. Wenn $r(h) \in \mathbb{R}^m$ und $s(h) \in \mathbb{R}^l$ zwei Größen sind, die beide von einer Variablen $h \in \mathbb{R}^n$ abhängen, so schreiben wir

$$r(h) = o\big(s(h)\big) \quad \text{für} \quad h \to 0 \qquad \text{bzw.} \qquad r(h) = O\big(s(h)\big) \quad \text{für} \quad h \to 0\,,$$

sofern

$$\lim_{h\to 0} \frac{\big|r(h)\big|}{\big|s(h)\big|} = 0 \qquad \text{bzw.} \qquad \limsup_{h\to 0} \frac{\big|r(h)\big|}{\big|s(h)\big|} < \infty$$

gilt, d.h. wenn im Limes $h \to 0$ der Betrag von $r(h)$ *schneller* bzw. *nicht langsamer* gegen 0 konvergiert als der Betrag von $s(h)$. Oftmals ist $s(h)$ eine Potenz von $|h|$, d.h. es gilt $l = 1$ sowie $s(h) = |h|^q$ für einen Exponenten $q > 0$, und in diesem Fall schreiben wir verkürzt

$$r(h) = o\big(\,|h|^q\,\big) \qquad \text{bzw.} \qquad r(h) = O\big(\,|h|^q\,\big)$$

und meinen

$$\lim_{h\to 0} \frac{\big|r(h)\big|}{|h|^q} = 0 \qquad \text{bzw.} \qquad \limsup_{h\to 0} \frac{\big|r(h)\big|}{|h|^q} < \infty\,.$$

Die Landau-Symbole sind ausgesprochen nützlich und erlauben es, kleine Fehlerterme in einer eleganten und intuitiven Weise zu behandeln. Es gelten aber auch einige auf den ersten Blick seltsam anmutende „Rechenregeln“. Zum Beispiel kann aus der richtigen Formel

$$O\big(|h|^2\big) + O\big(|h|^3\big) = O\big(|h|^2\big)$$

nicht $O\big(|h|^3\big) = 0$ geschlossen werden, sondern nur, dass die Summe eines Fehlerterms der Ordnung $|h|^2$ und eines anderen Fehlerterms der Ordnung $|h|^3$ insgesamt einen Fehler der Ordnung $|h|^2$ ergibt.

Merkregel: Die Symbole $o(|h|^q)$ und $O(|h|^q)$ bezeichnen keine konkreten Terme, sondern sie sind Platzhalter für eine ganze Klasse von Größen.

Klarstellung: Im Rahmen dieser Vorlesung ist die Benutzung der Landau-Symbole o und O rein fakultativ.

2.6 lokale Extrema in inneren Punkten

Definition Sei $f : U \to \mathbb{R}$ eine skalare Funktion.[44] Wir nennen einen Punkt $x_* \in U$ einen globalen Minimierer bzw. einen globalen Maximierer von f, falls

$$f(x_*) \leq f(x) \qquad \text{bzw.} \qquad f(x_*) \geq f(x)$$

für alle $x \in U$ gilt. Wir sprechen hingegen von einem lokalen Minimierer bzw. einem lokalen Maximierer, sofern ein Radius $\varepsilon > 0$ existiert, sodass die Ungleichung nicht unbedingt für alle $x \in U$, aber doch für alle $x \in U$ mit $|x - x_*| < \varepsilon$ erfüllt ist.

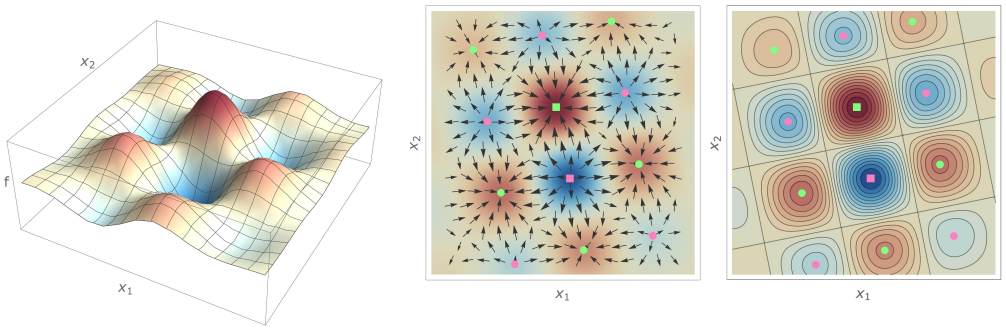

Abbildung Beispiel für eine Funktion $f : \mathbb{R}^2 \to \mathbb{R}$, wobei Blau bzw. Rot für kleine bzw. große Funktionswerte stehen und die Pfeile im mittleren Bild das Gradientenfeld illustrieren. Rosa bzw. Hellgrün repräsentieren Minimierer bzw. Maximimierer, wobei die lokalen bzw. globalen Extremstellen durch runde bzw. eckige Punkte dargestellt werden. Nicht eingezeichnet wurden die Sattelpunkte. Alle Extrema sind hier strikt und werden außerdem in inneren Punkten angenommen. Der sichtbare Rand ist nur die Grenze des Darstellungsbereiches, nicht des Definitionsbereiches von f.

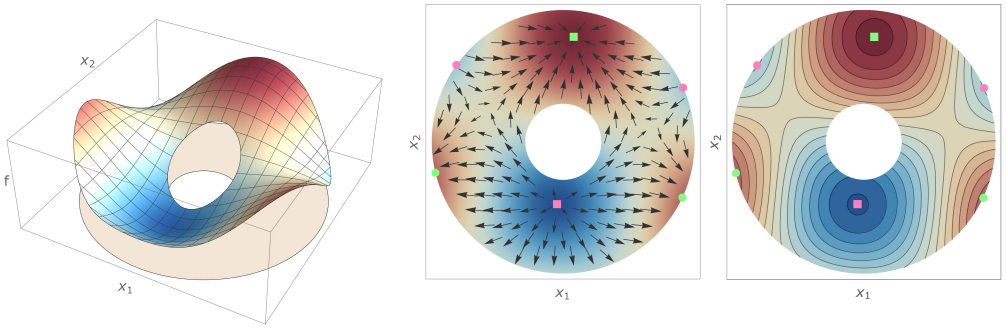

Abbildung Beispiel für eine Funktion $f : U \to \mathbb{R}$ auf einem abgeschlossenen Kreisring U, wobei es hier Extremstellen im Inneren sowie lokale Extremstellen auf dem Rand gibt. Letztere werden nicht durch die Theoreme in diesem Abschnitt abgedeckt, können aber mit der weiter unten erklärten *Multiplikatorenregel* berechnet werden. Die Farbkodierung ist wie im vorangegangenen Bild.

Bemerkung

1. Bei strikten Minimierern bzw. Maximierern (lokal oder global) gilt die jeweilige Ungleichung mit $<$ statt $\leq$ bzw. $>$ statt $\geq$ in allen Punkten $x \neq x_*$.

[44] Da im $\mathbb{R}^m$ mit $m > 1$ keine sinnvolle Ordnung existiert, haben die Begriffe *Minimum* und *Maximum* nur im Kontext skalarer Funktionen einen Sinn.

2. Ist x_* ein lokaler Minimierer, so wird $f(x_*)$ das entsprechende lokale Minimum genannt.[45] Minimierer werden oftmals auch Minimalstelle genannt, da in diesen Punkten ein Minimum angenommen wird. Die Begriffe globales Minimum, lokales Maximum, globales Maximum werden analog eingeführt.

3. Sowohl Minimierer als auch Maximierer werden oftmals als Extremstellen bezeichnet, wobei dann der entsprechende Funktionswert Extremum genannt wird. Bei globalen Extremstellen werden auch die Notationen

$$f(x_*) = \min f\,, \qquad x_* = \operatorname{argmin} f\,, \qquad f(x_*) = \max f\,, \qquad x_* = \operatorname{argmax} f$$

verwendet, wobei die Minimierer bzw. Maximierer aber nicht unbedingt eindeutig bestimmt sind, denn das globale Minimum bzw. Maximum kann in mehreren Punkten angenommen werden.

4. In der obigen Definition ist es zunächst nicht wichtig, ob $U \subset \mathbb{R}^n$ offen oder abgeschlossen oder weder noch ist, und im Prinzip können lokale oder globale Extrema in inneren Punkten oder in Randpunkten von U angenommen werden. Die jeweilige Theorie ist aber anders und wir werden uns in diesem Abschnitt auf Minima und Maxima konzentrieren, die in inneren Punkten angenommen werden. Extrema am Rand sind hingegen sehr viel schwieriger zu finden und zu klassifizieren, wobei wir erste Resultate weiter unten diskutieren werden.

Vereinbarung Im Folgenden ist $U \subset \mathbb{R}^n$ wieder offen und f eine skalare Funktion auf U. Insbesondere gibt es gar keine Randpunkte und jede Extremstelle von f ist automatisch ein innerer Punkt von U.

eindimensionaler Fall In *Analysis 1* hatten wir lokale Extrema für Funktionen $f : I \subseteq \mathbb{R} \to \mathbb{R}$ studiert und $f'(x_*) = 0$ als notwendige Bedingung für ein lokales Extremum in einem *inneren* Punkt x_* von I hergeleitet. Anschließend hatten wir mithilfe der eindimensionalen Variante des Satzes von Taylor die folgenden Aussagen abgeleitet:

1. Gilt $f''(x_*) > 0$, so nimmt f in x_* ein striktes lokales Minimum an.

2. Gilt $f''(x_*) < 0$, so nimmt f in x_* ein striktes lokales Maximum an.

3. Im Entartungsfall $f''(x_*) = 0$ ist keine einfache Entscheidung möglich bzw. die richtige Antwort hängt von den Eigenschaften höherer Ableitungen von f in x_* ab.

Wir hatten auch schon in Analysis 1 gesehen, dass in Extremstellen am Rand die Ableitung $f'(x_*)$ im Allgemeinen nicht verschwindet, sondern nur gewissen Vorzeichenbedingungen genügt.

Strategie Auch für $n \geq 2$ liefert der Satz von Taylor sowohl notwendige als auch hinreichende Kriterien für lokale Extrema, aber es gibt nicht nur eine erste und eine zweite Ableitung, sondern n bzw. n^2 partielle Ableitungen erster bzw. zweiter Ordnung.

[45] In der Literatur wird oftmals nicht sauber zwischen den Konzepten *Minimierer* und *Minimum* unterschieden, aber meist wird durch den Kontext klar, wovon gerade die Rede ist.

Wir hatten schon oben gesehen, dass diese in natürlicher Weise im Gradienten und in der quadratischen Hesse-Matrix

$$\operatorname{grad} f(x_*) = \begin{pmatrix} \partial_{x_1} f(x_*) \\ \vdots \\ \partial_{x_n} f(x_*) \end{pmatrix}, \qquad \operatorname{Hess} f(x_*) = \begin{pmatrix} \partial_{x_1}^2 f(x_*) & \dots & \partial_{x_n}\partial_{x_1} f(x_*) \\ \vdots & & \vdots \\ \partial_{x_1}\partial_{x_n} f(x_*) & \dots & \partial_{x_n}^2 f(x_*) \end{pmatrix}$$

gesammelt werden können. Sofern der Satz von Schwarz gilt, ist die Hesse-Matrix symmetrisch und kann daher im Reellen diagonalisiert werden.

notwendige Bedingung und erste Ableitungen

Lemma (**notwendige Bedingung für lokale Extremstellen**) Ist $f : U \to \mathbb{R}$ stetig differenzierbar, so gilt

$$\operatorname{grad} f(x_*) = 0\,,$$

sofern $x_* \in U$ ein lokaler Minimierer oder Maximierer ist.

Beweis *Vorbereitung*: Wir nehmen an, dass $v_* := \operatorname{grad} f(x_*)$ nicht der Nullvektor ist, d.h. dass $|v_*| > 0$ gilt, wählen $\varepsilon_* > 0$ mit $B_{\varepsilon_*}(x_*) \subset U$ und betrachten im Folgenden Punkte $x = x_* + h$ mit $0 < |h| < \varepsilon_*$. Der Satz von Taylor garantiert

$$f(x_* + h) = f(x_*) + v_* \cdot h + r_*(h)\,,$$

wobei der Fehlerterm $r_*(h) = R_{f,\,1,\,x_*}(x_* + h)$ der Konvergenzaussage

$$\frac{\big|r_*(h)\big|}{|h|} \quad \xrightarrow{h\to 0} \quad 0$$

genügt. Durch eventuelles Verkleinern von ε_* können wir daher sicherstellen, dass

$$\frac{\big|r_*(h)\big|}{|h|} \le \tfrac{1}{2}\,|v_*| \qquad \text{und damit} \qquad -\tfrac{1}{2}\,|v_*|\,|h| \le r_*(h) \le +\tfrac{1}{2}\,|v_*|\,|h|$$

für alle $h \in \mathbb{R}^n$ mit $|h| < \varepsilon_*$ gilt.

Hauptargument: Für alle $0 < \varepsilon < \varepsilon_*$ und mit der speziellen Wahl $h = +\varepsilon\, v_*/|v_*|$ erhalten wir

$$f(x_* + \varepsilon\, v_*/|v_*|) \ge f(x_*) + \varepsilon\,\frac{v_* \cdot v_*}{|v_*|} - \tfrac{1}{2}\,\varepsilon\,\frac{|v_*|^2}{|v_*|} = f(x_*) + \tfrac{1}{2}\,\varepsilon\,|v_*| > f(x_*)\,,$$

wohingegen sich

$$f(x_* - \varepsilon\, v_*/|v_*|) \le f(x_*) - \varepsilon\,\frac{v_* \cdot v_*}{|v_*|} + \tfrac{1}{2}\,\varepsilon\,\frac{|v_*|^2}{|v_*|} = f(x_*) - \tfrac{1}{2}\,\varepsilon\,|v_*| < f(x_*)$$

mit $h = -\varepsilon\, v_*/|v_*|$ ergibt (beachte, dass $v_* \cdot v_* = |v_*|^2$). Da die erste bzw. die zweite Abschätzung für alle hinreichend kleinen ε gilt, kann x_* kein lokaler Maximierer bzw. kein lokaler Minimierer von f sein.

Bemerkung

1. Im Lemma ist es nicht wichtig, ob es sich bei x_* um eine strikte oder eine nicht-strikte Extremstelle handelt.

2. Die Beweisidee kann wie folgt zusammengefasst werden: Wenn wir uns ausgehend von einem nicht-kritischen Punkt x_* mit Gradient $v_* = \operatorname{grad} f(x_*) \neq 0$ ein kleines Stück in Richtung $+v_*$ bzw. $-v_*$ bewegen, so wird der Wert von f mit Sicherheit zu- bzw. abnehmen. Dieselbe Idee hatten wir übrigens schon weiter oben benutzt, als wir gezeigt haben, dass der Gradient die Richtung des steilsten Anstiegs liefert.

3. Mit den Landau-Symbolen können die letzten beiden Formeln im Beweis auch als

$$f(x_* \pm \varepsilon\, v_*/\,|v_*|) = f(x_*) \pm \varepsilon\; |v_*| + o(\varepsilon)$$

   geschrieben werden und wir erkennen auch hier sehr schön den Kern des Arguments: Wegen $|v_*| \neq 0$ dominiert für kleine ε der erste Ordnungsterm alle Beiträge höherer Ordnung. Ist f sogar zweimal stetig differenzierbar, so können wir $o(\varepsilon)$ durch $O(\varepsilon^2)$ ersetzen und das Argument wird noch ein bisschen klarer.

4. Sowohl für das rigorose als auch das informelle Argument ist es sehr wichtig, dass die Extremstelle x_* im Inneren des Definitionsbereiches von U liegt (wobei sich dies bei uns sofort aus der angenommenen Offenheit von U ergibt). In der Tat, nur in einem inneren Punkt ist sichergestellt, dass wir uns ausgehend von x_* ein kleines Stück in jede denkbare Richtung bewegen können. Es kann natürlich passieren, dass eine Funktion f ein lokales oder gar globales Extremum in einem Randpunkt ihres Definitionsbereichs annimmt, aber dann muss der Gradient nicht mehr unbedingt verschwinden. Wir werden dieses Szenario weiter unten ausführlicher diskutieren.

Definition Wir nennen $x_* \in U$ einen kritischen Punkt von f, falls der Gradient von f in diesem Punkt verschwindet. Wir sprechen darüber hinaus von einem Sattelpunkt, wenn x_* weder lokaler Minimierer noch Maximierer von f ist.

Bemerkungen

1. Nicht jeder kritische Punkt ist eine lokale Extremstelle. Durch das Lösen der Gleichung $\operatorname{grad} f(x_*) = 0$ erhalten wir also nur *Kandidaten* für Extremstellen, wobei wir in einem zweiten Schritt durch Betrachtung der Hesse-Matrix oftmals entscheiden können, ob x_* ein lokaler Minimierer, ein lokaler Maximierer, oder ein Sattelpunkt ist. Die Details werden weiter unten erklärt.

2. Ein kritischer Punkt x_* ist genau dann Sattelpunkt, wenn für jeden Radius $\varepsilon > 0$ mindestens zwei Punkte x_- und x_+ in U existieren, sodass die Ungleichungen

$$f(x_-) < f(x_*) < f(x_+)\,, \qquad |x_- - x_*| < \varepsilon, \qquad |x_+ - x_*| < \varepsilon$$

   erfüllt sind.

über lokale Extremstellen am Rand Lokale Extrema in Randpunkten sind insofern anders, als dass dort der Gradient von f nicht verschwinden muss. Man kann geometrische Bedingungen an den Gradienten herleiten (siehe zum Beispiel die Diskussion der Multiplikatorenregel weiter unten). Für den Moment möchten wir mit einem Beispiel illustrieren, dass diese Gradientenbedingungen zumindest für $n = 2$ geometrisch sehr einleuchtend sind und mit einfachen „dynamischen“ Argumenten motiviert werden können.

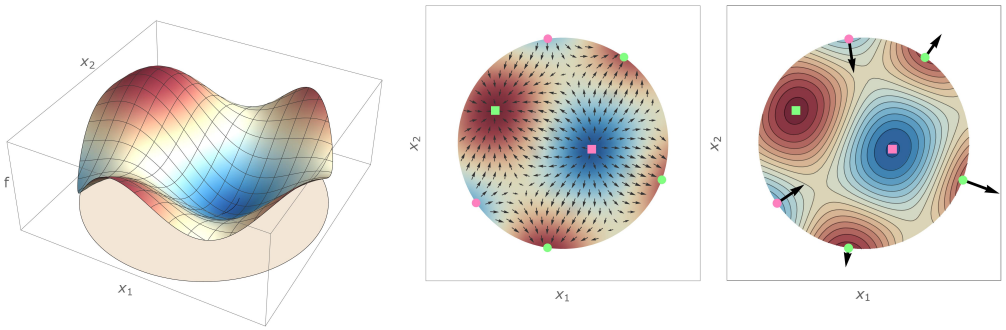

Abbildung Beispiel für eine skalare Funktion f auf einer *abgeschlossenen* Kreisscheibe, die lokale Minima und Maxima im Inneren und auf dem Rand annimmt, wobei im rechten Bild die jeweiligen Gradientenvektoren stark vergrößert dargestellt sind. Bei einer Maximalstelle auf dem Rand (hellgrüne runde Punkte) steht der Gradient senkrecht auf der Randkurve und zeigt nach außen. Denn würde er zum Beispiel nach innen zeigen, so könnten wir den Wert von f erhöhen, indem wir ein kleines Stück in Richtung des Gradienten ins Innere von D laufen. Und wenn der Gradient zwar nach außen zeigen, aber nicht senkrecht stehen würde, so könnten wir ihn in einen *normalen* und einen *tangentialen* Anteil zerlegen und den Wert von f dadurch erhöhen, dass wir uns entlang des Randes von D ein kleines Stück in tangentialer Richtung bewegen. Analog folgt, dass der Gradient in Minimierern auf dem Rand (rosa runde Kreise) auch senkrecht auf der Randkurve steht, aber diesmal nach innen zeigen wird.

Intermezzo

Resultate aus der linearen Algebra Jede symmetrische reelle $n\times n$-Matrix H besitzt n *reelle* Eigenwerte $\lambda_1, \dots, \lambda_n$ (die nicht unbedingt verschieden sein müssen) sowie eine Basis aus dazugehörigen *reellen* Eigenvektoren $e_1, \dots, e_n$, sodass die Formeln

$$H\,e_j = \lambda_j\,e_j\,, \qquad e_i \cdot e_j = \delta_i^j = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{für } i \neq j \end{cases}$$

für alle $i, j \in \{1, \dots, n\}$ erfüllt sind, wobei diese Gleichungen auch in matrizenwertiger Form als

$$E^T H\,E = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \qquad E^T E = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}, \qquad E := \begin{pmatrix} | & & | \\ e_1 & \dots & e_n \\ | & & | \end{pmatrix}$$

geschrieben werden können. Die orthogonale Matrix E beschreibt den Übergang in die *orthonormale Eigenbasis* von H und erfüllt $E^T = E^{-1}$. Insbesondere gelten die *Basis-Darstellungsformeln*

$$h = \sum_{j=1}^{n} \big(h \cdot e_j\big)\,e_j\,, \qquad H\,h = \sum_{j=1}^{n} \lambda_j\,\big(h \cdot e_j\big)\,e_j$$

sowie

$$h^T h = h \cdot h = |h|^2 = \sum_{j=1}^{n} \big(h \cdot e_j\big)^2 , \qquad h^T H\, h = h \cdot H\, h = \sum_{j=1}^{n} \lambda_j \big(h \cdot e_j\big)^2$$

für jeden Vektor $h \in \mathbb{R}^n$, wobei $\cdot$ das Skalarprodukt im $\mathbb{R}^n$ meint und h^T der zum Spaltenvektor h transponierte Zeilenvektor ist. Die Eigenwerte λ_j können als Nullstellen des charakteristischen Polynoms $\chi(\lambda) = \det(\lambda\,\mathsf{1} - H)$ berechnet werden und e_j liegt im Kern von $\lambda_j\,\mathsf{1} - H$.[46]

Definitheit symmetrischer Matrizen Mit den obigen Notationen und den Basis-Darstellungsformeln können wir leicht (Übungsaufgabe) die logischen Äquivalenzen

$$\lambda_j \geq +\mu \quad \text{für alle } \; j \in \{1, \ldots, n\} \quad \Longleftrightarrow \quad h^T H\, h \geq +\mu\, |h|^2 \quad \text{für alle } \; h \in \mathbb{R}^n$$

sowie

$$\lambda_j \leq -\mu \quad \text{für alle } \; j \in \{1, \ldots, n\} \qquad \Longleftrightarrow \qquad h^T H\, h \leq -\mu\, |h|^2 \quad \text{für alle } \; h \in \mathbb{R}^n$$

etablieren, wobei μ eine nicht-negative reelle Zahl bezeichnet. Gelten die beiden Teile der ersten/zweiten Äquivalenz für ein $\mu > 0$ bzw. für $\mu = 0$, so nennen wir H positiv definit/negativ definit bzw. positiv semidefinit/negativ semidefinit. In allen anderen Fällen besitzt H sowohl positive als auch negative Eigenwerte und wir sprechen von einer indefiniten Matrix.

Lokale Extrema und Eigenwerte der Hesse-Matrix Als Vorbereitung für den allgemeinen Fall wollen wir für ein quadratisches Polynom in zwei Variablen untersuchen, unter welchen Bedingungen an die zweiten Ableitungen ein kritischer Punkt als lokaler Minimierer oder lokaler Maximierer klassifiziert werden kann. Dazu betrachten wir eine reelle symmetrische Matrix

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} , \qquad H_{12} = H_{21}$$

sowie das quadratische Polynom $p : \mathbb{R}^2 \to \mathbb{R}$ mit

$$p(x) = c + \tfrac{1}{2}\, x^T H\, x = c + \tfrac{1}{2}\, H_{11}\, x_1^2 + H_{12}\, x_1\, x_2 + \tfrac{1}{2}\, H_{22}\, x_2^2 .$$

Der Ursprung $(0,\, 0)$ ist wegen $\operatorname{grad} p(0,\, 0) = 0$ immer ein kritischer Punkt, wobei $H = \operatorname{Hess} p(0,\, 0)$ gerade die entsprechende Hesse-Matrix ist.[47]

Spezialfall Diagonalmatrix: Unter der Annahme $H_{12} = H_{21} = 0$ sind $H_{11} = \lambda_1$ und $H_{12} = \lambda_2$ die beiden Eigenwerte von H und wir können leicht die folgenden Aussagen ableiten:

Standardfall $[--]$: Sind λ_1 und λ_2 beide negativ, so gilt

$$p(x_1,\, x_2) < p(0,\, 0)$$

für alle $x \in \mathbb{R}^2$ und p nimmt im Ursprung ein striktes lokales Maximum an.

[46] Alle diese Aussagen wurden in *Lineare Algebra* bewiesen, wobei dort sicher eine andere Notation verwendet wurde. Beachte auch, dass es bei einer symmetrischen Matriz keine Jordan-Eigenwerte geben kann und dass bei mehrfachen Eigenwerten die Eigenbasis nicht eindeutig bestimmt ist.

[47] Es gilt sogar $H = \operatorname{Hess} p(x_1,\, x_2)$, aber uns interessiert hier nur die Hesse-Matrix im Ursprung.

Standardfall $[++]$: Analog folgt

$$p(x_1, x_2) > p(0, 0)$$

falls λ_1 und λ_2 beide positiv sind, d.h. p besitzt im Ursprung ein striktes Maximum.

Standardfall $[-+]$: Im Fall von $\lambda_1 < 0 < \lambda_2$ gilt

$$p(x_1, 0) < p(0, 0) < p(0, x_2)\,,$$

d.h. der Ursprung ist weder Minimierer noch Maximierer, sondern ein Sattelpunkt. Dasselbe gilt im Fall $\lambda_2 < 0 < \lambda_1$.

Entartungsfälle $[-0]$, $[00]$ *und* $[0+]$: Gilt $\lambda_1 = 0$ und/oder $\lambda_2 = 0$, so besitzt die Diagonalmatrix H den (einfachen oder doppelten) Eigenwert 0 und wir sprechen von einem Entartungsfall. Für quadratische Polynome können wir zwar immer noch alles explizit ausrechnen, aber die Ergebnisse können nicht mehr so einfach auf den allgemeinen Fall übertragen werden. Beachte auch, dass hier $\det(H) = 0$ gilt.

allgemeiner Fall: Gilt $H_{12} = H_{21} \neq 0$, so müssen wir die symmetrische Matrix H zunächst diagonalisieren, indem wir ihre Eigenwerte und Eigenvektoren berechnen. Der Satz über die Hauptachsentransformation reeller symmetrischer Matrizen liefert uns eine orthogonale 2×2-Matrix E mit

$$E^T H\, E =: \tilde{H} = \begin{pmatrix} \tilde{H}_{11} & 0 \\ 0 & \tilde{H}_{22} \end{pmatrix}, \qquad E^T = E^{-1}\,,$$

wobei $\lambda_1 = \tilde{H}_{11}$ und $\lambda_2 = \tilde{H}_{22}$ die Eigenwerte von H sind und die entsprechenden *normalisierten* Eigenvektoren die Spalten von E bilden. Führen wir nun durch

$$x = E\,\tilde{x} \qquad \text{bzw.} \qquad \tilde{x} = E^T x$$

neue Koordinaten $\tilde{x}_1$ und $\tilde{x}_2$ ein, erhalten wir

$$p(x) = c + \tfrac{1}{2}\,\tilde{x}^T E^T H\, E\,\tilde{x} = c + \tfrac{1}{2}\,\tilde{x}^T \tilde{H}\,\tilde{x} = c + \tfrac{1}{2}\,\tilde{H}_{11}\,\tilde{x}_1^2 + \tfrac{1}{2}\,\tilde{H}_{22}\,\tilde{x}_2^2$$

und können analog zu oben die Standardfälle $[--]$, $[-+]$ und $[++]$ diskutieren. Es gibt natürlich auch wieder die Entartungsfälle mit $\det(H) = \det(\tilde{H}) = 0$.

Diskussion

1. Analoge Rechnungen können auch für $n \geq 3$ durchgeführt werden. Für $n = 4$ ergibt sich zum Beispiel im Standardfall $[----]$ bzw. $[++++]$ wieder ein lokales Maximum bzw. Minimum, wohingegen die Standardfälle $[---+]$, $[--++]$ und $[-+++]$ jeweils zu einem Sattelpunkt gehören.

2. Man könnte meinen, dass unsere Rechnungen nur spezielle quadratische Polynome abdecken. Der Satz von Taylor besagt aber, dass wir lokal jede (zweimal stetig differenzierbare) Funktion f durch ein quadratisches Polynom approximieren können. Genauer gesagt: Wird das zweite Taylor-Polynom von f in einem *kritischen* Punkt $x_* \in U$ ausgewertet, so verschwinden alle Monome vom Grad

1 (wegen grad $f(x_*) = 0$) und es bleiben nur die Monome nullten oder zweiten Grades stehen. Insbesondere gilt

$$f(x) \approx T_{f,2,x_*}(x) = c_* + \tfrac{1}{2}(x - x_*)^T H_* (x - x_*)$$

mit $c_* = f(x_*)$ und $H_* = \operatorname{Hess} f(x_*)$.

3. Die Idee ist, dass das Studium der zweiten Taylor-Approximation $T_{f,2,x_*}$ uns im Prinzip alle lokalen Informationen über f in der Nähe eines kritischen Punktes $x_* \in U$ liefert und dass die Beiträge der höheren Ordnungsterme vernachlässigt werden können (siehe das folgende Theorem). Dies ist allerdings nur „fast immer“ richtig, denn die Entartungsfälle mit $\det(H_*) = 0$ stellen die Ausnahme von der Regel dar. In diesen Fällen reichen die zweifachen partiellen Ableitungen eben nicht aus, um das lokale Verhalten von f vollständig zu charakterisieren.

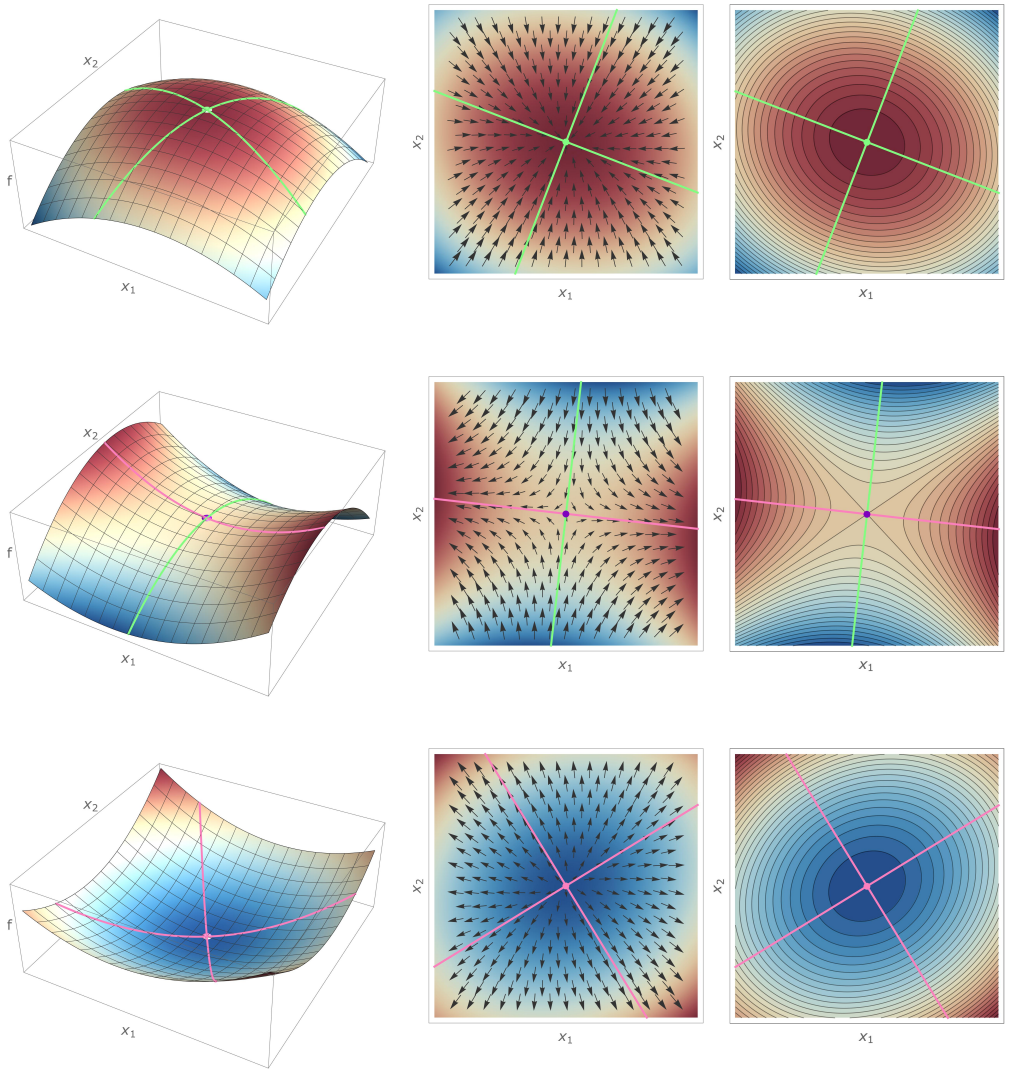

Abbildung Die drei Standardfälle in zwei Dimensionen, die das lokale Verhalten einer skalaren Funktion f in der Nähe eines nicht-entarteten kritischen Punktes x_* beschreiben. Die farbigen Kurven repräsentieren die *Hauptachsen*, d.h. die beiden Eigenrichtungen der regulären Hesse-Matrix $\operatorname{Hess} f(x_*)$, wobei rosa bzw. hellgrün den minimierenden bzw. maximierenden Richtungen und damit den Eigenvektoren zu positiven bzw. negativen Eigenwerten entsprechen. *Oben*: Ein striktes lokales Maximum mit zwei negativen Eigenwerten ($[--]$). *Mitte*: Ein Sattelpunkt mit einem negativen und einem positiven Eigenwert ($[-+]$). *Unten*: Ein striktes lokales Minimum mit zwei positiven Eigenwerten ($[++]$).

hinreichende Bedingung und zweite Ableitungen

Theorem (**hinreichende Bedingung für den Standardfall**) Sei $f : U \to \mathbb{R}$ zweimal stetig differenzierbar und sei $x_* \in U$ ein kritischer Punkt von f, der der Nicht-Entartungsbedingung

$$\det(H_*) \neq 0 \qquad \text{mit} \qquad H_* := \operatorname{Hess} f(x_*)$$

genügt. Dann ist x_*

1. ein strikter lokaler Minimierer, falls alle Eigenwerte von H_* positiv sind,
2. ein strikter lokaler Maximierer, falls alle Eigenwerte von H_* negativ sind,
3. ein Sattelpunkt, falls H_* sowohl positive als auch negative Eigenwerte besitzt.

Im Entartungsfall $\det(H_*) = 0$ verschwindet mindestens ein Eigenwert von H_* und x_* kann ohne weitere Informationen über das lokale Verhalten von f nicht klassifiziert werden.

Beweis *Vorbereitung*: Wie schon im Beweis des Lemmas wählen wir $\varepsilon_* > 0$ mit $B_{\varepsilon_*}(x_*) \subset U$ und betrachten Punkte $x = x_* + h$ mit $|h| < \varepsilon_*$. Wir verwenden auch wieder den Satz von Taylor, aber diesmal die zweite Ordnungsapproximation. Wegen $\operatorname{grad} f(x_*) = 0$ kann diese als

$$f(x_* + h) = f(x_*) + \tfrac{1}{2}\, h^T H_*\, h + r_*(h)\,,$$

geschrieben werden, wobei

$$\frac{\big|r_*(h)\big|}{|h|^2} \quad \xrightarrow{h\to 0} \quad 0$$

für den entsprechenden Fehlerterm $r_*(h) = R_{f,\,2,\,x_*}(x_* + h)$ gilt. Durch eventuelles Verkleinern von ε_* können wir daher sicherstellen, dass

$$\frac{\big|r_*(h)\big|}{|h|^2} \leq \tfrac{1}{4}\,\mu_* \qquad \text{und damit} \qquad -\tfrac{1}{4}\,\mu_*\,|h|^2 \leq r_*(h) \leq +\tfrac{1}{4}\,\mu_*\,|h|^2$$

für alle $h \in \mathbb{R}^n$ mit $|h| < \varepsilon_*$ gilt, wobei $\mu_* > 0$ den minimalen Betrag eines Eigenvektors von H_* gezeichnet.

Hauptargument 1: Nach Voraussetzung und aufgrund der Definitheitseigenschaften gilt $h^T H_*\, h \geq \mu_*\,|h|^2$ für alle $h \in \mathbb{R}^n$ und mit $0 < |h| < \varepsilon_*$ erhalten wir

$$f\big(x_* + h\big) \geq f(x_*) + \tfrac{1}{2}\,\mu_*\,|h|^2 - \tfrac{1}{4}\,\mu_*\,|h|^2 = f(x_*) + \tfrac{1}{4}\,\mu_*\,|h|^2 > f(x_*)\,.$$

Insbesondere ist x_* ein strikter Minimierer von f auf der Kugel $B_\varepsilon(x_*)$.

Hauptargument 2: Die Behauptung ergibt sich aus dem ersten Teil, sofern dieser nicht auf f, sondern auf die Funktion $-f$ angewendet wird. Alternativ können wir analog zu oben argumentieren, wobei diesmal $h^T H_*\, h \leq -\mu_*\,|h|^2$ für alle $h \in \mathbb{R}^n$ gilt.

Hauptargument 3: Nach Voraussetzung existieren zwei Eigenvektoren e_- bzw. e_+ von H_* mit $|e_-| = 1 = |e_+|$ sowie

$$e_-^T\, H_*\, e_- \leq -\mu_* \qquad \text{bzw.} \qquad e_+^T\, H_*\, e_+ \leq +\mu_*\,.$$

Mit analogen Abschätzungen zu oben und mit der speziellen Wahl $h = \varepsilon\, e_-$ bzw. $h = \varepsilon\, e_+$ zeigen wir, dass

$$f(x_* + \varepsilon\, e_-) \leq f(x_*) - \tfrac{1}{4}\,\varepsilon^2 \mu_*\,, \qquad \text{bzw.} \qquad f(x_* + \varepsilon\, e_+) \geq f(x_*) + \tfrac{1}{4}\,\varepsilon^2\,\mu_*$$

für alle $0 < \varepsilon < \varepsilon_*$ gilt, und schließen insgesamt, dass jede noch so kleine Umgebung von x_* Punkte enthält, in denen der Wert von f kleiner bzw. größer als $f(x_*)$ ist. □

Bemerkungen

1. Auch in Entartungsfällen gelingt oftmals eine Klassifikation mithilfe einer Taylor-Entwicklung von f in x_*. Allerdings müssen dann nicht nur die zweiten, sondern auch höhere (also dritte, vierte usw.) Ableitungen zu Rate gezogen werden. In *Analysis 1* hatten wir dies schon für eindimensionale Funktionen diskutiert und analoge Resultate gibt es auch für skalare Funktionen auf Teilmengen des $\mathbb{R}^n$. Die Details sind aber deutlich komplizierter, da wesentlich mehr Fallunterscheidungen zu treffen sind.

2. In der Literatur findet sich oftmals eine äquivalente Klassifikation, die auf dem Konzept der Definitheit beruht. Siehe dazu weiter oben.

3. Die Klassifikation nicht-entarteter kritischer Punkte erfolgt in der Praxis meist durch die (exakte oder approximative) Berechnung der Eigenwerte der Hesse-Matrix H_*, obwohl es durchaus andere Möglichkeiten gibt, zum Beispiel das *Kriterium von Sylvester und Hurwitz*. Allerdings setzt dieses die Kenntnis der Hauptminoren der Matrix H_* voraus und ist damit in der Regel auch nicht leicht auszuwerten. Für $n = 2$ gibt es jedoch erstaunlich einfache Kriterien, die nur die Determinante und die Spur von H_* involvieren und die wir im Anschluss diskutieren.

Klassifikationsregeln für $n = 2$ Aus den Resultaten der linearen Algebra ergibt sich unmittelbar das folgende Schema:[48]

(1) $\det(H_*) < 0$: *Sattelpunkt* (ein positiver und ein negativer Eigenwert)

(2) $\det(H_*) = 0$: *entartet* (mindestens ein Eigenwert verschwindet)

(3) $\det(H_*) > 0$: (beide Eigenwerte haben dasselbe Vorzeichen)

  (3.1) $\operatorname{tr}(H_*) < 0$: *striktes Maximum* (beide Eigenwerte sind negativ)

  (3.2) $\operatorname{tr}(H_*) > 0$: *striktes Minimum* (beide Eigenwerte sind positiv)

Geometrie von Sattelpunkten Ist x_* ein nicht-entarteter Sattelpunkt von f, so beschreibt jeder Eigenvektor e zu einem negativen bzw. positiven Eigenwert λ von H_* eine maximierende bzw. eine minimierende Richtung, wobei dies wie folgt verstanden werden kann: Die parametrisierte Kurve $\gamma : I \to \mathbb{R}^n$ sowie die Funktion $g : I \to \mathbb{R}$ mit

$$\gamma(t) := x_* + t\,e\,, \qquad g(t) := f\big(\gamma(t)\big)$$

sind auf einem offenen Intervall $(-\varepsilon, \varepsilon)$ wohldefiniert und der allgemeine Satz von Taylor liefert

$$\begin{aligned} g(t) = f\big(x_* + t\,e\big) &= f(x_*) + \operatorname{grad} f(x_*) \cdot (t\,e) + \tfrac{1}{2}\,(t\,e)^T \operatorname{Hess} f(x_*)(t\,e) + o(|t\,e|^2) \\ &= f(x_*) + \tfrac{1}{2}\,t^2\,e^T H_* e + o(t^2) = f(x_*) + \tfrac{1}{2}\,\lambda\,t^2 + o(t^2) \end{aligned}$$

als eindimensionale Taylor-Entwicklung von g. Insbesondere gilt

$$g(0) = f(x_*)\,, \qquad \dot g(0) = 0\,, \qquad \ddot g(0) = \lambda\,,$$

[48] Die wesentliche Beobachtung ist, dass für jede symmetrische 2×2-Matrix H mit Eigenwerten λ_1 und λ_2 die Formeln $\det(H) = \lambda_1\,\lambda_2$ und $\operatorname{tr}(A) = \lambda_1 + \lambda_2$ gelten.

d.h. g besitzt für $\lambda < 0$ bzw. $\lambda > 0$ in $t = 0$ ein lokales Maximum bzw. Minimum. Oder anders gesagt: Wenn wir f auf das Bild der Kurve γ einschränken, dann ist x_* ein lokaler Maximierer bzw. ein Minimierer.

Bemerkung: In einem nicht-entarteten Sattelpunkt summieren sich die Anzahlen der maximierenden und der minimierenden Richtungen immer zu n. Nicht-entartete Minimierer bzw. Maximierer besitzen in diesem Sinne genau n minimierende bzw. maximierende Richtungen.

Beispiel Wir betrachten $f : \mathbb{R}^2 \to \mathbb{R}$ mit

$$f(x_1,\, x_2) = (x_1 + x_2)^3 + (x_1 - x_2)^3 - 6\, x_1$$

und berechnen zunächst den Gradienten

$$\operatorname{grad} f(x_1,\, x_2) = \begin{pmatrix} 3\,(x_1 + x_2)^2 + 3\,(x_1 - x_2)^2 - 6 \\ 3\,(x_1 + x_2)^2 - 3\,(x_1 - x_2)^2 \end{pmatrix} = \begin{pmatrix} 6\, x_1^2 + 6\, x_2^2 - 6 \\ 12\, x_1\, x_2 \end{pmatrix}$$

sowie die Hesse-Matrix

$$\operatorname{Hess} f(x_1,\, x_2) = \begin{pmatrix} 12\, x_1 & 12\, x_2 \\ 12\, x_2 & 12\, x_1 \end{pmatrix} .$$

Die nichtlinearen Gleichungen für die kritischen Punkte lauten

$$0 \overset{!}{=} 6\, x_1^2 + 6\, x_2^2 - 6\,, \qquad 0 \overset{!}{=} 12\, x_1\, x_2$$

und können in diesem Beispiel leicht gelöst werden. Wir erhalten die vier kritischen Punkte

$$(1.)\ \ x_* = (-1,\, 0)\,, \quad (2.)\ \ x_* = (+1,\, 0)\,, \quad (3.)\ \ x_* = (0,\, -1)\,, \quad (4.)\ \ x_* = (0,\, +1)\,,$$

die wir nun der Reihe nach klassifizieren wollen. Wir wollen dabei aber nicht das Spur-Determinanten-Kriterium verwenden, sondern direkt mit den Eigenwerten und -vektoren argumentieren.

Kritische Punkte 1 und 2: Es gilt

$$H_* = \begin{pmatrix} \mp 12 & 0 \\ 0 & \mp 12 \end{pmatrix}$$

und wegen der Diagonalstruktur können wir die Eigenwerte sofort ablesen: Wir erhalten den jeweils doppelten Eigenwert -12 bzw. $+12$ und schließen, dass f im ersten bzw. zweiten kritischen Punkt ein striktes lokales Maximum bzw. Minimum annimmt.

Kritische Punkte 3 und 4: Diesmal erhalten wir keine Diagonalmatrix und müssen in einer Nebenrechnung die Matrix diagonalisieren. Dies liefert

$$H_* = \begin{pmatrix} 0 & \mp 12 \\ \mp 12 & 0 \end{pmatrix} , \qquad E^{-1}\, H_*\, E = \begin{pmatrix} -12 & 0 \\ 0 & +12 \end{pmatrix} ,$$

wobei die normalisierten Eigenvektoren spaltenweise aus der orthogonalen Übergangsmatrix

$$E = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \qquad \text{bzw.} \qquad E = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$$

abgelesen werden können. Aufgrund der verschiedenen Vorzeichen der Eigenwerte schließen wir, dass es sich sowohl beim dritten als auch beim vierten kritischen Punkt jeweils um einen Sattelpunkt handelt.

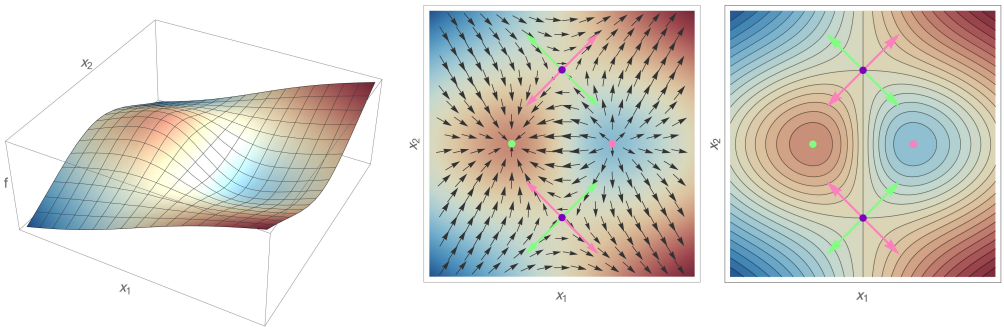

Abbildung Die (leicht verzerrt dargestellten) Plots zum gerechneten Beispiel. Die kritischen Punkte sind farbig markiert: rosa bzw. hellgrün für den lokalen Minimierer bzw. Maximierer, lila für die beiden Sattelpunkte). Die Pfeile in den Sattelpunkten repräsentieren die jeweiligen Hauptachsen und damit auch die minimierenden und maximierenden Richtungen.

Gegenbeispiele Die drei Formeln

$$f(x_1, x_2) = x_1^4 + x_2^4\,, \qquad f(x_1, x_2) = x_1^2\,x_2 - x_1\,x_2^2\,, \qquad f(x_1, x_2) = x_1^3 + x_2^4$$

beschreiben jeweils eine Funktion f, für die der Ursprung $(0, 0)$ ein kritischer Punkt ist. Da aber die dazugehörige Hesse-Matrix jeweils die Nullmatrix ist, liefert das Theorem keine entsprechenden Klassifikationen. Durch Auswertung der dritten und vierten partiellen Ableitungen (oder mit anderen Methoden) kann jedoch gezeigt werden, dass der Ursprung im ersten Fall ein strikter lokaler Minimierer, in den beiden anderen Fällen aber ein (jeweils entarteter) Sattelpunkt ist.

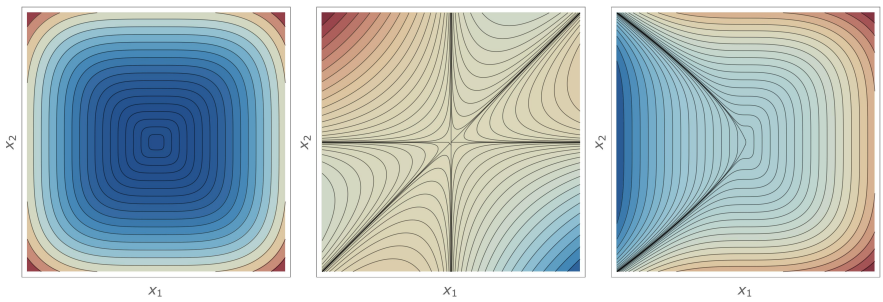

Abbildung Die Konturplots der drei Gegenbeispiele, wobei der entartete kritische Punkt $(0, 0)$ jeweils der Mittelpunkt des dargestellten Bereichs ist.

Klarstellung Durch das Studium von Ableitungen kann *generell nicht* entschieden werden, ob es sich um lokale oder globale Extremstellen handelt, denn Ableitungen charakterisieren nur das *lokale Verhalten* von f in der Nähe eines kritischen Punktes. Die Suche nach globalen Extremstellen ist ein globales Problem und kann sehr schnell sehr aufwendig werden. Es sei denn, die Funktion f und oder die Menge U erfüllen sehr einschränkende Bedingungen.

Extrema unter Nebenbedingungen und Multiplikatorenregel

Vorbemerkung Wir betrachten wieder eine skalare Funktion $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$, suchen aber diesmal Extremstellen von f unter m Gleichungsbedingungen

$$g_1(x) = c_1\,, \qquad \dots\,, \qquad g_m(x) = c_m\,,$$

die durch eine vektorwertige Funktion $g : U \to \mathbb{R}^m$ sowie einen Vektor $c \in \mathbb{R}^m$ festgelegt werden. Wir können in diesem Abschnitt noch keine Theoreme beweisen, wollen aber eine wichtige Methode vorstellen, mit der solche Probleme in der Praxis gelöst werden können. Beim *Satz über implizite Funktionen* werden wir dann einige der theoretischen Grundlagen bereitstellen.

Bemerkung

1. Das Finden von Minima und Maxima ist das prototypische Grundproblem der *Mathematischen Optimierung*. Die Menge

   $$G := N_c(g) = \big\{x \in U \,:\, g_i(x) = c_i\,,\;\; i \in \{1, \dots, m\}\big\}\,,$$

   ist dabei die Menge der zulässigen Punkte, f wird Zielfunktion genannt und die Gleichungen $g_i(x) = c_i$ sind die Nebenbedingungen.

2. In aller Regel gilt $m < n$, denn jede skalare Gleichung der Form $g_i(x) = c_i$ reduziert – zumindest auf einer heuristischen oder informellen Ebene – die Anzahl der „Freiheitsgrade“ um 1.

3. Analog zur Diskussion im vorangegangenen Abschnitt können wir über lokale und globale Maxima und Minima sowie die entsprechenden Minimierer und Maximierer der Funktion f auf der Menge G reden. Allerdings wird G in der Regel keine inneren Punkte besitzen. Wir können daher nicht erwarten, dass der Gradient von f in lokalen Extremstellen verschwindet.

Prinzip (**Multiplikatorenregel**) Sei x_* eine lokale Extremstelle von f in G. Dann existiert ein $\lambda_* \in \mathbb{R}^m$, sodass

$$\operatorname{grad} f(x_*) = \sum_{i=1}^{m} \lambda_{*,i} \operatorname{grad} g_i(x_*) \qquad \text{bzw.} \qquad \operatorname{Jac} f(x_*) = \lambda_*^T \operatorname{Jac} g(x_*)$$

gilt. Insbesondere kann im Punkt x_* der Gradient von f als Linearkombination der Gradienten der Funktionen g_i dargestellt werden.

Bemerkung

1. Die Komponenten von λ_* werden Lagrange-Multiplikatoren genannt. In der Literatur wird die Gradientenformel oftmals als $\operatorname{grad}_x L(x_*, \lambda_*) = 0$ angegeben, wobei $L(x, \lambda) := f(x) - \lambda \cdot g(x)$ die Lagrange-Funktion ist und $\operatorname{grad}_x L(x, \lambda)$ den n-dimensionalen Vektor bezeichnet, der aus den partiellen Ableitungen $\partial_{x_j} L$ besteht.

2. Im Fall $m = 1$ (eine skalare Nebenbedingung) lautet die Multiplikatorenregel

   $$\operatorname{grad} f(x_*) = \lambda_* \operatorname{grad} g(x_*)$$

   und besitzt eine unmittelbare geometrische Interpretation. Siehe dazu das Bild.

3. Die Multiplikatorenregel liefert in Kombination mit den Nebenbedingungen insgesamt $n + m$ Gleichungen für die n Komponenten von x_* sowie die m Komponenten von λ_*. Diese Gleichungen sind in der Regel nichtlinear und damit nicht unbedingt einfach zu lösen. Außerdem kann es mehrere Lösungen geben.

4. Die Multiplikatorenregel formuliert eine *notwendige Bedingung* und liefert daher zunächst nur *Kandidaten* für die Extremstellen von f in G. Die mathematischen Resultate zur systematischen Klassifikation können wir hier nicht besprechen, aber in der Praxis ist es oftmals gar nicht so schwierig zu entscheiden, ob es sich um Minimierer, Maximierer oder Sattelpunkte handelt.

5. Die Multplikatorenregel kann im Prinzip auch verwendet werden, wenn einige oder alle der Nebenbedingungen Ungleichungen sind. Siehe dazu auch die Diskussion des einfachen Spezialfalls weiter unten. In der Optimierung spricht man dann meist von den *Karush-Kuhn-Tucker-* oder *KKT-Bedingungen.*

6. In der rigorosen Theorie wird das Prinzip durch gewisse Differenzierbarkeit- und Nicht-Entartungsannahmen an die Funktionen f und g_i sowie ihre Ableitungen ergänzt.

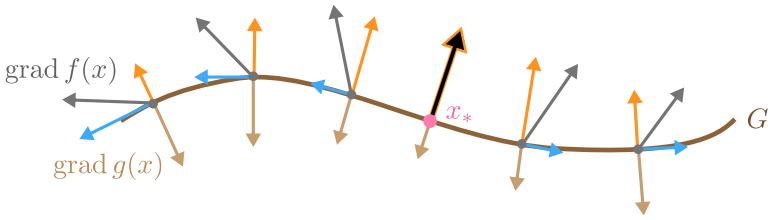

Abbildung Geometrische Bedeutung der Multiplikatorenregel für $n = 2$ und $m = 1$: In jedem Punkt $x \in G = N_g(c)$ steht $\operatorname{grad} g(x)$ (hellbraun) senkrecht auf G (dunkelbraun) und $\operatorname{grad} f(x)$ (grau) kann in einen *tangentialen* (blau) und einen *normalen* Anteil (orange) zerlegt werden. Die Multiplikatorenregel besagt, dass in einer lokalen Extremstelle x_* von f in G (rosa) der tangentiale Anteil von f verschwindet bzw. dass die Gradienten von f und g parallel sind und sich nur durch den skalaren Faktor λ_* unterscheiden. Ist diese Bedingung verletzt, so kann x_* weder Minimierer noch Maximierer sein, denn wir können den Wert von f dadurch vergrößern bzw. verkleinern, indem wir uns ein kleines Stück mit bzw. gegen den tangentialen Vektor entlang von G bewegen.

Beispiel Für $n = 2$, $m = 1$ und $U = \mathbb{R}^2$ betrachten wir die Funktionen

$$f(x_1, x_2) = x_1 x_2\,, \qquad g(x_1, x_2) = x_1^2 + x_2^2\,,$$

d.h. für $c > 0$ ist G die Kreislinie vom Radius $\sqrt{c}$ um den Ursprung (für $c \leq 0$ ist das Problem entartet). Die Multiplikatorenregel liefert die vektorwertige Gleichung

$$\begin{pmatrix} x_{*,2} \\ x_{*,1} \end{pmatrix} = \operatorname{grad} f(x_{*,1}, x_{*,2}) = \lambda_* \operatorname{grad} g(x_{*,1}, x_{*,2}) = \begin{pmatrix} 2\,\lambda_* x_{*,1} \\ 2\,\lambda_* x_{*,2} \end{pmatrix},$$

und außerdem soll

$$x_{*,1}^2 + x_{*,2}^2 = g(x_{*,1}, x_{*,2}) = c$$

gelten. Es gibt also *drei* nichtlineare skalare Gleichungen für die *drei* Unbekannten $x_{*,1}$, $x_{*,2}$ und λ_*, die wir wie folgt berechnen können: Wir bemerken zunächst, dass $x_{*,1} \neq 0$ gelten muss, da andernfalls die erste Gleichung $x_{*,2} = 0$ und damit einen Widerspruch zur dritten Gleichung liefern würde. Analog zeigen wir $x_{*,2} \neq 0$. Setzen wir die erste Gleichung in die zweite ein, so erhalten wir $x_{*,1} = 4\,\lambda_*^2\, x_{*,1}$ und damit $\lambda_* = \pm\frac{1}{2}$. Die

erste und die zweite Gleichung implizieren nun $x_{*,2} = \pm x_{*,1}$ und die Nebenbedingung liefert jeweils $2\,x_{*,1}^2 = 2\,x_{*,2}^2 = c$. Insgesamt erhalten wir die folgenden vier Kandidaten für die Extremstellen von f in G:

$$\text{(1)}\ \ \lambda_* = -\tfrac{1}{2},\quad x_{*,1} = -\sqrt{\tfrac{1}{2}\,c},\quad x_{*,2} = +\sqrt{\tfrac{1}{2}\,c} \qquad \Big(\text{mit } f(x_{*,1},\, x_{*,2}) = -\tfrac{1}{2}\,c\Big)$$

$$\text{(2)}\ \ \lambda_* = -\tfrac{1}{2},\quad x_{*,1} = +\sqrt{\tfrac{1}{2}\,c},\quad x_{*,2} = -\sqrt{\tfrac{1}{2}\,c} \qquad \Big(\text{mit } f(x_{*,1},\, x_{*,2}) = -\tfrac{1}{2}\,c\Big)$$

$$\text{(3)}\ \ \lambda_* = +\tfrac{1}{2},\quad x_{*,1} = -\sqrt{\tfrac{1}{2}\,c},\quad x_{*,2} = -\sqrt{\tfrac{1}{2}\,c} \qquad \Big(\text{mit } f(x_{*,1},\, x_{*,2}) = +\tfrac{1}{2}\,c\Big)$$

$$\text{(4)}\ \ \lambda_* = +\tfrac{1}{2},\quad x_{*,1} = +\sqrt{\tfrac{1}{2}\,c},\quad x_{*,2} = +\sqrt{\tfrac{1}{2}\,c} \qquad \Big(\text{mit } f(x_{*,1},\, x_{*,2}) = +\tfrac{1}{2}\,c\Big)$$

Aus der Multiplikatoren-Regel können wir aber weder ablesen, ob es sich um Minimierer oder Maximierer handelt, noch ob es sich um lokale oder globale Versionen handelt.

Bemerkungen:

1. Im konkreten Fall ist G eine kompakte Menge und f eine stetige Funktion. Daher muss es ein globales Minimum und ein globales Maximum geben, aber dieses Existenzargument ist unabhängig von der Multiplikatorenregel. Außerdem werden in den vier Kandidaten nur zwei verschiedene Funktionswerte angenommen. Insgesamt folgt, dass die Formeln (1) und (2) beide einem globalen Minimum, die Formeln (3) und (4) jeweils einem globalen Maximum entsprechen.

2. Wir haben hier die Gleichungen in $x_{*,1}$, $x_{*,2}$ und λ_* formuliert. In der Praxis lässt man den Index $*$ meist weg und rechnet mit x_1, x_2 und λ.

Alternative Lösung: In diesem Beispiel ist G das Bild der parametrisierten Kurve

$$\gamma(t) = \begin{pmatrix} \sqrt{c}\,\cos(t) \\ \sqrt{c}\,\sin(t) \end{pmatrix}, \qquad t \in [0,\, 2\pi]$$

und anstelle von f können wir die eingeschränkte Funktion $\tilde{f} : [0,\, 2\pi] \to \mathbb{R}$ mit

$$\tilde{f}(t) = f\big(\gamma(t)\big) = c\cos(t)\sin(t) = \tfrac{1}{2}\,c\sin(2\,t)$$

mithilfe einer eindimensionalen Kurvendiskussionen untersuchen. Diese zeigt, dass $\tilde{f}$ in $t_* = \frac{1}{4}\pi$ und $t_* = \frac{5}{4}\pi$ ein globales Maximum annimmt, wohingegen $t_* = \frac{3}{4}\pi$ und $t_* = \frac{7}{4}\pi$ globale Minimierer sind (siehe das Bild).

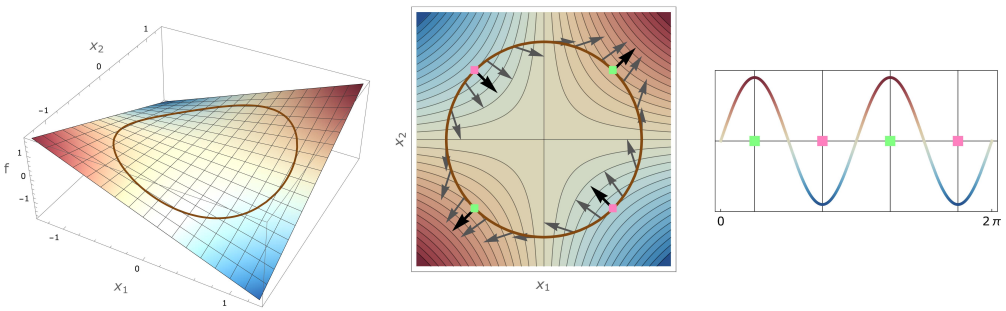

Abbildung Die Bilder zum Beispiel. *Links*: Flächenplot von f. *Mitte*: Konturplot von f und Menge G (braun). Die grauen bzw. schwarzen Pfeile stellen grad $f(x)$ in ausgewählten Punkten aus G bzw. in den vier Extremstellen dar. Die schwarzen Pfeile stehen gemäß der Multiplikatorenregel senkrecht auf G. *Rechts*: Graph der reduzierten Funktion $\tilde{f}$.

Bemerkung Ganz allgemein gilt: Ist eine Parametrisierung γ der Menge G (als Kurve oder Fläche oder Mannigfaltigkeit) bekannt, so kann durch Einführung der reduzierten Funktion $\tilde{f} = f \circ \gamma$ das Optimierungsproblem mit Nebenbedingung in ein Problem ohne Nebenbedingung überführt werden. Insbesondere können die Extremstellen von $\tilde{f}$ mithilfe von grad $\tilde{f}$ und Hess $\tilde{f}$ gesucht und charakterisiert werden. Allerdings ist eine explizite Parametrisierung von G oftmals entweder nicht verfügbar oder produziert sehr komplizierte Ausdrücke für $\tilde{f}$.

über Nebenbedingungen in Form von Ungleichungen* Als einfachsten Fall betrachten wir eine Funktion f in zwei Variablen auf der Menge

$$V := \{x \in \mathbb{R}^2 \,:\, g(x) \leq c\},$$

wobei $g : \mathbb{R}^2 \to \mathbb{R}$ eine hinreichend gute Funktion ist. Insbesondere ist dann die Menge $G = \{x \in \mathbb{R}^2 \,:\, g(x) = c\}$ gerade der Rand von V. Für jede Extremstelle x_* von f auf der Menge V wird die Multiplikatorenregel

$$\operatorname{grad} f(x_*) = \lambda_* \operatorname{grad} g(x_*) \qquad \text{mit}$$

für ein $\lambda_* \in \mathbb{R}$ erfüllt sein, aber diese Aussage kann in zweifacher Weise ergänzt werden:

1. $g(x_*) < c$ impliziert $\lambda_* = 0$, denn wenn die Ungleichung in der Extremstelle x_* strikt ist, so ist x_* ein innerer Punkt von V und wir hatten schon gesehen, dass dann grad $f(x_*) = 0$ gelten muss. Liegt x_* jedoch am Rand von V und damit in der Menge G, so müssen die Gradienten von f und g in x_* parallel sein und λ_* wird in aller Regel nicht verschwinden.

2. Ist x_* ein Minimierer bzw. Maximierer von f am Rand von V, so gilt $\lambda_* \leq 0$ bzw. $\lambda_* \geq 0$. Geometrisch meint dies, dass in einer Extremstelle $x_* \in G$ der Gradient von f aus Sicht von V immer nach innen bzw. nach außen zeigt, denn der Gradient von g zeigt immer in das Außengebiet von V.

Diese nützlichen Erweiterung der Multiplikatorenregel gelten nur bei Ungleichungsnebenbedingungen und können auch wieder mit „dynamischen“ Argumenten begründet und verstanden werden. Siehe dazu das folgende Bild sowie das Beispiel oben zu Extremstellen am Rand.

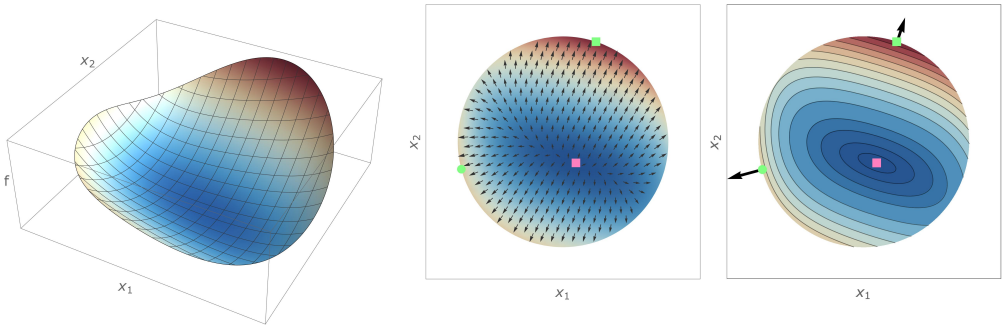

Abbildung Eine skalare Funktion f auf einer abgeschlossenen Kreisscheibe V, die ihr globales Minimum (rosa) in einem inneren Punkt annimmt, in dem der Gradient von f notwendigerweise verschwindet. Die Funktion besitzt außerdem einen lokalen sowie einen globalen Maximierer auf dem Rand (jeweils hellgrün), wobei der Gradient von f in jedem dieser Punkte senkrecht auf dem Rand von V steht und *nach außen* zeigt (siehe die nicht-maßstabsgerechten Pfeile im rechten Bild). Insbesondere gilt in jeder der drei Extremstellen die oben beschriebene erweiterte Multiplikatorenregel. Ein weiteres Beispiel für eine Funktion auf einer abgeschlossenen Kreisscheibe hatten wir bereits oben diskutiert.

Vorlesungswoche 09

Einschub: Funktionen von Matrizen und Differenzierbarkeit*

Vorbemerkung In diesem Abschnitt stellen wir einige nützliche Resultate zur Differenzierbarkeit von Funktionen F zusammen, die auf einer offenen Teilmenge $W \subseteq \mathbb{M}^{n\times n}$ definiert sind und Werte in $\mathbb{M}^{n\times n}$ oder in $\mathbb{R}$ annehmen, wobei $\mathbb{M}^{n\times n}$ den Vektorraum aller reellen $n\times n$-Matrizen bezeichnet.[49] Aus theoretischer Sicht sind alle Aussagen durch die vorherigen Abschnitte abgedeckt, denn wegen $\mathbb{M}^{n\times n} \cong \mathbb{R}^{n^2}$ kann jede solche Funktion F auch als Abbildung f auf einer Teilmenge des $\mathbb{R}^{n^2}$ interpretiert werden, aber aus der Existenz der Matrizenmultiplikation ergeben sich einige Besonderheiten. Insbesondere zeigt es sich, dass die totale Ableitung von F meist viel einfacher berechnet werden kann als die Gesamtheit aller partiellen Ableitungen von f.

Erinnerung $F : W \to \mathbb{M}^{n\times n}$ ist genau dann total differenzierbar, wenn für jedes $X \in W$ eine lineare Abbildung $\mathrm{D}F(X) : \mathbb{M}^{n\times n} \to \mathbb{M}^{n\times n}$ existiert, sodass

$$\frac{\big|F(X+H) - F(X) - \big(\mathrm{D}F(X)\big)(H)\big|}{|H|} \quad \xrightarrow{H\to 0} \quad 0$$

gilt, wobei H eine von Null verschiedene $n\times n$-Matrix repräsentiert und $|\cdot|$ sich im Zähler und im Nenner auf den euklidischen Betrag von Matrizen bezieht. Eine analoge Aussage betrifft die totale Differenzierbarkeit skalarer Funktionen $F : W \to \mathbb{R}$, wobei dann $\mathrm{D}F(X)$ dann eine lineare Abbildung von $\mathbb{M}^{n\times n}$ nach $\mathbb{R}$ ist.

Beispiel Die totale Ableitungsformel

$$\big(\mathrm{D}F(X)\big)(H) = X\,H + H\,X \qquad \text{für} \qquad F(X) = X^2$$

kann mit Matrizenmultiplikation direkt nachgerechnet werden.[50] Insbesondere ergibt sich die Restgliedformel

$$R(X) = (X+H)^2 - X^2 - \big(X\,H + H\,X\big) = H^2$$

und der Kompatibilitätssatz garantiert $|R(H)| = |H^2| \leq |H|^2$.

Lemma (**Ableitung der Determinante**) Die Funktion $\det : \mathbb{M}^{n\times n} \to \mathbb{R}$ ist für jedes $n \in \mathbb{N}$ total differenzierbar und damit auch stetig. Für $n = 2$ bzw. $n = 3$ gilt

$$\big(\mathrm{D}\det(X)\big)(H) = \det\begin{pmatrix} H_{11} & X_{12} \\ H_{21} & X_{22} \end{pmatrix} + \det\begin{pmatrix} X_{11} & H_{12} \\ X_{21} & H_{22} \end{pmatrix}$$

bzw.

$$\big(\mathrm{D}\det(X)\big)(H) = \det\begin{pmatrix} H_{11} & X_{12} & X_{13} \\ H_{21} & X_{22} & X_{23} \\ H_{31} & X_{32} & X_{33} \end{pmatrix} + \det\begin{pmatrix} X_{11} & H_{12} & X_{13} \\ X_{21} & H_{22} & X_{23} \\ X_{31} & H_{32} & X_{33} \end{pmatrix} + \det\begin{pmatrix} X_{11} & X_{12} & H_{13} \\ X_{21} & X_{22} & H_{23} \\ X_{31} & X_{32} & H_{33} \end{pmatrix}$$

[49] Die Menge $\mathbb{M}^{n\times n}$ ist natürlich nicht nur Vektorraum, sondern sogar Algebra. Sie ist aber für $n > 1$ kein Körper, da nicht jede von Null verschiedene quadratische Matrix invertiert werden kann.

[50] Wenn wir die konkrete Funktion F (Quadrierung einer Matrix) als Abbildung $f : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ schreiben, so wird die entsprechende Jacobi-Matrix n^4 Einträge umfassen. Für $n = 2/3/4/5$ sind das bereits 16/81/256/625 verschiedene partielle Ableitungen!

und für jeden anderen Wert von n gibt es eine analoge Formel mit einer Summe von n Determinanten auf der rechten Seite, wobei in jeder der auftretenden Matrizen genau eine Spalte von H, die anderen aber von X stammen.

Beweis Alle Behauptungen ergeben sich unmittelbar aus der Tatsache, dass die Determinante einer Matrix eine Summe von Produkten aus jeweils n Faktoren ist und linear von jeder Spalte abhängt.[51] □

Lemma (**Ableitung der Inversionsabbildung**) Die Menge

$$W := \big\{ X \in \mathbb{M}^{n\times n} \,:\, \det(X) \neq 0 \big\}$$

ist offen und die Abbildung $\operatorname{inv} : W \to W$ mit $\operatorname{inv}(X) := X^{-1}$ ist sowohl stetig als auch total differenzierbar, wobei

$$\big(\mathrm{D}\operatorname{inv}(X)\big)(H) = -\operatorname{inv}(X)\, H\, \operatorname{inv}(X) = -X^{-1} H X^{-1}$$

die Formel der totalen Ableitung ist.

Beweis Die Stetigkeit der Determinantenabbildung impliziert die Offenheit von W und in Kombination mit der Cramerschen Regel auch die Stetigkeit von inv.[52] Um die Formel für $D\operatorname{inv}(X)$ zu beweisen, fixieren wir X mit $\det X \neq 0$ und betrachten für alle H mit hinreichend kleinem Betrag das Restglied

$$\begin{aligned} R(H) &:= (X+H)^{-1} - X^{-1} + X^{-1} H X^{-1} \\ &= (X+H)^{-1}\Big(\mathsf{1} - (X+H)X^{-1} + (X+H)X^{-1}HX^{-1}\Big) \\ &= (X+H)^{-1}\Big(\mathsf{1} - XX^{-1} - HX^{-1} + XX^{-1}HX^{-1} + HX^{-1}HX^{-1}\Big) \\ &= (X+H)^{-1} H X^{-1} H X^{-1}, \end{aligned}$$

wobei wir bei den Umformungen nur die Rechenregeln der Matrizenmultiplikation sowie $(X+H)^{-1}(X+H) = \mathsf{1} = XX^{-1}$ benutzt haben. Mit dem Kompatibilitätssatz des euklidischen Betrags erhalten wir

$$\frac{\big|R(H)\big|}{|H|} \leq \frac{\big|(X+H)^{-1}\big|\big|H\big|\big|X^{-1}\big|\big|H\big|\big|X^{-1}\big|}{|H|} = \big|(X+H)^{-1}\big|\big|X^{-1}\big|^2 |H| \quad \xrightarrow{H\to 0} \quad 0$$

und damit die Formel für die totale Differenzierbarkeit. □

Bemerkungen

1. Die angegebenen Ableitungsformeln implizieren in einem zweiten Schritt, dass die Funktionen det und inv sogar unendlich oft stetig differenzierbar sind.

2. Für matrizenwertige Funktionen kann eine Produktregel für totale Ableitungen formuliert und bewiesen werden (Übungsaufgabe).

[51] Für $n = 2$ und $n = 3$ können wir die Behauptung natürlich auch direkt nachrechnen.

[52] Die Cramersche Regel wird in der *Linearen Algebra* bewiesen und liefert für jede invertierbare Matrix $X \in \mathbb{M}^{n\times n}$ eine komplizierte, aber explizite Formel für $\operatorname{inv}(X) = X^{-1}$, in der nur Determinanten auftauchen. Sie lautet

$$\big(\operatorname{inv}(X)\big)_{ij} = \frac{(-1)^{i+j}\det\big(X_{[ji]}\big)}{\det(X)},$$

wobei die Zahl links gerade in der i-ten Zeile und j-ten Spalte von $\operatorname{inv}(X)$ steht und $X_{[ji]}$ die $(n-1)\times(n-1)$-Matrix ist, die aus X durch Streichung der j-ten Zeile und i-ten Spalte entsteht.

2.7 lokaler Umkehrsatz

Vorbemerkung Eine wichtige und immer wiederkehrende Frage ist, ob sich ein gegebenes Gleichungssystem der Form

$$y_i = f_i(x_1, \ldots, x_n) \qquad \text{mit} \qquad i \in \{1, \ldots, n\}$$

nach den x_j auflösen lässt bzw. ob die entsprechende Funktion $f : \mathbb{R}^n \to \mathbb{R}^n$ invertiert werden kann. Für lineare oder affine Abbildungen liefert die *Lineare Algebra* erschöpfende Antworten, aber im nichtlinearen Fall ist dieses Problem deutlich komplizierter. In diesem Abschnitt formulieren und beweisen wir einen der zentralen Sätze der Differentialrechnung, der die Existenz einer *lokalen Umkehrfunktion* in der Nähe eines fixierten Punktes $x_* \in \mathbb{R}^n$ garantiert, sofern die Jacobi-Matrix von f in x_* regulär ist. Dieses Resultat besitzt sehr viele Anwendungen, denn die hinreichende Bedingung kann oftmals mit wenig Aufwand verifiziert werden. Die Frage nach der Existenz einer globalen Umkehrfunktion ist deutlich anspruchsvoller und kann *generell nicht* durch das Studium von partiellen Ableitungen in einzelnen Punkten entschieden werden.

Theorem (**lokaler Umkehrsatz**) Seien $U \subseteq \mathbb{R}^n$ offen, $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und x_* ein Punkt in U, in dem die Nicht-Entartungsbedingung

$$\det\big(\operatorname{Jac} f(x_*)\big) \neq 0$$

erfüllt ist. Dann existieren offene Mengen V, $W \subseteq \mathbb{R}^n$ mit $x_* \in V \subseteq U$ und $f(x_*) \in W$, sodass V von f bijektiv auf W abgebildet wird. Insbesondere ist die lokale Umkehrabbildung $f^{-1} : W \to V$ wohldefiniert und stetig differenzierbar, wobei

$$\operatorname{Jac} f^{-1}\big(f(x)\big) = \big(\operatorname{Jac} f(x)\big)^{-1}$$

für alle $x \in U$ gilt.

Beweis* *Vereinfachungen*: Der Einfachheit halber betrachten wir nur den Spezialfall $U = \mathbb{R}^2$, aber alle Argumente dieses Beweises können auf allgemeine offene Mengen U übertragen werden, sofern an einigen Stellen die technischen Details sorgfältiger formuliert werden. Ohne Beschränkung der Allgemeinheit können wir außerdem

$$x_* = 0\,, \qquad f(x_*) = 0\,, \qquad \operatorname{Jac} f(x) = 1$$

annehmen, denn andernfalls betrachten wir $\breve{f} : \mathbb{R}^n \to \mathbb{R}^n$ mit

$$\breve{f}(\breve{x}) := \big(\operatorname{Jac} f(x_*)\big)^{-1}\big(f(x_* + \breve{x}) - f(\breve{x})\big)$$

und wenden alle nachfolgenden Betrachtungen auf diese Funktion und $\breve{x}_* = 0$ an.

Vorbereitungen: Die Abbildungen

$$x \in \mathbb{R}^n \mapsto \det\big(\operatorname{Jac} f(x)\big)\,, \qquad x \in \mathbb{R}^n \mapsto \big|1 - \operatorname{Jac} f(x)\big|$$

sind beide stetig[53] und daher sind

$$\big\{x \in \mathbb{R}^n : \det\big(\operatorname{Jac} f(x)\big) \neq 0\big\}\,, \qquad \big\{x \in \mathbb{R}^n : |1 - \operatorname{Jac} f(x)| < \tfrac{1}{2}\big\}$$

[53] Wir benutzen hier den Kompositionssatz sowie die Tatsache, dass die Funktion $J \mapsto |J|$ und $J \mapsto \det(J)$ stetig auf $\mathbb{M}^{n\times n}$, dem Raum aller $n\times n$-Matrizen, sind.

zwei offene Mengen, die nach Voraussetzung den Punkt 0 und damit auch jeweils eine kleine Kugel um 0 enthalten. Wir können daher einen Radius $\varepsilon > 0$ wählen, sodass

$$\det\big(\operatorname{Jac} f(x)\big) \neq 0 \qquad \text{und} \qquad \big|1 - \operatorname{Jac} f(x)\big| < \tfrac{1}{2}$$

für alle $x \in B_\varepsilon(0)$ gilt und dies impliziert zum einen, dass $\operatorname{Jac} f(x)$ für jeden dieser Punkte eine invertierbare Matrix ist.[54] Die zweite wichtige Konsequenz ist, dass der Mittelwertsatz der Differentialrechnung angewendet auf die Funktion $q : \mathbb{R}^n \to \mathbb{R}^n$ mit

$$q(x) := x - f(x) \qquad \text{und} \qquad \operatorname{Jac} q(x) = 1 - \operatorname{Jac} f(x)$$

die lokale Lipschitz-Abschätzung

$$\big|q(x) - q(\tilde{x})\big| = \big|x - \tilde{x} - f(x) + f(\tilde{x})\big| \leq \tfrac{1}{2}\,|x - \tilde{x}|$$

für alle x, $\tilde{x} \in B_\varepsilon(0)$ garantiert.

Existenz von f^{-1} via Fixpunktargument: Wir setzen $A := \overline{B}_{\varepsilon/2}(0)$, fixieren $\tilde{y}$ mit $|\tilde{y}| \leq \varepsilon/4$ beliebig und betrachten die stetig differenzierbare Funktion $\tilde{q} : A \to \mathbb{R}^n$ mit

$$\tilde{q}(x) := \tilde{y} + q(x) = \tilde{y} + x - f(x)\,,$$

wobei $\tilde{y}$ die Rolle eines Parameters spielt und $\tilde{q}(0) = \tilde{y}$ gilt. Die Lipschitz-Abschätzung für q impliziert

$$\big|\tilde{q}(x) - \tilde{q}(\tilde{x})\big| \leq \tfrac{1}{2}\,|x - \tilde{x}|$$

für alle x, $\tilde{x} \in A$ und in Kombination mit der Dreiecksungleichung erhalten wir

$$\big|\tilde{q}(x)\big| \leq \big|\tilde{q}(x) - \tilde{q}(0)\big| + \big|\tilde{q}(0)\big| \leq \tfrac{1}{2}\,|x - 0| + |\tilde{y}| \leq \tfrac{1}{4}\,\varepsilon + \tfrac{1}{4}\,\varepsilon = \tfrac{1}{2}\,\varepsilon$$

für alle $x \in A$. Wir schließen, dass $\tilde{q}$ die abgeschlossene Menge A kontraktiv in sich abbildet und nach dem Banachschen Fixpunktsatz einen eindeutigen Fixpunkt $\tilde{x} \in A$ besitzt. Dieser erfüllt

$$\tilde{q}(\tilde{x}) = \tilde{x} \qquad \text{bzw.} \qquad f(\tilde{x}) = \tilde{y}$$

und hängt natürlich von der Wahl von $\tilde{y}$ ab. Insgesamt haben wir gezeigt, dass die lokale Umkehrfunktion $f^{-1} : \overline{B}_{\varepsilon/4}(0) \to \overline{B}_{\varepsilon/2}(0)$ wohldefiniert ist.

Stetigkeit von f^{-1}: Seien nun y, $\tilde{y}$ zwei beliebige, aber verschiedene Punkte in der offenen Menge $W := B_{\varepsilon/4}(0)$ und seien x, $\tilde{x}$ die entsprechenden Punkte in $\overline{B}_{\varepsilon/2}(0)$, d.h.

$$x = f^{-1}(y)\,, \qquad \tilde{x} = f^{-1}(\tilde{y})\,, \qquad y = f(x)\,, \qquad \tilde{y} = f(\tilde{x})\,.$$

Aus den Definitionen, der Dreiecksungleichung sowie der Lipschitz-Abschätzung für q ergibt sich

$$\big|x - \tilde{x}\big| \leq \big|q(x) - q(\tilde{x})\big| + \big|f(x) - f(\tilde{x})\big| \leq \tfrac{1}{2}\,|x - \tilde{x}| + \big|y - \tilde{y}\big|$$

und mit einfachen Termumstellungen folgt via

$$\big|f^{-1}(y) - f^{-1}(\tilde{y})\big| = \big|x - \tilde{x}\big| \leq 2\,|y - \tilde{y}|$$

[54] Aus der *Linearen Algebra* quadratischer Matrizen wissen wir, dass $\det(J) \neq 0$ genau dann gilt, wenn J^{-1} existiert.

eine Lipschitz-Abschätzung für f^{-1}. Mit dieser können wir zeigen (Übungsaufgabe), dass f^{-1} die Menge W stetig nach $V := f^{-1}(W) \subset \overline{B}_{\varepsilon/2}(0)$ abbildet und dass V eine offene Teilmenge des $\mathbb{R}^n$ ist.

Differenzierbarkeit von f^{-1}: Seien y, $\tilde{y}$ und x, $\tilde{x}$ wie im letzten Beweisschritt, wobei wir nun y als fixiert und $\tilde{y}$ als variabel betrachten. Nach Voraussetzung ist f auch total differenzierbar in x, d.h. es gilt

$$f(\tilde{x}) - f(x) = \operatorname{Jac} f(x)(\tilde{x} - x) + r(\tilde{x},\, x) \qquad \text{mit} \qquad \frac{\big|r(\tilde{x},\, x)\big|}{|\tilde{x} - x|} \quad \xrightarrow{\tilde{x}\to x} \quad 0\,.$$

Nach Anwendung der inversen Matrix zu $J := \operatorname{Jac} f(x)$ sowie einfacher Umformungen erhalten wir

$$f^{-1}(\tilde{y}) - f^{-1}(y) = J^{-1}\big(\tilde{y} - y\big) + s(\tilde{y},\, y)\,, \qquad s(\tilde{y},\, y) := -J^{-1}\, r\big(\tilde{x},\, x\big)$$

und das entsprechende Restglied kann durch

$$\frac{\big|s(\tilde{y},\, y)\big|}{|\tilde{y} - y|} \le \big|J^{-1}\big|\, \frac{\big|r\big(x,\, \tilde{x}\big)\big|}{|\tilde{x} - x|}\, \frac{|\tilde{x} - x|}{|\tilde{y} - y|} \le 2\,\big|J^{-1}\big|\, \frac{\big|r\big(\tilde{x},\, x\big)\big|}{|\tilde{x} - x|}$$

abgeschätzt werden, wobei wir die Lipschitz-Stetigkeit von f^{-1} ausgenutzt haben. Diese impliziert außerdem, dass $\tilde{x} \to x$ für $\tilde{y} \to y$ gilt und wir schließen, dass f^{-1} im Punkt y total differenzierbar ist, wobei J^{-1} gerade die entsprechende Jacobi-Matrix ist.

Stetigkeit der Ableitungen von f^{-1}: Auf der Menge W gilt

$$\operatorname{Jac} f^{-1} = \operatorname{inv} \circ \operatorname{Jac} f \circ f^{-1}$$

und die Stetigkeit von $\operatorname{Jac} f^{-1} : W \to \mathbb{M}^{n\times n}$ ergibt sich aus dem Kompositionssatz, wobei inv die Inversionsabbildung quadratischer Matrizen ist, deren Stetigkeit (und totale Differenzierbarkeit) wir oben gezeigt haben. □

Bemerkungen

1. Den Zusammenhang zwischen der Jacobi-Matrix von f in x und der von f^{-1} in $f(x)$ haben wir bereits weiter oben aus der Kettenregel abgeleitet, aber dort hatten wir die Existenz und die Differenzierbarkeit von f^{-1} vorausgesetzt. Der Umkehrsatz zeigt, dass beides schon aus der Invertierbarkeit von $\operatorname{Jac} f(x_*)$ folgt.

2. Die im Beweis angegebenen Mengen V und W sind nicht optimal, d.h. die Aussage des Theorems gilt meist auf viel größeren Mengen. Für die innermathematischen Anwendungen ist es aber oftmals sehr wichtig, dass V und W offen sind bzw. so gewählt werden können.

3. Im Entartungsfall $\det\big(\operatorname{Jac} f(x_*)\big) = 0$ können wir im Allgemeinen keine Aussage treffen, d.h. es kann sein, dass f trotzdem eine lokale Umkehrfunktion besitzt oder dass dem nicht so ist. Man kann allerdings weitere hinreichende Bedingungen formulieren, aber diese sind deutlich komplizierter und benötigen höhere partielle Ableitungen von f in x_*.

4. Wenn $f : \mathbb{R}^n \to \mathbb{R}^n$ eine affine Abbildung ist, so gilt

   $$\operatorname{Jac} f(x) = \operatorname{Jac} f(0)\,, \qquad f(x) = f(0) + \operatorname{Jac} f(0)\, x$$

   für alle $x \in \mathbb{R}^n$ und das Theorem folgt mit $x_* = 0$ und $V = W = \mathbb{R}^n$ aus einem bekannten Resultat der *Linearen Algebra*, wobei die Notwendigkeit von $\det\big(\operatorname{Jac} f(x_*)\big) \neq 0$ sich unmittelbar aus dem Dimensionssatz ergibt. Für eine nichtlineare Abbildung f kann der Umkehrsatz auch wie folgt verstanden werden:

Prinzip: Wenn die affine Approximation $T_{f,\,1,\,x_*} : \mathbb{R}^n \to \mathbb{R}^n$ mit

$$T_{f,\,1,\,x_*}(x) := f(x_*) + \operatorname{Jac} f(x_*)\,(x - x_*)$$

invertierbar ist, so ist auch f in der Nähe von x_* invertierbar.

Diese Aussage liefert den Schlüssel zu vielen anderen Resultaten der Mathematik. Beachte auch, dass die i-te Komponente von $T_{f,\,1,\,x_*}$ gerade das Taylor-Polynom erster Ordnung der skalaren Funktion f_i ist.

5. Das Theorem liefert die Existenz einer lokalen Umkehrfunktion, jedoch keine explizite Formel für deren Berechnung. Unser Beweis mit dem Banachschen Fixpunktsatz impliziert aber für jedes $y \in \mathbb{R}^n$, das hinreichend nah bei $y_* = f(x_*)$ liegt, dass die durch

$$x_0 := x_*\,, \qquad x_{k+1} := x_k + \big(\operatorname{Jac} f(x_*)\big)^{-1}\big(y - f(x_k)\big)$$

rekursiv definierte Folge $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ für $k \to \infty$ gegen einen Grenzvektor x_∞ mit $f(x_\infty) = y$ konvergiert, wobei dieser in der Nähe von x_* liegen wird. Damit können Näherungswerte für $f^{-1}(y)$ auf einem Computer berechnet werden.[55]

6. Mit etwas mehr Aufwand können wir das Theorem wie folgt verschärfen: Ist f sogar K-mal stetig differenzierbar, so besitzt auch f^{-1} diese Eigenschaft.

7. Eine bijektive Abbildung zwischen zwei offenen Mengen wird Diffeomorphismus genannt, wenn sie und ihre Umkehrabbildung beide stetig differenzierbar sind. Im Kontext des Umkehrsatzes trifft dies sowohl auf $f : V \to W$ als auch auf $f^{-1} : W \to V$ zu.

Beispiel Wir betrachten die Funktion $f : \mathbb{R}^2 \to \mathbb{R}^2$ mit

$$f_1(x_1,\,x_2) := x_1 + x_2\,, \qquad f_2(x_1,\,x_2) := x_1\,x_2\,, \qquad \operatorname{Jac} f(x_1,\,x_2) = \begin{pmatrix} 1 & 1 \\ x_2 & x_1 \end{pmatrix}.$$

Für jeden Punkt $(x_{*,1},\,x_{*,2})$ mit $x_{*,1} \neq x_{*,2}$ existiert damit eine lokale Umkehrfunktion, aber für $x_{*,1} = x_{*,2}$ können wir das Theorem nicht anwenden. Außerdem gilt

$$f(-\xi,\,+\xi) = \big(0,\,-\xi^2\big) = f(+\xi,\,-\xi)$$

für jedes $\xi \in \mathbb{R}$, d.h. f ist nicht injektiv auf $\mathbb{R}^2$ und damit nicht global invertierbar.

Bemerkung: Durch das Auflösen der quadratischen Gleichungen $y_j = f_j(x_1,\,x_2)$ nach x_1 und x_2 erhalten wir mit

$$g_{\pm,\,1}(y_1,\,y_2) = \tfrac{1}{2}\,y_1 \pm \tfrac{1}{2}\sqrt{y_1^2 - 4\,y_2}\,, \qquad g_{\pm,\,2}(y_1,\,y_2) = \tfrac{1}{2}\,y_1 \mp \tfrac{1}{2}\sqrt{y_1^2 - 4\,y_2}$$

[55] Die eng verwandte Rekursionsvorschrift

$$x_{k+1} = x_k + \big(\operatorname{Jac} f(x_k)\big)^{-1}\big(y - f(x_k)\big)$$

heißt *Newton-Verfahren* und spielt in der *Numerischen Mathematik* eine wichtige Rolle. Der Vorteil ist, dass x_* gar nicht mehr explizit auftaucht, aber der Nachteil besteht darin, dass die Jacobi-Matrix von f in jedem Schritt in einem anderen Punkt ausgewertet werden muss.

zwei lokale Umkehrfunktionen $g_\pm : W \to V_\pm$ von f, die beide auf der Menge

$$W := \big\{(y_1,\, y_2) \;:\; y_2 < \tfrac{1}{4}\, y_1^2\big\}$$

definiert sind, aber Werte in

$$V_- := \big\{(x_1,\, x_2) \;:\; x_1 > x_2\big\} \qquad \text{bzw.} \qquad V_+ := \big\{(x_1,\, x_2) \;:\; x_1 < x_2\big\}\,.$$

annehmen. Für jeden Randpunkt mit $y_2 = \frac{1}{4}\, y_1^2$ gilt $g_-(y_1,\, y_2) = g_+(y_1,\, y_2)$, aber die partiellen Ableitungen werden singulär. Diese Formeln liefern natürlich mehr Informationen als der lokale Umkehrsatz, aber die Existenz und Kenntnis expliziter Invertierungsformeln ist die Ausnahme.

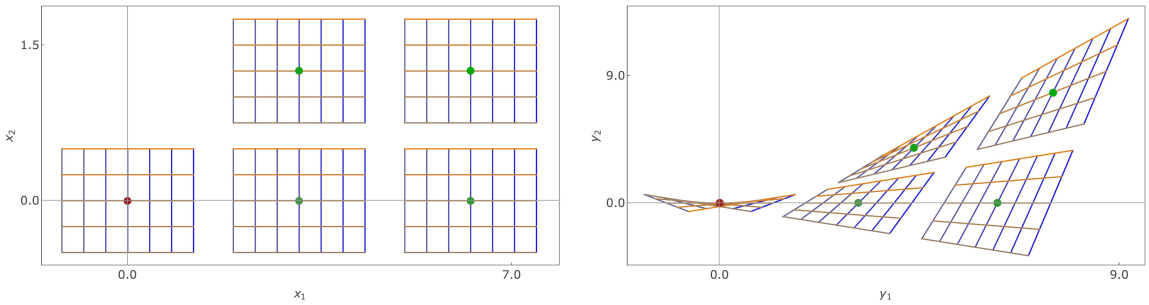

Abbildung Illustration des eben gerechneten Beispiels. In den vier grünen Punkten kann der lokale Umkehrsatz angewendet werden und das Bild zeigt jeweils eine konsistente, aber nicht maximale Wahl der offenen Mengen V (als $(x_1,\, x_2)$-Rechteck) und W (in diesem Fall ein $(y_1,\, y_2)$-Trapez). Im rot markierten Punkt $(0,\, 0)$ ist die Nicht-Entartungsbedingung jedoch verletzt und man kann zeigen, dass keine lokale Umkehrfunktion existiert.

Beispiel Wir betrachten die punktierte Ebene $U := \mathbb{R}^2 \setminus \{(0,\, 0)\}$ sowie die Funktion $f : U \to U$ mit

$$f_1(x_1,\, x_2) = x_1^2 - x_2^2\,, \qquad f_2(x_1,\, x_2) = 2\, x_1\, x_2\,.$$

Wegen

$$\operatorname{Jac} f(x_1,\, x_2) = \begin{pmatrix} +2\, x_1 & -2\, x_2 \\ +2\, x_2 & +2\, x_1 \end{pmatrix}\,, \quad \det\big(\operatorname{Jac} f(x_1,\, x_2)\big) = 4\,\big(x_1^2 + x_2^2\big) > 0$$

können wir f in jedem Punkt aus U lokal invertieren, aber wegen $f(0,\, +1) = f(0,\, -1)$ kann es keine globale Umkehrfunktion geben.

Achtung: Dieses Beispiel illustriert, dass aus der lokalen Invertierbarkeit in jedem Punkt im Allgemeinen **nicht** die Existenz einer globalen Umkehrfunktion folgt.

Bemerkung: Die Funktion f entspricht der komplexen Quadrierung, denn es gilt

$$\big(x_1 + \mathtt{i}\, x_2\big)^2 = f_1(x_1,\, x_2) + \mathtt{i}\, f_2(x_1,\, x_2)\,.$$

Insbesondere ist f surjektiv, aber nicht injektiv, und jede lokale Umkehrfunktion liefert einen *Ast der komplexen Wurzel*. Wir werden dies in der Vorlesung *Funktionentheorie* genauer untersuchen und verstehen.

Beispiel Durch die Formeln

$$y_1 = \frac{-2\,x_2}{(1-x_1)^2 + x_2^2}\,, \qquad y_2 = \frac{1 - x_1^2 - x_2^2}{(1-x_1)^2 + x_2^2}$$

wird in Physikernotation — also via $y_i = f_i(x_1,\, x_2)$ — eine stetig differenzierbare Funktion $f : U \to H$ definiert, wobei

$$U := \{(x_1,\, x_2)\, : x_1^2 + x_2^2 < 1\} \qquad \text{bzw.} \qquad H := \{(y_1,\, y_2)\, :\, y_2 > 0\}$$

die Einheitskreisscheibe bzw. die obere Halbebene ist. Die Formel

$$\frac{\partial y}{\partial x} = \frac{1}{\big((1-x_1)^2 + x_2^2\big)^2} \begin{pmatrix} -4\,(1-x_1)\,x_2 & -2\,(1-x_1^2) + 2\,x_2^2 \\ +2\,(1-x_1^2) - 2\,x_2^2 & -4\,(1-x_1)\,x_2 \end{pmatrix}$$

impliziert

$$\det\left(\frac{\partial y}{\partial x}\right) = \frac{\partial y_1}{\partial x_1}\,\frac{\partial y_2}{\partial x_2} - \frac{\partial y_1}{\partial x_2}\,\frac{\partial y_2}{\partial x_1} = \frac{4}{\big((1-x_1)^2 + x_2^2\big)^2} > 0$$

und der Umkehrsatz kann in jedem Punkt aus U angewendet werden. Auch in diesem Beispiel können wir viel bessere Aussagen treffen, denn direkte Rechnungen zeigen, dass durch die Formeln

$$x_1 = \frac{y_1^2 + y_2^2 - 1}{y_1^2 + (1+y_2)^2}\,, \qquad x_2 = \frac{-2\,y_1}{y_1^2 + (1+y_2)^2}$$

die globale Umkehrfunktion $g = f^{-1} : H \to U$ gegeben ist.

Bemerkung: In komplexer Schreibweise gilt

$$y_1 + \mathtt{i}\,y_2 = \mathtt{i}\,\frac{1 + (x_1 + \mathtt{i}\,x_2)}{1 - (x_1 + \mathtt{i}\,x_2)}\,, \qquad x_1 + \mathtt{i}\,x_2 = \frac{(y_1 + \mathtt{i}\,y_2) - \mathtt{i}}{(y_1 + \mathtt{i}\,y_2) + \mathtt{i}}\,,$$

d.h. f und f^{-1} sind die reellen Gegenstücke zu speziellen *Möbius-Transformationen*, wobei f^{-1} die *Cayley-Transformation* ist.

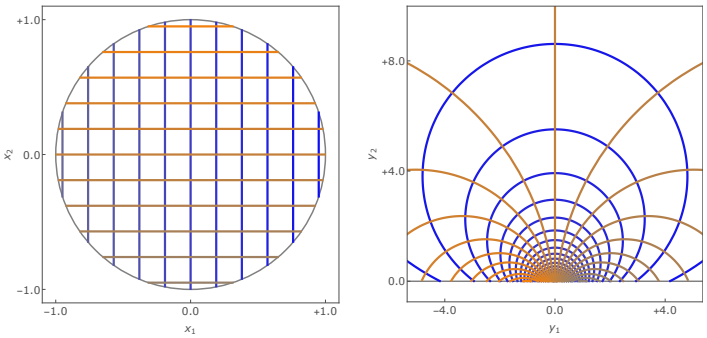

Abbildung Die Transformationen aus dem letzten Beispiel, wobei links die offene Kreisscheibe U und rechts ein Ausschnitt der Halbebene V gezeigt wird. Blau und Orange entsprechen den Niveaulinien von x_1 bzw. x_2.

2.8 Satz über implizite Funktionen

Vorbemerkung Im letzten Abschnitt hatten wir Systeme aus n Gleichungen mit n Unbekannten sowie die Invertierbarkeit der entsprechenden Funktionen untersucht. In diesem Abschnitt studieren wir unterbestimmte Systeme mit m Gleichungen für $k > m$ Variablen. Auf einer heuristischen Ebene erwarten wir, dass es bei gegebener rechter Seite mehrere Lösungen gibt und dass wir mithilfe der m Gleichungen genau m Variablen eliminieren und durch die anderen $n = k - m$ Variablen ausdrücken können. Wir beginnen mit dem zentralen, aber relativ abstrakten Resultat und diskutieren anschließend, wie dieses in der mathematischen Praxis verwendet werden kann.

Bezeichnungen Im Folgenden setzen wir $k = n + m$ mit $n \in \mathbb{N}$, bezeichnen Punkte aus dem $\mathbb{R}^k \cong \mathbb{R}^n \times \mathbb{R}^m$ als

$$(x, y) = (x_1, \ldots, x_n, y_1, \ldots, y_m)$$

und betrachten vektorwertige Funktionen f mit $f(x, y) \in \mathbb{R}^m$ sowie die quadratische $m \times m$-Matrix

$$\operatorname{Jac}_y f(x, y) = \begin{pmatrix} \partial_{y_1} f_1(x, y) & \ldots & \partial_{y_m} f_1(x, y) \\ \vdots & & \vdots \\ \partial_{y_1} f_m(x, y) & \ldots & \partial_{y_m} f_m(x, y) \end{pmatrix},$$

die nur die partiellen Ableitungen von f nach den y_i enthält und aus den letzten m Spalten der Jacobi-Matrix von f besteht. Die zugrunde liegende Motivation ist ein Gleichungssystem der Form

$$f_i(x_1, \ldots, x_n, y_1, \ldots, y_m) = c_i\,, \qquad i \in \{1, \ldots, m\}\,,$$

dass wir für fixiertes $c \in \mathbb{R}^m$ nach den y_i auflösen möchten. Wir wollen also jede Variable y_i eliminieren bzw. durch eine entsprechende Funktion g_i in den Variablen $x_1, \ldots, x_n$ ersetzen.

Theorem (**Satz über implizite Funktionen**) Sei $f : M \to \mathbb{R}^m$ eine stetig differenzierbare Funktion auf der offenen Menge $M \subseteq \mathbb{R}^k$ und sei (x_*, y_*) ein Punkt in M, sodass die Nicht-Entartungsbedingung

$$\det\big(\operatorname{Jac}_y f(x_*, y_*)\big) \neq 0$$

erfüllt ist. Dann existieren offene Mengen $U \in \mathbb{R}^n$, $V \in \mathbb{R}^m$ mit $x_* \in U$, $y_* \in V$ sowie eine stetig differenzierbare Funktion $g : U \to V$ mit $y_* = g(x_*)$, sodass die logische Äquivalenz

$$f(x, y) = f(x_*, y_*) \quad \Longleftrightarrow \quad y = g(x)$$

für alle $x \in U$ und alle $y \in V$ gilt.

Beweis Wir betrachten die Funktion $F : M \to \mathbb{R}^{n+m}$

$$F(x, y) := \big(x, f(x, y)\big)\,,$$

die nach Konstruktion und aufgrund der Voraussetzungen an f stetig differenzierbar ist. Die entsprechende Jacobi-Matrix kann als Blockmatrix

$$\operatorname{Jac} F(x,\,y) = \left(\begin{array}{c|c} 1 & 0 \\ \hline \operatorname{Jac}_x f(x,\,y) & \operatorname{Jac}_y f(x,\,y) \end{array}\right)$$

geschrieben werden, wobei $\operatorname{Jac}_x f(x,\,y)$ bzw. $\operatorname{Jac}_y f(x,\,y)$ eine $m\times n$- bzw. eine $m\times m$-Matrix ist, 1 für die n-dimensionale Einheitsmatrix steht und 0 die Nullmatrix mit n Zeilen und m Spalten meint. Die Rechenregeln für Determinanten[56] liefern

$$\det\big(\operatorname{Jac} F(x,\,y)\big) = \det\big(\operatorname{Jac}_y f(x,\,y)\big)$$

und wir können daher den lokalen Umkehrsatz für die Funktion F im Punkt $(x_*,\,y_*)$ auswerten, der unter F auf $(x_*,\,z_*)$ mit $z_* := f(x_*,\,y_*)$ abgebildet wird. Dadurch erhalten wir zwei offene Mengen $P,\,Q \subseteq \mathbb{R}^{n+m}$ mit $(x_*,\,y_*) \in P$ und $(x_*,\,z_*) \in Q$ sowie eine lokale Umkehrfunktion $H : Q \to P$ von F, die stetig differenzierbar ist. Die spezielle Struktur von F — für die ersten n Komponenten gilt $F_j(x,\,y) = x$ — impliziert die Darstellungsformel

$$H(x,\,z) = \big(x,\,h(x,\,z)\big) \qquad \text{und damit} \qquad f\big(x,\,h(x,\,z)\big) = z$$

für alle $(x,\,z) \in Q$, wobei die Funktion $h : Q \to \mathbb{R}^m$ aus den letzten m-Komponenten von H gebildet wird und damit auch stetig differenzierbar ist. Wir setzen

$$U := B_\varrho(x_*)\,, \qquad V := B_\sigma(y_*)\,, \qquad W := B_\tau(z_*)\,,$$

wobei wir die positiven Radien $\varrho,\,\sigma,\,\tau$ so klein wählen, dass die $n{+}m$-dimensionale Kugel von Radius $\sqrt{\varrho^2+\sigma^2}$ bzw. $\sqrt{\varrho^2+\tau^2}$ um $(x_*,\,y_*)$ bzw. $(x_*,\,z_*)$ ganz in P bzw. Q liegt, und dies impliziert (Nachrechnen!), dass auch $U\times V$ bzw. $U\times W$ eine Teilmenge von P bzw. Q ist. Abschließend definieren wir

$$g(x) := h(x,\,z_*) \qquad \text{für} \quad x \in U\,,$$

wobei dies $f\big(x,\,g(x)\big) = f\big(x,\,h(x,\,z_*)\big) = z_*$ impliziert. Sei nun umgekehrt $(x,\,y)$ ein beliebiger Punkt in $U\times V$ mit $f(x,\,y) = z_*$. Wegen $F(x,\,y) = \big(x,\,z_*\big) \in Q$ ergibt sich

$$(x,\,y) = H(x,\,z_*) = \big(x,\,h(x,\,z_*)\big) = \big(x,\,g(x)\big)$$

und damit auch $y = g(x)$. Hieraus folgt auch $y_* = g(x_*)$. □

Bemerkungen

1. Die Aussage des Theorems kann auch so verstanden werden: In der Nähe des Punktes $(x_*,\,y_*)$ sieht die Niveaumenge

$$N_f(c) := \big\{(x,\,y) \in M \,:\, f(x,\,y) = c\big\}\,, \qquad c := f(x_*,\,y_*) \in \mathbb{R}^m$$

aus wie der Graph der Funktion g. Man sagt auch, g sei implizit durch die Gleichung $f(x,\,y) = c$ definiert.

[56] Zum Beispiel die sukzessive Anwendung des Laplaceschen Entwicklungssatzes auf die ersten n Spalten.

2. Wir haben den Satz über implizite Funktionen mithilfe des Umkehrsatzes bewiesen. Es ist auch möglich, den Umkehrsatz aus dem Satz über implizite Funktionen herzuleiten (Übungsaufgabe). Beide Theoreme sind also äquivalent.

3. Der verallgemeinerte Umkehrsatz impliziert die folgende Verschärfung: Ist f sogar K-mal stetig differenzierbar, so besitzt auch g diese Eigenschaft.

4. Die Mengen U und V sowie die Funktion g hängen natürlich von (x_*, y_*) ab. Die im Beweis angegebenen Mengen U und V sind nicht optimal, d.h. die Aussage des Theorems gilt meist auf größeren Mengen. Die Beispiele zeigen aber, dass die Funktion g im Allgemeinen nicht global, sondern nur lokal existiert.

5. Das Theorem liefert *hinreichende* Bedingungen für die Existenz einer impliziten Funktion g. Im Entartungsfall $\det\big(\operatorname{Jac}_y f(x_*, y_*)\big) = 0$ können wir das Theorem nicht anwenden und g kann, aber muss nicht existieren. Wir werden unten sehen, dass wir im Entartungsfall durch Auswertung der zweiten Ableitungen von f oftmals wertvolle Informationen erhalten.

6. Die Kettenregel impliziert

$$\operatorname{Jac}_x f\big(x, g(x)\big) + \operatorname{Jac}_y f\big(x, g(x)\big)\ \operatorname{Jac} g(x) = 0$$

und damit insbesondere

$$\operatorname{Jac} g(x_*) = -\big(\operatorname{Jac}_y f\big(x_*, y_*\big)\big)^{-1} \operatorname{Jac}_x f\big(x_*, y_*\big)\,.$$

Beide Gleichungen sind nützlich und werden weiter unten für $k = 2$ und $m = 1$ genauer untersucht.

Achtung Es gibt sowohl in der theoretischen als auch in der anwendungsorientierten Literatur viele unterschiedliche Varianten des Satzes über implizite Funktionen. Diese sind zwar alle mehr oder weniger äquivalent, unterscheiden sich aber zum Teil stark in den verwendeten Bezeichnungen, Symbolen und Formulierungen.

1 Gleichung für 2 Variablen bzw. Kurven in der Ebene

Überblick Als prototypische Anwendung des Theorems betrachten wir eine stetig differenzierbare Funktion $f : \mathbb{R}^2 \to \mathbb{R}$, wobei wir die beiden Variablen mit x und y bezeichnen, sowie die Lösungsmenge $N_f(c)$ der Gleichung

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

für einen gegebenen Wert $c \in \mathbb{R}$. Der Satz über implizite Funktionen kann nun wie folgt angewendet werden:

1. Unter den Voraussetzungen

$$f(x_*, y_*) = c\,, \qquad \partial_y f(x_*, y_*) \neq 0$$

existieren Intervalle I und J mit $x_* \in I$ und $y_* \in J$ sowie eine Funktion $g : I \to J$, sodass die Implikation

$$f(x, y) = c \qquad \Longleftrightarrow \qquad y = g(x)$$

für alle $(x, y) \in I \times J$ gilt. Insbesondere kann $N_f(c)$ lokal als Graph der Funktion g in der Variablen x betrachtet werden.

2. Unter den Voraussetzungen

$$f(x_*, y_*) = c\,, \qquad \partial_x f(x_*, y_*) \neq 0$$

existieren Intervalle K und L mit $x_* \in K$ und $y_* \in L$ sowie eine Funktion $h : L \to K$, sodass die Implikation

$$f(x, y) = c \qquad \Longleftrightarrow \qquad x = h(y)$$

für alle $(x, y) \in K \times L$ gilt. Insbesondere kann $N_f(c)$ lokal als Graph der Funktion h in der Variablen y betrachtet werden.

Beachte, dass in der zweiten Formulierung die Rollen von x und y gerade vertauscht sind und dass g, I, J bzw. h, K, L von x_* bzw. y_* abhängen.

Beispiel Für die Funktion

$$f(x, y) := x^2 + y^2\,, \qquad \partial_x f(x, y) = 2\,x\,, \qquad \partial_y f(x, y) = 2\,y$$

und jeden gegebenen Punkt (x_*, y_*) mit $f(x_*, y_*) = c > 0$ deckt das Theorem die folgenden vier Fälle ab, wobei in diesem einfachen Beispiel die impliziten Funktionen direkt angegeben werden können:

1. Für $y_* > 0$ gilt lokal $y = g(x) = +\sqrt{c - x^2}$.
2. Für $y_* < 0$ gilt lokal $y = g(x) = -\sqrt{c - x^2}$.
3. Für $x_* > 0$ gilt lokal $x = h(y) = +\sqrt{c - y^2}$.
4. Für $x_* < 0$ gilt lokal $x = h(y) = -\sqrt{c - y^2}$.

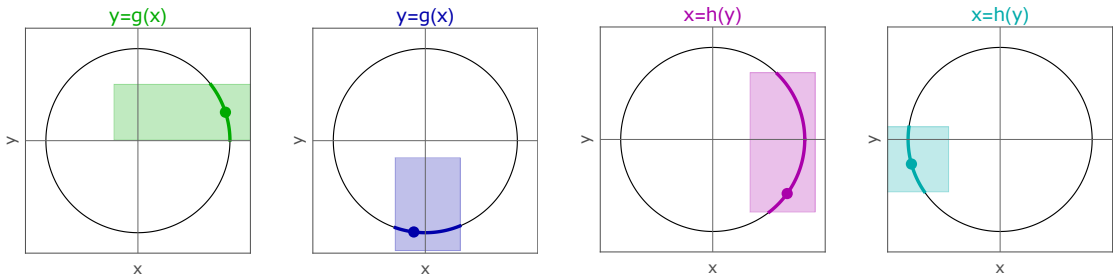

Abbildung Die vier Anwendungen des Satzes über implizite Funktionen aus dem letzten Beispiel, wobei der farbige Punkt die Wahl von (x_*, y_*) illustriert. Die farbige Box repräsentiert die Menge $I \times J$ bzw. $K \times L$, in der das jeweilige Kreissegment ein Graph ist, d.h. dort kann x als Funktion von y oder y als Funktion von x geschrieben werden. Beachte, dass der Satz über implizite Funktionen keine Aussage über die Größe der Boxen bzw. der Segmente macht, sondern nur deren Existenz garantiert. Im konkreten Fall ist klar, dass das jeweils maximale Segment die nördliche, südliche, östliche bzw. westliche Halbkreislinie ist. Die dargestellten Segmente sind jedoch nicht maximal.

Beispiel Die Niveaumenge $N_f(0)$ der Funktion

$$f(x, y) := \left(x^2 + y^2\right)^3 - \left(x^2 - y^2\right)^2$$

beschreibt ein vierblättriges Kleeblatt, wobei direkte Rechnungen

$$\partial_x f(x, y) = 6\,x\left(x^2 + y^2\right)^2 - 4\,x\left(x^2 - y^2\right), \quad \partial_y f(x, y) = 6\,y\left(x^2 + y^2\right)^2 + 4\,y\left(x^2 - y^2\right)$$

liefern. Wir wollen diesmal zunächst die Menge $N_f(0) \cap N_{\partial_y f}(0)$ berechnen, d.h. alle Punkte $(x_*,\, y_*) \in \mathbb{R}^2$ mit

$$f(x_*,\, y_*) = 0\,, \qquad \partial_y f(x_*,\, y_*) = 0\,,$$

in denen wir nach dem Satz über implizite Funktionen *nicht* nach y auflösen können, d.h. für die das Theorem keine entsprechende Funktion g liefert. Im Fall von $y_* = 0$ gilt dann

$$x_* = -1 \quad \text{oder} \quad x_* = 0 \quad \text{oder} \quad x_* = +1$$

und für $y_* \neq 0$ erhalten wir nach kleineren Rechnungen zunächst

$$x_*^2 + y_*^2 = +\frac{4}{9}\,, \qquad x_*^2 - y_*^2 = -\frac{8}{27}$$

und anschließend

$$x_*^2 = \frac{2}{27}\,, \qquad y_*^2 = \frac{10}{27}\,.$$

Wir erhalten damit insgesamt die 7 Punkte

$$N_f(0) \cap N_{\partial_y f}(0) = \Big\{ \big(0,\, 0\big)\,,\ \big(\pm 1,\, 0\big)\,,\ \big(\pm\tfrac{\sqrt{2}}{3\sqrt{3}},\, \pm\tfrac{\sqrt{2}\sqrt{5}}{3\sqrt{3}}\big) \Big\}$$

und analog (bzw. durch Vertauschung von x und y) zeigen wir

$$N_f(0) \cap N_{\partial_x f}(0) = \Big\{ \big(0,\, 0\big)\,,\ \big(0,\, \pm 1\big)\,,\ \big(\pm\tfrac{\sqrt{2}\sqrt{5}}{3\sqrt{3}},\, \pm\tfrac{\sqrt{2}}{3\sqrt{3}}\big) \Big\}$$

für die Menge der Punkte, in denen wir nach Satz über implizite Funktionen *nicht* nach x auflösen können bzw. für die es keine Funktion h gibt.

Zusammenfassung: Insgesamt erhalten wir die folgenden Aussagen:

1. In allen Punkten $(x_*,\, y_*) \in N_f(0)$, die weder zu $N_{\partial_y f}(0)$ noch zu $N_{\partial_y f}(0)$ gehören, können wir lokal sowohl nach x als auch nach y auflösen. Insbesondere kann die Niveaumenge $N_f(0)$ in der Nähe eines solchen Punktes sowohl in der Form $y = g(x)$ als auch in der Form $x = h(y)$ geschrieben werden.

2. Im Koordinatenursprung kann der Satz über implizite Funktionen überhaupt nicht angewendet werden, wobei wir im konkreten Fall einsehen können, dass dies kein Defizit des Theorems ist, sondern geometrische Eigenschaften von $N_f(0)$ widerspiegelt (siehe das Bild).

3. Es gibt weitere zwölf Punkte, in denen wir wenigstens nach x oder nach y auflösen können.

Bemerkung: Wir können die Kleeblatt-Menge $N_f(0)$ auch als Bild der parametrisierten Kurve

$$\gamma(t) = \cos(2\,t) \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}\,, \qquad t \in [0,\, 2\pi]$$

betrachten, wobei sich dann die *globalen* Formeln

$$x = \cos(2\,t)\,\cos(t)\,, \qquad y = \cos(2\,t)\,\sin(t)$$

ergeben, die in jedem Punkt aus $N_f(0)$ gelten. Allerdings ist die Kenntnis einer expliziten Parametrisierung der Niveaumenge $N_f(0)$ die Ausnahme.

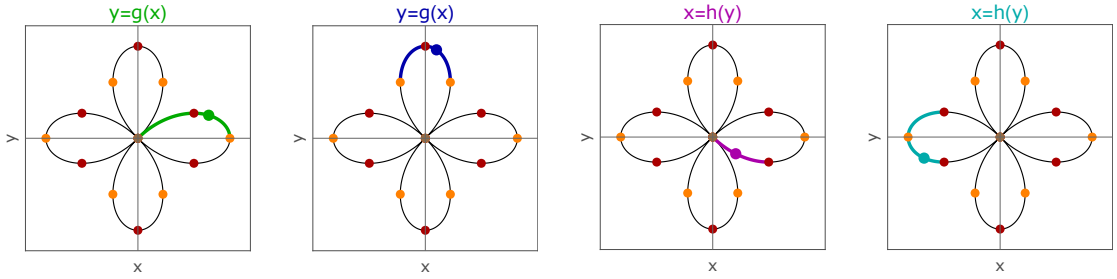

Abbildung Auch die Kleeblattkurve kann lokal als Graph betrachtet werden, wobei wir diesmal jeweils die maximalen Kurvensegmente gezeichnet haben. In den roten bzw. orangen Punkten können wir mit dem Satz über implizite Funktionen nach y bzw. x, aber nicht nach x bzw. y auflösen, denn jeder dieser Punkte entspricht einer lokalen Extremstelle von g bzw. f (siehe auch die Diskussion weiter unten). Im Koordinatenursprung (braun) kann weder nach x noch nach y aufgelöst werden, denn es handelt sich um einen Mehrfachpunkt der Kurve γ bzw. der Menge $\Gamma = N_f(0)$.

1 Gleichung für 3 Variablen bzw. Flächen im Raum

Überblick Wir betrachten skalare Funktionen f in drei Variablen (die wir x, y, z nennen) und studieren die Lösungen der Gleichung

$$f(x,\, y,\, z) = c$$

für gegebenes $c \in \mathbb{R}$. Die entsprechende Niveaumenge $N_f(c)$ ist diesmal in der Regel eine *zweidimensionale* Fläche und der Satz über implizite Funktionen kann wie folgt formuliert werden:

1. Unter den Voraussetzungen

$$f(x_*,\, y_*,\, z_*) = c\,, \qquad \partial_z f(x_*,\, y_*,\, z_*) \neq 0$$

existieren eine Menge $W \subset \mathbb{R}^2$ sowie ein Intervall J mit $(x_*,\, y_*) \in W$ und $z_* \in J$ sowie eine Funktion $g : W \to J$, sodass die Implikation

$$f(x,\, y,\, z) = c \qquad \Longleftrightarrow \qquad z = g(x,\, y)$$

für alle $(x,\, y,\, z)$ mit $(x,\, y) \in W$ und $z \in J$ gilt. Insbesondere sieht $N_f(c)$ lokal wie der Graph von g aus.

2. Analoge Resultate ergeben sich durch Vertauschung der Variablen.

Beispiel Die Gleichung

$$1 = f(x,\, y,\, z) = x^2 + y^2 + z^2$$

beschreibt die Einheitssphäre und damit eine der gekrümmten Standardflächen im $\mathbb{R}^3$. Nach Berechnung aller partiellen Ableitungen von f schließen wir aus dem Satz über implizite Funktionen, dass wir in der Nähe eines Punktes mit $f(x_*,\, y_*,\, z_*) = 1$ und

$$0 \neq x_* \qquad \text{bzw.} \qquad 0 \neq y_* \qquad \text{bzw.} \qquad 0 \neq z_*$$

die Sphäre lokal als Graph betrachten können, indem wir die Gleichung nach x bzw. y bzw. z auflösen. Natürlich können in diesem einfachen Beispiel die entsprechenden impliziten Funktionen direkt angegeben werden.

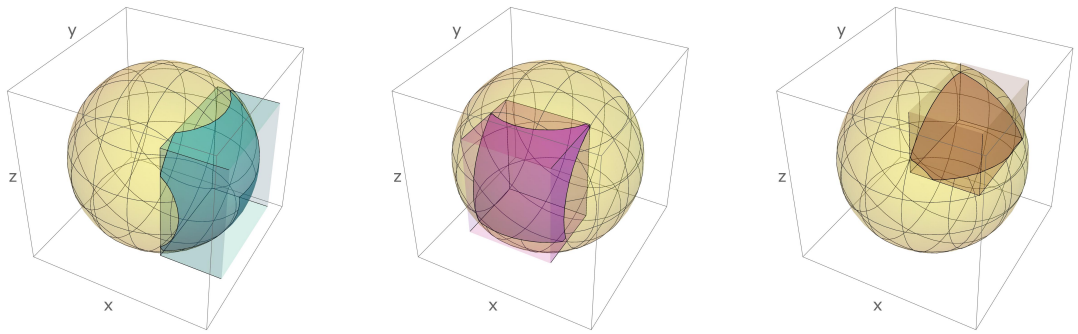

Abbildung Zur Anwendung des Satzes über implizite Funktionen auf der Sphäre, wobei das Segment der Farbe Türkis bzw. Lila bzw. Braun als Graph einer Funktion in den Variablen (y, z) bzw. (x, z) bzw. (x, y) betrachtet werden kann.

2 Gleichungen für 3 Variablen bzw. Kurven im Raum

Überblick Diesmal geht es um zwei skalare Funktionen f_1, f_2 in den drei Variablen x_1, x_2, x_3 sowie die Punktmenge

$$N_f(c) = N_{f_1}(c_1) \cap N_{f_2}(c_2)\,,$$

die als Schnittmenge zweier Flächen in der Regel aus (endlich vielen) Kurven besteht. Nach dem Satz über implizite Funktionen kann $N_f(c)$ lokal als Graph einer vektorwertigen Funktion mit einer Variablen geschrieben werden. Insbesondere können wir hier immer nach zwei Variablen auflösen, wobei es mehrere Möglichkeiten gibt, diese zu wählen, und wir auch immer die Details der Notation anpassen dürfen:

1. Unter den Voraussetzungen

   $$f_1(x_{*,1},\, x_{*,2},\, x_{*,3}) = c_1\,, \qquad f_2(x_{*,1},\, x_{*,2},\, x_{*,3}) = c_2$$

   und

   $$\det \begin{pmatrix} \partial_{x_2} f_1(x_{*,1},\, x_{*,2},\, x_{*,3}) & \partial_{x_3} f_1(x_{*,1},\, x_{*,2},\, x_{*,3}) \\ \partial_{x_2} f_2(x_{*,1},\, x_{*,2},\, x_{*,3}) & \partial_{x_3} f_2(x_{*,1},\, x_{*,2},\, x_{*,3}) \end{pmatrix} \neq 0$$

   existieren drei Intervalle I_j mit $x_{*,j} \in I_j$ sowie zwei Funktionen $g_2 : I_1 \to I_2$ und $g_3 : I_1 \to I_3$, sodass die Implikation

   $$f_1(x_1,\, x_2,\, x_3) = c_1\,,\;\; f_2(x_1,\, x_2,\, x_3) = c_2 \qquad \Longleftrightarrow \qquad x_2 = g_2(x_1)\,,\;\; x_3 = g_3(x_1)$$

   für alle $(x_1,\, x_2,\, x_3) \in I_1 \times I_2 \times I_3$ gilt.

2. Analoge Resultate ergeben sich durch Vertauschung der Variablen.

Beispiel Mit

$$f_1(x_1,\, x_2,\, x_3) = x_1^2 + x_2^2 + x_3^3\,, \quad f_2(x_1,\, x_2,\, x_3) = x_1^2 + (x_1 - x_2)^2\,, \quad c_1 = 1\,, \quad c_2 = \tfrac{1}{4}$$

ist die Niveaumenge $N_{f_1}(c_1)$ wieder die Einheitssphäre im $\mathbb{R}^3$, aber $N_{f_2}(c_2)$ beschreibt einen Zylinder mit elliptischem Querschnitt und Achse in x_3-Richtung. Die Schnittmenge dieser beiden Flächen besteht aus zwei geschlossenen Kurven. Der Satz über implizite Funktionen liefert nun Bedingungen, unter denen man zwei der drei Variablen

lokal eliminieren kann. Im konkreten Fall kann man wieder explizite Formeln durch das Lösen quadratischer Gleichungen ableiten. Zum Beispiel

$$g_2(x_1) = x_1 \pm \sqrt{\tfrac{1}{4} - x_1^2}\,, \qquad g_3(x_1) = \pm\sqrt{1 - x_1^2 - \left(x_1 \pm \sqrt{\tfrac{1}{4} - x_1^2}\right)^2}\,,$$

wobei der Punkt $(x_{*,1},\, x_{*,2},\, x_{*,3})$ jeweils festlegt, welches Vorzeichen in jeder der $\pm$-Alternativen zu wählen ist.

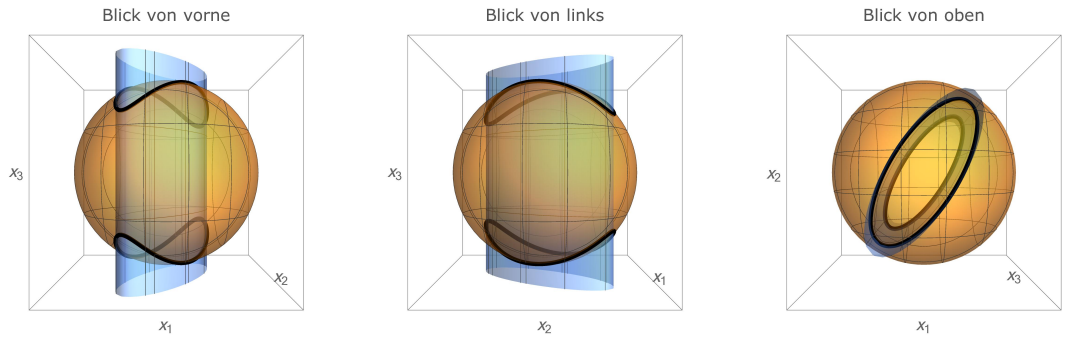

Abbildung Ein Zylinder (Blau) und eine Sphäre (Gelb) schneiden sich in zwei geschlossenen Kurven (Schwarz). Jede dieser Kurven kann lokal als Graph einer vektorwertigen Funktion mit einer Variablen beschrieben werden (hier nicht dargestellt).

Differentiation impliziter Funktionen

Vorbemerkung Man kann auch mit impliziten Funktionen rechnen. Wir beschränken uns der Einfachheit halber auf den Fall $k = 2$ und $m = 1$ und betrachten für eine skalare Funktion f in den Variablen x, y und eine fixierte Zahl c die entsprechende Niveaumenge $N_f(c)$.

Wir nennen einen Punkt $(x_*,\, y_*) \in N_f(c)$ singulär oder (wie schon oben eingeführt) kritisch, wenn $\operatorname{grad} f(x_*,\, y_*) = (0,\, 0)$ gilt, und andernfalls regulär. Insbesondere gilt in jedem regulären Punkt

$$\partial_y f(x_*,\, y_*) \neq 0 \qquad \text{und/oder} \qquad \partial_x f(x_*,\, y_*) \neq 0\,,$$

d.h. wir können nach dem Satz über implizite Funktionen nach y und/oder x auflösen und $N_f(c)$ lokal via $y = g(x)$ und/oder $x = h(y)$ als Graph einer Funktion ansehen.

Ableitungen von g und h in regulären Punkten Wenn wir in der impliziten Definition

$$c = f\big(x,\, g(x)\big) \qquad \text{bzw.} \qquad c = f\big(h(y),\, y\big)$$

auf beiden Seiten nach x bzw. y differenzieren, so erhalten wir nach Anwendung der Kettenregel

$$0 = \frac{\mathrm{d}}{\mathrm{d}x} f\big(x,\, g(x)\big) = \partial_x f\big(x,\, g(x)\big) + \partial_y f\big(x,\, g(x)\big)\, g'(x)$$

bzw.

$$0 = \frac{\mathrm{d}}{\mathrm{d}y} f\big(h(y),\, y\big) = \partial_x f\big(h(y),\, y\big)\, h'(y) + \partial_y f\big(h(y),\, y\big)\,.$$

Diese Gleichungen entsprechen der abstrakten Formel von oben und implizieren die *gewöhnlichen Differentialgleichungen*

$$g'(x) = -\frac{\partial_x f\big(x,\, g(x)\big)}{\partial_y f\big(x,\, g(x)\big)} \qquad \text{bzw.} \qquad h'(y) = -\frac{\partial_y f\big(h(y),\, y\big)}{\partial_x f\big(h(y),\, y\big)}\,.$$

Wir können diese zum Beispiel benutzen, um einen alternativen Existenzbeweis für g bzw. h zu führen oder um Approximationsformeln für g bzw. h zu berechnen. Wir werden dies später genauer diskutieren und verstehen. Im Punkt $(x_*,\, y_*)$ können wir die Ableitungen jedoch schon jetzt berechnen, denn es gilt

$$g'(x_*) = -\frac{\partial_x f(x_*,\, y_*)}{\partial_y f(x_*,\, y_*)} \qquad \text{bzw.} \qquad h'(y_*) = -\frac{\partial_y f(x_*,\, y_*)}{\partial_x f(x_*,\, y_*)}$$

wegen $g(x_*) = y_*$ bzw. $h(y_*) = x_*$.

Bemerkung: In Physikernotation werden die Differentialgleichungen oftmals als

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{\partial f}{\partial x} \Big/ \frac{\partial f}{\partial y} \qquad \text{bzw.} \qquad \frac{\mathrm{d}x}{\mathrm{d}y} = -\frac{\partial f}{\partial y} \Big/ \frac{\partial f}{\partial x}$$

angegeben. Diese Kurzschreibweise hilft bei direkten Rechnungen, aber fördert nicht unbedingt das Verständnis.

lokale Extrema von g und h Eine direkte Konsequenz der soeben abgeleiteten Formeln betrifft die Fälle

$$0 = \partial_x f(x_*,\, y_*) \neq \partial_y f(x_*,\, y_*) \qquad \text{bzw.} \qquad 0 = \partial_y f(x_*,\, y_*) \neq \partial_x f(x_*,\, y_*)\,,$$

also reguläre Punkte aus $N_f(c) \cap N_{\partial_x f}(0)$ bzw. $N_f(c) \cap N_{\partial_y f}(0)$, in denen wir nach dem Satz über implizite Funktionen nicht nach x bzw. y, sondern nur nach y bzw. x auflösen können. Wir hatten diese Punkte im Kleeblattbeispiel rot bzw. orange markiert und die nun folgenden Rechnungen können dort sehr einfach überprüft bzw. interpretiert werden. Für diese Punkte gilt

$$g'(x_*) = 0 \qquad \text{bzw.} \qquad h'(y_*) = 0$$

d.h. x_* bzw. y_* ist eine Nullstelle der Ableitung von g bzw. h. Durch zweifache Differentiation der impliziten Definition von g bzw. h nach x bzw. y und Auswertung in $x = x_*$ bzw. $y = y_*$ (Nachrechnen!) erhalten wir

$$g''(x_*) = -\frac{\partial_x^2 f(x_*,\, y_*)}{\partial_y f(x_*,\, y_*)} \qquad \text{bzw.} \qquad h''(y_*) = -\frac{\partial_y^2 f(x_*,\, y_*)}{\partial_x f(x_*,\, y_*)}\,,$$

d.h. wir können x_* bzw. y_* mittels der ersten und zweiten partiellen Ableitungen von f in $(x_*,\, y_*)$ klassifizieren. Im Standardfall $g''(x_*) \neq 0$ bzw. $h''(y_*) \neq 0$ ist x_* bzw. y_* eine strikte Extremstelle von g bzw. h und wir können definitiv nicht lokal nach x bzw. y auflösen. Insbesondere verstehen wir nun, warum der Satz über implizite Funktionen keine Aussage über die entsprechende Auflösbarkeit machen kann. Im Entartungsfall $g''(x_*) = 0$ bzw. $h''(x_*) = 0$ können wir jedoch nicht so einfach argumentieren, sondern müssen zusätzlich höhere Ableitungen zu Rate ziehen.

über singuläre Punkte* Im Fall von $\operatorname{grad} f(x_*, y_*) = (0, 0)$ können wir — zumindest mit dem Satz über implizite Funktionen — weder nach y noch nach x auflösen. Wir können aber — wie schon weiter oben bei der Klassifikation kritischer Punkte — die Funktion f durch ihr quadratisches Taylor-Polynom approximieren und dann durch explizite Rechnungen das qualitative lokale Verhalten in der Nähe von (x_*, y_*) beschreiben. Insbesondere ergeben sich dadurch die folgenden Aussagen:

1. Gilt $\det\big(\operatorname{Hess} f(x_*, y_*)\big) > 0$, so ist (x_*, y_*) eine strikte lokale Extremstelle von f und damit ein isolierter Punkt von $N_f(c)$.
2. Gilt $\det\big(\operatorname{Hess} f(x_*, y_*)\big) < 0$, so ist (x_*, y_*) ein Sattelpunkt von f sowie ein Doppelpunkt von $N_f(c)$.
3. Der Fall $\det\big(\operatorname{Hess} f(x_*, y_*)\big) = 0$ ist wieder entartet und eine Klassifikation braucht dritte oder noch höhere Ableitungen von f. Hier können *Umkehrpunkte*, *Mehrfachpunkte* oder andere Degenerierungen auftreten, siehe zum Beispiel den Koordinatenursprung im Kleeblattbeispiel.

Beispiel Für die Funktion

$$f(x, y) = (x-1)\,y^2 + x^2\,(x+p)$$

mit Parameter $p \in \mathbb{R}$ gilt

$$f(0, 0) = 0\,, \qquad \operatorname{grad} f(0, 0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}\,, \qquad \operatorname{Hess} f(0, 0) = \begin{pmatrix} 2\,p & 0 \\ 0 & -2 \end{pmatrix}\,,$$

d.h. der Koordinatenursprung $(x_*, y_*) = (0, 0)$ gehört für jeden Wert von p zu $N_f(0)$, ist aber immer ein singulärer Punkt. Für $p < 0$ handelt es sich um ein lokales Maximum von f und damit einen isolierten Punkt von $N_f(0)$. Für $p > 0$ ist der Koordinatenursprung jedoch ein Sattelpunkt von f sowie ein Doppelpunkt von $N_f(0)$. Der Wechsel findet gerade bei $p = 0$ (Entartungsfall) statt und entspricht hier einem Umkehrpunkt.

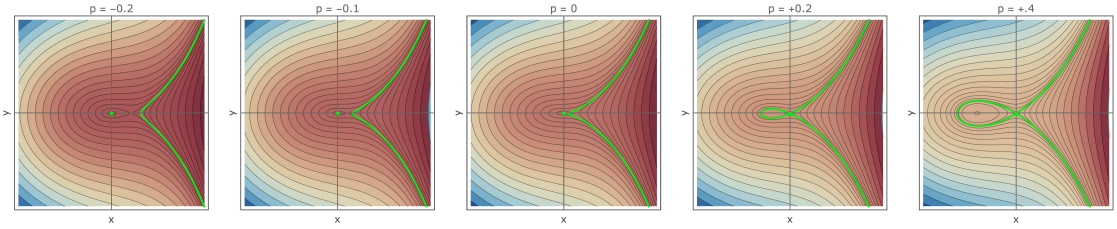

Abbildung Die Niveaumenge $N_f(0)$ (grün) aus dem gerade gerechneten Beispiel für verschiedene Werte des Parameters p. Den Wert $p = 0$ werden wir später *Bifurkationswert* nennen, da hier ein fundamentaler, d.h. qualitativer Wechsel in der Gestalt von $N_f(0)$ auftritt.

Kapitel 3

Differentialgleichungen

Vorlesungswoche 10

Motivation Viele Probleme in den Natur- und Anwendungswissenschaften werden in natürlicher Weise durch *Differentialgleichungen* beschrieben bzw. modelliert. In diesem Kapitel stellen wir die Grundlagen der entsprechenden mathematischen Theorie zusammen und werden insbesondere die zentralen Konzepte der *Existenz*, *Eindeutigkeit* und *Stabilität* von Lösungen einführen.

3.1 Grundbegriffe

Setting Im Folgenden seien U eine offene Menge in $\mathbb{R}\times\mathbb{R}^n$ und $f : U \to \mathbb{R}^n$ eine stetige Funktion. Dabei bezeichnen wir die Punkte in U sowie die Argumente von f mit $(t,\, x)$, wobei t eine reelle Zahl und $x = (x_1,\, \ldots,\, x_n)$ ein n-dimensionaler Vektor ist.

Definition Wir nennen

$$\dot{x}(t) = f\big(t,\, x(t)\big)$$

die Differentialgleichung erster Ordnung zu f. Eine Lösung dieser Differentialgleichung ist eine stetig differenzierbare Kurve $x : I \to \mathbb{R}^n$ auf einem Intervall I, sodass für jedes $t \in I$ der Punkt $\big(t,\, x(t)\big)$ in U liegt und außerdem die Gleichung erfüllt ist.

Beispiele

1. Im Fall von $n = 1$, $U = \mathbb{R}\times\mathbb{R}$ und $f(t,\, x) = \alpha\, x$ mit Konstante $\alpha \in \mathbb{R}$ lautet die Differentialgleichung $\dot{x}(t) = \alpha\, x(t)$. Diese besitzt (Nachrechnen!) die Lösungen

   $$x(t) = C\, \exp\left(\alpha\, t\right),$$

   wobei die Konstante C und das Definitionsintervall I beliebig gewählt werden können. Unten werden wir verstehen, dass es wirklich keine weiteren Lösungen gibt, aber im Moment ist das nicht klar.

2. Die Differentialgleichung für $n = 1$, $U = \mathbb{R}\times\mathbb{R}$ und $f(t,\, x) = x^2$ besitzt auch unendlich viele Lösungen, nämlich

   $$x(t) = \frac{1}{C - t} \qquad \text{mit} \qquad t \in I = (-\infty,\, C) \qquad \text{oder} \qquad t \in I = (C,\, +\infty)\,,$$

wobei C wieder eine beliebige Konstante ist und die Lösungseigenschaft einfach nachgerechnet werden kann. Beachte aber, dass wegen

$$\lim_{t \nearrow C} x(t) = +\infty \qquad \text{bzw.} \qquad \lim_{t \searrow C} x(t) = -\infty$$

keine dieser Lösungen auf ganz $\mathbb{R}$ definiert ist. Man spricht auch von *Blowup*-Lösungen (siehe Bild).

3. Durch Nachrechnen zeigen wir, dass die Formel

$$x(t) = \begin{pmatrix} +\cos(t) & +\sin(t) \\ -\sin(t) & +\cos(t) \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}, \qquad t \in I = \mathbb{R}$$

für jede Wahl der Konstanten C_1 und C_2 eine Lösung der zweidimensionalen Differentialgleichung

$$\dot{x}(t) = A\,x(t)\,, \qquad A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

liefert, wobei hier $n = 2$, $U = \mathbb{R}\times\mathbb{R}^2$ sowie $f(t,\, x) = A\,x$ gilt.

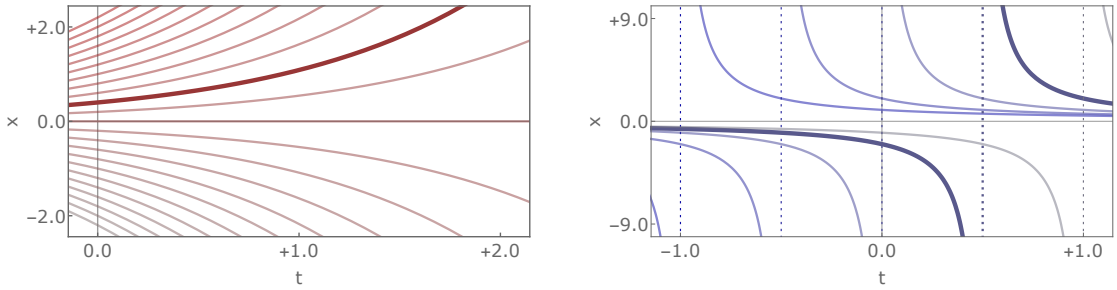

Abbildung *Links*: Einige Lösungen zum ersten Beispiel, die jeweils einer anderen Wahl von C entsprechen, wobei eine dieser Lösungen zur besseren Anschauung herausgestellt wurde. Beachte, dass für jedes feste t die Lösungen eindeutig durch ihre Werte $x(t)$ unterschieden werden können. Oder anders gesagt: Es gibt im Bild *keinen* Schnittpunkt zwischen zwei Lösungen. *Rechts*: Die analogen Bilder zum zweiten Beispiel. Für jeden Wert von C gibt es nun zwei Lösungen (auf jeweils disjunkten Intervallen), die links bzw. rechts der entsprechenden gestrichelten Hilfslinie einen *Blowup* aufweisen.

Bemerkungen

1. Für eine gegebene Differentialgleichung kann es sehr schwierig sein, Lösungen zu finden oder zu konstruieren. Es ist aber in der Regel sehr einfach zu entscheiden, ob eine geratene oder sonst wie erhaltene Funktion $x : I \to \mathbb{R}^n$ eine Lösung ist oder nicht, denn wir müssen nur nach t differenzieren und alle Terme in die Gleichung einsetzen.

2. Wir werden immer versuchen, den Definitionsbereich I einer Lösung so groß wie möglich zu wählen, d.h. wir suchen eigentlich sogenannte maximale Lösungen. Dabei kann es vorkommen, dass maximale Lösungen via $I = \mathbb{R}$ für alle t definiert sind oder dass das maximale Definitionsintervall nur halb-unendlich oder gar nur endlich ist (*Blowup*-Lösungen).

3. Sehr viele Gesetze der Naturwissenschaften können als Differentialgleichungen formuliert werden und die Aufgabe besteht oftmals darin, alle oder wenigstens spezielle Lösungen zu finden bzw. deren Eigenschaften zu verstehen.

4. Wir wollen in diesem Kapitel wieder t als *Zeit* interpretieren, da sehr viele Differentialgleichungen wirklich *dynamische Gesetze* bzw. die *Evolution* eines Systems beschreiben. Wir benutzen daher auch wieder *Punkt* (statt *Strich*) für die Ableitung nach t. Es gibt auch Differentialgleichungen, bei denen t keine Zeit, sondern vielleicht eine räumliche Variable oder ein sonstiger Parameter ist. Dann verwendet man gerne andere Buchstaben und schreibt zum Beispiel $z'(y) = -z(y) + \sin(y)$ statt $\dot{x}(t) = -x(t) + \sin(t)$.

5. Wir betrachten hier nur die *gewöhnlichen* (oder auch *ordinären*) Differentialgleichungen, bei denen es nur *eine skalare unabhängige Größe* (bei uns standardmäßig t genannt), aber im Allgemeinen mehrere *abhängige Größen* (die skalaren Komponenten x_j von x) gibt. Differentialgleichungen, in denen die unabhängige Größe selbst vektorwertig ist bzw. mehrere skalare Komponenten besitzt, nennt man *partiell*. Die entsprechende Theorie und Praxis ist *deutlich* anspruchsvoller und kann hier nicht diskutiert werden.

6. Bei theoretischen Betrachtungen werden wir die abhängigen Größen in der Regel zu einem Vektor $x \in \mathbb{R}^n$ zusammenfassen. Wir schreiben Differentialgleichungen auch oftmals verkürzt als

$$\dot{x} = f\big(t, x\big),$$

wobei dann immer mitgedacht werden muss, dass entlang einer Lösung x von t abhängen wird.

Achtung In vielen Lehrbüchern zur *Analysis* werden andere Notationen verwendet, wobei es neben der Wahl anderer Buchstaben auch konzeptionelle Unterschiede gibt.[1] Die Theorie ist am Ende natürlich immer dieselbe, aber gerade am Anfang sollten Sie Formeln aus verschiedenen Quellen nicht unbesehen miteinander vermischen.

Vorgabe von Anfangswerten Bei einer n-dimensionalen Differentialgleichung erster Ordnung gibt es in aller Regel insgesamt n unabhängige Freiheitsgrade. Insbesondere sind selbst maximale Lösungen nicht eindeutig, sondern werden n freie Konstanten enthalten, die wir standardmäßig mit $C_1, \ldots, C_n$ bezeichnen. Deswegen können wir neben der Differentialgleichung auch noch Anfangsbedingungen der Form

$$x(t_*) = x_*$$

stellen, wobei $t_* \in \mathbb{R}$ und $x_* \in \mathbb{R}^n$ mit $(t_*, x_*) \in U$ gegeben sind. Die Kombination aus Differentialgleichung und Anfangsbedingung wird Anfangswertproblem genannt. Der Satz von Picard-Lindelöf (siehe unten) garantiert für eine *stetig differenzierbare* Funktion f, dass es zu jedem (t_*, x_*) genau eine maximale Lösung gibt. Oder anders gesagt: Durch die Angabe von Anfangswerten werden die Freiheitsgrade der allgemeinen Lösung eliminiert.

[1] Ein alternatives — und vor allen bei Mathematikern beliebtes — Lösungskonzept kann mit unseren Schreibweisen wie folgt formuliert werden:

> Eine Lösung der Differentialgleichung $\dot{x} = f(t, x)$ ist eine Kurve $\xi : I \to \mathbb{R}^n$, sodass die Formeln $\big(t, \xi(t)\big) \in U$ und $\dot{\xi}(t) = f\big(t, \xi(t)\big)$ für alle $t \in I$ gelten.

Die *Unterscheidung* zwischen x und ξ hat zwar einige Vorteile, bringt aber auch viele Nachteile mit sich. In dieser Vorlesung benutzen wir den Buchstaben x sowohl für die nicht-zeitlichen Variablen von f als auch für die Komponenten der Lösungskurve, da wir damit viele Formeln und Ergebnisse einfacher verstehen und herleiten können.

Beispiele Für die bereits oben diskutierten drei Beispiele lautet die Lösungsformel für das entsprechende Anfangswertproblem wie folgt:

1. Es gilt
$$x(t) = x_* \exp(t - t_*),$$
denn die Anfangsbedingung legt via $x_* = C \exp(t_*)$ die Konstante C fest.

2. Durch einfache Rechnungen können wir wieder C eliminieren und erhalten
$$x(t) = \frac{x_*}{1 + x_*(t_* - t)}$$
mit
$$t \in I = (-\infty,\, x_*^{-1} + t_*) \qquad \text{oder} \qquad t \in I = (x_*^{-1} + t_*,\, +\infty).$$
Für $x_* = 0$ ergibt sich die stationäre Lösung $x(t) = 0$ für $t \in \mathbb{R}$.

3. Kombinieren wir die allgemeine Lösungsformel von oben mit der Anfangsbedingung, so ergibt sich
$$x_* = \begin{pmatrix} x_{*,1} \\ x_{*,2} \end{pmatrix} = \begin{pmatrix} +\cos(t_*) & +\sin(t_*) \\ -\sin(t_*) & +\cos(t_*) \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$$
und nach elementaren Rechenschritten (Invertierung einer zweidimensionalen Drehmatrix, Ausnutzung von Additionstheoremen) erhalten wir
$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} +\cos(t - t_*) & +\sin(t - t_*) \\ -\sin(t - t_*) & +\cos(t - t_*) \end{pmatrix} \begin{pmatrix} x_{*,1} \\ x_{*,2} \end{pmatrix}$$
als Lösungsformel für das Anfangswertproblem.

Sprechweisen

1. Gilt $f(t, x) = A(t)\,x + b(t)$ mit gegebener $n \times n$-Matrix $A(t)$ und gegebenem Vektor $b(t) \in \mathbb{R}^n$, so sprechen wir von einer linearen, andernfalls von einer nichtlinearen Differentialgleichung 1. Ordnung. Gilt bei einer linearen Gleichung $b(t) = 0$ für alle t, so wird die Gleichung auch homogen genannt, andernfalls inhomogen.

2. Hängt f gar nicht von t ab, so wird die Differentialgleichung autonom genannt und wir schreiben dann die Differentialgleichung verkürzt als $\dot{x} = f(x)$. Beachte aber, dass die Lösungen sehr wohl von t abhängen werden.

3. Bei autonomen Gleichungen kann es stationäre Lösungen geben, d.h. Lösungen der Bauart $x(t) = x_*$ für alle $t \in \mathbb{R}$. Diese sind besonders wichtig und durch die Bedingung $f(x_*) = 0$ charakterisiert.

4. Die Fälle $n = 1$ bzw. $n > 1$ nennt man wieder skalar bzw. vektorwertig (oder vektoriell), wobei eine autonome Gleichung in zwei Dimensionen ($n = 2$) auch planar genannt wird. Eine vektorwertige Gleichung kann immer *komponentenweise* als
$$\begin{pmatrix} \dot{x}_1(t) \\ \vdots \\ \dot{x}_n(t) \end{pmatrix} = \begin{pmatrix} f_1\big(t,\, x_1(t),\, \ldots,\, x_n(t)\big) \\ \vdots \\ f_n\big(t,\, x_1(t),\, \ldots,\, x_n(t)\big) \end{pmatrix}$$
geschrieben werden, wobei man diese Form auch als ein *System von (gekoppelten skalaren) Differentialgleichungen* bezeichnet.

Achtung Um Missverständnisse zu vermeiden, wollen wir jetzt schon festhalten:

1. Jede lineare und autonome Differentialgleichung kann — zumindest im Prinzip — exakt gelöst werden, obwohl die entsprechenden Formeln sehr unhandlich sein können. Lineare, aber nicht-autonome Gleichungen sind (zumindest für $n > 1$) deutlich schwieriger und in der Regel nicht mehr exakt lösbar.

2. Auch für nichtlineare Differentialgleichungen gibt es (selbst im autonomen Fall) nur in den seltensten Fällen geschlossene Lösungsformeln. Trotzdem kann die Mathematik sehr viele qualitative Aussagen über nichtlineare Gleichungen bzw. das Verhalten ihrer Lösung machen. Es wird sich insbesondere zeigen, dass wir für $n = 1$ und $n = 2$ die Eigenschaften nichtlinearer autonomer Gleichungen erster Ordnung auch ohne Lösungsformel sehr gut und fast vollständig charakterisieren können. Der Fall $n \geq 3$ ist aber wesentlich anspruchsvoller, insbesondere weil es dann *chaotische Effekte* geben kann.

3. Wir können im Prinzip jede Differentialgleichung (skalar oder vektorwertig, linear oder nichtlinear, autonom oder nicht) approximativ auf dem Computer lösen und es gibt viele entsprechende Softwarepakete. Eine Diskussion geeigneter numerischer Verfahren sowie ihrer Vor- und Nachteile erfordert aber eine eigene Vorlesung.

Gleichungen höherer Ordnung Es gibt natürlich auch Differentialgleichungen höherer Ordnung und vor allem die Gleichungen zweiter Ordnung spielen neben den Gleichungen erster Ordnung eine herausragende Rolle in der theoretischen Physik.

Merksatz Jede Differentialgleichung höherer Ordnung kann in eine äquivalente, aber höherdimensionale Gleichung erster Ordnung transformiert werden, indem durch einen *Universaltrick* neue Variablen eingeführt werden. Damit reicht es, die Theorie der Gleichungen erster Ordnung zu entwickeln bzw. zu studieren. Die Transformation offenbart außerdem, wie viele und welche Anfangsbedingungen bei einer Gleichung höherer Ordnung vorgeschrieben werden können.

Beispiele

1. Ein einfaches, aber prototypisches Beispiel zweiter Ordnung ist die Gleichung

$$\ddot{y}(t) + \beta\,\dot{y}(t) + \gamma\,y(t) = 0$$

mit den Parametern β und γ, wobei wir die abhängige Größe diesmal als $y(t)$ geschrieben haben. Wir können nun durch

$$x_1(t) = y(t)\,, \quad x_2(t) = \dot{y}(t)$$

zwei neue Variablen x_1 und x_2 einführen und die skalare Differentialgleichung zweiter Ordnung für $y(t)$ als

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} = \begin{pmatrix} x_2(t) \\ -\beta\,x_2(t) - \gamma\,x_1(t) \end{pmatrix}$$

schreiben und damit in ein *System von zwei gekoppelten skalaren* Gleichungen für die $x_j(t)$ überführen. In der Tat, die erste transformierte Gleichung spiegelt die

Definition der $x_j(t)$ wider, wohingegen die zweite die Ursprungsgleichung kodiert. Wir können durch direkte Rechnungen überprüfen, dass beide Formulierungen äquivalent sind (Übungsaufgabe). Insbesondere können für die Gleichung zweiter Ordnung und jedes feste $t_* \in \mathbb{R}$ zwei Anfangsbedingungen in der Form

$$x_{*,1} = x_1(t_*) = y(t_*)\,, \qquad x_{*,2} = x_2(t_*) = \dot{y}(t_*)$$

gestellt werden.

2. Wir betrachten

$$\dddot{y}(t) = \ddot{y}(t) + \phi\big(\dot{y}(t)\big)\, y(t) + \psi\big(y(t)\big)$$

als ein Beispiel für eine skalare Gleichung dritter Ordnung, wobei ψ und ϕ zwei gegebene skalare Funktionen auf $\mathbb{R}$ bezeichnen. Diesmal besteht der Trick darin, die drei Größen

$$x_1(t) = y(t)\,, \qquad x_2(t) = \dot{y}(t)\,, \qquad x_3(t) = \ddot{y}(t)\,,$$

einzuführen und die Ursprungsgleichung durch das äquivalente, dreidimensionale System

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{pmatrix} = \begin{pmatrix} x_2(t) \\ x_3(t) \\ x_3(t) + \phi\big(x_2(t)\big)\, x_1(t) + \psi\big(x_1(t)\big) \end{pmatrix}$$

zu ersetzen. Die ersten beiden Gleichungen ergeben sich wieder direkt aus der Definition der neuen Variablen $x_j(t)$ und insgesamt können wir drei Anfangswerte $x_* \in \mathbb{R}^3$ vorgeben.

3.2 Anwendungsbeispiele*

Newtonsche Abkühlung Die skalare Differentialgleichung

$$\dot{T}(t) = \lambda \left(T_{\text{umg}}(t) - T(t)\right)$$

beschreibt für einen homogenen Körper, wie sich seine räumlich gemittelte Temperatur $T(t)$ an die vorgegebene Außentemperatur $T_{\text{umg}}(t)$ anpasst, wobei λ eine materialabhängige Konstante ist. Wir werden unten verstehen, wie man die Lösungsformel

$$T(t) = T_* \exp\big(-\lambda\,(t - t_*)\big) + \lambda \int_{t_*}^{t} \exp\left(-\lambda\,(t-\sigma)\right) T_{\text{umg}}(\sigma)\,\mathrm{d}\sigma$$

ableitet, wobei $T_* = T(x_*)$ der Anfangswert ist. Ist die Umgebungstemperatur konstant, d.h. gilt $T_{\text{umg}}(t) = T_\#$, so vereinfacht sich die Formel zu

$$T(t) = T_* \exp\big(-\lambda\,(t - t_*)\big) + T_\#\big(1 - \exp\big(-\lambda\,(t - t_*)\big)\big)$$

und impliziert

$$T(t) - T_\# = \big(T_* - T_\#\big)\, \exp\big(-\lambda\,(t - t_*)\big)\,.$$

Insbesondere wird die Temperaturdifferenz nun exponentiell abklingen, wobei λ gerade die Stärke dieses Prozesses beschreibt und $(\ln 2)/\lambda$ die entsprechende *Halbwertszeit* ist (in dieser Zeitspanne halbiert sich immer die Temperaturdifferenz). Durch ähnliche Gleichungen können übrigens sehr viele Wachstums- oder Zerfallsprozesse beschrieben werden.

Ausbreitung einer Pandemie Ein sehr einfaches Modell ist das sogenannte *SIR-System*

$$\begin{aligned} \dot{S}(t) &= -\ \mu\, S(t)\, I(t)\,, \\ \dot{I}(t) &= +\ \mu\, S(t)\, I(t) \quad -\ \eta\, I(t)\,, \\ \dot{R}(t) &= \qquad\qquad\quad +\ \eta\, I(t)\,, \end{aligned}$$

wobei $S(t)$ bzw. $I(t)$ bzw. $R(t)$ die *Konzentration* der noch nicht infizierten (*susceptible*) bzw. der gerade infizierten (*infected*) bzw. der (durch Tod oder erlangte Immunität) entfernten (*removed*) Mitglieder einer Population beschreibt. Insbesondere sind $S(t)$, $I(t)$ und $R(t)$ hier immer reelle Zahlen aus dem Intervall $[0,\,1]$. Die Parameter μ bzw. η repräsentieren die Infektionsrate bzw. die Summe aus Sterbe- und Genesungsrate und werden im klassischen Modell als konstant angenommen. Insbesondere kann dieses Modell nicht die Wirkung von Quarantäne-Maßnahmen oder den unterschiedlichen Verlauf in Teilpopulationen beschreiben, denn dafür müsste man sehr viel mehr Größen und Parameter berücksichtigen.

Die Lösung dieses nichtlinearen, aber autonomen Differentialgleichungssystems hängt natürlich von den Werten der Parameter μ und η sowie von den Anfangsdaten

$$S(t_*)\,, \qquad I(t_*)\,, \qquad R(t_*)$$

ab. Eine wichtige und spezielle Eigenschaft ist der *Erhaltungssatz*

$$\dot{S}(t) + \dot{I}(t) + \dot{R}(t) = 0 \qquad \text{bzw.} \qquad S(t) + I(t) + R(t) = \text{const} \qquad \text{für alle} \quad t \geq t_*\,.$$

Dieser beschreibt, dass jedes Mitglied der Population zu genau einer der drei Gruppen gehört und erlaubt es, eine der drei abhängigen Größen auf einfache Weise zu eliminieren.

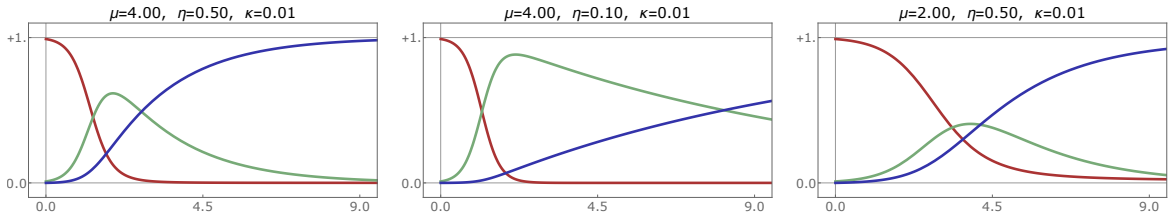

Abbildung Verschiedene Lösungen zum SIR-Modell, wobei die Anfangsdaten $S(0) = 1-\kappa$, $I(0) = \kappa$ und $R(0) = 0$ verwendet wurden und die rote bzw. grüne bzw. blaue Kurve den zeitlichen Verlauf von S bzw. I bzw. R darstellt. Alle numerischen Werte sind bzgl. gewisser Referenzwerte zu interpretieren.

Bemerkung Die Differentialgleichungssysteme in den Anwendungswissenschaften sind nur *Modelle der Wirklichkeit*, die unter gewissen (meist sehr vielen) vereinfachten Annahmen hergeleitet wurden und die reale Welt nicht exakt beschreiben. In der Regel gibt es ganze Modell-Hierarchien, wobei die zu Grunde liegenden Gleichungen immer komplexer werden und daher auch schwieriger zu lösen sind bzw. immer mehr Parameter enthalten, die man durch Messungen oder Analogieschlüsse bestimmen muss. In der Epidemiologie werden neben dem obigen SIR-Modell viele weitere Differentialgleichungen verwendet, zum Beispiel das SIS- oder das SEIS-Modell. Nur innerhalb der Mathematik gibt es universelle bzw. ewige Gleichungen, aber diese beziehen sich letztlich auch auf eine idealisierte Welt. Eine perfekte Sphäre existiert zum Beispiel nur in unserer Vorstellung, aber nicht in der Natur.

Elektrischer Schwingkreis Wir betrachten einen RLC-Schwingkreis (siehe Bild) mit gegebener Spuleninduktivität L, Ohmschem Widerstand R und Kondensatorkapazität C. Wird nun eine zeitabhängige Spannung $U(t)$ angelegt, zum Beispiel $U(t) = A\sin(\Omega t)$, so wird ein zeitabhängiger Strom $I(t)$ fließen, den man mithilfe einer Differentialgleichung charakterisieren und anschließend mit analytischen oder numerischen Methoden ausrechnen kann.

Wir wollen kurz skizzieren, wie man diese Gleichung ableiten kann. Die Kirchoffsche Maschenregel postuliert

$$U_R(t) + U_L(t) + U_C(t) = U(t)$$

für die Teilspannungen und die Gesetze der drei (idealisierten) Bauelemente können als

$$U_R(t) = R\,I(t)\,, \qquad U_L(t) = L\,\dot I(t)\,, \qquad I(t) = C\,\dot U_C(t)$$

geschrieben werden. Wir differenzieren nun die ersten drei Gleichungen (aber nicht die vierte) nach der Zeit t und erhalten nach Einsetzen und Umgruppieren der Terme

$$L\,\ddot I(t) + R\,\dot I(t) + C^{-1}\,I(t) = \dot U(t)$$

als Differentialgleichung für die zu bestimmende Stromstärke $I(t)$, wobei wir alle Terme mit bzw. ohne die Unbekannte links bzw. rechts versammelt haben. Es handelt sich um eine lineare, nicht-autonome Differentialgleichung zweiter Ordnung, die man zum Beispiel um die Anfangsbedingungen für $I(t_*)$ und $\dot I(t_*)$ ergänzen kann. Wir werden weiter unten sehen, wie man exakte Lösungsformeln ableiten kann und wie viele wesentliche Parameter es eigentlich gibt.

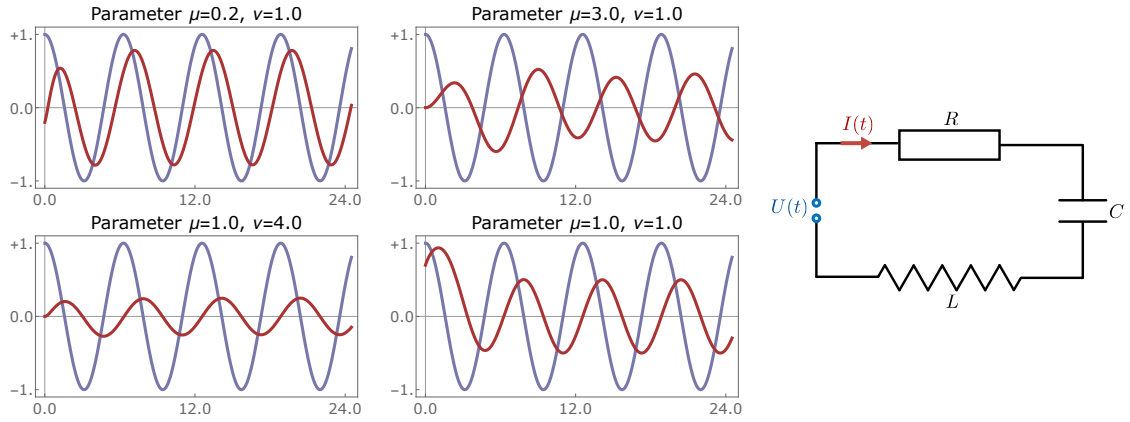

Abbildung Vier Lösungen (links) des elektrischen RLC-Schwingkreises (rechts) für verschiedene Parameterwerte und Anfangsdaten, wobei der zeitliche Verlauf der Spannung U bzw. der Stromstärke I in blau bzw. rot gezeichnet sind. Wir sehen, dass sich bei dieser Differentialgleichung und periodischem Input nach relativ kurzer Zeit eine (fast) periodische Lösung einstellt. Auch hier beziehen sich die numerischen Zahlenwerte an den Achsen auf geeignet gewählte Referenzgrößen und die Parameter sind ebenfalls in entdimensionalisierter Form gegeben (siehe unten).

über physikalische Dimensionen, Einheiten und Referenzgrößen** Die Differentialgleichungen der Natur- und Ingenieurwissenschaften enthalten meist sehr viele Parameter, wohingegen Mathematiker sehr sparsam mit Konstanten umgehen. Zum Beispiel enthält die Gleichung für den RLC-Schwingkreis mit anliegender Wechselspannung

$$U(t) = A\,\sin\big(\Omega\,t\big)\,, \qquad \dot U(t) = A\,\Omega\,\cos\big(\Omega\,t\big)$$

insgesamt fünf Parameter, nämlich R, L, C von den Bauelementen sowie A und Ω von der Anregung, wobei wir der Einfachheit halber mögliche Anfangsdaten in unseren Betrachtungen ignorieren wollen. Alle diese Parameter haben natürlich ihre physikalische Bedeutung, aber wir können uns fragen, ob sie denn wirklich alle wesentlich sind oder ob man nicht zumindest einige vernachlässigen darf. Dies ist nicht nur von theoretischem Interesse, sondern betrifft auch die Frage, wie viele numerische Simulationen man eigentlich durchführen muss, um die Lösungen einer gegebenen Differentialgleichung möglichst vollständig zu verstehen. Der Schlüssel für eine Antwort liegt in den physikalischen Dimensionen (bzw. den Einheiten) aller beteiligten Parameter sowie der abhängigen und unabhängigen Größen. Für den Schwingkreis ergeben sich die folgenden Dimensionsexponenten:

Größe	Zeit	Masse	Länge	Stromstärke	SI-Einheit
R	-3	$+1$	$+2$	-2	$\mathsf{s}^{-3}\,\mathsf{kg}\,\mathsf{m}^2\,\mathsf{A}^{-2}$
L	-2	$+1$	$+2$	-2	$\mathsf{s}^{-2}\,\mathsf{kg}\,\mathsf{m}^2\,\mathsf{A}^{-2}$
C	$+4$	-1	-2	$+2$	$\mathsf{s}^{4}\,\mathsf{kg}^{-1}\,\mathsf{m}^{-2}\,\mathsf{A}^{2}$
Ω	-1	0	0	0	s^{-1}
A	-3	$+1$	$+2$	-1	$\mathsf{s}^{-3}\,\mathsf{kg}\,\mathsf{m}^2\,\mathsf{A}^{-1}$
t	$+1$	0	0	0	s
I	0	0	0	$+1$	A
$\dot{I}$	-1	0	0	$+1$	$\mathsf{s}^{-1}\,\mathsf{A}$
$\ddot{I}$	-2	0	0	$+1$	$\mathsf{s}^{-2}\,\mathsf{A}$
$\dot{U}$	-4	$+1$	$+2$	-1	$\mathsf{s}^{-4}\,\mathsf{kg}\,\mathsf{m}^2\,\mathsf{A}^{-1}$

Der Punkt ist, dass es nur zwei unabhängige dimensionslose Produkte von Potenzen der fünf Parameter gibt und dass es daher nur zwei wesentliche Parameter im Schwingkreisproblem gibt. In der angewandten Mathematik wird eine solche *Dimensionsanalyse* auch *Buckinghamsches Prinzip* oder *Buckinghamsches* Π-*Theorem* genannt. Es gibt mehrere Möglichkeiten, die unabhängigen dimensionslosen Parameter zu wählen, die aber letztlich alle äquivalent sind. Im konkreten Fall können wir zum Beispiel

$$\mu = \Omega^2\,L\,C \qquad \text{und} \qquad \nu = \Omega\,R\,C$$

verwenden und aus mathematischer Sicht ist die Schwingkreis-Gleichung mit ihren fünf dimensionsbehafteten Parametern äquivalent zur *entdimensionalisierten* Differentialgleichung

$$\mu\,J''(s) + \nu\,J'(s) + J(s) = \cos(s)\,,$$

die nur die zwei wesentlichen Parameter μ und ν enthält und beschreibt, wie sich der entdimensionalisierte Strom

$$J(s) := A^{-1}\,C^{-1}\,\Omega^{-1}\,I\big(\Omega^{-1}\,s\big)$$

in Abhängigkeit von der entdimensionalisierten Zeit $s := \Omega\,t$ ändert. Hat man diese Gleichung für die konkreten Werte von μ und ν gelöst (analytisch oder numerisch), so ergibt sich der gesuchte Strom via $I(t) = A\,C\,\Omega\,J(\Omega\,t)$.

*Übungsaufgabe**: Rechnen Sie nach, dass jede Lösung $J(s)$ der entdimensionalisierten Gleichung eine Lösung $I(t)$ der dimensionsbehafteten Gleichung liefert und umgekehrt.

*Bemerkung**: Man kann die Entdimensionalisierung auch als Wahl geeigneter Referenzgrößen oder spezieller Einheiten interpretieren. Zum Beispiel sind Ω^{-1} bzw. $A\,C\,\Omega$ die

Referenzzeit bzw. die Referenzstromstärke und s bzw. J quantifizieren für t bzw. I die entsprechenden dimensionslosen Anteile. Die Referenzspannung ist gerade A.

*Merkregel**: Sowohl aus theoretisch-analytischer als auch aus praktisch-numerischer Sicht ist es immer sinnvoll, alle in einer Differentialgleichung auftauchenden Größen zu entdimensionalisieren.

Räuber-Beute-Modelle Die *Lotka-Volterra*-Gleichung

$$\dot{B}(t) = +\alpha\, B(t) - \beta\, B(t)\, R(t)\,, \qquad \dot{R}(t) = -\gamma\, R(t) + \delta\, B(t)\, R(t)\,,$$

beschreibt die Wechselwirkung zwischen einer Beute- und einer Räuberpopulation, wobei $B(t)$ bzw. $R(t)$ positive reelle Zahlen sind und als Anzahlen interpretiert werden können. Die zu Grunde liegenden Modelierungsannahmen sowie die Bedeutung der vier Parameter α, β, γ, δ werden auf WIKIPEDIA erklärt. In diesem planaren Differentialgleichungssystem gibt es genau eine stationäre Lösung (das *biologische Gleichgewicht*), aber wenn dieses — zum Beispiel durch äußere Einflüsse — gestört wird, so findet das System nicht zurück, sondern weist ein zeitperiodisches Verhalten auf (siehe das Bild). Ähnliche Gleichungen werden auch in den Wirtschaftswissenschaften verwendet (wobei $B(t)$ bzw. $R(t)$ dann das *Angebot* bzw. die *Nachfrage* ist) und erlauben es, die Existenz von Konjunkturzyklen zu verstehen.

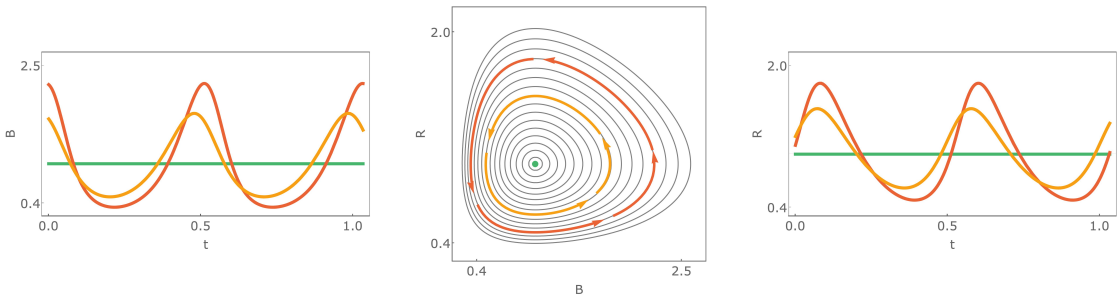

Abbildung Verschiedene Lösungen des Lotka-Volterra-Models für den Parametersatz $\alpha = \beta = 2$ und $\gamma = \delta = 1$. Das linke und das rechte Bild zeigen den zeitlichen Verlauf von $x_1(t)$ und $x_2(t)$ für drei verschiedene Lösungen, die jeweils einer anderen Farbe entsprechen. In der Mitte ist für mehrere Lösungen der entsprechende *Orbit*, d.h. das Bild der parametrisierten Kurve $t \mapsto x(t)$, dargestellt (*Phasenporträt*). Die grüne Lösung ist stationär und stellt das Gleichgewicht dar. Jede andere Lösung oszilliert um dieses Gleichgewicht.

zwei Modelle für das Fadenpendel Die skalaren (und entdimensionalisierten) Differentialgleichungen

$$\ddot{y}(t) = -y(t) \qquad \text{bzw.} \qquad \ddot{y}(t) = -\sin\big(y(t)\big)$$

sind zweiter Ordnung und werden *physikalisches* bzw. *mathematisches* Pendel genannt, wobei die Größe $y(t)$ den zeitabhängigen Auslenkungswinkel darstellt. Die erste Gleichung stellt eigentlich eine Vereinfachung der zweiten dar — denn nach dem Satz von Taylor gilt $\sin(y(t)) \approx y(t)$ für alle kleinen $y(t)$ — und beide können via

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} = \begin{pmatrix} +x_2(t) \\ -x_1(t) \end{pmatrix} \qquad \text{bzw.} \qquad \begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} = \begin{pmatrix} +x_2(t) \\ -\sin\big(x_1(t)\big) \end{pmatrix}$$

als planares System erster Ordnung geschrieben werden. In der Mechanik werden beide Pendelmodelle unter vereinfachenden Annahmen aus den Newtonschen Axiomen oder durch den Lagrangeschen Formalismus abgeleitet, wobei

$$E(x_1,\, x_2) = \tfrac{1}{2}\, x_1^2 + \tfrac{1}{2}\, x_2^2 \qquad \text{bzw.} \qquad E(x_1,\, x_2) = 1 - \cos\big(x_1\big) + \tfrac{1}{2}\, x_2^2$$

die entsprechende Energiefunktion $E : \mathbb{R}^2 \to \mathbb{R}$ ist. Mit der Kettenregel können wir nachrechnen, dass

$$\frac{\mathrm{d}}{\mathrm{d}t} E\big(x_1(t),\, x_2(t)\big) = 0$$

für jede Lösung der Differentialgleichung gilt, d.h die *Energie bleibt längs jeder Lösung erhalten*. Insbesondere ist der entsprechende *Orbit*, d.h. die Menge $\{x(t) : t \in \mathbb{R}\} \subset \mathbb{R}^2$ immer in einer Niveaumenge von E enthalten. Neben den vielen Gemeinsamkeiten gibt es aber auch wichtige Unterschiede. Das physikalische Pendel kann als lineare Gleichung explizit gelöst werden, wobei wir oben schon

$$x_1(t) = y(t) = +C_1 \cos(t) + C_2 \sin(t)\,, \qquad x_2(t) = \dot{y}(t) = -C_1 \sin(t) + C_2 \cos(t)$$

verifiziert hatten. Für das mathematische Pendel gibt es jedoch keine analogen Formeln. Außerdem besitzt diese Gleichung neben zeitperiodischen auch unbeschränkte Lösungen, wobei wir dies sehr gut im Konturplot der Energiefunktion erkennen können. Beachte auch, dass wir jede Lösung als *Parametrisierung* einer Niveaulinie $N_E(e)$ interpretieren können, wobei der Wert $e \in \mathbb{R}$ via $e = E(x(t_*))$ durch Anfangswerte festgelegt ist.

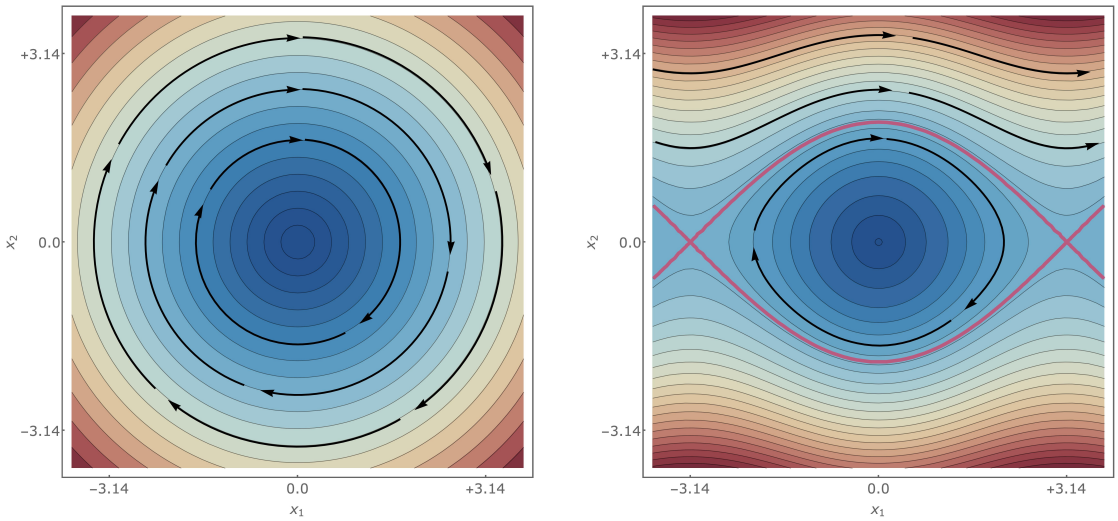

Abbildung Das Phasenportrait des physikalischen (links) und des mathematischen (rechts) Pendels, wobei jeweils die Niveaumengen der Energiefunktion dargestellt sind und drei ausgewählte Orbits hervorgehoben wurden. Rechts ist außerdem in lila die *Separatrix* gezeichnet, die die periodischen Orbits von den unbeschränkten Orbits trennt. Sie ist auch eine Niveaulinie der Energie E und verbindet die Sattelpunkte von E.

Geometrische Differentialgleichungen Auch innerhalb der Mathematik treten viele Differentialgleichungen auf und beschreiben zum Beispiel *Geodäten*, d.h. die kürzesten Verbindungen auf gekrümmten Flächen.

3.3 elementare Lösungsmethoden

Ziel Wir stellen in diesem Abschnitt zwei bekannte Methoden vor, mit denen gewisse Differentialgleichungen exakt gelöst werden können. Es sei aber noch einmal betont, dass es für viele Differentialgleichungen keine geschlossene Lösungsformel gibt.

Trennung der Veränderlichen

Vorbemerkung Wir betrachten eine *skalare* Differentialgleichung *erster* Ordnung der Bauart

$$\dot{x}(t) = g\big(x(t)\big)\, h(t)\,,$$

d.h. unter der Strukturannahme $f(x,\, t) = g(x)\, h(t)$. Eine solche Differentialgleichung nennt man auch *separierbar*.

Heuristische Herleitung der Lösungsformel In vereinfachter Notation gilt

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x,\, t) = g(x)\, h(t)$$

und formal können wir dies auch als

$$\frac{\mathrm{d}x}{g(x)} = h(t)\,\mathrm{d}t$$

schreiben, wobei dies an die infinitisimale Variante der Transformationsregel für eindimensionale Integrale erinnert. Wir ergänzen auf beiden Seiten ein unbestimmtes Integralzeichen und spendieren auch noch eine Integrationskonstante C. Wir erhalten dadurch

$$\int \frac{\mathrm{d}x}{g(x)} = \int h(t)\,\mathrm{d}t + C\,,$$

wobei die linke bzw. rechte Seite nach Definition der unbestimmten Integrale eine Funktion in x bzw. t ist. Bezeichnen wir nun mit G bzw. H eine entsprechende Stammfunktion, so erhalten wir

$$G(x) = H(t) + C$$

bzw.

$$x(t) = G^{-1}\big(H(t) + C\big)$$

als geschlossene Lösungsformel. Hierbei bezeichnet G^{-1} die *Umkehrfunktion* von G, der Stammfunktion von $1/g$. Der Wert der Integrationskonstanten C kann durch Anfangsdaten bestimmt werden. Mit $x_* = x(t_*)$ ergibt sich zum Beispiel $C = G(x_*) - H(t_*)$ durch konsistentes Einsetzen.

Bemerkung Uns braucht hier nicht zu interessieren, ob bzw. wie wir unsere formalen Rechenschritte mathematisch rigoros begründen können (das geht, braucht aber die Theorie von *Differentialformen*), sondern wir machen uns einfach nachträglich klar, dass die Methode bei richtiger Anwendung eine sinnvolle Lösung liefert. In der Tat, wenn wir

$$G(x(t)) = C + H(t)$$

nach t differenzieren, so liefert die Kettenregel

$$G'(x(t))\,\dot{x}(t) = \dot{H}(t) \qquad \text{bzw.} \quad \frac{\dot{x}(t)}{g\big(x(t)\big)} = h(t)\,,$$

wobei wir die Definition von G und H, d.h. die Formeln

$$G'(x) = \frac{1}{g(x)}\,, \qquad \dot{H}(t) = h(t)$$

benutzt haben. Insbesondere erhalten wir für jeden Wert der Konstante C eine Lösung der Differentialgleichung und das Theorem von Picard-Lindelöf (siehe unten) garantiert, dass es auch keine weiteren Lösungen gibt.

Beispiele

1. Um das Anfangswertproblem

$$\dot{x}(t) = \big(x(t)\big)^2 \sin(t)\,, \qquad x(t_*) = x_* > 0$$

zu lösen, schreiben wir die Differentialgleichung im Zwischenschritt als

$$\frac{\mathrm{d}x}{x^2} = \sin(t)\,\mathrm{d}t \qquad \text{bzw.} \qquad \int \frac{\mathrm{d}x}{x^2} = \int \sin(t)\,\mathrm{d}t + C\,.$$

In diesem Fall können wir nun sowohl links als auch rechts leicht eine Stammfunktion angeben und erhalten

$$-\frac{1}{x} = -\cos(t) + C \qquad \text{bzw.} \qquad x = \frac{1}{\cos(t) - C}\,,$$

wobei wir die Integrationskonstante C durch die Auswertung der Anfangsbedingung, d.h. via

$$C = \cos(t_*) - \frac{1}{x_*}$$

eliminieren können. Insgesamt haben wir damit die Lösungsformel

$$x(t) = \frac{1}{\cos(t) - \cos(t_*) + \dfrac{1}{x_*}} = \frac{x_*}{x_*\cos(t) - x_*\cos(t_*) + 1}$$

abgeleitet. Wir können auch wieder die Probe machen, d.h. durch Differentiation nach t leicht nachprüfen, dass diese Formel wirklich eine Lösung liefert.

Bemerkung: Unsere Rechnungen und die Endformel sind nur sinnvoll, solange $x(t) \neq 0$ gilt. Für $x_* \neq 0$ gilt dies zumindest für Zeiten $t \approx t_*$, aber es kann

zu endlichen Zeiten einen Blowup geben (nämlich dann, wenn der Nenner in der Lösungsformel den Wert 0 annimmt). Das passiert zum Beispiel bei den Anfangsdaten $t_* = 0$, $x_* = 2$ für $\cos(t) = \frac{1}{2}$, d.h. wir können die Lösungsformel in diesem Fall nur auf dem Zeitintervall $(-\pi/3, +\pi/3)$ verwenden. Für die Anfangsdaten $t_* = 0$, $x_* < 1/2$ existiert die Lösung aber global in der Zeit. Für $x_* = 0$ gibt es außerdem offensichtlich die triviale Lösung $x(t) = 0$ für alle $t \in \mathbb{R}$.

2. Für das Anfangswertproblem

$$\dot{x}(t) = \sqrt{x(t)}\, t\,, \qquad x(0) = x_* > 0$$

erhalten wir als Zwischenschritte zunächst

$$\frac{\mathrm{d}x}{\sqrt{x}} = t\,\mathrm{d}t \qquad \text{bzw.} \qquad \int \frac{\mathrm{d}x}{\sqrt{x}} = \int t\;\mathrm{d}t + C$$

sowie

$$2\sqrt{x} = \tfrac{1}{2}t^2 + C \qquad \text{bzw.} \qquad x = \left(\tfrac{1}{4}\,t^2 + \tfrac{1}{2}\,C\right)^2 .$$

Eine Auswertung der Anfangsbedingung zur Zeit $t_* = 0$ liefert $C = 2\sqrt{x_*}$ und damit die Lösungsformel

$$x(t) = \left(\sqrt{x}_* + \tfrac{1}{4}t^2\right)^2 .$$

Bemerkung: Im Fall von $x_* < 0$ gibt es diesmal keine sinnvolle reelle Lösung des Anfangswertproblems und unsere formalen Rechnungen machen letztlich keinen Sinn. Für Anfangswerte $x_* > 0$ existiert die Lösung jedoch global in der Zeit, wobei $x(t) > 0$ für $t \in \mathbb{R}$ gilt und sicherstellt, dass $\sqrt{x(t)}$ immer eine wohldefinierte reelle Zahl ist. Im Grenzfall $x_* = 0$ gibt es wieder die triviale Lösung $x(t) = 0$ für alle t aber auch die nicht-triviale Lösung $x(t) = \frac{1}{16}\,t^4$. Dieses seltsame Verhalten hat wieder damit zu tun, dass die Wurzelfunktion in 0 zwar definiert, aber nicht differenzierbar ist.

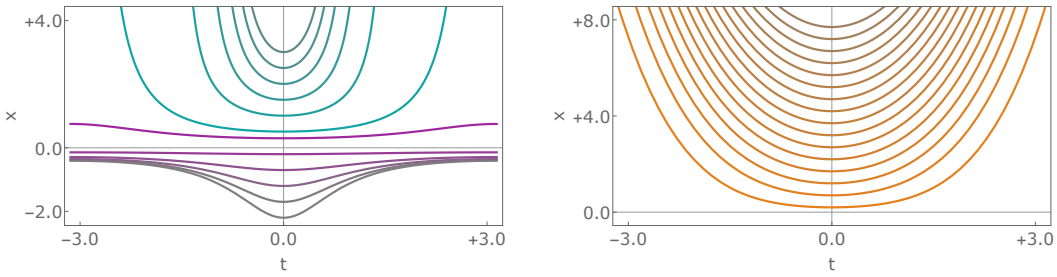

Abbildung Einige Lösungen der eben gerechneten Beispiele mit $t_* = 0$. *Links*: Das erste Beispiel besitzt Lösungen mit (türkis) als auch ohne (lila) Blowup. *Rechts*: Im zweiten Beispiel existieren alle Lösungen global in der Zeit, aber nur für nichtnegative Anfangsdaten.

Bemerkung

1. Die soeben beschriebene Lösungsmethode wird auch Separation der Variablen genannt. Auch hier gilt: Merken Sie sich nicht die Formeln, sondern das Prinzip ihrer Herleitung!

2. In der Praxis kann die Methode auch versagen, nämlich dann, wenn die Stammfunktionen G, H oder die Umkehrfunktion G^{-1} nicht berechnet werden können.

Nichtlineare autonome Gleichungen Ein wichtiger Spezialfall ist die autonome skalare Gleichung

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

mit gegebener nichtlinearer Funktion $f : \mathbb{R} \to \mathbb{R}$. In diesem Fall können wir $g(x) = f(x)$ und $h(x) = 1$ wählen und erhalten die Lösungsformel

$$x(t) = F^{-1}\big(t + C\big) \qquad \text{bzw.} \qquad x(t) = F^{-1}\big(t - t_* + F(x_*)\big)$$

für die Differentialgleichung bzw. das Anfangswertproblem, wobei F eine Stammfunktion zu $1/f$ ist, zum Beispiel

$$F(x) = \int_{x_*}^{x} \frac{\mathrm{d}\tilde{x}}{f(\tilde{x})} \qquad \text{oder} \qquad F(x) = \int_{0}^{x} \frac{\mathrm{d}\tilde{x}}{f(\tilde{x})} \,.$$

Wir hatten schon in *Analysis 1* gesehen, dass es viele Stammfunktionen gibt, die sich aber alle nur in additiven Konstanten unterscheiden. Die Lösungsformeln hängen aber nicht davon ab, welche Stammfunktion F gewählt wird.

Bemerkung Wir werden unten sehen, dass man für autonome skalare Gleichungen das qualitative Lösungsverhalten auch ohne explizite Formeln, sondern allein mit graphischen Mitteln vollständig charakterisieren kann.

Beispiele

1. Im Fall $f(x) = \exp(-p\,x)$ mit Parameter $p \neq 0$ ergeben sich via

   $$\exp(p\,x)\,\mathrm{d}x = \mathrm{d}t\,, \qquad p^{-1}\,\exp(p\,x) = t + C$$

   die Lösungsformeln

   $$x(t) = p^{-1}\,\ln(p\,t + p\,C)\,, \qquad x(t) = p^{-1}\ln\big(\exp(p\,x_*) - p\,t_* + p\,t\big)\,,$$

   die wir wieder mit einer einfachen Probe verifizieren können oder durch Einsetzen von

   $$F(x) = p^{-1}\,\exp(p\,x)\,, \qquad F^{-1}(z) = p^{-1}\,\ln(p\,z)$$

   aus der abstrakten Formel von oben ableiten können. Beachte, dass auch hier die Lösungen nicht für alle Zeiten $t \in \mathbb{R}$ existieren, sondern nur solange das Argument im Logarithmus positiv bleibt, d.h. für $t > t_* - p^{-1}\,\exp(p\,x_*)$. Für $p = 0$ können wir die Formeln nicht verwenden. In diesem Fall gilt $\dot{x}(t) = 1$ für alle t und

   $$x(t) = t + C\,,$$

   ist die entsprechende allgemeine Lösung.

2. Für $f(x) = x^p$ mit Parameter $p \neq +1$ können wir

   $$F(x) = \frac{x^{1-p}}{1-p}\,, \qquad F'(x) = x^{-p} = \frac{1}{f(x)}\,, \qquad F^{-1}(z) = \big((1-p)\,z\big)^{1(1-p)}$$

wählen und erhalten die Lösungsformeln

$$x(t) = \big((1-p)\,(t+C)\big)^{1/(1-p)}\,, \qquad x(t) = \big((1-p)\,(t-t_*) + x_*^{1-p}\big)^{1/(1-p)}\,.$$

Alternativ können wir diese wieder direkt aus

$$\frac{\mathrm{d}x}{x^p} = \mathrm{d}t$$

durch formale Integration beider Seiten ableiten. Der Term $x(t)$ muss natürlich immer wohldefiniert sein, d.h. in Abhängigkeit von p kann es wieder nicht-globale Lösungen geben. Den Spezialfall $p = 2$ hatten wir schon am Anfang des Kapitels betrachtet.

3. Für $f(x) = p\,x$ kann man analoge Rechnungen durchführen, muss aber wegen $\int \mathrm{d}x/x = \ln|x|$ eine Fallunterscheidung bzgl. des Vorzeichens von x bzw. x_* treffen. Am Ende ergibt sich

$$x(t) = \pm\exp(p\,t + p\,C)\,, \qquad x(t) = x_*\,\exp\big(p\,(t-t_*)\big)\,,$$

wobei die zweite Formel für alle x_* und p verwendet werden kann. Die erste kann man auch als

$$x(t) = \tilde{C}\,\exp(p\,t)$$

schreiben, wobei dann die Substitution $\tilde{C} = \pm\exp(p\,C)$ zu Grunde liegt.

Lineare nichtautonome Gleichungen Eine weitere Anwendung der Separationsmethode sind Gleichungen der Bauart

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

mit gegebener Funktion a, wobei diese zu der in diesem Unterabschnitt betrachteten Klasse von Differentialgleichungen gehören, da wir $g(x) = x$ und $h(t) = a(t)$ setzen können. Unter der Annahme $x > 0$ erhalten wir $G(x) = \ln(x)$ und wollen für h die spezielle Stammfunktion $H(t) = \int_{t_*}^{t} a(\tau)$ wählen (sodass $H(t_*) = 0$ gilt). Damit erhalten wir

$$x(t) = x_*\exp\left(\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau\right),$$

als Lösungsformel, die sogar für alle $x_* \in \mathbb{R}$ (und nicht nur für $x_* > 0$) verwendet werden kann. In der Tat, es gilt $x(t_*) = x_*$ und die Kettenregel sowie der Hauptsatz der Differential- und Integralrechnung implizieren

$$\dot{x}(t) = x_*\exp\left(\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau\right) a(t) = a(t)\,x(t)$$

für alle $t \in \mathbb{R}$. Diese Formel ist ausgesprochen wichtig und nützlich. Im Spezialfall $a(t) = \alpha$ erhalten wir wieder $\dot{x}(t) = x_*\exp\big(\alpha\,(t-t_*)\big)$.

Variation der Konstanten – einfachster Fall

Vorbemerkung In diesem Abschnitt lösen wir das lineare Anfangswertproblem

$$\dot{x}(t) = a(t)\,x(t) + b(t)\,, \qquad x(t_*) = x_*$$

für gegebene Funktionen a und b.

Heuristische Herleitung der Lösungsformel Im homogenen Fall gilt $b(t) = 0$ für alle t und wir hatten bereits gesehen, dass

$$x_{\mathrm{hom}}(t) = C \exp\left(\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau\right),$$

die entsprechende allgemeine Lösungsformel ist, wobei C eine freie Konstante ist. Der Trick zur Bestimmung einer Lösung des inhomogenen Problems mit rechter Seite $b(t)$ besteht darin, die Konstante C zu „variieren“. Oder anders gesagt, wir suchen Lösungen mittels des Ansatzes

$$x(t) = C(t) \exp\left(\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau\right),$$

wobei wir das Gesetz für die „zeitabhängige Konstante“ $C(t)$ noch finden müssen. Differenzieren wir die letzte Formel nach t, so erhalten wir

$$\dot{x}(t) = \Big(\dot{C}(t) + C(t)\,a(t)\Big) \exp\left(\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau\right) = \dot{C}(t) \exp\left(\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau\right) + a(t)\,x(t)$$

und nach Einsetzen in die Differentialgleichung sowie kleineren Umstellungen ergibt sich

$$\dot{C}(t) = \exp\left(-\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau\right) b(t)\,.$$

Die wesentliche Beobachtung ist, dass in dieser Gleichung zwar noch $\dot{C}(t)$, aber kein $C(t)$ mehr auftaucht und dass wir daher beide Seiten der Gleichung integrieren können. Dies liefert[2]

$$C(t) = C(t_*) + \int_{t_*}^{t} \exp\left(-\int_{t_*}^{\sigma} a(\tau)\,\mathrm{d}\tau\right) b(\sigma)\,\mathrm{d}\sigma$$

und anschließend

$$\begin{aligned} x(t) &= C(t_*) \exp\left(\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau\right) + \int_{t_*}^{t} \exp\left(\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau\right) \exp\left(-\int_{t_*}^{\sigma} a(\tau)\,\mathrm{d}\tau\right) b(\sigma)\,\mathrm{d}\sigma \\ &= x_* \exp\left(\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau\right) + \int_{t_*}^{t} \exp\left(\int_{\sigma}^{t} a(\tau)\,\mathrm{d}\tau\right) b(\sigma)\,\mathrm{d}\sigma\,, \end{aligned}$$

[2]Wir haben die Integrationsvariable hier mit σ bezeichnet, hätten aber auch jeden anderen, noch nicht verwendeten Buchstaben wählen können.

wobei wir benutzt haben, dass unser Ansatz $C(t_*) = x_*$ impliziert und dass

$$\int_{t_*}^{t} a(\tau)\,\mathrm{d}\tau \; - \int_{t_*}^{\sigma} a(\tau)\,\mathrm{d}\tau = \int_{t}^{\sigma} a(\tau)\,\mathrm{d}\tau$$

aus der Gebietsadditivität bestimmter Integrale folgt. Insgesamt haben wir damit die gesuchte Lösungsformel gefunden. Diese sieht wegen der ineinander verschachtelten Integrale kompliziert aus, aber wenn wir die Integrale berechnen können (was oftmals, aber bei weitem nicht immer gelingt), erhalten wir explizite Ausdrücke für $x(t)$.

Bemerkungen

1. Auch hier gilt wieder: Die Herleitung ist wichtiger als die Endformel.
2. Wir werden weiter unten dieselben Ideen und Methoden für eine sehr viel größere Klasse von linearen Differentialgleichungen anwenden (*Duhamel-Prinzip*).
3. Wir können leicht nachrechnen (Übungsaufgabe), dass

$$x_{\mathrm{part}}(t) := \int_{t_*}^{t} \exp\left(- \int_{\sigma}^{t} a(\tau)\,\mathrm{d}\tau \right) b(\sigma)\,\mathrm{d}\sigma$$

eine spezielle bzw. partikuläre Lösung der Differentialgleichung ist, die im konkreten Fall der Anfangsbedingung $x(t_*) = 0$ genügt. Insbesondere kann die allgemeine Lösung der inhomogenen Differentialgleichung als

$$x(t) = x_{\mathrm{hom}}(t) + x_{\mathrm{part}}(t)\,,$$

geschrieben werden und wir werden sehen, dass ein analoges Resultat auch für lineare Differentialgleichungen in höheren Dimensionen gilt, wobei wir dann auf der linken Seite $x_{\mathrm{inhom}}(t)$ anstelle von $x(t)$ schreiben werden.

Beispiele

1. Für die Gleichung

$$\dot{x}(t) = \alpha\, x(t) + b(t)$$

kann die abstrakte Lösungsformel mit $a(t) = \alpha$ und wegen $\int_{t_*}^{t} \alpha\,\mathrm{d}\tau = \alpha\,(t - t_*)$ als

$$x(t) = x_* \exp\big(\alpha\,\big(t - t_*\big)\big) + \int_{t_*}^{t} \exp\big(\alpha\,(t - \sigma)\big)\, b(\sigma)\,\mathrm{d}\sigma$$

geschrieben werden. Gilt sogar $b(t) = \beta$, so können wir auch das verbleibende Integral ausrechnen und erhalten

$$x(t) = x_* \exp\big(\alpha\,\big(t - t_*\big)\big) + \alpha^{-1}\,\beta\,\Big(\exp\big(\alpha\,\big(t - t_*\big)\big) - 1\Big)\,.$$

Die Formeln für die Newtonsche Abkühlung (siehe oben) ergeben sich durch den Notationswechsel $x(t) = T(t)$, $\alpha = -\lambda$ und $b(t) = \lambda\, T_{\mathrm{umg}}(t)$ bzw. $\beta = \lambda\, T_{\#}$.

2. Wir wollen die verschiedenen Varianten der Methode an dem sehr einfachen Anfangswertproblem

$$\dot{x}(t) + x(t) = \sin(t)\,, \qquad x(0) = x_*$$

illustrieren.

(a) Wenn wir (durch Raten oder sonstwie) wissen, dass

$$x_{\text{part}}(t) = \tfrac{1}{2}\big(1 + \sin(t) - \cos(t)\big)$$

eine partikuläre Lösung der Differentialgleichung ist, so müssen wir nur noch die allgemeine Lösung der homogenen Gleichung addieren. Im konkreten Fall meint das

$$x(t) = C\,\exp(-t) + x_{\text{part}}(t)$$

und eine Auswertung der Anfangsbedingung (für die inhomogene, nicht die homogene Lösung) liefert $C = x_* + x_{\text{part}}(0) = x_*$ und damit

$$x(t) = \big(x_* + \tfrac{1}{2}\big)\,\exp(-t) + \tfrac{1}{2}\big(\sin(t) - \cos(t)\big)$$

als Lösungsformel für das inhomogene Anfangswertproblem.

(b) Wir können natürlich die abstrakten Lösungsformeln mit $a(t) = -1$, $t_* = 0$ und $b(t) = \sin(t)$ verwenden und erhalten

$$x(t) = x_*\,\exp(-t) + \int_0^t \exp(-t+\sigma)\,\sin(\sigma)\,\mathrm{d}\sigma\,.$$

Die Berechnung des Integrals gelingt via

$$\int_0^t \exp(-t+\sigma)\,\sin(\sigma)\,\mathrm{d}\sigma = \tfrac{1}{2}\big(\exp(-t) + \sin(t) - \cos(t)\big)$$

und nach Einsetzen sowie mit elementaren Termumformungen ergibt sich dasselbe Ergebnis wie oben.

(c) Wir lösen zuerst das homogene Problem und *variieren die Konstanten*, d.h. wir benutzen den Ansatz

$$x(t) = C(t)\,\exp(-t)\,, \qquad \dot{x}(t) = \dot{C}(t)\,\exp(-t) - C(t)\,\exp(-t)$$

und setzen ihn in die inhomogene Gleichung ein. Dies ergibt

$$\dot{C}(t)\,\exp(-t) = \sin(t)$$

wobei die Terme mit $C(t)$ sich gerade aufheben (andernfalls haben wir was falsch gemacht). Der Hauptsatz der Differential- und Integralrechnung garantiert

$$C(t) = C(0) + \int_0^t \exp(\sigma)\sin(\sigma)\,\mathrm{d}\sigma = C(0) + \tfrac{1}{2} + \tfrac{1}{2}\exp(t)\big(\sin(t) - \cos(t)\big)\,,$$

wobei $C(0)$ die Rolle der freien Integrationskonstante übernimmt. Insgesamt erhalten wir

$$x(t) = \big(C(0) + \tfrac{1}{2}\big)\,\exp(-t) + \tfrac{1}{2}\big(\sin(t) - \cos(t)\big)$$

und nach Einsetzen von $x_* = C(0)$ wiederum die gleiche Lösungsformel.

Bemerkung: Der dritte Lösungsweg ist natürlich der beste, weil wir uns hier nur eine Idee, aber keine komplizierten Formeln merken müssen.

3. Wir wollen das Anfangswertproblem

$$\dot{x}(t) - t^{-1}\, x(t) = t^3\,, \qquad x(1) = 1$$

lösen, wobei diesmal natürlich $t > 0$ gelten muss, da für $t = 0$ die Differentialgleichung keinen Sinn hat. Die allgemeine homogene Lösung ist in diesem Fall durch

$$x_{\mathrm{hom}}(t) = C\, \exp\left(\int_1^t \frac{\mathrm{d}\tau}{\tau} \right) = C\, \exp\big(\ln t\big) = C\, t$$

gegeben.

(a) Wenn wir durch Scharfes Hinsehen die partikuläre Lösung

$$x_{\mathrm{part}}(t) = \tfrac{1}{3}\, t^4$$

finden, so ergibt sich

$$x(t) = C\, t + \tfrac{1}{3}\, t^4 = \tfrac{2}{3}\, t + \tfrac{1}{3}\, t^4\,,$$

wobei wir $C = \frac{2}{3}$ durch Auswertung der Anfangsbedingung gewonnen haben. Beachte, dass die partikuläre Lösung nicht der Anfangsbedingung genügt. Das ist aber auch nicht notwendig, sondern wird am Ende durch die richtige Wahl von C kompensiert.

(b) Wir können diese Lösungsformel alternativ auch aus dem allgemeinen Resultat mit $a(t) = 1/t$, $b(t) = t^3$ und $t_* = 1$, $x_* = 1$ ableiten, da alle Integrale via

$$\begin{aligned} x(t) &= \exp\left(\int_1^t \frac{\mathrm{d}\tau}{\tau} \right) + \int_1^t \exp\left(\int_\sigma^t \frac{\mathrm{d}\tau}{\tau} \right) \sigma^3\, \mathrm{d}\sigma \\ &= t + \int_1^t \frac{t}{\sigma}\, \sigma^3\, \mathrm{d}\sigma = t + t\, \big[\tfrac{1}{3}\, \sigma^3 \big]_{\sigma=1}^{\sigma=t} \end{aligned}$$

berechnet werden können.

(c) Variation der Konstanten meint diesmal

$$x(t) = C(t)\, t\,, \qquad \dot{C}(t) = t^2$$

und wir erhalten wegen

$$C(t) = C(1) + \int_1^t \sigma^2\, \mathrm{d}\sigma = C(1) + \tfrac{1}{3}\, t^3 - \tfrac{1}{3}$$

wieder die obige Formel, da die Anfangsbedingung $C(1) = 1$ impliziert.

Vorlesungswoche 11

3.4 Satz von Picard-Lindelöf

Vorbemerkung Wir beweisen in diesem Abschnitt den *Hauptsatz über gewöhnliche Differentialgleichungen*, nämlich dass für jede hinreichend gute Funktion f und alle Anfangsdaten $(t_*,\, x_*)$ das jeweilige Anfangswertproblem genau eine maximale Lösung besitzt. Das entsprechende Theorem wird *Satz von Picard-Lindelöf* genannt und mithilfe des Banachschen Fixpunktsatzes bewiesen. Die zentrale Beobachtung ist dabei, dass jede Differentialgleichung auch als Integralgleichung interpretiert werden kann.

Vereinbarung In diesem Abschnitt betrachten wir die Differentialgleichung

$$\dot{x}(t) = f\big(t,\, x(t)\big)\,,$$

wobei $f : U \to \mathbb{R}^n$ eine gegebene stetige Funktion auf der offenen Menge $U \subseteq \mathbb{R} \times \mathbb{R}^n$ ist, sowie die Anfangsbedingung

$$x_* = x(t_*)\,,$$

mit beliebig fixierten Daten $(t_*,\, x_*) \in U$.

Lemma (**Umformulierung als Integralgleichung**) Eine Kurve $x : I \to \mathbb{R}^n$ erfüllt genau dann das Anfangswertproblem, wenn die Integralgleichung

$$x(t) = x_* + \int_{t_*}^{t} f\big(\tau,\, x(\tau)\big)\,\mathrm{d}\tau$$

für alle $t \in I$ erfüllt ist.

Beweis Die Behauptung ergibt sich unmittelbar aus dem Hauptsatz der Differential- und Integralrechnung (siehe *Analysis 1*), sofern dieser komponentenweise angewendet wird. □

Bemerkungen

1. Auf diesem Lemma beruht letztlich die gesamte mathematische Lösungstheorie gewöhnlicher Differentialgleichungen. Es besagt, dass jedes Anfangswertproblem in ein äquivalentes Fixpunktproblem mit einem geeignet definierten Integraloperator überführt werden kann.

2. Man kann für jede stetige Funktion f zeigen, dass die Integralgleichung im Lemma mindestens einen Fixpunkt besitzt (*Existenzsatz von Peano* bzw. *Fixpunktsatz von Schauder*), aber es kann im Allgemeinen viele Lösungen geben (siehe das folgende Gegenbeispiel).

3. Unter etwas schärferen Voraussetzungen an f wird es aber auf jedem hinreichend kleinen Intervall I um t_* genau einen Fixpunkt der Integralgleichung und damit genau eine Lösung des entsprechenden Anfangswertproblems geben. Wir werden dies nun im Detail diskutieren.

Gegenbeispiel Wir betrachten die stetige Funktion $f : \mathbb{R} \to \mathbb{R}$ mit

$$f(x) = x^{1/3} = \begin{cases} +\sqrt[3]{|x|} & \text{für } x > 0, \\ 0 & \text{für } x = 0, \\ -\sqrt[3]{|x|} & \text{für } x < 0, \end{cases}$$

sowie das dazugehörige autonome Anfangswertproblem

$$\dot{x}(t) = x(t)^{1/3}\,, \qquad x(0) = x_*$$

mit der Anfangszeit $t_* = 0$. Im Fall von $x_* = 0$ existieren unendlich viele maximale Lösungen[3], nämlich

$$x(t) = \pm \begin{cases} 0 & \text{für } t \le T, \\ \left(\frac{2}{3}\,(t-T)\right)^{3/2} & \text{für } t \ge T, \end{cases}$$

wobei $T > 0$ ein freier Parameter ist und auch das Vorzeichen beliebig gewählt werden kann.

Bemerkungen:

1. Die Funktion f ist zwar auf ganz $\mathbb{R}$ definiert und stetig, aber sie ist auf *keinem* Intervall K mit $0 \in K$ Lipschitz-stetig.[4] Daher können wir für $x_* = 0$ den Satz von Picard-Lindelöf (siehe unten) *nicht* anwenden.

2. Auf $\mathbb{R} \setminus \{0\}$ ist f aber stetig differenzierbar und damit auch lokal Lipschitz-stetig. Insbesondere können wir für $x_* < 0$ bzw. $x_* > 0$ die Funktion f auf der Menge $V = (-\infty, 0)$ bzw. $V = (0, +\infty)$ betrachten und der Satz von Picard-Lindelöf wird die Existenz und Eindeutigkeit einer maximalen Lösung implizieren. Durch Trennung der Veränderlichen können wir diese zu

$$x(t) = +\left(\tfrac{2}{3}\,t + |x_*|^{2/3}\right)^{3/2} \text{ bzw. } x(t) = -\left(\tfrac{2}{3}\,t + |x_*|^{2/3}\right)^{3/2} \text{ für } t > -\tfrac{3}{2}\,|x_*|^{2/3}$$

berechnen, wobei sie via $x(t) \to 0$ für $t \searrow -\frac{3}{2}\,|x_*|^{2/3}$ den Rand von V bzw. den kritischen Wert $x = 0$ erreicht.

Merkregel: Bei Differentialgleichungen mit Wurzeln — oder vergleichbaren, nicht-differenzierbaren Singularitäten — kann es selbst bei vorgeschriebenen Anfangsdaten mehrere Lösungen geben.

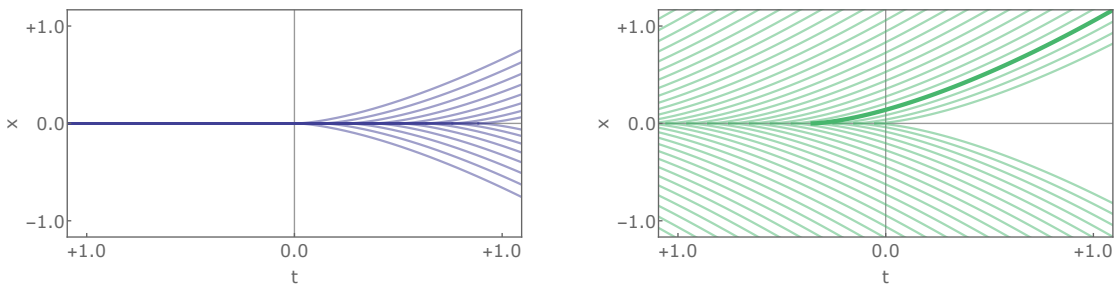

Abbildung Verschiedene Lösungen zum Anfangswertproblem aus dem Gegenbeispiel. *Links*: Für $x_* = 0$ gibt es unendlich viele maximale Lösungen. *Rechts*: Für jedes $x_* \neq 0$ gibt es hingegen genau eine maximale Lösung, wobei diese für eine Wahl von x_* zur besseren Darstellung hervorgehoben wurde.

[3] Wir können leicht nachrechnen, dass die angegebene Formel wirklich eine Lösung des betrachtetem Anfangswertproblems liefert. Da sie außerdem auf ganz $\mathbb{R}$ definiert ist, handelt es sich auch um eine maximale Lösung.

[4] Diese Aussage kann mit dem Mittelwertsatz aus *Analysis 1* sowie der uneigentlichen Grenzwertaussage $\lim_{x\to 0} 1/\sqrt[3]{x^2} = +\infty$ begründet werden.

Definition Die Funktion f erfüllt die lokale Lipschitz-Bedingung, wenn für jede kompakte Teilmenge $K \subset U$ eine Konstante L existiert (die von K abhängen darf), sodass

$$\big|f(t,\, x) - f(t,\, \tilde{x})\big| \le L\big|x - \tilde{x}\big|$$

für alle $(t,\, x)$, $(t,\, \tilde{x}) \in K$ gilt. Kann L unabhängig von K gewählt werden, so sagen wir, f genügt der globalen Lipschitz-Bedingung auf ihrem Definitionsbereich.

Bemerkungen

1. Die Lipschitz-Bedingung fordert gerade, dass f nicht nur stetig bzgl. t und x ist, sondern sogar lokal Lipschitz-stetig bzgl. x ist.

2. Die erste Version des Mittelwertsatzes der Differentialrechnung garantiert, dass jede stetig differenzierbare Funktion f die Lipschitz-Eigenschaft besitzt, wobei dann

   $$L := \max\big\{\big|\mathrm{Jac}_x\, f(t,\, x)\big| \,:\, (t,\, x) \in K\big\}$$

   gewählt werden kann. In der mathematischen Praxis wird f in aller Regel stetig differenzierbar sein und der unten formulierte Existenz- und Eindeutigkeitssatz von Picard-Lindelöf kann angewendet werden.

3. Die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ mit $f(t,\, x) = \cos(t)\,|x|$ ist zwar nicht differenzierbar, aber sie erfüllt via

   $$\big|f(t,\, x) - f(t,\, \tilde{x})\big| = \big|\cos(t)\big|\big||x| - |\tilde{x}|\big| \le \big||x| - |\tilde{x}|\big| \le \big|x - \tilde{x}\big|$$

   die Lipschitz-Bedingung mit der globalen Konstanten $L = 1$.

Vereinbarung Im Rest dieses Abschnitts setzen wir stets voraus, dass f der lokalen Lipschitz-Bedingung genügt!

Theorem (**Satz von Picard-Lindelöf : eindeutige lokale Lösung**) Für jedes $(t_*,\, x_*) \in U$ existiert ein $\varepsilon > 0$, sodass das Anfangswertproblem eine eindeutige Lösung auf dem Intervall $(t_* - \varepsilon,\, t_* + \varepsilon)$ besitzt.

Beweis *Vorbereitungen*: Wir wählen $\delta > 0$ sowie $\varrho > 0$, sodass die abgeschlossene sowie beschränkte (und damit auch kompakte) Menge $[t_* - \delta,\, t_* + \delta] \times \overline{B}_\varrho(x_*)$ ganz in U liegt sowie eine entsprechende Lipschitz-Konstante L von f. Wir definieren außerdem

$$C := \max\big\{\big|f(t,\, x_*)\big| \,:\, t \in [t_* - \delta,\, t_* + \delta]\big\}\,, \qquad \varepsilon := \min\left\{\delta,\, \frac{1}{2\,L}\,,\, \frac{\varrho}{2\,C}\right\}$$

und setzen $I := (t_* - \varepsilon,\, t_* + \varepsilon)$.

Fixpunktargument: Wir betrachten den Funktionenraum[5]

$$\mathcal{X} := \mathsf{BC}(I) = \big\{x : I \to \mathbb{R}^n \text{ stetig und beschränkt}\big\}$$

[5] Wir könnten in diesem Beweis auch mit Funktionen auf dem abgeschlossenen Intervall $\overline{I}$ arbeiten, aber die Behauptung im Theorem wird traditionell mit einem offenen Intervall formuliert.

mit der Norm

$$\|x\|_\infty = \sup\big\{\big|x(t)\big| \;:\; t \in I\big\}$$

sowie die Teilmenge

$$\mathcal{A} := \big\{x \in \mathcal{X} \;:\; x(t) \in \overline{B}_\varrho(x_*) \text{ für alle } t \in I\big\}\,,$$

wobei unsere Resultate aus dem ersten Kapitel implizieren, dass $\mathcal{X}$ ein vollständiger normierter Raum ist und dass $\mathcal{A}$ eine abgeschlossene Teilmenge von $\mathcal{X}$ ist. Des Weiteren definieren wir die Abbildung $\mathcal{F} : \mathcal{A} \to \mathcal{X}$ durch

$$\mathcal{F}(x)(t) := x_* + \int_{t_*}^{t} f\big(\tau,\, x(\tau)\big)\,\mathrm{d}\tau\,,$$

wobei die rechte Seite stetig von $t \in I$ abhängt und Werte in $\mathbb{R}^n$ annimmt. Für je zwei Funktionen $x,\, \tilde{x} \in \mathcal{A}$ und jedes $t \in I$ ergibt sich insbesondere

$$\begin{aligned}
\big|\mathcal{F}(x)(t) - \mathcal{F}(\tilde{x})(t)\big| &= \left|\int_{t_*}^{t} \Big(f\big(\tau,\, x(\tau)\big) - f(\tau,\, \tilde{x}(\tau))\Big)\,\mathrm{d}\tau\right| \\
&\le \int_{t_*}^{t} \Big|f\big(\tau,\, x(\tau)\big) - f(\tau,\, \tilde{x}(\tau))\Big|\,\mathrm{d}\tau \le L \int_{t_*}^{t} \big|x(\tau) - \tilde{x}(\tau)\big|\,\mathrm{d}\tau \\
&\le L\,\|x - \tilde{x}\|_\infty \int_{t_*}^{t} \mathrm{d}\tau \le L\,\|x - \tilde{x}\|_\infty\,|t - t_*| = L\,\varepsilon\,\|x - \tilde{x}\|_\infty \\
&\le \tfrac{1}{2}\,\|x - \tilde{x}\|_\infty
\end{aligned}$$

aus der Lipschitz-Eigenschaft von f und die Supremumsbildung über $t \in I$ liefert

$$\big\|\mathcal{F}(x) - \mathcal{F}(\tilde{x})\big\|_\infty \le \tfrac{1}{2}\big\|x - \tilde{x}\big\|_\infty\,,$$

d.h. $\mathcal{F}$ erfüllt eine Kontraktionsabschätzung. Die spezielle Wahl $\tilde{x} = x_*$ (konstante Funktion auf I) impliziert

$$\big\|\mathcal{F}(x) - \mathcal{F}(x_*)\big\|_\infty \le \tfrac{1}{2}\,\big\|x - x_*\big\|_\infty$$

und weil auch $\|\mathcal{F}(x_*) - x_*\|_\infty \le \frac{1}{2}\,\varrho$ wegen

$$\big|\mathcal{F}(x_*)(t) - x_*\big| \le \int_{t_*}^{t} \big|f(\tau,\, x_*)\big|\,\mathrm{d}\tau \le C\,\big|t - t_*\big| \le \varepsilon\,C \le \tfrac{1}{2}\,\varrho$$

gilt, ergibt sich für jedes $x \in \mathcal{A}$ aus der Dreiecksungleichung die Abschätzung

$$\big\|\mathcal{F}(x) - x_*\big\|_\infty \le \big\|\mathcal{F}(x) - \mathcal{F}(x_*)\big\|_\infty + \big\|\mathcal{F}(x_*) - x_*\big\|_\infty \le \tfrac{1}{2}\,\varrho + \tfrac{1}{2}\,\varrho = \varrho\,.$$

Insgesamt schließen wir, dass $\mathcal{F}$ die Menge $\mathcal{A}$ kontraktiv in sich abbildet und der Satz von Banach garantiert die Existenz und Eindeutigkeit eines Fixpunktes $x \in \mathcal{A}$ mit $x = \mathcal{F}(x)$. Insbesondere ist die Kurve $x : I \to \overline{B}_\varrho(x_*)$ die einzige Lösung

des Anfangswertproblems auf dem Intervall I, die ausschließlich Werte innerhalb von $\overline{B}_\varrho(x_*)$ annimmt.

Verbessertes Eindeutigkeitsresultat: Sei nun $\tilde{x} : I \to \mathbb{R}^n$ eine weitere Lösung des Anfangswertproblems auf dem Intervall I. Wir wollen nun indirekt zeigen, dass die Lösung auch in $\mathcal{A}$ liegt und nehmen daher an, dass ein $\tilde{t} \in I$ mit $\bigl|\tilde{x}(\tilde{t}) - x_*\bigr| > \varrho$ existiert. Wir diskutieren zunächst den Unterfall $t_* < \tilde{t} < t_* + \varepsilon$. Da $\tilde{x}$ eine stetige Funktion auf I ist, gilt $\bigl|\tilde{x}(t) - x_*\bigr| < \varrho$ für alle $t > t_*$, die hinreichend nahe bei t_* liegen. Deshalb muss es ein $t_\# \in (t_*, \tilde{t})$ geben, sodass $\bigl|\tilde{x}(t_\#) - x_*\bigr| = \varrho$ gilt und außerdem $\bigl|\tilde{x}(t) - x_*\bigr| < \varrho$ für alle $t \in (t_*, t_\#)$ erfüllt ist.[6] Analog zu oben zeigen wir

$$\begin{aligned}\bigl|\mathcal{F}(\tilde{x})(t_\#) - x_*\bigr| &\le \bigl|\mathcal{F}(\tilde{x})(t_\#) - \mathcal{F}(x_*)(t_\#)\bigr| + \bigl|\mathcal{F}(x_*)(t_\#) - x_*\bigr| \\ &\le \int_{t_*}^{t_\#} L\,\bigl|\tilde{x}(\tau) - x_*\bigr|\,\mathrm{d}\tau + \int_{t_*}^{t_\#} C\,\mathrm{d}\tau \le L\,\varrho\,(t_\# - t_*) + C\,(t_\# - t_*) \\ &< L\,\varepsilon\,\varrho + C\,\varepsilon = \tfrac{1}{2}\,\varrho + \tfrac{1}{2}\,\varrho = \varrho\,,\end{aligned}$$

wobei wir benutzt haben, dass $0 < t_\# - t_* < \varepsilon$ gilt. Andererseits ist $\tilde{x}$ eine Lösung des Anfangswertproblems, d.h. es gilt $\tilde{x}(t_\#) = \mathcal{F}(\tilde{x}(t_\#))$ und wir erhalten via

$$|\tilde{x}(t_\#) - x_*| = \bigl|\mathcal{F}(\tilde{x})(t_\#) - x_*\bigr| < \varrho = |\tilde{x}(t_\#) - x_*|$$

einen Widerspruch. Im Unterfall $t_* - \varepsilon < \tilde{t} < t_*$ argumentieren wir analog und schließen, dass $\tilde{t}$ und $t_\#$ nicht existieren. Insbesondere gehört $\tilde{x}$ zu $\mathcal{A}$ und stimmt mit der durch den Fixpunktsatz bereitgestellten Lösung x überein. □

Bemerkungen

1. Der Beweis ist einer der wichtigsten Anwendungen des Banachschen Fixpunktsatzes, wobei die Lipschitz-Bedingung an f eine zentrale Rolle spielt.

2. Wir werden weiter unten eine zweite Fassung des Theorems herleiten, die nicht nur die Existenz und Eindeutigkeit lokaler Lösungen auf hinreichend kleinen Intervallen garantiert, sondern die sogenannten maximalen Lösungen charakterisiert.

3. Der im Beweis angegebene Wert für ε ist nicht optimal. Beachte auch, dass der optimale Wert für ε im Allgemeinen nicht nur von f, sondern auch von den Anfangsdaten (t_*, x_*)abhängen wird.

4. Gilt $U = \mathbb{R} \times \mathbb{R}^n$ und erfüllt f sogar die globale Lipschitz-Bedingung, so kann der Beweis so modifiziert werden, dass er mit deutlich weniger technischen Details eine wesentlich bessere Aussage liefert. Siehe dazu die Übungen.

Folgerung (**Eindeutigkeitssatz**) Für je zwei Lösungen $x : I \to \mathbb{R}^n$ und $\tilde{x} : \tilde{I} \to \mathbb{R}$ der Differentialgleichung gilt: Wenn die Gleichung $x(t) = \tilde{x}(t)$ für ein $t \in I \cap \tilde{I}$ erfüllt ist, so gilt sie schon für alle $t \in I \cap \tilde{I}$.

[6] Die reelle Zahl $t_\#$ ist also die kleinste Zeit zwischen t_* und $t_* + \varepsilon$, für die $x(t)$ im Rand der Kugel $\overline{B}_\varrho(x_*)$ liegt.

Beweis Wir betrachten das Intervall $J := I \cap \tilde{I}$ und können o.B.d.A. annehmen, dass dieses nichtleer ist (denn andernfalls ist nichts zu zeigen). Wir schreiben $J = (t_-, t_+)$, wobei $t_- = -\infty$ und/oder $t_+ = +\infty$ zugelassen ist, und definieren die Menge

$$T_* := \{t \in J \ : \ x(t) = \tilde{x}(t)\},$$

die aufgrund der Voraussetzung mindestens ein (und von nun an fixiertes) $t_0 \in J$ enthält und damit nichtleer ist. Insbesondere sind

$$t_{*,-} := \inf\big\{t \in (t_-, t_0] \ : \ (t, t_0] \subseteq T_*\big\} \in [t_-, t_0]$$

und

$$t_{*,+} := \sup\big\{t \in [t_0, t_+) \ : \ [t_0, t) \subseteq T_*\big\} \in [t_0, t_+]$$

wohldefiniert (siehe das nachfolgende Bild). Wir wollen nun durch zwei Widerspruchsargumente zeigen, dass sowohl $t_{*,-} = t_-$ als auch $t_{*,+} = t_+$ gilt. Wir nehmen zunächst an, dass $t_- < t_{*,-} \leq t_0$ gilt, und bemerken, dass dann $t_{*,\pm} \in T_*$ aus der Stetigkeit von x und $\tilde{x}$ folgt. Wir wenden außerdem das Theorem auf die Anfangsbedingung $(t_{*,-}, x(t_{*,-})) = (t_{*,-}, \tilde{x}(t_{*,-}))$ an und erhalten ein $\varepsilon_- > 0$ sowie eine entsprechende Lösung $x_- : (t_{*,-} - \varepsilon_-, t_{*,-} + \varepsilon_-) \to \mathbb{R}^n$ der Differentialgleichung. Durch eventuelles Verkleinern von ε_- können wir sicherstellen, dass $t_- < t_{*,-} - \varepsilon_-$ gilt und die lokale Eindeutigkeit von x_- garantiert $x_-(t) = x(t) = \tilde{x}(t)$ für alle Zeiten t zwischen $t_{*,-} - \varepsilon_-$ und $t_{*,-}$ gilt. Dies impliziert via

$$x(t) = \tilde{x}(t) \qquad \text{für alle} \quad t \in (t_{*,-} - \varepsilon_-, t_0]$$

einen Widerspruch zur minimalen Wahl von $t_{*,-}$ und wir haben $t_{*,-} = t_-$ gezeigt. Mit analogen Argumenten führen wir die Annahme $t_{*,+} < t_+$ zum Widerspruch. □

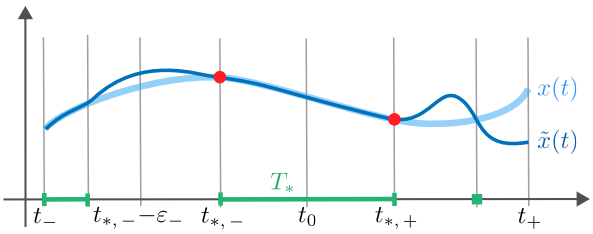

Abbildung Illustration der hypothetischen Situation im Beweis des Eindeutigkeitssatzes, wobei die Widersprüche durch Anwendung des Satzes von Picard-Lindelöf bzgl. der rot markierten Anfangsdaten entstehen.

Theorem (**Satz von Picard-Lindelöf : eindeutige maximale Lösung**) Für jedes $(t_*, x_*) \in U$ existieren genau ein offenes Intervall $I = (t_-, t_+)$ sowie genau eine Kurve $x : I \to \mathbb{R}^n$, sodass die beiden folgenden Aussagen erfüllt sind:

1. Für jede Lösung $\tilde{x} : \tilde{I} \to \mathbb{R}^n$ des Anfangswertproblems gilt

$$\tilde{I} \subseteq I \qquad \text{sowie} \qquad \tilde{x}(t) = x(t) \quad \text{für alle} \quad t \in \tilde{I}\,.$$

2. Für jede kompakte Menge $K \subset U$ mit $(t_*, x_*) \in K$ existieren Zeiten τ_- und τ_+ mit $t_- < \tau_- \leq \tau_+ < t_+$, sodass $\big(t, x(t)\big) \in U \setminus K$ für alle $t \in (t_-, \tau_-)$ sowie alle $t \in (\tau_+, \tau_-)$ gilt.

Diese Kurve x wird die eindeutige maximale Lösung des Anfangswertproblems genannt und man sagt, sie verlässt jedes Kompaktum.

Beweis *Teil 1*: Nach dem ersten Teil des Theorems ist die Menge aller Lösungen des Anfangswertproblems nichtleer und im Folgenden bezeichnen wir Lösungen mit $x^{[s]} : I^{[s]} \to \mathbb{R}^n$, wobei s zur Menge S gehört und $I^{[s]} = (t_-^{[s]}, t_+^{[s]})$ ein offenes Intervall ist, dass t_* enthält.[7] Wir definieren[8]

$$t_- := \inf \left\{ t_-^{[s]} \ : \ s \in S \right\}, \qquad t_+ := \sup \left\{ t_+^{[s]} \ : \ s \in S \right\},$$

sowie

$$x(t) := x^{[s]}(t) \qquad \text{sofern} \qquad t \in I_s \,,$$

für alle $t \in I$, der Eindeutigkeitssatz sicherstellt, dass der Wert von $x(t)$ nicht von der konkreten Wahl von $s \in S$ abhängt (siehe auch das nachfolgende Bild). Die Kurve $x : I \to \mathbb{R}^n$ ist nach Konstruktion die eindeutige maximale Lösung des Anfangswertproblems.

Teil 2: Angenommen, es gibt kein τ_- mit der geforderten Eigenschaft. Dann existiert mindestens eine monoton fallende Folge $(t_n)_{n\in\mathbb{N}}$ in I mit

$$t_n \quad \xrightarrow{n\to\infty} \quad t_- \qquad \text{sowie} \qquad \big(t_n, x(t_n)\big) \in K \quad \text{für alle} \quad n \in \mathbb{N} \,.$$

Nach dem Satz von Bolzano-Weierstrass existiert eine strikt monoton wachsende Indexfolge $(n_j)_{j\in\infty}$, sodass $\big(t_{n_j}, x(t_{n_j})\big)$ für $j \to \infty$ gegen einen Grenzwert in $K \subset \mathbb{R}^{n+1}$ konvergiert, wobei dieser nach Konstruktion als (t_-, x_-) mit $x_- := \lim_{j\to\infty} x(t_{n_j}) \in \mathbb{R}^n$ geschrieben werden kann.[9] Wir wenden die lokale Version des Theorems auf die Anfangswerte (t_-, x_-) an und erhalten eine Lösung x_- der Differentialgleichung, die auf einem Intervall $(t_- - \varepsilon_-, t_- + \varepsilon_-)$ definiert ist. Der Eindeutigkeitssatz garantiert, dass durch

$$\tilde{x}_-(t) := \begin{cases} x_-(t) & \text{für } t \in (t_- - \varepsilon_-, t_-] \\ x(t) & \text{für } t \in (t_-, t_+) \end{cases}$$

eine Lösung der Differentialgleichung gegeben ist, für die sowohl $\tilde{x}_-(t_-) = x_-$ als auch $\tilde{x}_-(t_*) = x_*$ gilt. Das ist aber ein Widerspruch zur Maximalität von x und wir schließen, dass unsere Annahme der Nichtexistenz von τ_- falsch war. Analog zeigen wir die Existenz von τ_+.

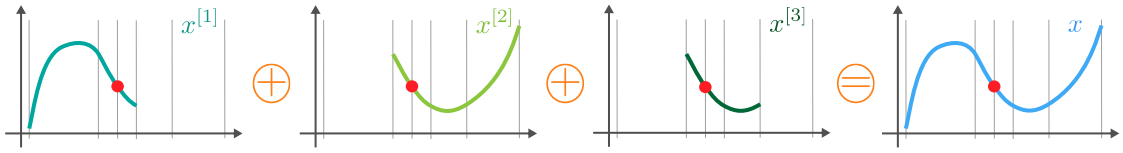

Abbildung Schematische Illustration der ersten Beweisidee in der zweiten Fassung des Satzes von Picard-Lindelöf: Die maximale Lösung (blau) entsteht durch „Verkleben“ aller Lösungen (grün), wobei der Eindeutigkeitssatz sicherstellt, dass dies möglich ist, und wir hier beispielhaft nur drei Lösungen betrachten. Der rote Punkte repräsentiert die Anfangsdaten (t_*, x_*).

[7] Die Indexmenge S enthält genauso viele Elemente, wie es Lösungen des Anfangswertproblems gibt, und wird daher eine überabzählbare Menge sein.

[8] Etwas vereinfacht kann man sagen: t_- bzw. t_+ ist die kleinste untere bzw. die größte obere Grenze aller Existenzintervalle.

[9] Beachte, dass eine Folge von Vektoren genau dann konvergiert, wenn sie komponentenweise konvergiert.

Bemerkungen

1. Das maximale Existenzintervall (t_-, t_+) hängt im Allgemeinen von den Anfangswerten (t_*, x_*) ab. Siehe dazu die Beispiele zu Beginn des Kapitels.

2. Der Satz von Picard-Lindelöf liefert die Existenz einer eindeutigen Lösung sowohl in der „Zukunft“ (also für $t_* < t < t_+$) als auch in der „Vergangenheit“ (also für $t_- < t < t_*$). Das ist aus mathematischer Sicht sehr wichtig, obwohl wir in vielen Anwendungen nur an der maximalen Lösung für $t > t_*$ interessiert sind. [10]

3. Der zweite Teil des Theorems wird oftmals als Rand-zu-Rand-Eigenschaft bezeichnet (siehe die nachfolgenden Erklärungen), obwohl die Formulierung mittels kompakter Mengen $K \subset U$ deutlich präziser und weniger missverständlich ist.

stetige Abhängigkeit von den Anfangsdaten Mit mehr Aufwand können wir zeigen, dass die maximale Lösung stetig — und oftmals sogar in differenzierbarer Weise — von den Anfangsdaten sowie von eventuell vorhandenen Parametern in f abhängt. An dieser Stelle verzichten wir auf eine präzise mathematische Formulierung und einen entsprechenden Beweis, werden aber diesen Aspekt weiter unten im Zusammenhang mit Sensitivitäten noch einmal diskutieren.

Verhalten maximaler Lösungen In vielen Fällen gilt $U = \mathbb{R} \times V$, wobei V eine offene, aber beschränke Teilmenge des $\mathbb{R}^n$ ist (zum Beispiel eine offene Kugel oder ein offener Quader). In diesem Fall impliziert der zweite Teil des Theorems die Aussage

> *Entweder gilt $t_+ = +\infty$ (Existenz der maximalen Lösung für alle zukünftigen Zeiten) oder $x(t)$ erreicht zur Zeit t_+ mit $t_* < t_+ < \infty$ den Rand der Menge V (Blowup der maximalen Lösung in endlicher Zeit).*

sowie einen analogen Lehrsatz für t_-.

Heuristische Interpretation: Die Menge $\mathbb{R} \times V \subset \mathbb{R}^{n+1}$ kann als unendlich ausgedehnter Zylinder mit Querschnittsfläche V interpretiert werden, wobei $\mathbb{R} \times \partial V$ gerade die *Mantelfläche* ist und $\{-\infty\} \times V$ bzw. $\{+\infty\} \times V$ als unendlich ferne *Grund-* bzw. *Deckfläche* betrachtet werden kann. Damit ergeben sich die folgenden vier Fälle:

1. $t_- = -\infty$ und $t_+ = +\infty$: Die maximale Lösung verbindet die Grundfläche mit der Deckfläche.

2. $t_- = -\infty$ und $t_+ < +\infty$: Die maximale Lösung startet in der Grundfläche und endet in der Mantelfläche.

3. $t_- > -\infty$ und $t_+ = +\infty$: Die maximale Lösung startet in der Mantelfläche, aber endet in der Deckfläche.

4. $t_- > -\infty$ und $t_+ < +\infty$: Die maximale Lösung verbindet Punkte in der Mantelfläche.

Im Fall $U = \mathbb{R}^{n+1}$ (d.h. $U = \mathbb{R} \times V$ mit $V = \mathbb{R}^n$) können die beiden Alternativen

$$\textit{Es gilt } t_+ = +\infty \textit{ oder } \lim_{t \nearrow t_+} \big|x(t)\big| = \infty.$$

[10] Daher kommt auch die Bezeichnung *Anfangsdaten*.

sowie

$$\textit{Es gilt } t_- = -\infty \textit{ oder } \lim\nolimits_{t \nearrow t_-} \big|x(t)\big| = \infty.$$

aus dem Theorem abgeleitet werden.

Heuristische Interpretation: Die Menge U entsteht aus den Zylindern $\mathbb{R} \times B_\varrho(0)$ im Grenzübergang $\varrho \to 0$ und in diesem Sinne kann der Rand von V als eine unendlich ferne Spähre von Radius ∞ um den Mittelpunkt 0 interpretiert werden.[11]

Merkregel Die maximale Lösung eines Anfangswertproblems existiert nur dann nicht für alle Zeiten, wenn sie in endlicher Zeit den Rand des Definitionsbereiches von f erreicht oder in endlicher Zeit „explodiert“, wobei beide Phänomene als *Blowup* bezeichnet werden. Oder sehr salopp gesprochen: Maximale Lösungen hören nur dann auf zu existieren, wenn es dafür einen zwingenden Grund gibt.

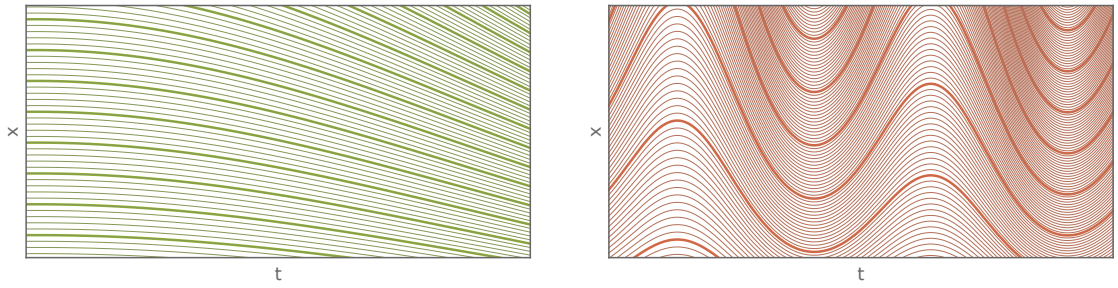

Abbildung Zwei typische Beispiele für die Lösungen der Differentialgleichung mit einer stetig differenzierbaren Funktion $f : U \to \mathbb{R}$, wobei U ein offenes Rechteck im $\mathbb{R}^2$ ist (es gilt also $n = 1$) und für ausgewählte Lösungen der Graph der jeweiligen Funktion x als dünne oder als dicke Linie dargestellt ist. Nach dem Satz von Picard-Lindelöf verläuft durch jeden Punkt $(t_*, x_*) \in U$ genau eine maximale Lösung der Differentialgleichung, wobei sich zwei verschiedene Lösungen nach dem Eindeutigkeitssatz *niemals* kreuzen oder berühren. Beachte auch, dass jede maximale Lösung von Rand zu Rand läuft, wobei der Rand ∂U in der t-Richtung (also links oder rechts im Bild) oder in der x-Richtung (oben oder unten) erreicht werden kann. Das analoge Bild für $n = 2$ ist dreidimensional, wobei dann jede Lösung der Differentialgleichung einer Raumkurve entspricht. Siehe auch die andere graphische Darstellung für autonome Differentialgleichungen mit $n = 2$.

weitere Betrachtungen

Picard-Iteration* Der Banachsche Fixpunktsatz liefert nicht nur die Existenz und Eindeutigkeit der Lösungen von Anfangswertproblemen, sondern auch das Rekursionsschema

$$x^{[0]}(t) = x_*\,, \qquad x^{[k+1]}(t) = x_* + \int_{t_*}^{t} f\big(\tau,\, x^{[k+1]}(\tau)\big)\,\mathrm{d}\tau\,,$$

mit dem ausgehend von der konstanten Funktion $x^{[0]}$ schrittweise immer bessere Approximationen $x^{[k]}$ gewonnen werden können.[12]

[11] Beachte aber, dass der topologische Rand von $\mathbb{R}^n$ die leere Menge ist.

[12] Der obere Index $[k]$ bezieht sich hier weder auf Komponenten noch auf Ableitungen, sondern ist der Index in der Rekursionsfolge.

Bemerkungen:

1. Die Rekursionsvorschrift kann auch als
$$\dot{x}^{[k+1]}(t) = f\big(t,\, x^{[k}(t)\big)\,, \qquad x^{[k+1]}(t_*) = x_*$$
geschrieben werden.

2. Der Beweis der lokalen Version des Satzes von Picard-Lindelöf garantiert, dass die entstehenden Funktionenfolge $\big(x^{[k]}\big)_{k\in\mathbb{N}}$ zumindest auf hinreichend kleinen Intervallen um t_* für $k \to \infty$ gegen eine Grenzfunktion $x^{[\infty]}$ konvergiert, die das Anfangswertproblem löst.

3. Die Picard-Iteration ist zwar von großem theoretischen Interesse, aber in der Praxis meist wenig brauchbar, da die auftretenden Integrale nur schwer berechnet werden können. Bei der numerischen Lösung von Anfangswerten auf dem Computer benutzt man ganz andere Approximationsschemata, die ohne explizite Berechnung von Integralen auskommen.

Beispiel Wir wollen die Picard-Iteration beispielhaft für das Anfangswertproblem
$$\dot{x}(t) = x(t)\,, \qquad x(0) = 1$$
durchführen. Wir starten mit der konstanten Funktion
$$x^{[0]}(t) = 1$$
und berechnen im ersten Schritt die Funktion $x^{[1]}$ aus der Funktion $x^{[0]}$ via
$$x^{[1]}(t) = 1 + \int_0^t x^{[0]}(\tau)\,\mathrm{d}\tau = 1 + \int_0^t 1\,\mathrm{d}\tau = 1 + t\,.$$
Analog erhalten wir
$$x^{[2]}(t) = 1 + \int_0^t x^{[1]}(\tau)\,\mathrm{d}\tau = 1 + \int_0^t (1+\tau)\,\mathrm{d}\tau = 1 + t + \tfrac{1}{2}\,t^2$$
sowie
$$x^{[3]}(t) = 1 + \int_0^t x^{[2]}(\tau)\,\mathrm{d}\tau = 1 + \int_0^t \big(1 + \tau + \tfrac{1}{2}\tau^2\big)\,\mathrm{d}\tau = 1 + t + \tfrac{1}{2}\,t^2 + \tfrac{1}{3}\,t^3$$
im zweiten bzw. dritten Schritt und können nun sukzessive fortfahren. Mit vollständiger Induktion über k können wir die explizite Formel
$$x^{[k]}(t) = \sum_{m=0}^{k} \frac{1}{m!}\,t^m = 1 + t + \ldots + \tfrac{1}{(k-1)!}\,t^{k-1} + \tfrac{1}{k!}\,t^k$$
ableiten, wobei wir den Induktionsanfang bereits verifiziert haben und der Induktionsschritt $k \rightsquigarrow k{+}1$ aus dem Nachrechnen von
$$1 + \int_0^t \Big(1 + \tau + \ldots + \tfrac{1}{(k-1)!}\,\tau^{k-1} + \tfrac{1}{k!}\,\tau^k\Big)\,\mathrm{d}\tau = 1 + t + \tfrac{1}{2}\,t + \ldots + \tfrac{1}{k!}\,t^k + \tfrac{1}{(k+1)!}\,t^{k+1}$$

besteht. Insbesondere sehen wir, dass

$$x^{[k]}(t) \quad \xrightarrow{k\to\infty} \quad x^{[\infty]}(t) = \sum_{m=0}^{\infty} \frac{1}{m!}\, t^m = \exp(t)$$

gilt, d.h. dass die durch die Picard-Iteration gewonnenen Funktionen $x^{[k]}$ im Limes $k \to \infty$ wirklich die Lösung des Anfangswertproblems reproduzieren.

Lemma (**Lemma von Gronwall, einfachste Version**) Sei $\xi : [t_1,\, t_2) \to \mathbb{R}$ eine stetige und skalare Funktion, sodass

$$\xi(t) \le \gamma_0 + \gamma_1 \int_{t_1}^{t} \xi(\tau)\, \mathrm{d}\tau$$

für geeignete Konstanten $\gamma_0,\, \gamma_1 > 0$ und alle $t \in [t_1,\, t_2)$ erfüllt ist. Dann gilt

$$\xi(t) \le \gamma_0 \, \exp\big(\gamma_1\,(t - t_1)\big)$$

für alle $t \in [t_1,\, t_2)$.

Beweis Wir betrachten die Funktion $\eta : [t_1,\, t_2) \to \mathbb{R}$ mit

$$\eta(t) := \gamma_1 \, \exp\big(-\gamma_1\,(t - t_1)\big) \int_{t_1}^{t} \xi(\tau)\, \mathrm{d}\tau$$

und erhalten mit den Rechenregeln für Ableitungen aus *Analysis 1* sowie aufgrund der Voraussetzung die Abschätzung

$$\dot\eta(t) = \gamma_1 \, \exp\big(-\gamma_1\,(t - t_1)\big) \left(\xi(t) - \gamma_1 \int_{t_1}^{t} \xi(\tau)\, \mathrm{d}\tau \right) \le \gamma_0\, \gamma_1 \, \exp\big(-\gamma_1\,(t - t_1)\big)\,,$$

wobei auf der rechten Seite weder η noch ξ auftaucht. Wir ersetzen t durch τ und integrieren beide Seiten über $\tau \in [t_1,\, t]$ mit $t \in [t_1,\, t_2)$. Der Fundamentalsatz der Analysis sowie direkte Rechnungen liefern

$$\eta(t) \le \int_{t_1}^{t} \dot\eta(\tau)\, \mathrm{d}\tau = \gamma_0 \, \Big(1 - \exp\big(-\gamma_1\,(t - t_1)\big)\Big)\,,$$

wobei wir auch $\eta(t_1) = 0$ benutzt haben. In Kombination mit der Definition von η ergibt sich

$$\gamma_1 \int_{t_1}^{t} \xi(\tau)\, \mathrm{d}\tau = \exp\big(\gamma_1\,(t - t_1)\big)\, \eta(t) \le \gamma_0 \, \exp\big(\gamma_1\,(t - t_1)\big) - \gamma_0$$

und die Behauptung folgt nach Einsetzen in die Voraussetzung. □

Bemerkungen

1. Das Lemma ist der einfachste Vertreter einer ganzen Klasse von integralen oder differentiellen *Vergleichsprinzipien* und stellt ein ausgesprochen wichtiges, aber eher technisches Hilfsmittel in der Theorie gewöhnlicher Differentialgleichungen dar.

2. Es gibt alternative und weniger willkürlich erscheinende Beweise. Diese benutzen *Ober-* und *Unterlösungen* für Anfangswertprobleme, aber diese Konzepte können wir in unserer Vorlesung nicht einführen.

Korollar (**lineares Wachstum impliziert Existenz für alle Zeiten**) Sei f auf ganz $\mathbb{R} \times \mathbb{R}^n$ definiert und gelte

$$\bigl|f(t,\, x)\bigr| \leq c_0(t) + c_1(t)\,|x|$$

für alle $(t,\, x) \in \mathbb{R} \times \mathbb{R}^n$, wobei $c_0,\, c_1 : \mathbb{R} \to [0,\, \infty)$ zwei stetige und nichtnegative Funktionen sind. Dann existiert für alle Anfangsdaten $(t_*,\, x_*)$ die entsprechende maximale Lösung auf ganz $\mathbb{R}$.

Beweis Nach dem Satz von Picard-Lindelöf existiert die eindeutige maximale Lösung $x : I \to \mathbb{R}^n$ auf einem offenen Intervall $I = (t_-,\, t_+)$ mit $t_- < t_* < t_+$.

Existenz für $t \geq t_*$: Die Integralgleichung aus dem Lemma zur Umformulierung impliziert für alle $t \in (t_*,\, t_+)$ die Abschätzung

$$\bigl|x(t)\bigr| \leq \bigl|x_*\bigr| + \int_{t_*}^{t} \bigl|f\bigl(\tau,\, x(\tau)\bigr)\bigr|\,\mathrm{d}\tau \leq x_* + \int_{t_*}^{t} c_0(\tau)\,\mathrm{d}\tau + \int_{t_*}^{t} c_1(\tau)\bigl|x(\tau)\bigr|\,\mathrm{d}\tau\,,$$

wobei wir die Dreiecksungleichung sowie die Eigenschaften eindimensionaler Integrale verwendet haben. Unter der Annahme $t_+ < \infty$ sind

$$\gamma_0 := |x_*| + \int_{t_*}^{t_+} c_0(\tau)\,\mathrm{d}\tau\,, \qquad \gamma_1 := \max\{c_1(\tau) \;:\; \tau \in [t_*,\, t_+]\}$$

wohldefinierte reelle Zahlen und das Lemma von Gronwall — angewendet mit $t_1 = t_*$, $t_2 := t_+$ und $\xi(t) = \bigl|x(t)\bigr|$ liefert

$$\bigl|x(t)\bigr| \leq \gamma_0\,\exp\bigl(\gamma_1\,\bigl(t - t_*\bigr)\bigr) \leq \gamma_0\,\exp\bigl(\gamma_1\,\bigl(t_+ - t_*\bigr)\bigr) < \infty$$

für alle $t \in [t_*,\, t_+)$, aber dies widerspricht der Rand-zu-Rand-Eigenschaft maximaler Lösungen. Daher gilt $t_+ = +\infty$.

Existenz für $t \leq t_*$: Wir betrachten die reparametrisierte Kurve $\tilde{x} : \tilde{I} \to \mathbb{R}^n$ mit

$$\tilde{x}(t) := x\bigl(2\,\tilde{t}_* - t\bigr) \qquad \tilde{I} := (2\,t_* - t_+,\, 2\,t_* - t_-)$$

und zeigen mit direkten Rechnungen, dass diese das Anfangswertproblem

$$\dot{\tilde{x}}(t) = \tilde{f}\bigl(t,\, \tilde{x}(t)\bigr)\,, \qquad \tilde{x}(t_*) = x_*$$

löst,[13] wobei die rechte Seite $\tilde{f}\bigl(t,\, x\bigr) = f\bigl(2\,t_* - t,\, x\bigr)$ der Abschätzung

$$\bigl|\tilde{f}\bigl(t,\, x\bigr)\bigr| \leq \tilde{c}_0(t) + \tilde{c}_1(t)\,\bigl|x\bigr| \qquad \text{mit} \qquad \tilde{c}_j(t) := c_j(2\,t_* - t)$$

genügt. Wir können nun alle Argumente aus dem ersten Beweisteil wiederholen und zeigen, dass $2\,t_* - t_- = +\infty$ und damit $t_- = -\infty$ gilt. □

[13] Wir benutzen hier eine sogenannte *Zeitumkehrung*. Dies ist ein Standardtrick um mathematische Ergebnisse über zukünftige Zeiten auf vergangene Zeiten zu übertragen.

Beispiele

1. Jede Lösung der linearen Differentialgleichung

$$\dot{x}(t) = A(t)\,x(t) + b(t)$$

mit stetigen Funktionen $A : \mathbb{R} \to \mathbb{M}^{n\times n}$ und $b : \mathbb{R} \to \mathbb{R}^n$ existiert für alle Zeiten, denn wir können das Korollar mit $c_0(t) = \big|b(t)\big|$ und $c_1(t) = \big|A(t)\big|$ verwenden. Insbesondere kann kein Blowup auftreten. Im Rahmen unserer Vorlesung ist dies die wichtigste Konsequenz des Lemmas von Gronwall.

2. Jede Lösung der nichtlinearen Gleichung

$$\dot{x}(t) = a(t)\,\sin\big(b(t)\,x(t)\big)$$

existiert für alle Zeiten, sofern die Koeffizientenfunktionen a, $b : \mathbb{R} \to \mathbb{R}$ stetig sind.

3. Erfüllt f die *globale Lipschitz-Bedingung* auf $\mathbb{R} \times \mathbb{R}^n$, so gilt

$$\big|f(t,\,x)\big| \le \big|f(t,\,0)\big| + \big|f(t,\,x) - f(t,\,0)\big| \le f(t,\,x) + L\,\big|x - 0\big|$$

und das Korollar kann angewendet werden.

Besonderheiten autonomer Gleichungen

Vereinbarung In diesem Abschnitt studieren wir die Differentialgleichung

$$\dot{x}(t) = f\big(x(t)\big)\,,$$

bei der die rechte Seite nicht explizit von t abhängt, sondern durch ein *Vektorfeld* $f : V \subseteq \mathbb{R}^n \to \mathbb{R}^n$ gegeben ist. Dabei setzen wir immer voraus, dass V offen ist und dass f der lokalen Lipschitz-Bedingung genügt, sodass der Satz von Picard-Lindelöf mit $U = \mathbb{R} \times V$ angewendet werden kann.

Definition Für jede Lösung $x : I \to \mathbb{R}^n$ nennen wir die Menge

$$X := \big\{x(t) \,:\, t \in I\big\}$$

die entsprechende Trajektorie (oder den entsprechenden Orbit), wobei es sich gerade um das *Bild* von x im Sinne der Kurventheorie handelt.[14]

Bemerkungen

1. Eine parametrisierte Kurve $x : I \to \mathbb{R}^n$ kann nur dann eine Lösung der Differentialgleichung sein, wenn sie Werte in V annimmt. Insbesondere ist die jeweilige Trajektorie immer eine Teilmenge von V und x beschreibt, wie diese in Abhängigkeit von der Zeit t durchlaufen wird.

2. Jede Lösung $x : I \to V$ ist eine *Integralkurve* an das Vektorfeld f, denn die Differentialgleichung besagt, dass der Tangentialvektor $\dot{x}(t)$ zu jeder Zeit $t \in I$ durch $f\big(x(t)\big)$ gegeben ist. Siehe auch die nachfolgenden Bilder.

[14] Es ist nicht ungewöhnlich, dass identische Konstrukte in verschiedenen Bereichen der Mathematik anders bezeichnet werden.

3. Die lokale Lipschitz-Bedingung meint im autonomen Fall, dass für jede kompakte Teilmenge $K \subset V$ eine Konstante L existiert, sodass

$$\big|f(x) - f(\tilde{x})\big| \leq \big|x - \tilde{x}\big|$$

für alle x, $\tilde{x} \in K$ gilt. Eine hinreichende Bedingung ist — wie schon weiter oben diskutiert — die stetige Differenzierbarkeit von f.

4. Die von Rand-zu-Rand-Bedingung meint im autonomen Fall für jede maximale Lösung $x : (t_-, t_+) \to V$ das Folgende: Entweder gilt $t_+ = +\infty$ (bzw. $t_- = -\infty$) oder $x(t)$ verlässt für $t \nearrow t_+$ (bzw. $t \searrow t_-$) jedes Kompaktum $K \subset V$.

5. Viele Differentialgleichungen in den Natur- und Ingenieurwissenschaften sind autonom, denn die fundamentalen Naturgesetze ändern sich nicht mit der Zeit. Differentialgleichungen, die Systeme mit sich verändernder äußerer Anregung modellieren, sind aber in aller Regel nicht-autonom.

6. Jede Nullstelle $\xi \in V$ von f definiert via $x(t) = \xi$ eine stationäre (also konstante) Lösung mit $\dot{x}(t) = 0$ für alle t.[15] In anwendungsrelevanten Differentialgleichungen beschreiben die stationären Lösungen oftmals die *Gleichgewichte* des zugrunde liegenden physikalischen, biologischen oder technischen Systems. Sie spielen außerdem eine wichtige Rolle bei der qualitativen Untersuchung von Differentialgleichungen, vor allem, wenn ihre *Stabilitätseigenschaften* bekannt sind (siehe dazu weiter unten).

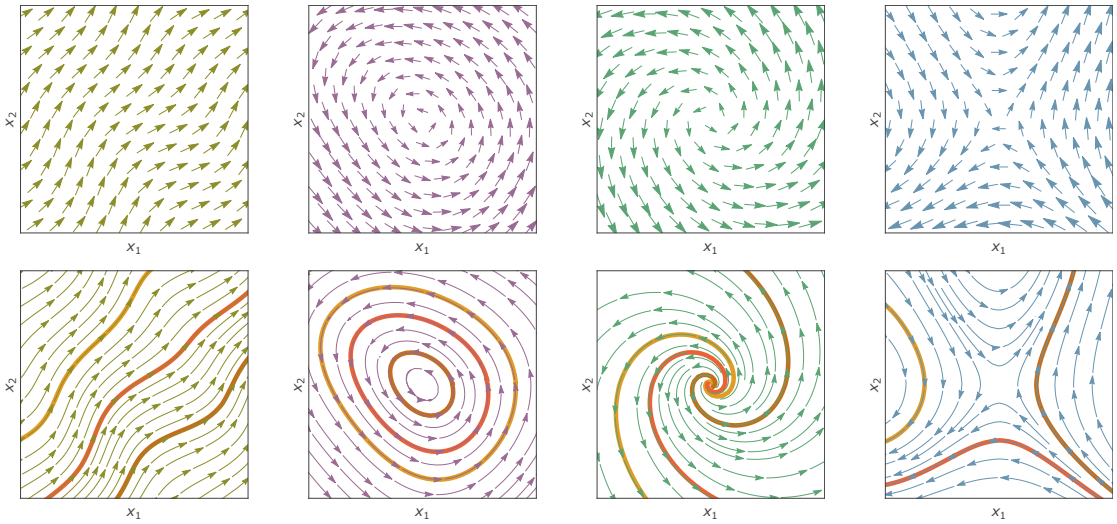

Abbildung *Oben*: Vier Beispiele für Vektorfelder auf einem offenen Quadrat $V \subset \mathbb{R}^2$. Weitere Beispiele finden sich weiter unten. *Unten*: Visualisierung ausgewählter Integralkurven (bzw. von Lösungen der entsprechenden autonomen Differentialgleichung), wobei jeweils drei hervorgehoben wurden. Beachte, dass immer nur die Trajektorie, d.h. das Bild der parametrisierten Lösungskurve dargestellt ist. Wie schnell diese durchlaufen wird, kann aus der graphischen Darstellung nicht (bzw. nur indirekt durch die Länge der Pfeile) abgelesen werden.

[15] Umgekehrt kann leicht gezeigt werden, das jede stationäre Lösung einer Nullstelle von f entspricht.

Lemma (**Shift-Invarianz bei autonomen Gleichungen**) Seien $x : I \to V$ und $\tilde{x} : \tilde{I} \to V$ zwei maximale Lösungen der autonomen Differentialgleichung, die via

$$x\big(t_*\big) = x_* = \tilde{x}\big(\tilde{t}_*\big)$$

zu unterschiedlichen Zeiten durch denselben Punkt $x_* \in V$ laufen. Dann gehört t genau dann zu I, wenn $t - t_* + \tilde{t}_*$ Element von $\tilde{I}$ ist, und es gilt

$$x\big(t\big) = \tilde{x}\big(t - t_* + \tilde{t}_*\big)$$

für alle $t \in I$.

Beweis Durch Nachrechnen verifizieren wir, dass die Kurve $\hat{x} : \hat{I} \to V$ mit

$$\hat{I} := \big\{t \in \mathbb{R} \,:\, t - t_* + \tilde{t}_* \in \tilde{I}\big\}\,, \qquad \hat{x}(t) := \tilde{x}\big(t - t_* + \tilde{t}_*\big)$$

sowie die Kurve $\check{x} : \check{I} \to V$ mit

$$\check{I} := \big\{t \in \mathbb{R} \,:\, t - \tilde{t}_* + t_* \in I\big\}\,, \qquad \check{x}(t) := x\big(t - \tilde{t}_* + t_*\big)$$

beide die Differentialgleichung lösen, wobei $\hat{x}(t_*) = \tilde{x}(\tilde{t}_*) = x_* = x(t_*) = \check{x}(\tilde{t}_*)$ nach Konstruktion erfüllt ist. Insbesondere gilt $\hat{x}(t_*) = x(t_*)$ bzw. $\check{x}(\tilde{t}_*) = \tilde{x}(\tilde{t}_*)$ und der Satz von Picard-Lindelöf liefert $\hat{I} \subseteq I$ und $\hat{x}(t) = x(t)$ für alle $t \in \hat{I}$ bzw. $\check{I} \subseteq \tilde{I}$ und $\check{x}(t) = \tilde{x}(t)$ für alle $t \in \check{I}$. Die Behauptungen ergeben sich nun nach einfachen Termumstellungen. □

Bemerkungen

1. Die dem Beweis zugrunde liegende Idee ist eigentlich ganz einfach: Wird eine Lösung der autonomen Gleichung in der Zeit verschoben, so entsteht wieder eine Lösung. Diese ist auf einem anderen Zeitintervall definiert, besitzt jedoch dieselbe Trajektorie (im Lemma gilt also $X = \tilde{X}$).

2. Aufgrund der zeitlichen Shift-Invarianz ist bei Anfangswertproblemen autonomer Differentialgleichungen der Wert von t_* nicht wirklich wichtig (die Wahl von x_* hingegen schon). Oftmals wird daher $t_* = 0$ gesetzt.

3. Für nicht-autonome Gleichungen gibt es in der Regel keine analoge Aussage.

   Gegenbeispiel: Die beiden Kurven $x(t) = t-1$ und $\tilde{x}(t) = \exp(-t)+t-1$ erfüllen

   $$\dot{x}(t) = t - x(t) \quad \text{mit} \quad x(1) = 0 \qquad \text{sowie} \qquad \dot{\tilde{x}}(t) = t - \tilde{x}(t) \quad \text{mit} \quad \tilde{x}(0) = 0\,,$$

   aber sie können nicht durch eine Zeitverschiebung ineinander überführt werden.

Theorem (**autonome skalare Differentialgleichungen**) Sei $f : \mathbb{R} \to \mathbb{R}$ stetig differenzierbar und seien ξ_1 und ξ_2 zwei aufeinanderfolgende Nullstellen von f mit

$$f'(\xi_1) > 0 > f'(\xi_2) \text{ und } f(x) > 0 \text{ für } \xi_1 < x < \xi_2 \qquad (\textit{blau im Bild})$$

bzw.

$$f'(\xi_1) < 0 < f'(\xi_2) \text{ und } f(x) < 0 \text{ für } \xi_1 < x < \xi_2 \qquad (\textit{gelb im Bild}).$$

Dann existiert für jedes (t_*, x_*) mit $x_* \in (\xi_1, \xi_2)$ die entsprechende maximale Lösung für alle $t \in \mathbb{R}$ und ist

strikt monoton wachsend in t mit $\lim_{t\to-\infty} x(t) = \xi_1$ und $\lim_{t\to+\infty} x(t) = \xi_2$

bzw.

strikt monoton fallend in t mit $\lim_{t\to-\infty} x(t) = \xi_2$ und $\lim_{t\to+\infty} x(t) = \xi_1$.

Insbesondere „verbindet“ x die beiden stationären Punkte ξ_1 und ξ_2.

Beweis Wir zeigen nur die erste Behauptung; die zweite folgt dann mit analogen Argumenten.

untere und obere Schranke: Nach dem Satz von Picard-Lindelöf existiert eine eindeutige maximale Lösung $x : I \to \mathbb{R}$ mit $I = (t_-, t_+)$ des Anfangswertproblems. Die alles entscheidende Beobachtung ist, dass

$$\xi_1 =: x_1(t) < x(t) < x_2(t) := \xi_2$$

für alle $t \in I$ gilt, wobei $x_1, x_2 : \mathbb{R} \to \mathbb{R}$ zwei stationäre Lösungen der Differentialgleichung sind. In der Tat, wenn zum Beispiel eine Zeit $\check{t} \in I$ mit $x(\check{t}) = \xi_2$ existiert, so gibt es wegen der Stetigkeit von x sowie nach dem Zwischenwertsatz aus *Analysis 1* auch eine Zeit $\hat{t} \in (t_*, \check{t})$ mit $x(\hat{t}) = \xi_2 = x_2(\hat{t})$. Der Eindeutigkeitssatz impliziert dann aber $x(t) = x_2(t) = \xi_2$ für alle $t \in I$ und damit via $x_* = x(t_*) = \xi_2$ einen Widerspruch. Damit haben wir die obere Schranke für $x(t)$ etabliert und die untere Schranke kann analog hergeleitet werden.

globale Existenz und Monotonie: Die Schranken implizieren zusammen mit der Rand-zu-Rand-Eigenschaft maximaler Lösungen, dass $t_- = -\infty$ und $t_+ = +\infty$ gilt. Zusammen mit der Differentialgleichung liefern sie die Abschätzung

$$\dot{x}(t) = f\big(x(t)\big) > 0$$

für alle $t \in \mathbb{R}$. Insgesamt haben wir gezeigt, dass x auf ganz $\mathbb{R}$ definiert und strikt monoton wachsend ist.

asymptotisches Verhalten: Die Eigenschaften stetiger, monotoner und beschränkter Funktionen (siehe *Analysis 1*) garantieren, dass die beiden Grenzwerte

$$x(\pm\infty) := \lim_{t\to\pm\infty} x(t)$$

wohldefiniert sind, und dass $\xi_1 \le x(-\infty) < x_* < x(+\infty) \le \xi_2$ gilt. Wir nehmen nun an, es gelte $x(+\infty) < \xi_2$. Aufgrund der Eigenschaften von f existiert $\delta > 0$, sodass $\dot{x}(t) = f(x(t)) \ge \delta$ für alle $t \ge t_*$ gilt und der Fundamentalsatz der Analysis impliziert

$$x(t) = x_* + \int_{t_*}^{t} \dot{x}(\tau)\,\mathrm{d}\tau \ge x_* + \delta\,(t - t_*) \quad \xrightarrow{t\to+\infty} \quad +\infty\,.$$

Das ist aber ein Widerspruch zur oberen Schranke und daher gilt $x(+\infty) = \xi_2$. Mit ähnlichen Argumenten zeigen wir, dass $x(-\infty) = \xi_1$ gilt, da wir andernfalls Zeiten $t \le t_*$ betrachten und via

$$x(t) = x_* - \int_{t}^{t_*} \dot{x}(\tau)\,\mathrm{d}\tau \le x_* - \delta\,(t_* - t) \quad \xrightarrow{t\to+\infty} \quad -\infty$$

erneut einen Widerspruch konstruieren können. □

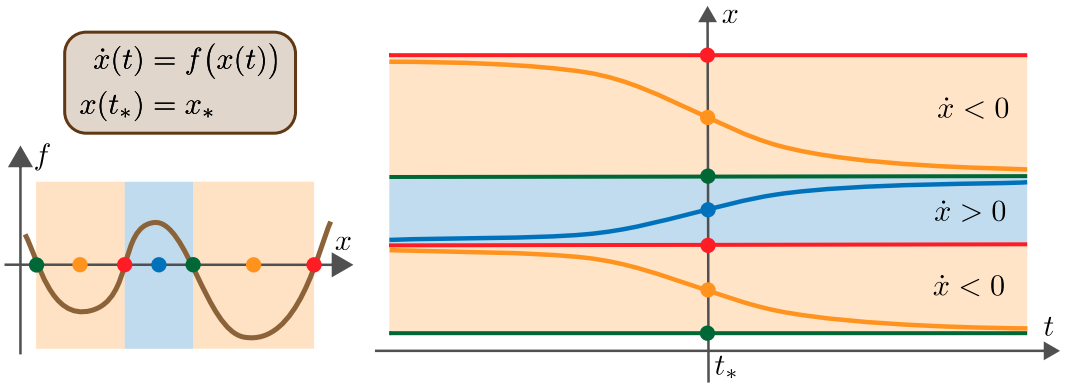

Abbildung Graphische Darstellung des soeben formulierten Theorems, wobei wir die grünen bzw. roten stationären Lösungen weiter unten als *stabil* bzw. *instabil* klassifizieren werden. Beachte, dass sich die Monotonie-Eigenschaften der Lösungen direkt aus der Differentialgleichung ablesen lassen und dass sich die Lösungen für verschiedene Werte von x_* im t-x-Diagramm niemals schneiden dürfen, da sonst im Schnittpunkt ein Widerspruch zum Satz von Picard-Lindelöf entstünde.

Bemerkungen

1. Das Theorem beschreibt nur die typischen Standardfälle und wir können mit ähnlich einfachen Argumenten auch entsprechende Resultate für mehrfache oder anders entartete Nullstellen — also für den Fall $f(\xi_j) = f'(\xi_j) = 0$ — herleiten. Außerdem gibt es Varianten mit $\xi_1 = -\infty$ oder $\xi_2 = +\infty$.

2. Mit mehr Aufwand können wir die exponentiellen Konvergenzabschätzungen
$$\big|x(t) - x(-\infty)\big| \le C \exp\left(+\lambda_- t\right), \qquad \big|x(t) - x(+\infty)\big| \le C \exp\left(-\lambda_+ t\right)$$
beweisen, wobei $\lambda_- = +f'(x(-\infty))$ und $\lambda_+ = -f'(x(+\infty))$ positiv sind.

3. Das Theorem illustriert, dass wir manchmal auch ohne explizite Lösungsformeln sehr genaue Informationen über die Lösungen von autonomen Differentialgleichungen erhalten können. Allerdings gilt das Theorem nur für $n = 1$, d.h. für autonome skalare Gleichungen. Im planaren Fall ($n = 2$) können wir das qualitative Verhalten von Lösungen auch sehr gut verstehen und visualisieren, obwohl dann die Komponentenfunktionen in aller Regel nicht mehr monoton sind und zum Beispiel auch periodisch sein können. Siehe dazu die folgenden Beispiele für sogenannte *Phasenportraits*. In drei oder mehr Dimensionen ($n \ge 3$) ist die qualitative Theorie autonomer Differentialgleichungen jedoch deutlich anspruchsvoller, vor allem weil dann auch chaotische Effekte auftreten können.

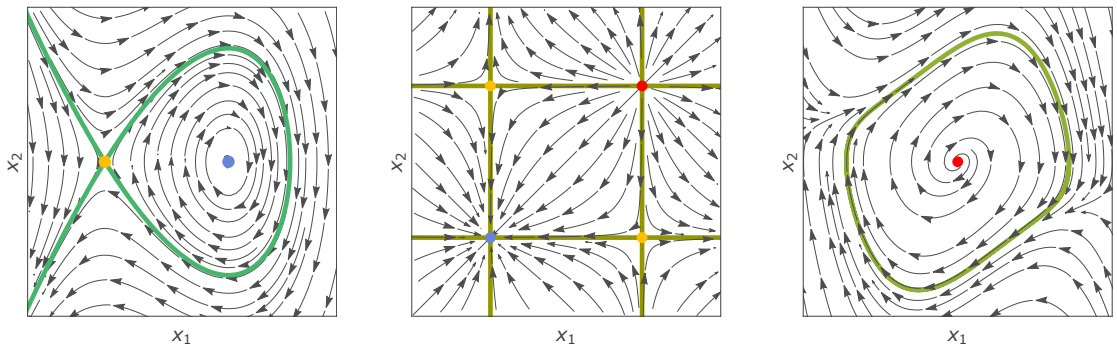

Abbildung Drei numerisch berechnete Beispiele für Phasenportraits planarer Vektorfelder, mit deren Hilfe das qualitative Verhalten der entsprechenden Differentialgleichung sehr gut verstanden werden kann. Besonders hervorgehoben sind stationäre Punkte, *hetero-* und *homokline* Orbits sowie *Grenzzyklen*.

Vorlesungswoche 12

3.5 lineare Differentialgleichungen

Vorbemerkung In diesem Abschnitt betrachten wir die lineare Differentialgleichung

$$\dot{x}(t) = A(t)\,x(t) + b(t)\,,$$

wobei $A : \mathbb{R} \to \mathbb{M}^{n\times n}$ und $b : \mathbb{R} \to \mathbb{R}^n$ gegebene stetige Funktionen sind. Wir wollen dabei zunächst keine Anfangswerte vorschreiben, sondern beliebige Lösungen der Differentialgleichung studieren, und beginnen mit dem *homogenen* Fall, in dem $b(t) = 0$ für alle $t \in \mathbb{R}$ gilt.

Erinnerung Wir hatten schon im Nachgang zum Satz von Picard-Lindelöf gezeigt, dass bei einer linearen Differentialgleichung (homogen oder nicht) alle maximalen Lösungen für alle Zeiten definiert sind. Wir werden daher im Folgenden immer Lösungen betrachten, die auf ganz $\mathbb{R}$ definiert sind.

homogene Gleichung

Theorem (**Hauptsatz über linear homogene Gleichungen**) Für die lineare und homogene Differentialgleichung

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

gelten die folgenden Aussagen:

1. (*Superpositionsprinzip*) Sind x, $\tilde{x} : \mathbb{R} \to \mathbb{R}^n$ zwei Lösungen und sind η, $\tilde{\eta}$ zwei reelle Zahlen, so ist $\eta\,x + \tilde{\eta}\,\tilde{x} : \mathbb{R} \to \mathbb{R}^n$ mit

   $$\big(\eta\,x + \tilde{\eta}\,\tilde{x}\big)(t) := \eta\,x(t) + \tilde{\eta}\,\tilde{x}(t)$$

   auch eine Lösung.

2. (*linearer Lösungsraum*) Die Menge aller Lösungen ist ein reeller Vektorraum der Dimension n, wobei für je n Lösungen $x^{[1]}$, ..., $x^{[n]}$ die folgenden drei Aussagen paarweise äquivalent sind:

   (a) Sie bilden eine Basis des Lösungsraumes.

   (b) Es existiert ein $t_* \in \mathbb{R}$, sodass $x^{[1]}(t_*)$, ..., $x^{[n]}(t_*)$ linear unabhängige Vektoren sind.

   (c) Die Vektoren $x^{[1]}(t)$, ..., $x^{[n]}(t)$ sind für jedes $t \in \mathbb{R}$ linear unabhängig.

Beweis Die erste Behauptung kann einfach nachgerechnet werden und die zweite ergibt sich aus den Ergebnissen des vorherigen Abschnitts (Übungsaufgabe). □

Bemerkungen

1. Das Superpositionsprinzip gilt für die Lösungen der (homogenen) Differentialgleichung, d.h. ohne Berücksichtigung von Anfangsdaten, denn bei Anfangswertproblemen ist die Lösung eindeutig. Für nichtlineare Gleichungen gibt es *kein* Superpositionsprinzip und für inhomogene lineare Gleichungen muss es etwas abgewandelt werden (siehe unten).

2. Die Äquivalenz im zweiten Teil impliziert, dass linear unabhängige Anfangsdaten zu linear unabhängigen Lösungen gehören und umgekehrt.

3. Wir können bei linear homogenen Gleichungen auch komplex rechnen, d.h. die Komponenten von $x(t)$ und $A(t)$ dürfen komplexe Zahlen sein und alle Aussagen gelten sinngemäß weiterhin. Die unabhängige Größe t muss aber immer reell sein.

Definition Sind $x^{[1]}$, ..., $x^{[n]}$ linear unabhängige Lösungen, so wird $X : \mathbb{R} \to \mathbb{M}^{n\times n}$ mit

$$X(t) := \begin{pmatrix} | & & | \\ x^{[1]}(t) & \dots & x^{[n]}(t) \\ | & & | \end{pmatrix} = \begin{pmatrix} x_1^{[1]}(t) & \dots & x_1^{[n]}(t) \\ \vdots & & \vdots \\ x_n^{[1]}(t) & \dots & x_n^{[n]}(t) \end{pmatrix}$$

als Fundamentalmatrix (der Differentialgleichung) bezeichnet und $\det X(t)$ wird die entsprechende Wronski-Determinante genannt.

Bemerkungen

1. Das Theorem impliziert, dass die Matrix $X(t)$ für jedes $t \in \mathbb{R}$ invertierbar ist.

2. Per Definition liefern die Spalten von $X(t)$ linear unabhängige Lösungen der Differentialgleichung. Für die Zeilen wird dies aber im Allgemeinen nicht gelten.

3. Ist $X(t)$ eine Fundamentalmatrix und $c \in \mathbb{R}^n$ ein beliebiger *Spaltenvektor*, so liefert

$$x(t) := X(t)\, c = c_1\, x^{[1]}(t) + \ \dots\ + c_n\, x^{[n]}(t)$$

   auch eine Lösung der Differentialgleichung und jede Lösung kann in dieser Form dargestellt werden. Beachte aber, dass für einen *Zeilenvektor* d das Produkt $d\, X(t)$ im Allgemeinen keine Lösung darstellt.

4. Fundamentalmatrizen sind nicht eindeutig, denn es gibt viele Basen im Lösungsraum. Bei gegebener Anfangszeit t_* werden wir häufig zusätzlich

$$x^{[1]}(t_*) = e^{[1]}\,, \qquad \dots\,, \qquad x^{[n]}(t_*) = e^{[n]}$$

   fordern, wobei $e^{[j]}$ der j-te Einheitsvektor im $\mathbb{R}^n$ ist. Dann gilt $X(t_*) = \mathsf{1}$, aber je nach Anwendung oder Kontext sind auch andere Wahlen für die Basislösungen sinnvoll. Siehe dazu auch die Formeln weiter unten zur Umrechnung zwischen verschiedenen Fundamentallösungen.

5. Wir werden im nächsten Abschnitt sehen, dass die Fundamentalmatrizen im autonomen Fall — also wenn $A(t) = A$ für alle $t \in \mathbb{R}$ gilt — explizit als Matrixexponential berechnet werden können. Bei nicht-autonomen Gleichungen ist dies im Allgemeinen nicht möglich, aber der Hauptsatz ist trotzdem sehr wichtig. Er garantiert, dass wir aus n geratenen oder sonstwie erhaltenen linear unabhängigen Lösungen alle anderen in einfacher Weise zusammenbauen können.

Beispiele

1. Für das autonome und planare System

$$\dot{x}(t) = A\,x(t)\,, \qquad A = \begin{pmatrix} 0 & +1 \\ -1 & 0 \end{pmatrix}$$

ist

$$X(t) = \begin{pmatrix} +\cos(t) & +\sin(t) \\ -\sin(t) & +\cos(t) \end{pmatrix} \qquad \text{mit} \qquad \det\big(X(t)\big) = 1$$

eine Fundamentalmatrix, die den zwei linear unabhängigen Lösungen

$$x^{[1]}(t) = \begin{pmatrix} +\cos(t) \\ -\sin(t) \end{pmatrix}, \qquad x^{[2]}(t) = \begin{pmatrix} +\sin(t) \\ +\cos(t) \end{pmatrix}$$

entspricht. Jede andere Lösung kann via

$$x(t) = c_1\,x^{[1]}(t) + c_2\,x^{[2]}(t) = X(t)\,c$$

mit $c \in \mathbb{R}^2$ als Superposition dieser Basislösungen dargestellt werden. Die Formel

$$\widetilde{X}(t) = \begin{pmatrix} +\cos(t-t_*) & +\sin(t-t_*) \\ -\sin(t-t_*) & +\cos(t-t_*) \end{pmatrix}$$

liefert für jedes feste t_* eine weitere Fundamentallösung und wir verizifieren

$$\widetilde{X}(t_*) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \widetilde{X}(t) = X(t) \begin{pmatrix} +\cos(t_*) & -\sin(t_*) \\ +\sin(t_*) & +\cos(t_*) \end{pmatrix}$$

mit direkten Rechnungen und trigonometrischen Additionstheoremen.

2. Die nicht-autonome Differentialgleichung

$$\dot{x}(t) = A(t)\,x(t)\,, \qquad A(t) = \begin{pmatrix} -2 & 0 \\ t & -1 \end{pmatrix}$$

besitzt die Fundamentalmatrix

$$X(t) = \begin{pmatrix} 0 & -\mathsf{e}^{-2\,t} \\ \mathsf{e}^{-t} & (t+1)\mathsf{e}^{-2\,t} \end{pmatrix}, \qquad \det\big(X(t)\big) = \mathsf{e}^{-3\,t}\,,$$

wobei wir leicht nachprüfen können, dass jede Spalte in der Tat einer Lösung der Differentialgleichung entspricht.

Korollar (**Lösungsformel mit Anfangsdaten**) Für jedes $(t_*, x_*) \in \mathbb{R} \times \mathbb{R}^n$ beschreibt die Kurve $x : \mathbb{R} \to \mathbb{R}^n$ mit

$$x(t) = X(t)\, X^{-1}(t_*)\, x_*$$

die eindeutige maximale Lösung des entsprechenden homogenen Anfangswertproblems, wobei $X : \mathbb{R} \to \mathbb{M}^{n\times n}$ eine beliebige Fundamentallösung ist.

Beweis Die dritte der vorherigen Bemerkungen — ausgewertet mit $c = X^{-1}(t_*)\, x_*$ — zeigt, dass x der homogenen Differentialgleichung genügt und die Anfangsbedingung ist via $x(t_*) = X(t_*)\, X^{-1}(t_*)\, x_* = x_*$ erfüllt. Die Eindeutigkeit ergibt sich aus dem Eindeutigkeitssatz. □

Bemerkungen

1. Die Matrix $X(t_*)\, X^{-1}(t_*)$ wird oftmals Übergangsmatrix genannt, denn sie beschreibt, wie $x(t)$ aus $x(t_*)$ berechnet werden kann. Das nichtlineare Analogon wird *Fluss der Differentialgleichung* genannt und später eingeführt.

2. Sind $X(t)$ und $\tilde{X}(t)$ zwei Fundamentalmatrizen, so implizieren das Korollar sowie der Eindeutigkeitssatz die Formel

   $$X(t)\, X^{-1}(t_*) = \widetilde{X}(t)\, \widetilde{X}^{-1}(t_*)$$

   und damit

   $$X(t) = \widetilde{X}(t)\, \widetilde{Y}\,, \qquad \widetilde{X}(t) = X(t)\, Y\,,$$

   wobei die Matrizen

   $$\widetilde{Y} = \widetilde{X}^{-1}(t_*)\, X(t_*)\,, \qquad Y = X^{-1}(t_*)\, \widetilde{X}(t_*)$$

   invers zueinander sind und nicht von t abhängen.[16] Oder anders gesagt: Zwei verschiedene Fundamentalmatrizen zur selben Differentialgleichung unterscheiden sich nur durch die rechtsseitige Multiplikation mit einer konstanten Matrix.

Dynamik von Fundamentalmatrizen Es gilt

$$\dot{X}(t) = A(t)X(t)$$

für alle Zeiten $t \in \mathbb{R}$, wobei wir diese *matrixwertige Version* der Differentialgleichung direkt nachrechnen können, indem wir die Gesetze der Matrizenmultiplikation sowie die spaltenweise Gültigkeit der *vektorwertigen Version* ausnutzen. Mit etwas mehr Aufwand (siehe die Übungen) können wir sogar zeigen, dass die Wronski-Determinante der *skalaren* Differentialgleichung

$$\tfrac{\mathrm{d}}{\mathrm{d}t} \det\big(X(t)\big) = \operatorname{tr}\big(A(t)\big)\, \det\big(X(t)\big)$$

[16] $\widetilde{Y}$ und Y hängen auch nicht von der Wahl von t_* ab, denn es gilt sogar

$$\widetilde{X}^{-1}(t)\, X(t) = \widetilde{X}^{-1}(t_*)\, X(t_*), \qquad X^{-1}(t)\, \widetilde{X}(t) = X^{-1}(t_*)\, \widetilde{X}(t_*)$$

für alle $t \in \mathbb{R}$ und alle $t_* \in \mathbb{R}$.

genügt[17] und daher durch

$$\det\big(X(t)\big) = \exp\left(\int_{t_*}^{t} \operatorname{tr}\big(A(\tau)\big)\,\mathrm{d}\tau\right) \det\big(X(t_*)\big)$$

gegeben ist.

Bemerkungen:

1. Die Formel für die Wronski-Determinante ist sehr bemerkenswert und wichtig, denn sie erlaubt es uns, die Determinante von $X(t)$ direkt aus den Anfangsdaten und ohne explizite Kenntnis von $X(t)$ zu berechnen.

2. Insbesondere gilt die folgende Aussage, die in vielen Bereichen der Mathematik nützlich ist: Ist $A(t)$ für alle Zeiten spurfrei (d.h. $\operatorname{tr}\big(A(t)\big)$ für alle $t \in \mathbb{R}$), so gilt $\det\big(X(t)\big) = \det\big(X(t_*)\big)$ für alle $t \in \mathbb{R}$ und die Wronski-Determinante ändert sich nicht mit der Zeit.

inhomogene Gleichung

Vorbemerkung Wir studieren nun die Lösungen der inhomogenen Gleichung und werden insbesondere sehen, dass die Methode *Variation der Konstanten* nicht nur für $n = 1$, sondern auch in höheren Dimensionen angewendet werden kann.

Lemma (**Charakterisierung des Lösungsraumes**) Ist $x_{\mathrm{part}} : \mathbb{R} \to \mathbb{R}^n$ eine gegebene inhomogene Lösung, so kann die allgemeine inhomogene Lösung als

$$x(t) = x_{\mathrm{hom}}(t) + x_{\mathrm{part}}(t)$$

geschrieben werden, wobei

$$x_{\mathrm{hom}}(t) = X(t)\,c$$

die allgemeine homogene Lösung mit freien Konstanten $c \in \mathbb{R}^n$ bezeichnet.

Beweis Da x_{part} nach Voraussetzung die inhomogene Gleichung erfüllt, können wir durch Nachrechnen leicht zeigen, dass x genau dann eine weitere inhomogene Lösung ist, wenn $x - x_{\mathrm{part}}$ der homogenen Gleichung genügt. Die Behauptung ergibt sich nun unmittelbar. □

Bemerkungen

1. x_{part} wird partikuläre Lösung genannt,[18] wobei wir diese durch geschicktes Raten, geeignete Ansätze oder die Berechnung gewisser Integrale (wie gleich beschrieben) ermitteln können. Man schreibt im Lemma oftmals auch $x_{\mathrm{inh}}(t)$ statt $x(t)$.

2. Wir können die Quintessenz des Lemmas auch so formulieren: Die Differenz zweier inhomogener Lösungen ist eine homogene Lösung. Insbesondere ist die Menge aller Lösungen der inhomogenen Gleichung kein linearer Vektorraum, sondern ein affiner Raum.

[17] Die Spur (englisch *trace*) der Matrix $A(t)$ ist

$$\operatorname{tr}\big(A(t)\big) = A_{11}(t) + \ \ldots\ + A_{nn}(t)\,,$$

d.h. die Summe ihrer Diagonaleinträge.

[18] Alternativ können wir auch von einer *speziellen Lösung* reden.

Theorem (**Hauptsatz über linear inhomogene Gleichungen**) Die allgemeine Lösung der inhomogenen Gleichung kann als

$$x(t) = X(t)\,c + \int_{t_*}^{t} X(t)\,X^{-1}(\sigma)\,b(\sigma)\,\mathrm{d}\sigma = X(t)\left(c + \int_{t_*}^{t} X^{-1}(\sigma)\,b(\sigma)\,\mathrm{d}\sigma\right)$$

geschrieben werden, wobei t_* eine beliebige Zeit ist (zum Beispiel die Anfangszeit) und das Integral über den vektorwertigen Integranden komponentenweise berechnet wird.

Beweis Variante 1: Wir zeigen, dass

$$x_{\mathrm{part}}(t) := X(t) \int_{t_*}^{t} X^{-1}(\sigma)\,b(\sigma)\,\mathrm{d}\sigma$$

eine Lösung der inhomogenen Gleichung ist, denn dann folgt die Behauptung aus dem vorangegangenen Lemma. Differentiation nach t liefert

$$\dot{x}_{\mathrm{part}}(t) \;=\; \tfrac{\mathrm{d}}{\mathrm{d}t} x_{\mathrm{part}}(t) \;=\; \dot{X}(t) \int_{t_*}^{t} X^{-1}(\sigma)\,b(\sigma)\,\mathrm{d}\sigma \;+\; X(t)\left(X^{-1}(t)\,b(t)\right),$$

wobei wir den Hauptsatz der Differential- und Integralrechnung aus *Analysis 1* komponentenweise angewendet haben. Der zweite Term auf der rechten Seite ist offensichtlich $b(t)$ und weil $X(t)$ der matrixwertige Version der Differentialgleichung genügt (siehe oben), erhalten wir schließlich

$$\dot{x}_{\mathrm{part}}(t) \;=\; A(t)\,X(t) \int_{t_*}^{t} X^{-1}(\sigma)\,b(\sigma)\,\mathrm{d}\sigma \;+\; b(t) \;=\; A(t)\,x_{\mathrm{part}}(t) \;+\; b(t)$$

und damit das gewünschte Ergebnis.

Variante 2: Wie im eindimensionalen Fall benutzen wir *Variation der Konstanten*, wobei der Ansatz diesmal (d.h. für $n > 1$) als

$$x(t) = X(t)\,c(t)$$

geschrieben wird und die Reihenfolge der Faktoren wieder wichtig ist, da die Matrizenmultiplikation nicht kommutativ ist. Differentiation nach t liefert

$$\dot{x}(t) = \dot{X}(t)\,c(t) + X(t)\,\dot{c}(t) = A(t)\,X(t)\,c(t) + X(t)\,\dot{c}(t) = A(t)\,x(t) + X(t)\,\dot{c}(t)\,,$$

wobei wir die Produktregel komponentenweise sowie die Differentialgleichung für $X(t)$ verwendet haben. Durch Einsetzen in die inhomogene Differentialgleichung für $x(t)$ erhalten wir

$$X(t)\,\dot{c}(t) = b(t) \qquad \text{bzw.} \qquad \dot{c}(t) = X^{-1}(t)\,b(t)$$

und der Hauptsatz der Differential- und Integralrechnung garantiert

$$c(t) = c(t_*) + \int_{t_*}^{t} X^{-1}(\sigma)\,b(\sigma)\,\mathrm{d}\sigma\,.$$

Die Behauptung folgt nun nach Multiplikation mit $X(t)$ und Einsetzen, sofern wir am Ende c statt $c(t_*)$ schreiben. □

Bemerkungen

1. Man nennt die Lösungsformel aus dem Hauptsatz auch das Duhamel-Prinzip bzw. die (höherdimensionale) Variation der Konstanten. Sie ist ausgesprochen wichtig und kann alternativ mit physikalisch motivierten Argumenten abgeleitet werden, wobei dann mit dem Superpositionsprinzip für die homogene Gleichung argumentiert wird.

2. Im eindimensionalen Fall $n = 1$ sind alle Vektoren und Matrizen reelle Zahlen und die entsprechende Formel aus dem Theorem hatten wir — mit $a(t)$ statt $A(t)$ — schon im Abschnitt über explizite Lösungsmethoden für skalare Gleichungen kennengelernt.

Auswertung der Lösungsformel Bei einer Anwendung des Theorems müssen für $n > 1$ im Allgemeinen die folgenden Teilprobleme behandelt werden:

1. Finde die allgemeine Lösung des homogenen Problems bzw. eine Fundamentalmatrix $X(t)$.

2. Berechne $X^{-1}(t)$ für jedes t, zum Beispiel durch Matrizeninversion.

3. Berechne das vektorwertige Integral in der Lösungsformel und aus diesem die entsprechende partikuläre Lösung des inhomogenen Problems.

Bei Anfangswertproblemen müssen die freien Kontanten $c \in \mathbb{R}^n$ noch als Lösung des linearen Gleichungssystems

$$x_* = x(t_*) = X(t_*)\, c$$

bestimmt werden, wobei die Invertierbarkeit der Matrix $X(t_*)$ sicherstellt, dass es genau eine Lösung gibt. Im Fall von $X(t_*) = \mathbb{1}$ vereinfacht sich dies zu $x_* = c$.

Bemerkungen:

1. Ist die homogene Gleichung autonom, so kann $X(t)$ als Matrixexponential explizit berechnet werden (siehe den nächsten Abschnitt). Im nicht-autonomen Fall gibt es aber keinen Algorithmus für die exakte Berechnung von Fundamentalmatrizen und man muss sich oftmals mit numerischen Approximationen begnügen.

2. Generell gilt: Beim Lösen linearer Differentialgleichungen mit $n > 1$ können die Rechnungen sehr schnell sehr kompliziert werden. Daher werden in den Natur- und Ingenieurwissenschaften oftmals auch andere Lösungsverfahren eingesetzt, die zumindest für gewisse Klassen von Differentialgleichungen schneller zum Ziel führen. Prominente Beispiele sind die *Fourier-* und die *Laplace-Transformation*, mit denen Differentialgleichungen in äquivalente algebraische Formeln überführt werden können.

Beispiel Die zweite Ordnungsgleichung

$$\ddot{y}(t) + y(t) = \cos(\omega\, t)$$

beschreibt einen *harmonischen Oszillator* mit periodischer Anregung der Frequenz ω. Sie kann via $x_1(t) = y(t)$ und $x_2(t) = \dot{y}(t)$ als planare inhomogene Gleichung mit

$$A = \begin{pmatrix} 0 & +1 \\ -1 & 0 \end{pmatrix}, \qquad b(t) = \begin{pmatrix} 0 \\ \cos(\omega\, t) \end{pmatrix}$$

geschrieben werden, und wir hatten schon gesehen, dass

$$X(t) = \begin{pmatrix} +\cos(t) & +\sin(t) \\ -\sin(t) & +\cos(t) \end{pmatrix} \qquad \text{mit} \qquad X^{-1}(t) = \begin{pmatrix} +\cos(t) & -\sin(t) \\ +\sin(t) & +\cos(t) \end{pmatrix}$$

eine mögliche Wahl der Fundamentalmatrix ist. Um die Lösungsformel auszuwerten, bestimmen wir zunächst das vektorwertige Integral, wobei wir $t_* = 0$ setzen und $\omega \neq \pm 1$ voraussetzen wollen. Wir erhalten

$$\begin{aligned} \int_0^t X^{-1}(\sigma)\, b(\sigma)\, \mathrm{d}\sigma &= \int_0^t \begin{pmatrix} -\sin(\sigma)\, \cos(\omega\,\sigma) \\ +\cos(\sigma)\, \cos(\omega\,\sigma) \end{pmatrix} \mathrm{d}\sigma = \begin{pmatrix} -\displaystyle\int_0^t \sin(\sigma)\, \cos(\omega\,\sigma)\, \mathrm{d}\sigma \\ +\displaystyle\int_0^t \cos(\sigma)\, \cos(\omega\,\sigma)\, \mathrm{d}\sigma \end{pmatrix} \\ &= \frac{1}{\omega^2 - 1} \begin{pmatrix} -\cos(t)\cos(\omega\, t) - \omega\, \sin(t)\sin(\omega\, t) + 1 \\ -\sin(t)\cos(\omega\, t) + \omega\, \cos(t)\, \sin(\omega\, t) \end{pmatrix} \end{aligned}$$

durch Berechnung der beiden skalaren Integrale (zum Beispiel mit zweifacher partieller Integration) und damit

$$x_{\mathrm{part}}(t) = X(t) \int_0^t X^{-1}(\sigma)\, b(\sigma)\, \mathrm{d}\sigma = \frac{1}{\omega^2 - 1} \begin{pmatrix} +\cos(t) - \cos(\omega\, t) \\ -\sin(t) + \omega\, \sin(\omega\, t) \end{pmatrix}$$

als partikuläre Lösung, für die nach Konstruktion $x_{\mathrm{part}}(0) = 0$ gilt. Die allgemeine inhomogene Lösung $x(t)$ entsteht, wenn wir zur partikulären Lösung die allgemeine homogene Lösung $X(t)\, c$ addieren, und insgesamt ergibt sich

$$y(t) = x_1(t) = \left(c_1 + \tfrac{1}{\omega^2 - 1}\right) \cos(t) + c_2\, \sin(t) - \tfrac{1}{\omega^2 - 1}\, \cos(\omega\, t)$$

sowie

$$\dot{y}(t) = x_2(t) = c_2\, \cos(t) - \left(c_1 + \tfrac{1}{\omega^2 - 1}\right) \sin(t) + \tfrac{\omega}{\omega^2 - 1}\, \sin(\omega\, t)\,,$$

wobei wir hier die Reihenfolge von Faktoren beliebig vertauschen dürfen, da alle Terme skalar sind. Die Konstanten c_1 und c_2 können durch die Vorgabe von zwei skalaren Anfangswerten bestimmt werden, wobei hier $c_1 = x_1(0) = y(0)$ und $c_2 = x_2(0) = \dot{y}(0)$ gilt.

Bemerkung: Unsere Formeln gelten offensichtlich nicht für $\omega = \pm 1$. In diesem Fall liefert die Berechnung der beiden skalaren Integrale andere Ausdrücke, nämlich

$$x_{\mathrm{part}}(t) = \tfrac{1}{2} \begin{pmatrix} t\, \sin(t) \\ t\, \cos(t) + \sin(t) \end{pmatrix}$$

sowie

$$y(t) = c_1 \cos(t) + c_2 \sin(t) + \tfrac{1}{2}\, t \sin(t)\,.$$

Das unterschiedliche Verhalten für $\omega^2 \neq 1$ und $\omega^2 = 1$ ist aus physikalischer Sicht mit *Resonanz* verbunden, denn die *Eigenfrequenzen* der Gleichung für $y(t)$ sind gerade ± 1 (äquivalent dazu ist, dass $\pm \mathrm{i}$ die Eigenwerte von A sind). Wird der harmonische Oszillator mit einer anderen Frequenz angeregt, so bleibt die Lösung $y(t)$ für alle Zeiten beschränkt. Bei $\omega = \pm 1$ ist das aber wegen des polynomiellen Vorfaktors anders und es tritt die sogenannte *Resonanzkatastrophe* ein.

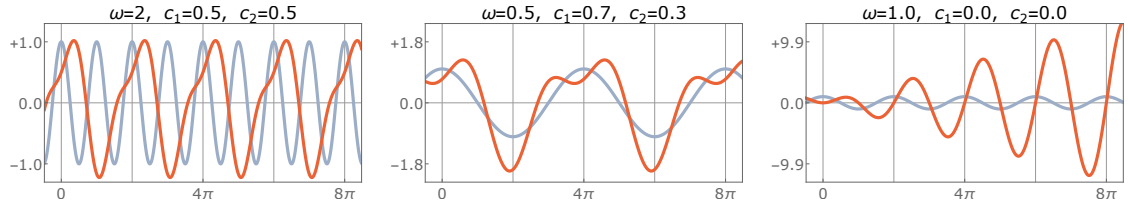

Abbildung Drei Lösungen des harmonischen Oszillators, wobei Rot bzw. Blau der Lösung bzw. der Anregung entspricht und rechts der resonante Fall dargestellt ist.

3.6 autonom homogene Differentialgleichungen

Vorbemerkung Das denkbar einfachste Anfangswertproblem in einer Dimension (also für $n = 1$) lautet

$$\dot{x}(t) = A\, x(t) \qquad x(t_*) = x_*$$

und besitzt die Lösung

$$x(t) = \exp\big((t - t_*)\, A\big)\, x_*\,,$$

wobei A, x_* und $x(t)$ jeweils reelle Zahlen sind. In diesem Abschnitt zeigen wir, dass eine analoge Formel auch für $n > 1$ gilt, d.h. wenn A eine zeit-unabhängige quadratische Matrix ist. Insbesondere werden wir zeigen, wie entsprechende Fundamentalmatrizen explizit berechnet werden können.

Matrixexponential

Theorem (**Existenz des Matrixexponentials**) Für jede quadratische Matrix $M \in \mathbb{M}^{n\times n}$ ist

$$\exp(M) = \sum_{k=0}^{\infty} \frac{M^k}{k!} = \mathsf{1} + M + \tfrac{1}{2}\, M^2 + \tfrac{1}{6}\, M^3 + \tfrac{1}{24}\, M^4 + \ldots$$

als absolut konvergente Reihe wohldefiniert,[19] wobei $M^0 = \mathsf{1}$ die Einheitsmatrix in $\mathbb{M}^{n\times n}$ ist und wie immer $0! = 1! = 1$ vereinbart sei. Außerdem gilt die Implikation

$$M = Q\, \widetilde{M}\, Q^{-1} \qquad \Longrightarrow \qquad \exp(M) = Q\, \exp\big(\widetilde{M}\big)\, Q^{-1}\,,$$

d.h. Basiswechsel und Berechnung des Matrixexponentials können vertauscht werden.

[19] Auch bei Matrizen dürfen wir e^M statt $\exp(M)$ schreiben.

Beweis *Teil 1*: Die Partialsummenfolge $(S_m)_{m\in\mathbb{N}} \subset \mathbb{M}^{n\times n}$ mit $S_m := \sum_{k=0}^{m} M^k/k!$ erfüllt

$$\big|S_l - S_m\big| = \left|\sum_{k=m+1}^{l} \frac{M^k}{k!}\right| \leq \sum_{k=m+1}^{l} \frac{\big|M^k\big|}{k!} \leq \sum_{k=m+1}^{l} \frac{\big|M\big|^k}{k!}$$

für alle $l > m$, wobei wir die Dreiecksungleichung sowie den Kompatibilitätssatz für die euklidische Norm von Matrizen verwendet haben. Die Konvergenz der skalaren Exponentialreihe (siehe *Analysis 1*) impliziert, dass

$$s_m := \sum_{k=0}^{m} \frac{|M|^k}{k!} \quad \xrightarrow{m\to\infty} \quad s_\infty := \sum_{k=0}^{\infty} \frac{|M|^k}{k!} = \exp\big(\,|M|\,\big)$$

im Sinne konvergenter und monoton wachsender Zahlenfolgen gilt. Wir erhalten damit zum einen

$$0 \leq \big|S_l - S_m\big| \leq s_l - s_m \leq s_\infty - s_m \quad \xrightarrow{m\to\infty} \quad 0$$

und schließen, dass $(S_m)_{m\in\mathbb{N}}$ eine Cauchy-Folge ist und daher wegen der Vollständigkeit von $\mathbb{M}^{n\times n}$ einen Grenzwert besitzt. Zum anderen ist die Konvergenz von $(s_m)_{m\in\mathbb{N}}$ per Definition äquivalent zur absoluten Konvergenz der matrixwertigen Exponentialreihe.

Teil 2: Unter der Annahme $M = Q\,\widetilde{M}\,Q^{-1}$ berechnen wir

$$M^2 = \big(Q\,\widetilde{M}\,Q^{-1}\big)\,\big(Q\,\widetilde{M}\,Q^{-1}\big) = Q\,\widetilde{M}\,\big(Q^{-1}\,Q\big)\,\widetilde{M}\,Q^{-1} = Q\,\widetilde{M}^2\,Q^{-1}$$

mit den Rechenregeln der Matrizenmultiplikation und durch vollständige Induktion über k zeigen wir $M^k = Q\,\widetilde{M}^k\,Q^{-1}$ für alle $k \in \mathbb{N}$. Dies impliziert $S_m = Q\,\widetilde{S}_m\,Q^{-1}$ für alle Partialsummen und die zweite Behauptung ergibt sich im Limes $m \to \infty$. □

Bemerkungen

1. Vereinfacht gesprochen kann die Beweisidee wie folgt zusammengefasst werden: Die matrixwertige Exponentialreihe konvergiert absolut, weil ihr reelles Analogon absolut konvergiert und weil $\mathbb{M}^{n\times n}$ vollständig ist.

2. Das Matrixexponential kann ganz analog für komplexwertige $n\times n$-Matrizen eingeführt werden. Beachte aber, dass es für nicht-quadratische Matrizen keine analoge Konstruktion gibt.

3. Matrixexponentiale können auch ohne unendliche Summenbildung berechnet werden, indem geeignete Basiswechsel und bekannte Formeln für Elementarbausteine miteinander kombiniert werden. Siehe dazu die Beispiele.

4. Mit den Methoden aus dem Beweis kann auch die Gültigkeit der nützlichen Formeln

$$\exp(M)\,M = M\exp(M)\,, \qquad \big|\exp(M)\big| \leq \exp\big(\,|M|\,\big)$$

gezeigt werden (Übungsaufgabe).

Beispiele

1. Für Diagonalmatrizen gilt

$$M = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix}, \quad M^2 = \begin{pmatrix} \mu_1^2 & 0 \\ 0 & \mu_2^2 \end{pmatrix}, \quad M^3 = \begin{pmatrix} \mu_1^3 & 0 \\ 0 & \mu_2^3 \end{pmatrix}, \qquad \dots$$

und damit

$$\exp(M) = \begin{pmatrix} \exp(\mu_1) & 0 \\ 0 & \exp(\mu_2) \end{pmatrix}.$$

2. Die Matrix

$$M = \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$$

kann mittels

$$\widetilde{M} = \begin{pmatrix} -2 & 0 \\ 0 & +4 \end{pmatrix}, \qquad Q = \frac{1}{\sqrt{2}} \begin{pmatrix} +1 & +1 \\ -1 & +1 \end{pmatrix}, \qquad Q^{-1} = \frac{1}{\sqrt{2}} \begin{pmatrix} +1 & -1 \\ +1 & +1 \end{pmatrix}$$

diagonalisiert werden, d.h. es gilt $M = Q\,\widetilde{M}\,Q^{-1}$ bzw. $\widetilde{M} = Q^{-1}\,M\,Q$. Daher können wir das gesuchte Matrixexponential via

$$\exp(M) = Q\,\exp\bigl(\widetilde{M}\bigr)\,Q^{-1} = \frac{1}{2} \begin{pmatrix} \mathsf{e}^{+4} + \mathsf{e}^{-2} & \mathsf{e}^{+4} - \mathsf{e}^{-2} \\ \mathsf{e}^{+4} - \mathsf{e}^{-2} & \mathsf{e}^{+4} + \mathsf{e}^{-2} \end{pmatrix}$$

berechnen.

3. Die verallgemeinerte Drehmatrix

$$M = \begin{pmatrix} +\mu & -\nu \\ +\nu & +\mu \end{pmatrix}$$

besitzt die beiden konjugiert komplexen Eigenwerte $\mu \pm \mathtt{i}\,\nu$ und diagonalisiert in $\mathbb{C}$ via

$$M = Q \begin{pmatrix} \mu - \mathtt{i}\,\nu & 0 \\ 0 & \mu + \mathtt{i}\,\nu \end{pmatrix} Q^{-1}, \qquad Q = \frac{1}{\sqrt{2}} \begin{pmatrix} +\mathtt{i} & +\mathtt{i} \\ +1 & -1 \end{pmatrix}, \qquad Q^{-1} = \overline{Q}^T.$$

Wir erhalten daher

$$\exp(M) = Q \begin{pmatrix} \mathsf{e}^{\mu - \mathtt{i}\,\nu} & 0 \\ 0 & \mathsf{e}^{\mu + \mathtt{i}\,\nu} \end{pmatrix} Q^{-1} = \mathsf{e}^{\mu} \begin{pmatrix} +\cos(\nu) & -\sin(\beta) \\ +\sin(\nu) & +\cos(\nu) \end{pmatrix},$$

nach elementaren Rechnungen und unter Verwendung der Euler-Formel.

4. Für die verallgemeinerte Jordan-Matrix

$$M = \begin{pmatrix} \mu & \eta \\ 0 & \mu \end{pmatrix}$$

verifizieren wir

$$M^2 = \begin{pmatrix} \mu^2 & 2\,\mu\,\eta \\ 0 & \mu^2 \end{pmatrix}, \quad M^3 = \begin{pmatrix} \mu^3 & 3\,\mu^2\,\eta \\ 0 & \mu^3 \end{pmatrix}, \quad \dots, \quad M^k = \begin{pmatrix} \mu^k & k\,\mu^{k-1}\,\eta \\ 0 & \mu^k \end{pmatrix}, \quad \dots$$

durch vollständige Induktion über k und

$$\exp(M) = \mathsf{e}^{\mu} \begin{pmatrix} 1 & \eta \\ 0 & 1 \end{pmatrix}$$

folgt nach Vergleich mit der reellen Exponentialreihe.

zur Summenformel Für Matrizen gilt im Allgemeinen

$$\exp\left(M+N\right) \neq \exp\left(M\right)\,\exp\left(N\right),$$

obwohl natürlich $\exp\left(m+n\right) = \exp\left(m\right)\,\exp\left(n\right)$ für reelle Zahlen m, n erfüllt ist. Der Grund ist, dass die Multiplikation von Matrizen — im Gegensatz zur Multiplikation von Zahlen — nicht kommutativ ist, und das Ungleichheitszeichen kann heuristisch wie folgt verstanden werden: Zum einen berechnen wir

$$\begin{aligned}\exp\left(M+N\right) &= \mathsf{1} + \left(M+N\right) + \tfrac{1}{2}\left(M+N\right)\left(M+N\right) + \ldots \\ &= \mathsf{1} + M + N + \tfrac{1}{2}\,M^2 + \tfrac{1}{2}\,MN + \tfrac{1}{2}\,NM + \tfrac{1}{2}\,N^2 + \ldots\end{aligned}$$

und erhalten zum anderen mithilfe des Cauchy-Produktes für Reihen die Formel

$$\begin{aligned}\exp\left(M\right)\,\exp\left(N\right) &= \left(\mathsf{1} + M + \tfrac{1}{2}\,M^2 + \ldots\right)\left(\mathsf{1} + N + \tfrac{1}{2}\,N^2 + \ldots\right) \\ &= \mathsf{1} + M + N + \tfrac{1}{2}\,M^2 + MN + \tfrac{1}{2}\,N^2 + \ldots\,.\end{aligned}$$

Im Allgemeinen gilt aber $MN \neq NM$, d.h. die gemischten quadratischen Terme in den beiden letzten Formeln sind *nicht identisch*. Analoges gilt für alle gemischten kubischen, quartischen und quintischen Terme.

Gegenbeispiel: Die Matrizen

$$M = \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}, \quad N = \begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}$$

kommutieren nicht (es gilt $MN = N \neq 0 = NM$) und die Formeln

$$\exp\left(M\right)\,\exp\left(N\right) = \begin{pmatrix}\mathsf{e} & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix} = \begin{pmatrix}\mathsf{e} & \mathsf{e}\\ 0 & 1\end{pmatrix} \neq \begin{pmatrix}\mathsf{e} & \mathsf{e}-1\\ 0 & 1\end{pmatrix} = \exp\left(M+N\right)$$

ergeben sich nach direkten Rechnungen.

Lemma (**bedingte Summenformel**) Die Formel

$$\exp\left(M\right)\,\exp\left(N\right) = \exp\left(M+N\right)$$

ist erfüllt, wenn M und N kommutieren, d.h. sofern $MN = NM$ gilt.

Beweis Die wesentliche Beobachtung ist, dass für zwei kommutierende Matrizen die verallgemeinerte binomische Formel

$$\left(M+N\right)^k = \sum_{l=0}^{k}\binom{k}{l}M^{k-l}\,N^l = \sum_{l=0}^{k}\frac{k!}{l!\,(k-l)!}\,M^l N^{k-l}$$

gilt, wobei diese leicht mit vollständiger Induktion über k bewiesen werden kann. Mithilfe des Cauchy-Produktes für absolut konvergente Reihen[20] ergibt sich

$$\left(\sum_{k=0}^{\infty}\frac{M^k}{k!}\right)\left(\sum_{l=0}^{\infty}\frac{N^l}{l!}\right) = \sum_{k=0}^{\infty}\sum_{l=0}^{k}\frac{M^l N^{k-l}}{l!\,(k-l)!} = \sum_{k=0}^{\infty}\frac{\left(M+N\right)^k}{k!}$$

und damit die Behauptung. □

[20] Wir hatten in *Analysis 1* das Cauchy-Produkt nur für absolut konvergente Zahlenreihen formuliert und bewiesen, aber alle Formeln und Beweisschritte können problemlos auf unendliche Summen von Matrizen übertragen werden.

Korollar (**Spezialfälle der Summenformel**) Für jede Matrix $M \in \mathbb{M}^{n\times n}$ und alle reellen Zahlen s, $\tilde{s}$ gilt

$$\exp\big(s\,M + \tilde{s}\,M\big) = \exp\big(s\,M\big)\,\exp\big(\tilde{s}\,M\big)\,, \qquad \exp\big(s\,M\big)\,\exp\big(-s\,M\big) = \mathsf{1}\,.$$

Insbesondere ist $\exp(s\,M)$ immer eine invertierbare Matrix und kommutiert mit $\exp(\tilde{s}\,M)$.

Beweis Die erste Formel ist eine Konzequenz des Lemmas, da $s\,M$ und $\tilde{s}\,M$ miteinander kommutieren. Die zweite Formel ergibt sich dann mit $\tilde{s} = -s$ und weil das Exponential der Nullmatrix die Einheitsmatrix ist. □

Blockdiagonalmatrizen Die elementaren Rechenregeln für Matrizen garantieren, dass Produkte von Blockdiagonalmatrizen selbst Blockdiagonalmatrizen sind, wobei die Multiplikation blockweise berechnet werden kann. Hieraus ergibt sich die Implikation

$$M = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} \quad\Longrightarrow\quad M^k = \begin{pmatrix} M_1^k & 0 \\ 0 & M_2^k \end{pmatrix}, \quad \exp(M) = \begin{pmatrix} \exp(M_1) & 0 \\ 0 & \exp(M_2) \end{pmatrix}$$

durch vollständige Induktion über k bzw. Grenzwertbildung (Übungsaufgabe), wobei M_1, M_2 beide quadratisch sind und 0 jeweils eine Rechtecksmatrix mit verschwindenden Einträgen bezeichnet. Analoge Aussagen gelten, wenn M aus mehr als zwei diagonalen Blöcken besteht.

Beispiel: Es gilt

$$\exp\begin{pmatrix} \mu_1 & 0 & 0 & 0 \\ 0 & \mu_2 & 0 & 0 \\ 0 & 0 & \mu_3 & \eta \\ 0 & 0 & 0 & \mu_3 \end{pmatrix} = \begin{pmatrix} \mathsf{e}^{\mu_1} & 0 & 0 & 0 \\ 0 & \mathsf{e}^{\mu_1} & 0 & 0 \\ 0 & 0 & \mathsf{e}^{\mu_3} & \eta\,\mathsf{e}^{\mu_3} \\ 0 & 0 & 0 & \mathsf{e}^{\mu_3} \end{pmatrix},$$

wobei wir die obigen Teilergebnisse für zwei 2×2-Blöcke zusammengesetzt haben.

nilpotente Matrizen Gilt $M^m = 0$ für einen Exponenten $m \in \mathbb{N}$, so reduziert sich das Matrixexponential via

$$\exp(M) = \mathsf{1} + M + \tfrac{1}{2}\,M^2 + \ldots + \frac{M^{m-1}}{(m-1)!}$$

auf eine endliche Summe und kann ohne Grenzwertbildung berechnet werden.

Beispiel: Für Vielfache von Jordan-Matrizen zum Eigenwert 0 erhalten wir

$$\exp\left(s\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + s\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$$

sowie

$$\exp\left(s\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + s\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} + \tfrac{1}{2}\,s^2\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & s & \frac{1}{2}\,s^2 \\ 0 & 1 & s \\ 0 & 0 & 1 \end{pmatrix}$$

und

$$\exp\left(s\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}\right) = \ldots = \begin{pmatrix} 1 & s & \frac{1}{2}\,s^2 & \frac{1}{6}\,s^3 \\ 0 & 1 & s & \frac{1}{2}\,s^2 \\ 0 & 0 & 1 & s \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Matrixexponential und Differentialgleichungen

Theorem (**fundamentale Eigenschaft des Matrixexponentials**) Für jedes $A \in \mathbb{M}^{n\times n}$ und jedes $t_* \in \mathbb{R}$ ist die durch

$$X(t) := \exp\big((t-t_*)\,A\big)$$

definierte parametrisierte Kurve $X : \mathbb{R} \to \mathbb{M}^{n\times n}$ stetig differenzierbar mit

$$\dot X(t) = X(t)\,A = A\,X(t) \qquad \text{für alle} \quad t \in \mathbb{R}$$

sowie $X(t_*) = \mathsf{1}$.

Beweis Aus der Summenformel ergibt sich

$$\begin{aligned}\frac{X(t+h)-X(t)}{h} &= \frac{\exp\big((t-t_*)\,A\big)\,\exp(h\,A) - \exp\big((t-t_*)\,A\big)}{h}\\ &= \exp\big((t-t_*)\,A\big)\big(A + R(h,\,t)\big) = X(t)\,\big(A + R(h,\,t)\big)\,,\end{aligned}$$

wobei der Fehlerterm

$$R(h,\,t) := \frac{\exp(h\,A)-\mathsf{1}}{h} - A = \tfrac{1}{2}\,h\,A^2 + \tfrac{1}{6}\,h^2\,A^3 + \tfrac{1}{24}\,h^3\,A^4 + \ldots = \sum_{k=0}^{\infty} \frac{h^{k+1}\,A^{k+2}}{(k+2)!}$$

für jedes t und jedes $h \neq 0$ eine absolut konvergente Reihe in $\mathbb{M}^{n\times n}$ darstellt, deren Betrag wir mit der Dreiecksungleichung, dem Kompatibilitätssatz und wegen $1/(k+2)! < 1/k!$ durch

$$\big|R(h,\,t)\big| \le \sum_{k=0}^{\infty} \frac{|h|^{k+1}\,|A|^{k+2}}{(k+2)!} \le |h|\,|A|^2 \sum_{k=0}^{\infty} \frac{|h|^k\,|A|^k}{k!} = |h|\,|A|^2\,\exp\big(\,|h|\,|A|\,\big)$$

abschätzen können. Insbesondere gilt

$$R(h,\,t) \quad \xrightarrow{h\to\infty} \quad 0\,, \qquad \frac{X(t+h)-X(t)}{h} \quad \xrightarrow{h\to\infty} \quad X(t)\,A$$

und dies liefert die Differenzierbarkeit von X sowie die erste Formel für $\dot X(t)$. Die zweite ergibt sich analog, da auch

$$\frac{X(h+t)-X(t)}{h} = \big(A + R(h,\,t)\big)\,X(t)$$

erfüllt ist. Schließlich gilt $X(t_*) = \exp(0) = \mathsf{1}$ per Definition. □

Korollar (**Lösung des autonomen und homogenen Anfangswertproblems**) Seien $A \in \mathbb{M}^{n\times n}$ sowie $t_* \in \mathbb{R}$ und $x_* \in \mathbb{R}^n$ beliebig fixiert. Dann ist die eindeutige maximale Lösung des Anfangswertproblems

$$\dot x(t) = A\,x(t)\,, \qquad x(t_*) = x_*$$

für alle Zeiten durch

$$x(t) = \exp\big((t-t_*)\,A\big)\,x_*$$

gegeben.

Beweis Mit dem Theorem kann einfach nachgerechnet werden, dass die angegebene Kurve $x : \mathbb{R} \to \mathbb{R}^n$ in der Tat eine Lösung darstellt und die Eindeutigkeit wird durch den Satz von Picard-Lindelöf sichergestellt. □

Bemerkungen

1. Alle Formeln gelten auch wieder im Komplexen. Beachte aber, dass die Reihenfolge der Terme in der Lösungsformel für $n > 1$ wichtig ist, da wir mit Zahlen, Vektoren und Matrizen operieren!

2. Das Theorem und das Korollar implizieren, dass

$$X(t) = \exp\big((t - t_*)\, A\big)$$

gerade die durch $X(t_*) = \mathsf{1}$ ausgezeichnete Fundamentalmatrix der autonomen und homogenen Differentialgleichung ist. Für zeitabhängige Matrizen gilt im Allgemeinen

$$X(t) \neq \exp\left(\int_{t*}^{t} A(\tau)\, \mathrm{d}\tau\right).$$

Das Matrixexponential auf der rechten Seite ist zwar immer noch wohldefiniert, liefert aber keine Lösungen der Differentialgleichung.

*Verallgemeinerung**: Wir können die Fundamentalmatrix einer homogenen, aber nicht-autonomen Differentialgleichung nur dann mit dem Matrixexponential berechnen, wenn die *sehr restriktive* Bedingung

$$A(t_1)\, A(t_2) = A(t_2)\, A(t_1) \qquad \text{für alle} \quad t_1, t_2 \in \mathbb{R}$$

erfüllt ist. Diese gilt immer im autonomen Fall (wegen $A(t_1) = A(t_2) = A$), aber nur selten im nicht-autonomen Fall.

Auswertung der Lösungsformel Das Theorem über die Jordansche Normalform (siehe die Vorlesung *Lineare Algebra*) besagt, dass für jedes $A \in \mathbb{M}^{n\times n}$ eine komplexwertige, aber invertierbare $n\times n$-Matrix Q existiert, sodass zum einen

$$A = Q\, \widetilde{A}\, Q^{-1}$$

gilt und zum anderen $\widetilde{A}$ eine Blockdiagonalmatrix ist, die nur aus Diagonal- und Jordan-Blöcken besteht. Die Berechnung von $\exp\big((t - t_*)\, \widetilde{A}\big)$ ist vergleichsweise einfach (siehe die Beispiele oben und unten) und liefert via

$$\exp\big((t - t_*)\, A\big) = Q\, \exp\big((t - t_*)\, \widetilde{A}\big)\, Q^{-1},$$

die Fundamentalmatrix, die zu A gehört.

Merkregel Matrixexponentiale werden in aller Regel dadurch bestimmt, dass das Exponential der entsprechenden Jodanschen Normalform blockweise berechnet wird und das Ergebnis anschließend mit der Basiswechselmatrix Q konjugiert wird.

Bemerkungen

1. Wir können den Basiswechsel auch wie folgt verstehen: Mit den Identifikationen

$$\tilde{x}(t) = Q^{-1}\, x(t)\,, \qquad x(t) = Q\, \tilde{x}(t)$$

ergibt sich die Äquivalenz

$$\dot{x}(t) = A\, x(t) \qquad \Longleftrightarrow \qquad \dot{\tilde{x}}(t) = \widetilde{A}\, \tilde{x}(t)\,,$$

d.h. die Lösungen der Differentialgleichungen mit A und $\widetilde{A}$ können sehr einfach mithilfe zeit-konstanter Koordinatenwechsel ineinander umgerechnet werden. Analog sind die jeweiligen Anfangsdaten durch $\tilde{x}_* = Q^{-1}\, x_*$ und $x_* = Q\, \tilde{x}_*$ gekoppelt.

2. Die Diagonaleinträge von $\widetilde{A}$ sind gerade die komplexen Eigenwerte von A und die Spalten von Q bilden eine entsprechende Basis von Eigenvektoren (wobei diese für Jordan-Eigenwerte ggf. in einem verallgemeinerten oder zyklischen Sinne zu verstehen sind). Die explizite Kenntnis der Eigenwerte von A ist auch für Stabilitätsuntersuchungen sehr wichtig. Siehe dazu den nächsten Abschnitt.

3. Wenn A reell bzw. komplex diagonalisierbar ist, so kann Q so gewählt werden, dass $Q^{-1} = Q^T$ bzw. $Q^{-1} = \overline{Q}^T$ gilt.

4. Wenn wir neben Diagonal- und Jordan-Matrizen auch verallgemeinerte Drehmatrizen als Blöcke in $\widetilde{A}$ zulassen, so können wir rein reell rechnen.

5. Das Matrixexponential eines Jordan-Blocks zum Eigenwert λ kann mit einem Zerlegungstrick berechnet werden. Bei einem Dreierblock gilt zum Beispiel

$$(t-t_*) \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} = \lambda\,(t-t_*) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + (t-t_*) \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

und damit

$$\begin{aligned} \exp\left((t-t_*) \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} \right) &= \mathsf{e}^{\lambda\,(t-t_*)} \exp\left((t-t_*) \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \right) \\ &= \mathsf{e}^{\lambda\,(t-t_*)} \begin{pmatrix} 1 & (t-t_*) & \frac{1}{2}\,(t-t_*)^2 \\ 0 & 1 & (t-t_*) \\ 0 & 0 & 1 \end{pmatrix}. \end{aligned}$$

Hierbei haben wir benutzt, dass ein Vielfaches der Einheitsmatrix in $\mathbb{M}^{n\times n}$ mit allen $n\times n$-Matrizen kommutiert und dass jedes Vielfache eines Jordan-Blocks zum Eigenwert 0 nilpotent ist, sodass das entsprechende Matrixexponential als endliche Summe berechnet werden kann.

Beispiele

1. Für eine zweidimensionale Diagonalmatrix A erhalten wir

$$\exp\left((t-t_*) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \right) = \begin{pmatrix} \mathsf{e}^{\lambda_1\,(t-t_*)} & 0 \\ 0 & \mathsf{e}^{\lambda_2\,(t-t_*)} \end{pmatrix}$$

und damit

$$x(t) = \begin{pmatrix} \mathsf{e}^{\lambda_1 (t-t_*)}\, x_{*,1} \\ \mathsf{e}^{\lambda_1 (t-t_*)}\, x_{*,2} \end{pmatrix} = \mathsf{e}^{\lambda_1 (t-t_*)} \begin{pmatrix} x_{*,1} \\ 0 \end{pmatrix} + \mathsf{e}^{\lambda_2 (t-t_*)} \begin{pmatrix} 0 \\ x_{*,2} \end{pmatrix}$$

als Lösungsformel für das Anfangswertproblem. In diesem entarteten Fall (die beiden skalaren Einzelgleichungen sind entkoppelt) hätten wir das Ergebnis natürlich auch direkt und ohne Matrixexponential ableiten können.

2. Ist A die zweidimensionale Jordan-Matrix zum Eigenwert λ, so ergibt sich

$$\begin{aligned} x(t) = \exp\left((t-t_*) \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}\right) \begin{pmatrix} x_{*,1} \\ x_{*,2} \end{pmatrix} &= \begin{pmatrix} \mathsf{e}^{\lambda (t-t_*)} & \mathsf{e}^{\lambda (t-t_*)} (t-t_*) \\ 0 & \mathsf{e}^{\lambda (t-t_*)} \end{pmatrix} \begin{pmatrix} x_{*,1} \\ x_{*,2} \end{pmatrix} \\ &= \mathsf{e}^{\lambda (t-t_*)} \begin{pmatrix} x_{*,1} \\ x_{*,2} \end{pmatrix} + (t-t_*)\, \mathsf{e}^{\lambda (t-t_*)} \begin{pmatrix} 0 \\ x_{*,2} \end{pmatrix} \end{aligned}$$

mit den oben beschriebenen Argumenten.

3. Verallgemeinerte Drehmatrizen sind ein Sonderfall, denn sie können zwar nicht im Reellen, aber über $\mathbb{C}$ diagonalisiert werden und besitzen jeweils ein Paar konjugiert komplexer Eigenwerte. Direkte Rechnungen (siehe auch das analoge Beispiel im Abschnitt zum Matrixexponential) liefern die reellen Formeln

$$\exp\left((t-t_*) \begin{pmatrix} +\alpha & -\beta \\ +\beta & +\alpha \end{pmatrix}\right) = \mathsf{e}^{\alpha (t-t_*)} \begin{pmatrix} +\cos\big(\beta (t-t_*)\big) & -\sin\big(\beta (t-t_*)\big) \\ +\sin\big(\beta (t-t_*)\big) & +\cos\big(\beta (t-t_*)\big) \end{pmatrix}$$

und

$$x(t) = \mathsf{e}^{\alpha (t-t_*)} \cos\big(\beta (t-t_*)\big) \begin{pmatrix} +x_{*,1} \\ +x_{*,2} \end{pmatrix} + \mathsf{e}^{\alpha (t-t_*)} \sin\big(\beta (t-t_*)\big) \begin{pmatrix} -x_{*,2} \\ +x_{*,1} \end{pmatrix},$$

wobei die beiden zeitabhängigen Vorfaktoren den komplex-exponentiellen Termen $\exp\big(\alpha (t-t_*) \pm \mathtt{i}\, \beta (t-t_*)\big)$ entsprechen.

4. Für die Matrix

$$A = \begin{pmatrix} -3 & 0 & 0 \\ -6 & 3 & -1 \\ -1 & 1 & 1 \end{pmatrix}$$

müssen wir im ersten Schritt die Jordansche Normalform bestimmen und erhalten

$$\widetilde{A} = \begin{pmatrix} -3 & 0 & 0 \\ 0 & +2 & 1 \\ 0 & 0 & +2 \end{pmatrix}, \qquad Q = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad Q^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & -1 \end{pmatrix}.$$

Im zweiten Schritt können wir die Formel

$$\exp\big(t\, \widetilde{A}\big) = \begin{pmatrix} \mathsf{e}^{-3t} & 0 & 0 \\ 0 & \mathsf{e}^{+2t} & t\, \mathsf{e}^{+2t} \\ 0 & 0 & \mathsf{e}^{+2t} \end{pmatrix} = \mathsf{e}^{-3t} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \mathsf{e}^{+2t} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix}$$

blockweise ablesen und die Konjugation mit Q liefert nach einfachen Rechnungen

$$\exp\big(t\, A\big) = \mathsf{e}^{-3t} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \mathsf{e}^{+2t} \begin{pmatrix} 0 & 0 & 0 \\ -t-1 & t+1 & -t \\ -t & +t & -t+1 \end{pmatrix},$$

wobei wir hier der Einfachheit halber $t_* = 0$ gesetzt haben. Die Formel für $x(t)$ ergibt sich schließlich im dritten Schritt nach Multiplikation mit x_* von rechts.

Gleichungen höherer Ordnung* Wir hatten in den Übungen gesehen, dass die allgemeine Lösung der Differentialgleichung

$$y^{(n)}(t) + \alpha_{n-1}\, y^{(n-1)}(t) + \quad \ldots \quad + \alpha_1\, y^{(1)}(t) + \alpha_0\, y^{(0)}(t) = 0$$

oftmals mittels eines exponentiellen Lösungsansatzes berechnet werden kann. Um dieses Prinzip zu begründen bzw. besser zu verstehen, führen wir die neuen Variablen

$$x_1(t) = y^{(0)}(t), \quad \ldots, \quad x_n(t) = y^{(n-1)}(t)$$

ein und betrachten die Differentialgleichung für $y(t)$ als autonom homogenes System erster Ordnung bzgl. x, wobei

$$A = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 1 \\ -\alpha_0 & -\alpha_1 & -\alpha_2 & \ldots & -\alpha_{n-3} & -\alpha_{n-2} & -\alpha_{n-1} \end{pmatrix}$$

die zugrunde liegende Matrix ist. Das charakteristische Polynom

$$p(\lambda) = \det\big(A - \lambda\, 1\!\!1\big) = (-1)^n \Big(\lambda^n + \alpha_{n-1}\, \lambda^{n-1} + \ \ldots \ + \alpha_1\, \lambda + \alpha_0\Big)$$

ist bis auf den Vorfaktor $(-1)^n$ gerade das Polynom, dessen Nullstellen $\lambda_1, \ldots, \lambda_n$ wir bei der Ansatzmethode bestimmen müssen.[21] Wir sehen nun, dass die λ_j gerade die Eigenwerte von A sind.

Standardfall: Sind die λ_j alle paarweise verschieden, so ist A diagonalisierbar, und die Berechnung der allgemeinen Lösung $x(t) = \exp(t\, A)\, c$ mittels des Matrixexponentials zeigt, dass jede Komponente von $x(t)$ eine Linearkombination der exponentiellen Terme $\exp(\lambda_j\, t)$ ist. Insbesondere liefern beide Methoden das gleiche Resultat, nämlich

$$y(t) = C_1\, \exp(\lambda_1\, t) \ + \ \ldots \ + \ C_n\, \exp(\lambda_n\, t)\,,$$

wobei die C_j gerade n freie Konstanten darstellen.

Entartungsfall: Man kann für die Matrix A zeigen, dass jede doppelte oder mehrfache Nullstelle ihres charakteristischen Polynoms einem Jordan-Eigenwert entspricht. Die Berechnung von $\exp(t\, A)$ wird daher auch polynomielle Korrekturfaktoren enthalten, wobei wir diese alternativ auch mit dem folgenden *Kochrezept* bestimmen können: Ist λ_j eine k_j-fache Nullstelle von p, so enthält die allgemeine Lösungsformel für $y(t)$ den Term

$$\big(C_{j,\,0} + C_{j,\,1}\, t \ + \ldots + C_{j,\,k_j-1}\, t^{k_j-1}\big)\, \mathrm{e}^{\lambda_j\, t}$$

anstelle von $C_j\, \mathrm{e}^{\lambda_j\, t}$. Insbesondere wird es am Ende insgesamt wieder n freie Konstanten in der allgemeinen Lösungsformel geben, denn die Vielfachheiten der verschiedenen Nullstellen summieren sich zu n.

[21] Die Berechnung von $\det\big(A - \lambda\, 1\!\!1\big)$ gelingt zum Beispiel durch eine n-fache Anwendung des Laplaceschen Entwicklungssatzes.

Beispiel Zu der Gleichung

$$y^{(4)}(t) + 2\,y^{(3)}(t) - 2\,y^{(1)}(t) - y^{(0)}(t) = 0$$

gehört das Polynom

$$p(\lambda) = \lambda^4 + 2\,\lambda^3 - 2\,\lambda - 1 = (\lambda - 1)\,(\lambda + 1)^3\,,$$

das die Dreifachnullstelle $\lambda_1 = -1$ sowie die Einfachnullstelle $\lambda_2 = +1$ besitzt.

Lösungsweg 1: Das Kochrezept liefert

$$y(t) = \big(C_{1,0} + C_{1,1}\,t + C_{1,2}\,t^2\big)\,\exp(-t) + C_2 \exp(+t)$$

als allgemeine Lösungsformel und wir können einfach nachrechnen, dass diese Formel für jede Wahl der insgesamt vier Konstanten eine Lösung der Differentialgleichung höherer Ordnung liefert.

Lösungsweg 2: Wir können nicht kochen und berechnen lieber das Matrixexponential $\exp(t\,A)$. Dazu bemerken wir, dass

$$A = \begin{pmatrix} 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & +1 \\ +1 & +2 & 0 & -2 \end{pmatrix} = Q\,\widetilde{A}\,Q^{-1}\,, \qquad \widetilde{A} = \begin{pmatrix} -1 & +1 & 0 & 0 \\ 0 & -1 & +1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & +1 \end{pmatrix}$$

gilt, wobei $\widetilde{A}$ ein dreidimensionales Jordan-Kästchen zum Eigenwert -1 enthält und der Wechsel in die Eigenbasis durch

$$Q = \begin{pmatrix} -1 & -3 & -6 & +1 \\ +1 & +2 & +3 & +1 \\ -1 & -1 & -1 & +1 \\ +1 & 0 & 0 & +1 \end{pmatrix}, \qquad Q^{-1} = \frac{1}{8}\begin{pmatrix} -1 & -3 & -3 & +7 \\ +6 & +10 & -6 & -10 \\ -4 & -4 & +4 & +4 \\ +1 & +3 & +3 & +1 \end{pmatrix}$$

beschrieben wird. Insgesamt ergibt sich

$$\exp\big(t\,A\big) = Q \begin{pmatrix} \mathsf{e}^{-t} & t\,\mathsf{e}^{-t} & \frac{1}{2}\,t^2\,\mathsf{e}^{-t} & 0 \\ 0 & \mathsf{e}^{-t} & t\,\mathsf{e}^{-t} & 0 \\ 0 & 0 & \mathsf{e}^{-t} & 0 \\ 0 & 0 & 0 & \mathsf{e}^{+t} \end{pmatrix} Q^{-1}$$

und wir erhalten nach einigen Rechnungen am Ende via

$$x(t) = \exp\big(t\,A\big)\,c$$

wieder die allgemeine Lösungsformel für $y(t) = x_1(t)$, wobei die Konstanten $C_{1,0}$, $C_{1,1}$, $C_{1,2}$ in linearer Weise von c_1, c_2, c_3, c_4 abhängen. Die resultierende Formel ist allerdings zu lang für diese Seiten.

Diskussion: Die Formeln des Kochrezeptes sind deutlich einfacher. Die Berechnung des Matrixexponentials ist aber allgemeiner und liefert auch die Begründung des Rezeptes.

Vorlesungswoche 13

3.7 Stabilität und Sensitivität*

Ziel In diesem Abschnitt betrachten wir wieder eine allgemeine Differentialgleichung erster Ordnung und wollen verstehen, wie sich Lösungen zu Anfangswertproblemen bei kleinen Änderungen der Anfangsdaten und/oder von Parametern in der Gleichung verhalten. Diese Frage ist nicht nur von großem theoretischen Interesse, sondern spielt auch in den Anwendungswissenschaften eine wichtige Rolle, da man die Anfangsdaten und Parameter oftmals nur gemessen hat und daher nicht exakt kennt. Wir beginnen mit Störungen von Anfangsdaten und werden erst später solche von Parametern berücksichtigen.

Stabilität von Lösungen

Vereinfachung Um die wesentlichen Ideen besser herauszuarbeiten, treffen wir die folgenden Vereinfachungen:

1. Die Funktion f (also die rechte Seite in der Differentialgleichung) ist auf ganz $\mathbb{R} \times \mathbb{R}^n$ definiert und außerdem stetig differenzierbar.

2. Die Anfangszeit ist immer $t_* = 0$.

3. Wir nehmen an, dass die betrachteten Lösungen der Differentialgleichung immer für alle Zeiten $t \geq 0$ existieren.

Notation Wir werden in diesem Abschnitt die Anfangsdaten nicht mehr mit x_*, sondern mit ξ bezeichnen, und außerdem nicht nur die Abhängigkeit von der Zeit, sondern auch die von den Anfangsdaten explizit angeben. Wir bezeichnen daher die Lösungen mit $x(t, \xi)$ und schreiben das Anfangswertproblem als

$$\partial_t x(t, \xi) = f\big(t, x(t, \xi)\big)\,, \qquad x(0, \xi) = \xi\,.$$

Die neue Schreibweise ist vollkommen äquivalent zur alten, macht aber besser deutlich, dass die Anfangsdaten eigentlich Variablen der Lösung sind.

Vorbemerkung Stabilität einer festgehaltenen Lösung meint, dass die Lösungen zu leicht gestörten Anfangsdaten von dieser Lösung „nicht weglaufen“, sondern „benachbart bleiben“ oder gar zu ihr „zurückkehren“. In der mathematischen Theorie gibt es einige Subtilitäten, die vor allem von Grenz- und Entartungsfällen stammen. Wir beginnen daher mit einer informellen Übersicht, was Stabilität bzw. Instabilität in einem echten Sinne meint, und diskutieren die leicht abweichende mathematische Definition später.

Beispiele

1. Für $n = 1$ beschreibt

$$\partial_t x(t, \xi) = \eta\, x(t, \xi)\,, \qquad x(0, \xi) = \xi$$

die einfachste lineare Differentialgleichung mit Parameter η (den wir hier als gegeben und fest ansetzen), Zeit t und variabler Anfangsbedingung ξ. Die Lösung ist offensichtlich durch

$$x(t, \xi) = \xi \exp(\eta t)$$

gegeben. Wir fixieren nun ein festes Anfangsdatum ξ_* sowie die entsprechende Lösung der Differentialgleichung $x(t, \xi_*)$, betrachten aber auch Lösungen $x(t, \xi)$ mit $\xi \neq \xi_*$, d.h. mit *gestörten* Anfangsdaten. Mithilfe der Lösungsformel können wir nun leicht die folgenden Aussagen ableiten:

(a) Für $\eta < 0$ gilt

$$\big|x(t, \xi) - x(t, \xi_*)\big| = \big|\xi - \xi_*\big| \exp(-|\eta| t) \quad \xrightarrow{t\to\infty} \quad 0\,,$$

d.h. jede gestörte Lösung nähert sich im Laufe der Zeit der festgehaltenen Lösung an. Dies ist gerade die Idee hinter *echter Stabilität*.

(b) Für $\eta > 0$ gilt

$$\big|x(t, \xi) - x(t, \xi_*)\big| = \big|\xi - \xi_*\big| \exp(+|\eta| t) \quad \xrightarrow{t\to\infty} \quad +\infty$$

d.h. jede gestörte Lösung wird sich von der vorgegebenen im Laufe der Zeit entfernen. Dieses Verhalten entspricht *echter Instabilität*.

(c) Für $\eta = 0$ gilt

$$\big|x(t, \xi) - x(t, \xi_*)\big| = \big|\xi - \xi_*\big| \qquad \text{für alle} \quad t \geq 0$$

d.h. die gestörte Lösung nähert sich der vorgegebenen nicht mehr an, läuft aber auch nicht von ihr weg. Insbesondere ist der Fall $\eta = 0$ gerade die Grenze zwischen echter Stabiliät und echter Instabilität und es würde sehr viel Sinn machen, ihn entweder mit keinem, oder mit allen beiden Konzepten in Verbindung zu setzen. Aus historischen Gründen hat sich allerdings eine andere Klassifikation durchgesetzt: Der Grenzfall η wird stabil, aber nicht instabil genannt (siehe unten).

2. Als Beispiel für $n = 2$ betrachten wir

$$\partial_t x(t, \xi) = A\, x(t, \xi)\,, \qquad x(0, \xi) = \xi\,, \qquad A = \frac{1}{2}\begin{pmatrix} \eta_1 + \eta_2 & \eta_1 - \eta_2 \\ \eta_1 - \eta_2 & \eta_1 + \eta_2 \end{pmatrix}.$$

Da die Matrix A diagonalisierbar ist mit Eigenwerten η_1 und η_2, ergibt sich (zum Beispiel nach Berechnung des Matrixexponentials) die Lösungsformel

$$\begin{pmatrix} x_1(t, \xi_1, \xi_2) \\ x_2(t, \xi_1, \xi_2) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \mathsf{e}^{\eta_1 t} + \mathsf{e}^{\eta_2 t} & \mathsf{e}^{\eta_1 t} - \mathsf{e}^{\eta_2 t} \\ \mathsf{e}^{\eta_1 t} - \mathsf{e}^{\eta_2 t} & \mathsf{e}^{\eta_1 t} + \mathsf{e}^{\eta_2 t} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix},$$

und damit

$$\begin{aligned} x_1(t, \xi) - x_1(t, \xi_*) &= \tfrac{1}{2}\,\mathsf{e}^{\eta_1 t}\big(\xi_1 + \xi_2 - \xi_{*,1} - \xi_{*,2}\big) + \tfrac{1}{2}\,\mathsf{e}^{\eta_2 t}\big(\xi_1 - \xi_2 - \xi_{*,1} + \xi_{*,2}\big) \\ x_2(t, \xi) - x_2(t, \xi_*) &= \tfrac{1}{2}\,\mathsf{e}^{\eta_1 t}\big(\xi_1 + \xi_2 - \xi_{*,1} - \xi_{*,2}\big) - \tfrac{1}{2}\,\mathsf{e}^{\eta_2 t}\big(\xi_1 - \xi_2 - \xi_{*,1} + \xi_{*,2}\big)\,. \end{aligned}$$

Insgesamt erhalten wir die folgenden Aussagen für jedes fixierte ξ_*:

(a) Für $\eta_1 < 0$ und $\eta_2 < 0$ klingen alle exponentiellen Terme für $t \to \infty$ ab. Insbesondere ist die Lösung $x(t, \xi_*)$ *echt stabil*, denn es gilt

$$\big|x(t, \xi) - x(t, \xi_*)\big| \quad \xrightarrow{t\to\infty} \quad 0$$

für alle ξ.

(b) Für $\eta_1 > 0$ und $\eta_2 > 0$ wachsen alle exponentiellen Terme. Dies impliziert

$$\big|x(t, \xi) - x(t, \xi_*)\big| \quad \xrightarrow{t\to\infty} \quad \infty$$

und damit *echte Instabilität* der Lösung $x(t, \xi_*)$.

(c) Für $\eta_1 < 0 < \eta_2$ bzw. $\eta_2 < 0 < \eta_1$ gibt es wachsende und fallende Terme. Unter der Bedingung $\xi_1 - \xi_2 - \xi_{*,1} + \xi_{*,2} = 0$ bzw. $\xi_1 + \xi_2 - \xi_{*,1} - \xi_{*,2} = 0$ ist der wachsende Term nicht aktiv und die beiden Lösungen $x(t, \xi)$ und $x(t, \xi_*)$ nähern sich an. Für *alle anderen* Anfangsdaten wird der Betrag der Differenz über alle Grenzen wachsen und insgesamt ist die Lösung mit Anfangsdatum ξ_* *echt instabil*.

(d) Auch hier gibt es Grenzfälle, zum Beispiel $\eta_1 < \eta_2 = 0$ oder $\eta_1 = \eta_2 = 0$, bei denen man eigentlich an der Grenze zwischen Stabilität und Instabilität ist.

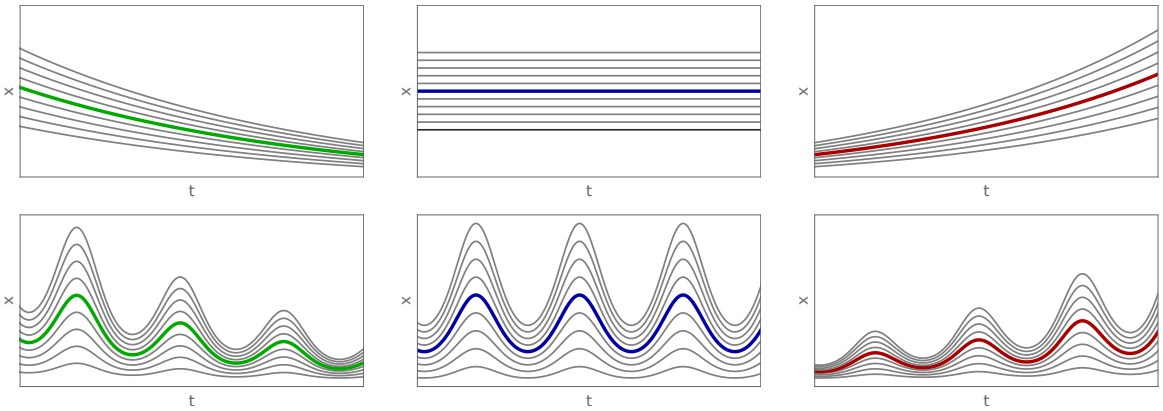

Abbildung Graphische Darstellung der informellen Konzepte für $n = 1$ sowie die Funktionen $f(t, x) = \eta\, x$ (oben) und $f(t, x) = \eta\, x + \sin(t)\, x$ (unten). Die grüne Lösung (links, $\eta = -1$) ist jeweils *echt stabil*, wohingegen die rote Lösung (rechts, $\eta = +1$) *echt instabil* ist. Die blaue Lösung ($\eta = 0$) entspricht dem entarteten Grenzfall.

Definition

1. Die Lösung $x(t, \xi_*)$ heißt stabil, falls benachbarte Lösungen sich für $t \to \infty$ nicht von ihr entfernen. Mathematisch wird dies als

$$\forall\, \varepsilon > 0 \quad \exists\, \delta > 0 \quad : \quad \big|\xi - \xi_*\big| < \delta \quad \Rightarrow \quad \sup_{t\geq 0} \big|x(t, \xi) - x(t, \xi_*)\big| < \varepsilon\,,$$

formuliert und meint, dass man für jedes (noch so kleine) $\varepsilon > 0$ ein (meist viel kleineres) $\delta > 0$ findet, sodass für alle Anfangsdaten ξ in der δ-Nähe von ξ_* und alle Zeiten $t \geq 0$ der Abstand von $x(t, \xi)$ und $x(t, \xi_*)$ kleiner als ε bleibt.

2. Laufen benachbarte Lösungen nicht nur nicht weg, sondern nähern sich für große Zeiten sogar an, so spricht man von asymptotischer Stabilität. In Formeln wird das durch die Zusatzbedingung

$$\exists\, \tilde\delta > 0 \quad : \quad \big|\xi - \xi_*\big| < \tilde\delta \quad \Rightarrow \quad \lim_{t\to\infty} \big|x(t, \xi) - x(t, \xi_*)\big| = 0\,,$$

ausgedrückt.

3. Eine nicht stabile Lösung wird instabil genannt.

Bemerkungen

1. Vereinfacht kann man sagen: Die asymptotische Stabilität in der mathematischen Definition ist das, was wir in den obigen Beispielen informell *echte Stabilität* genannt haben, wohingegen Instabilität ziemlich genau der *echten Instabilität* entspricht. Ist $x(t, \xi_*)$ jedoch nur stabil (aber nicht asymptotisch stabil), so handelt es sich eigentlich um einen Grenzfall und man sollte besser von „entartet stabil“ (oder von “gerade noch stabil“ bzw. „fast schon instabil“) reden. Allerdings hat sich diese Sprechweise nicht durchgesetzt.

2. Wir werden im Folgenden vor allem die Stabilität von stationären Lösungen untersuchen, für die jede der drei äquivalenten Bedingungen
$$x(t, \xi_*) = \xi_* \qquad \text{bzw.} \qquad \partial_t x(t, \xi_*) = 0 \qquad \text{bzw.} \qquad f(t, \xi_*) = 0$$
zu jeder Zeit $t \geq 0$ erfüllt ist. Stabilität ist aber etwas, was man für *jede* Lösung einer Differentialgleichung untersuchen kann.

3. Bei nichtlinearen Gleichungen gilt: Stabilität beschreibt das Verhalten unter *kleinen* Störungen von Anfangsdaten und macht keine Aussagen über *große* Störungen. Bei linear homogenen Differentialgleichungen ist das etwas anders, da es dort wegen der Linearität bzw. dem Superpositionsprinzip keinen Unterschied zwischen kleinen und großen Störungen gibt und weil die Lösungen daher entweder alle stabil oder alle instabil sind (siehe unten).

4. Unser Stabilitätskonzept bezieht sich immer auf zukünftige Zeiten $t \geq 0$. Man kann analoge Konzepte für vergangene Zeiten $t \leq 0$ einführen, aber diese spielen hier und in den meisten Anwendungen keine Rolle.

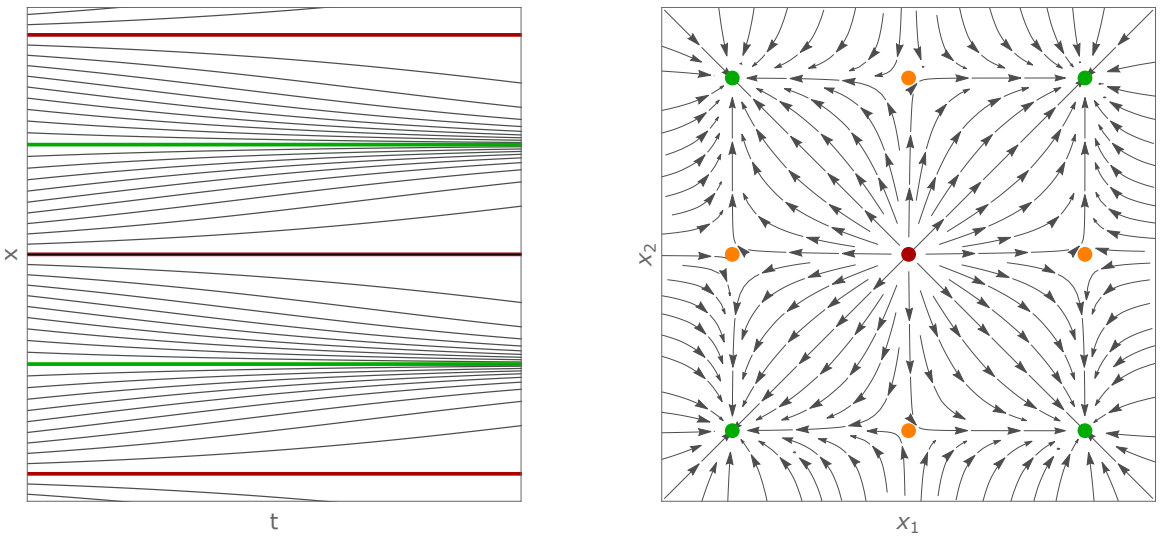

Abbildung Graphische Darstellung der Stabilität (grün) bzw. Instabilität (orange, rot) *stationärer* Lösungen für $n = 1$ (links, t-x-Diagramm) und $n = 2$ (rechts, Phasenportrait bzw. x_1-x_2-Diagramm), wobei die grünen Lösungen sogar asymptotisch stabil sind. Der Unterschied zwischen orange (*Sattel*) und rot (*Quelle*) manifestiert sich in veschiedenen Eigenwertsignaturen (siehe unten).

Stabilität bei linearen und homogenen Differentialgleichungen

Vorbemerkung Wir betrachten den linear homogenen Fall, den wir in diesem Abschnitt als
$$\partial_t x\big(t, \xi\big) = A(t)\, x\big(t, \xi\big)\,, \qquad x(0, \xi) = \xi$$

schreiben. Die Lösungen sind durch

$$x(t,\,\xi) = X(t)\,\xi$$

gegeben, wobei $X(t)$ eine Fundamentalmatrix ist (zum Beispiel die mit $X(0) = \mathsf{1}$). Insbesondere gibt es die triviale Lösung $x(t,\,0) = 0$ für alle $t \in \mathbb{R}$.

Lemma (**Stabilität bei einer linear homogenen Differentialgleichung**) Die triviale Lösung der Differentialgleichung ist

1. genau dann stabil, wenn es eine Konstante $C > 0$ gibt, sodass

$$\big|X(t)\big| \le C \qquad \text{für alle} \quad t \ge 0\,,$$

2. genau dann asymptotisch stabil, wenn

$$\big|X(t)\big| \quad \xrightarrow{t\to\infty} \quad 0\,.$$

Außerdem ist jede andere Lösung dann und nur dann stabil bzw. asymptotisch stabil, wenn die triviale Lösung diese Eigenschaft besitzt.

Beweis Alle Behauptungen ergeben sich aus der Definition einer Fundamentalmatrix, dem Superpositionsprinzip sowie den Eigenschaften quadratischer Matrizen. (Übungsaufgabe). □

Bemerkungen

1. Die Tatsache, dass die Stabilitätsbedingung bei linearen Differentialgleichungen (mit festen Parametern) immer für *alle* Lösungen erfüllt bzw. nicht erfüllt ist, ergibt sich aus dem Superpositionsprinzip. Insbesondere gilt stets

$$x(t,\,\xi) - x(t,\,\xi_*) = x(t,\,\xi - \xi_*)\,,$$

   d.h. die Differenz zweier Lösungen ist selbst Lösung. Im nichtlinearen Fall ist das anders und eine gegebene Differentialgleichung kann gleichzeitig stabile und instabile Lösungen besitzen.

2. Die Stabilitätsbedingungen im Lemma sind von eher theoretischem Interesse, da wir im nicht-autonomen Fall die Bedingungen an $X(t)$ nur schwer verifizieren können. Im autonomen Fall werden wir jedoch immer versuchen, das nachfolgende Theorem zu benutzen, da es wesentlich bessere Aussagen bereitstellt.

3. Ein wichtiger Spezialfall sind zeitperiodische Matrizen mit $A(t) = A(t + t_{\mathrm{per}})$, für die es die sogenannte *Floquet-Theorie* gibt.

Theorem (**Stabilität bei autonom homogenen Gleichungen**) Im autonomen Fall $A(t) = A$ gilt:

1. Besitzen alle Eigenwerte von A jeweils einen negativen Realteil, so ist jede Lösung asymptotisch stabil.

2. Hat auch nur ein Eigenwert von A einen positiven Realteil, so ist jede Lösung instabil.

Beachte, dass die Eigenwerte einer reellen Matrix komplexe Zahlen sein können.

Beweis Wir führen via

$$A = Q\,\widetilde{A}\,Q^{-1}\,, \qquad \widetilde{x}(t,\,\widetilde{\xi}) = Q^{-1}\,x(t,\,\xi)\,, \qquad \xi = Q\,\widetilde{\xi}$$

einen Koordinatenwechsel im $\mathbb{R}^n$ durch, wobei $\widetilde{A}$ die Jordansche Normalform von A ist und ihre Eigenwerte λ_j als Diagonaleinträge enthält. Die Differentialgleichung in den neuen Koordinaten lautet

$$\partial_t \widetilde{x}\big(t,\,\widetilde{\xi}\big) = \widetilde{A}\,\widetilde{x}\big(t,\,\widetilde{\xi}\big)\,, \qquad \widetilde{x}\big(0,\,\widetilde{\xi}\big) = \widetilde{\xi}$$

und die entsprechende Fundamentalmatrix $\widetilde{X}(t) = \exp(t\,\widetilde{A})$ kann explizit berechnet werden. Sie enthält die exponentiellen Terme $\mathtt{e}^{\lambda_j t}$ sowie ggf. polynomielle Korrekturfaktoren von den Jordan-Eigenwerten. Insbesondere kann nun die Behauptung mit $\widetilde{x}$ statt x einfach aus der Lösungsformel abgelesen werden. Da aber $X(t) = Q\,\widetilde{X}(t)$ für alle t gilt und die Basiswechselmatrizen Q und Q^{-1} nicht von t abhängen, folgt auch das gewünschte Resultat bzgl. x. □

Bemerkungen

1. Dieses Resultat ist ausgesprochen nützlich (vor allem in der Praxis), denn es stellt explizite Kriterien für die Stabilität bzw. Instabilität einer Lösung bereit, die nur von der Matrix A abhängen und die Kenntnis der Lösung nicht erfordern.

2. Besitzt einer oder mehrere der Eigenwerte von A einen verschwindenden Realteil, so handelt es sich um einen *Grenz-* bzw. *Entartungsfall* und wir müssen verschiedene Unterfälle unterscheiden. Dabei gilt:

   (a) Gibt es einen anderen Eigenwert mit positivem Realteil, so ist jede Lösung weiterhin instabil.

   (b) Gibt es einen Jordan-Eigenwert mit verschwindendem Realteil, so sind die Lösungen auch instabil.

   (c) Ist der Realteil aller Jordan-Eigenwerte negativ und der Realteil aller anderen Eigenwerte nichtpositiv, so sind die Lösungen der Differentialgleichung zwar noch stabil, aber nicht mehr asymptotisch stabil.

   Merkregel: Verschwindende Realteile sind kritisch für die Stabilität und man muss genauer hinschauen.

3. Eigenwerte mit Realteil 0 sind in der *Bifurkationsanalyse* wichtig, da sie einen Wechsel im qualitativen Lösungsverhalten andeuten bzw. nach sich ziehen können. Oder anders gesagt: Manchmal ist der Entartungsfall viel interessanter als der Standardfall.

Beispiele

1. Für die Matrix

   $$A = \begin{pmatrix} \eta & +1 \\ -1 & \eta \end{pmatrix}$$

   mit Parameter η berechnen wir die Eigenwerte zu

   $$\lambda_1 = \eta - \mathtt{i}\,, \qquad \lambda_2 = \eta + \mathtt{i}$$

und erhalten asymptotische Stabilität für $\eta < 0$ und Instabilität für $\eta > 0$, wobei man dieses entweder mit dem Theorem oder mittels der expliziten Lösungsformel (siehe oben) begründen kann. Im Entartungsfall $\eta = 0$ findet der Wechsel statt, wobei wir in diesem Beispiel mithilfe der expliziten Lösungen zeigen können, dass die Lösungen im Sinne der obigen Definition noch stabil (aber nicht mehr asymptotisch stabil) sind.

2. Die Jordan-Matrix

$$A = \begin{pmatrix} \eta & +1 & 0 \\ 0 & \eta & 1 \\ 0 & 0 & \eta \end{pmatrix}$$

besitzt den Eigenwert $\lambda = \eta$, dessen algebraische bzw. geometrische Vielfachheit 3 bzw. 1 ist. Beachte, dass diese Matrix schon in Normalform gegeben ist, d.h. es gilt $A = \widetilde{A}$ und $Q = Q^{-1} = \mathsf{1}$. Das Theorem liefert die asymptotische Stabilität bzw. Instabilität für $\eta < 0$ bzw. $\eta > 0$, wobei wir dies alternativ auch wieder direkt aus der Lösungsformel

$$\begin{aligned} x(t,\,\xi) &= \exp\left(t \begin{pmatrix} \eta & 1 & 0 \\ 0 & \eta & 1 \\ 0 & 0 & \eta \end{pmatrix} \right) \xi = \exp(\eta\, t) \begin{pmatrix} 1 & t & \frac{1}{2}\, t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix} \xi \\ &= \exp(\eta\, t) \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + t\, \exp(\eta\, t) \begin{pmatrix} \xi_2 \\ \xi_3 \\ 0 \end{pmatrix} + \tfrac{1}{2}\, t^2\, \exp(\eta\, t) \begin{pmatrix} \xi_3 \\ 0 \\ 0 \end{pmatrix} \end{aligned}$$

ableiten können. Diese Formel hatten wir schon weiter oben hergeleitet bzw. können ihre Gültigkeit einfach nachrechnen. Für $\eta = 0$ macht das Theorem keine Aussagen, aber wegen

$$x(t,\,\xi) = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + t \begin{pmatrix} \xi_2 \\ \xi_3 \\ 0 \end{pmatrix} + \tfrac{1}{2}\, t^2 \begin{pmatrix} \xi_3 \\ 0 \\ 0 \end{pmatrix}$$

schließen wir, dass die triviale Lösung für $\eta = 0$ instabil ist.

planarer Fall Wie wollen nun für eine reelle 2×2-Matrix A untersuchen, ob die triviale Lösung (und damit auch jede andere Lösung) stabil oder instabil ist. Dazu betrachten wir verschiedene Fälle und Unterfälle:

Standardfall: Die Matrix A besitzt zwei verschiedene und nicht verschwindende Eigenwerte, wobei diese entweder beide reell sind oder als ein Paar konjugiert komplexer Zahlen auftreten. Insbesondere ist A diagonalisierbar und wir können in natürlicher Weise sechs Unterfälle unterscheiden, die alle für Anwendungen wichtig sind (siehe das erste Bild). Dabei ist der Fall von zwei rein-imaginären Eigenwerten eigentlich auch entartet und nicht robust unter kleinen Störungen der Matrix.

Entartungsfälle, Teil 1: A besitzt nur einen Eigenwert, der entweder positiv oder negativ ist und die geometrische Vielfachheit 1 oder 2 haben kann (siehe das zweite Bild). In jedem dieser vier Unterfälle handelt es sich um einen entarteten Eigenwert, denn wenn wir die Einträge der Matrix ein bisschen stören, so werden in aller Regel zwei verschiedene Eigenwerte entstehen.

Entartungsfälle, Teil 2: Eine weitere Möglichkeit ist, dass A den Eigenwert 0 besitzt, wobei wir dann wieder verschiedene Unterfälle betrachten müssen, die wir hier aber nicht im Detail diskutieren wollen.

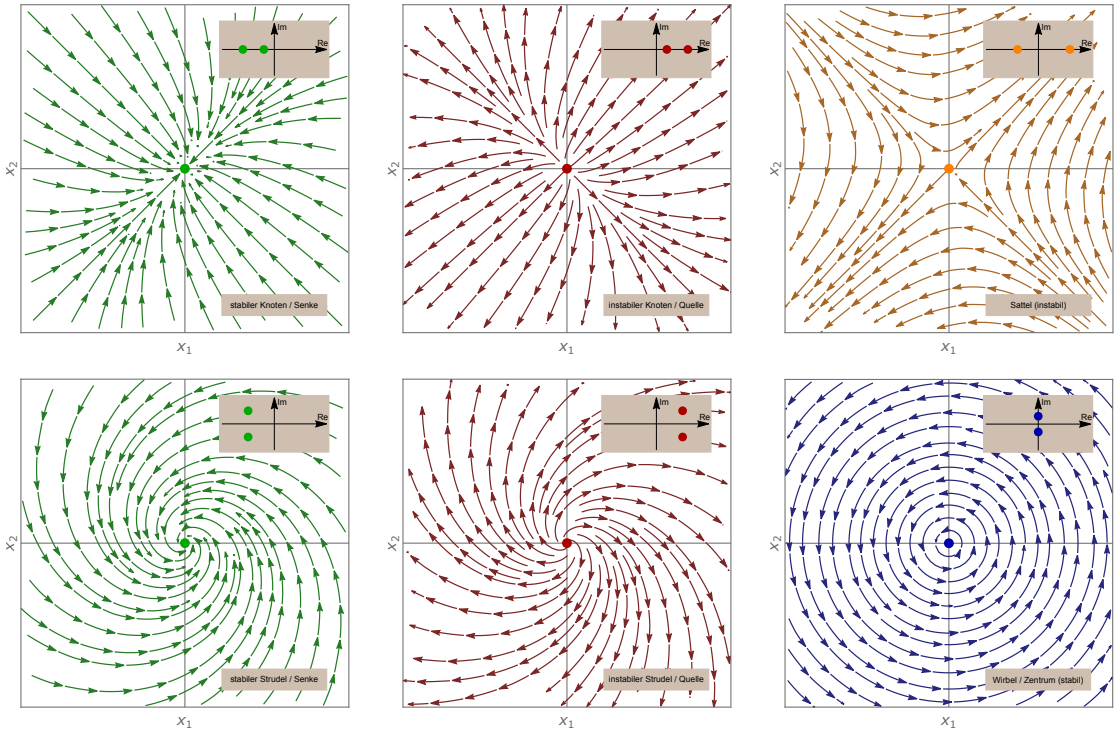

Abbildung Die sechs Möglichkeiten für eine 2×2-Matrix mit zwei verschiedenen Eigenwerten, die jeweils nicht 0 sind, wobei das Spektrum von A rechts oben in das jeweilige Phasenportrait gezeichnet wurde. Bei den ersten fünf handelt es sich um robuste Standardfälle, die auch vom Hauptsatz abgedeckt sind, wobei stabil bzw. instabil durch grün bzw. rot/orange kodiert wird. Der Fall rechts unten ist zwar wichtig, aber eigentlich entartet und nicht robust unter Störungen der Matrix. Die Bilder sind prototypisch in dem Sinne, dass bei gleichen Eigenwertsignaturen die Phasenportraits immer qualitativ gleich aussehen.

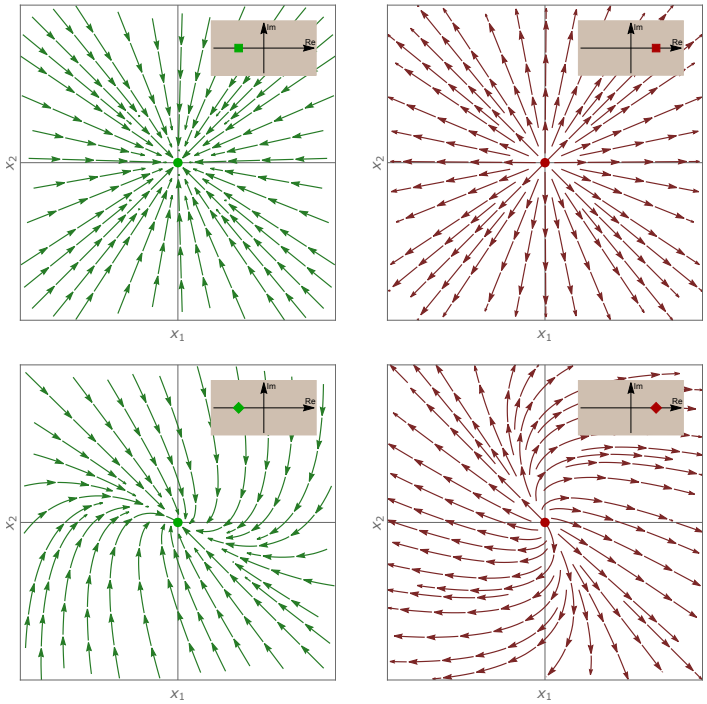

Abbildung Die vier Entartungsfälle der ersten Art mit einem doppelten, aber von Null verschiedenen Eigenwert, wobei oben bzw. unten die geometrische Vielfachheit 2 bzw. 1 ist. Nicht dargestellt sind die Entartungsfälle mit Eigenwert 0.

Bemerkung Für $n = 3$ oder gar $n > 3$ müssen wir natürlich sehr viel mehr Fälle unterscheiden. Die Standardsituation ist aber immer, dass A paarweise verschiedene Eigenwerte besitzt, die jeweils einen positiven oder einen negativen Realteil besitzen. Alle anderen Möglichkeiten sind auf die ein oder andere Weise entartet. Beachte auch, dass nur die Standardsituation robust unter kleinen Störungen der Matrix ist.

Stabilität bei nichtlinearen autonomen Differentialgleichungen

Setting Wir betrachten eine nichtlineare, aber autonome Differentialgleichung mit Vektorfeld $f : \mathbb{R}^n \to \mathbb{R}^n$, deren Lösungen durch das Anfangswertproblem

$$\partial_t x(t, \xi) = f\big(x(t, \xi)\big)\,, \qquad x(0, \xi) = \xi$$

charakterisiert sind. Des Weiteren sei $\xi_* \in \mathbb{R}^n$ mit $f(\xi_*) = 0$ gegeben, d.h. ein stationärer Punkt, der via $x(t, \xi_*) = \xi_*$ eine stationäre Lösung liefert.

Linearisierung Um das Verhalten der Lösungen in der Nähe des stationären Punktes zumindest näherungsweise zu verstehen, beginnen wir mit dem Ansatz

$$x(t, \xi) = \xi_* + z(t, \zeta)\,, \qquad \zeta = \xi - \xi_*\,,$$

d.h. wir betrachten neben der stationären Lösung eine weitere und bezeichnen mit $z(t, \zeta)$ bzw. ζ die Differenz zwischen beiden Lösungen bzw. den entsprechenden Anfangsdaten. Insbesondere gilt

$$\partial_t x(t, \xi) = \partial_t z(t, \zeta)$$

und der Satz von Taylor kombiniert mit $f(\xi_*) = 0$ impliziert

$$f\big(x(t, \xi)\big) = f\big(\xi_* + z(t, \zeta)\big) = \operatorname{Jac} f(\xi_*)\, z(t, \zeta) + O\big(\|z(t, \zeta)\|^2\big)\,.$$

Solange $\|z(t, \zeta)\|$ klein bleibt, erfüllt die Störung das approximative Problem

$$\partial_t z(t, \zeta) \approx \operatorname{Jac} f(\xi_*)\, z(t, \zeta)\,, \qquad z(0, \zeta) = \zeta\,,$$

wobei die entsprechende Differentialgleichung (mit $=$ anstelle von $\approx$) *linear, homogen und autonom* ist und die *Linearisierung* der nichtlinearen Gleichung im stationären Punkt ξ_* genannt wird. Die gegebene stationäre Lösung $x(t, \xi_*) = \xi_*$ entspricht dabei gerade der trivialen Lösung $z(t, 0) = 0$. Beachte, dass wir durch Vorgabe von ζ mit $\|\zeta\| \ll 1$ erreichen können, dass $\|z(t, \zeta)\|$ für alle hinreichend kleinen Zeiten $t \geq 0$ wirklich klein ist. Es kann aber sein, dass $\|z(t, \zeta)\|$ für große t immer größer wird und dies impliziert dann die Instabilität von ξ_*. Eine tiefe Erkenntnis der Mathematik ist nun, dass die Linearisierung viele qualitative Eigenschaften der nichtlinearen Gleichung in der Nähe von ξ_* oftmals richtig widerspiegelt.

Theorem (**Hauptsatz über Stabilität bei nichtlinearen Gleichungen**) Für die stationäre Lösung der nichtlinearen autonomen Gleichung gilt:

1. Sie ist asymptotisch stabil, sofern alle Eigenwerte von A_* einen negativen Realteil aufweisen.

2. Sie ist instabil, sofern mindestens ein Eigenwert von A_* einen positiven Realteil besitzt.

Dabei ist $A_* = \operatorname{Jac} f(\xi_*)$ die Jacobi-Matrix von f ausgewertet im stationären Punkt.

Bemerkungen

1. Der Beweis ist nicht einfach und benötigt einige technische Argumente. Wir wollen ihn daher hier nicht führen.

2. Der Hauptsatz wird auch *Prinzip der Linearisierten Stabilität* genannt, denn er erlaubt es (unter gewissen Voraussetzungen) die Stabilität oder Instabilität der nichtlinearen und stationären Lösung aus rein linearen Betrachtungen (Eigenwerte von Matrizen) abzuleiten. Es handelt sich um das *Standardverfahren* für die Untersuchung von Stabilität bei nichtlinearen Gleichungen und wird vor allem in den Anwendungswissenschaften sehr oft eingesetzt.

3. Der Hauptsatz macht keine Aussagen über die Stabilität der stationären Lösung, wenn einer oder mehrere der Eigenwerte von A_* einen verschwindenden Realteil aufweisen, wobei wir dies heuristisch wie folgt verstehen können: Die Eigenwerte mit Realteil 0 sind in der Linearisierung kritisch bzgl. Stabilität und Instabilität und stellen Grenzfälle dar. Die nichtlineare Stabilität bzw. Instabilität wird dann von den höheren Ordnungstermen in der Taylor-Approximation bzw. den höheren Ableitungen von f in ξ_* bestimmt, die wir aber in der Linearisierung vernachlässigen. Sind jedoch die Realteile aller Eigenwerte jeweils positiv oder negativ, so sind die linearen Terme dominant und der Hauptsatz garantiert, dass die höheren Ordnungsterme bei Stabilitätsuntersuchungen ignoriert werden können.

4. Wenn alle Eigenwerte von A_* einen nicht verschwindenden Realteil besitzen, so charakterisiert die Linearisierung nicht nur die Stabilität der stationären Lösung $x(t, \xi_*) = \xi_*$, sondern auch das qualitative Verhalten aller anderen Lösungen *in der Nähe* von ξ_*, d.h. für $x(t, \xi) \approx \xi_*$. Das ist die Aussage des *Theorems von Hartmann-Grobmann.*

Beispiele

1. Im eindimensionalen Fall ($n = 1$) ist das Vektorfeld eine Funktion $f : \mathbb{R} \to \mathbb{R}$ und die Nullstellen von f sind via $f(\xi_*) = 0$ gerade die stationären Punkte. Das linearisierte Anfangswertproblem ist auch skalar und durch

$$\partial_t z(t, \zeta) = a_* \, z(t, \zeta) \,, \qquad z(0, \zeta) = \zeta$$

gegeben. Es enthält den Parameter

$$a_* = f'(\xi_*) \,,$$

der gerade die einzige Komponente einer 1×1-Matrix ist. Der Hauptsatz kann damit wie folgt konkretisiert werden:

   (a) Für $a_* < 0$ ist ξ_* asymptotisch stabil.

   (b) Für $a_* > 0$ ist ξ_* instabil.

   (c) Für $a_* = 0$ können wir keine Aussage machen bzw. müssen höhere Ableitungen von f in ξ_* auswerten.

   Beachte, dass wir oben im Theorem über skalare und autonome Differentialgleichungen schon stärkere Aussagen für den Fall $a_* \neq 0$ hergeleitet hatten.

2. Wir betrachten die autonome Differentialgleichung zum planaren Vektorfeld

$$f(x_1, x_2) = \begin{pmatrix} -x_1 + x_2 \\ 2 - x_1^2 - x_2^2 \end{pmatrix}$$

und verifizieren durch direkte Rechnungen, dass es zwei stationäre Lösungen $x(t, \xi_*^{[j]}) = \xi_*^{[j]}$ gibt, nämlich

$$\xi_*^{[1]} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \qquad \xi_*^{[2]} = \begin{pmatrix} +1 \\ +1 \end{pmatrix},$$

für die auch $f(\xi^{[j]}) = 0$ gilt. Weitere Berechnungen liefern

$$A_*^{[1]} = \operatorname{Jac} f\big(\xi_*^{[1]}\big) = \begin{pmatrix} -1 & +1 \\ +2 & +2 \end{pmatrix}, \qquad \widetilde{A}_*^{[2]} = \tfrac{1}{2} \begin{pmatrix} 1 - \sqrt{17} & 0 \\ 0 & 1 + \sqrt{17} \end{pmatrix}$$

sowie

$$A_*^{[2]} = \operatorname{Jac} f\big(\xi_*^{[2]}\big) = \begin{pmatrix} -1 & +1 \\ -2 & -2 \end{pmatrix}, \qquad \widetilde{A}_*^{[2]} = \tfrac{1}{2} \begin{pmatrix} -3 - \mathtt{i}\sqrt{7} & 0 \\ 0 & -3 + \mathtt{i}\sqrt{7} \end{pmatrix},$$

wobei $A_*^{[j]} = Q^{[j]} \widetilde{A}_*^{[j]} \big(Q^{[j]}\big)^{-1}$ für geeignete Basiswechselmatrizen $Q^{[j]}$ gilt. Die Diagonaleinträge von $\widetilde{A}_*^{[j]}$ sind gerade die Eigenwerte von $A_*^{[j]}$ (die brauchen wir) und die Spalten von $Q^{[j]}$ bestehen aus Eigenvektoren von $A_*^{[j]}$ (die wir hier aber nicht berechnen müssen). Insbesondere klassifiziert der Hauptsatz $\xi_*^{[1]}$ bzw. $\xi_*^{[2]}$ als instabil bzw. stabil (beachte, dass $\sqrt{17} \approx 4.12$ und $\sqrt{7} \approx 2.65$).

3. Für die Differentialgleichung zum planaren Vektorfeld

$$f(x_1, x_2) = \begin{pmatrix} x_1^2 - x_2^2 \\ +x_1 x_2^2 - x_1 x_2 \end{pmatrix}$$

erhalten wir die drei stationären Lösungen bzw. Punkte

$$\xi_*^{[1]} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \xi_*^{[2]} = \begin{pmatrix} -1 \\ +1 \end{pmatrix}, \qquad \xi_*^{[3]} = \begin{pmatrix} +1 \\ +1 \end{pmatrix}$$

und berechnen

$$A_*^{[1]} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad A_*^{[2]} = \begin{pmatrix} -2 & -2 \\ 0 & -1 \end{pmatrix}, \qquad A_*^{[3]} = \begin{pmatrix} +2 & -2 \\ 0 & +1 \end{pmatrix}$$

sowie

$$\widetilde{A}_*^{[2]} = \begin{pmatrix} -2 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \widetilde{A}_*^{[3]} = \begin{pmatrix} +1 & 0 \\ 0 & +2 \end{pmatrix}.$$

Der Hauptsatz impliziert, dass $\xi_*^{[2]}$ bzw. $\xi_*^{[3]}$ stabil bzw. instabil ist, macht aber keine Aussage über $\xi_*^{[1]}$, da die entsprechende Jacobi-Matrix $A_*^{[1]}$ den Eigenwert 0 besitzt (sogar doppelt). Mit verfeinerten analytischen Argumenten (oder dem Plot des Phasenportraits) kann man aber zeigen, dass $\xi_*^{[1]}$ instabil ist.

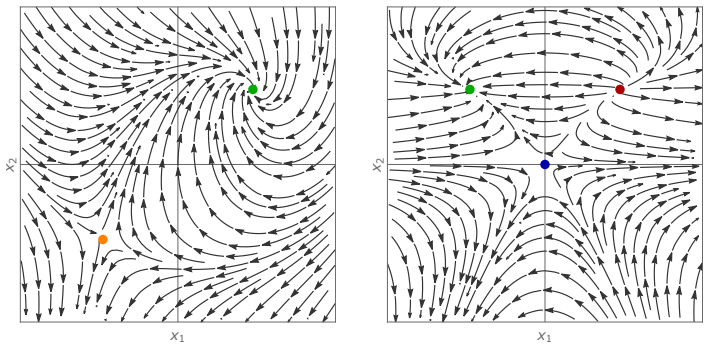

Abbildung Die Phasenportraits der gerade gerechneten nichtlinearen planaren Beispiele. Beachte, dass die Lösungen in der Nähe der grünen (asymptotisch stabil) bzw. orangen (instabil) stationären Punkte durch die jeweilige Linearisierung beschrieben werden können (vgl. dazu die Bilder der linearen Phasenportraits weiter oben). In der Nähe des entarteten blauen Punktes ist dies nicht möglich. Es handelt sich um einen *Affensattel*, für den es keine Entsprechung in der linearen Theorie gibt.

Sensitivitätsanalyse bei Differentialgleichungen

Ziel Wir wollen nun auch die Abhängigkeit der Lösungen von Parametern studieren und schreiben daher

$$\partial_t x(t,\,\xi,\,\eta) = f\big(t,\,x(t,\,\xi,\,\eta),\,\eta\big)\,, \qquad x(0,\,\xi,\,\eta) = \xi\,,$$

wobei $\eta \in \mathbb{R}^m$ für m skalare Parameter in der Differentialgleichung steht. Wir wollen außerdem voraussetzen, dass die (im Allgemeinen nichtlineare) Funktion $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ stetig differenzierbar ist, wobei ihre Variablen $t \in \mathbb{R}$, $x \in \mathbb{R}^n$ und $\eta \in \mathbb{R}^m$ sind.

Beispiele

1. Das einfachste Beispiel ist wieder die lineare und skalare Differentialgleichung

   $$\partial_t x(t,\,\xi,\,\eta) = \eta\, x(t,\,\xi,\,\eta)\,, \qquad x(0,\,\xi,\,\eta) = \xi$$

   mit Lösungsformel

   $$x(t,\,\xi,\,\eta) = \xi\, \exp\left(\eta\, t\right),$$

   wobei η nun nicht mehr eine festgehaltene Konstante ist, sondern als Variable betrachtet wird.

2. Die (bereits entdimensionalisierte) Differentialgleichung zweiter Ordnung

   $$\ddot{y}(t) = -\frac{1}{\big(1 + \eta\, y(t)\big)^2}$$

   modelliert die Ballistik eines vom Erdboden abgefeuerten Projektils (unter Vernachlässigung aller Reibungseffekte), wobei $y(t)$ die Flughöhe darstellt und η eine dimensionslose Größe bezeichnet, die üblicherweise sehr klein ist und in die verschiedene andere Konstanten (z. Bsp. der Abschusswinkel, eine Referenzgeschwindigkeit und die Erdbeschleunigung) einfließen. Für unsere Betrachtungen

transformieren wir die Gleichung in ein System erster Ordnung und schreiben das entsprechende Anfangswertproblem als

$$\partial_t \begin{pmatrix} x_1(t,\,\xi_1,\,\xi_2,\,\eta) \\ x_2(t,\,\xi_1,\,\xi_2,\,\eta) \end{pmatrix} = \begin{pmatrix} x_2(t,\,\xi_1,\,\xi_2,\,\eta) \\ -\dfrac{1}{\big(1+\eta\,x_2(t,\,\xi_1,\,\xi_2,\,\eta)\big)^2} \end{pmatrix}, \qquad x_j(0,\,\xi_1,\,\xi_2,\,\eta) = \xi_j\,,$$

wobei x_1 bzw. x_2 wieder für y bzw. $\dot{y}$ steht, wir aber diesmal die Abhängigkeit von η und den Anfangsdaten ξ_1, ξ_2 explizit schreiben. Für $\eta = 0$ reduziert sich die Gleichung auf $\ddot{y}(t) = -1$ und wir erhalten via

$$x_1(t,\,\xi_1,\,\xi_2,\,0) = \xi_1 + \xi_2\,t - \tfrac{1}{2}\,t^2\,, \qquad x_2(t,\,\xi_1,\,\xi_2,\,0) = \xi_2 - t$$

die klassische Wurfparabel. Diese Lösung existiert zwar für alle Zeiten, ist aber nur bis zur Aufprallzeit $t_\# = \xi_2 + \sqrt{2\,\xi_1 + \xi_2^2}$, d.h. für $0 \le t \le t_\#$, physikalisch sinnvoll. Eine naheliegende Frage ist nun, wie stark sich die Lösung ändert, wenn der Parameter η nicht mehr Null gesetzt wird. Für $\eta \neq 0$ gibt es zwar keine explizite Lösungsformel mehr, aber zumindest für kleine η kann man die Fragen mit einer Sensitivitätsanalyse gut beantworten.

Sensitivitäten Sind ξ_* gegebene Anfangsdaten sowie η_* ein gegebener Satz von Parametern, so impliziert der Satz von Taylor die Approximationsformel

$$x_k(t,\,\xi,\,\eta) - x_k(t,\,\xi_*,\,\eta_*) \approx \sum_{j=1}^{n} \partial_{\xi_j} x_k(t,\,\xi_*,\,\eta_*)(\xi_j - \xi_{*,j}) + \sum_{i=1}^{m} \partial_{\eta_i} x_k(t,\,\xi_*,\,\eta_*)(\eta_i - \eta_{*,i}),$$

sofern die partiellen Ableitungen von x_k nach ξ_j und η_i existieren und stetig sind. Wenn dies so ist, nennt man die Ableitungen $\partial_{\xi_j} x_k$ bzw. $\partial_{\eta_i} x_k$ die Sensitivität von x_k bzgl. ξ_j bzw. η_i. Die Sensitivitäten quantifizieren zu führender Ordnung, wie stark sich kleine Störungen der Anfangsdaten bzw. der Parameter auf die Lösungen der Differentialgleichung auswirken. Man kann nun folgende Fragen stellen:

1. Existieren die Sensitivitäten, d.h. hängt x wirklich in stetig differenzierbarer Weise von ξ und η ab?

2. Muss man für die Berechnung der Sensitivitäten eine explizite Lösungsformel verwenden oder geht das auch anders?

Theorem (**Hauptsatz über Sensitivitäten**) Die Sensitivitäten

$$V(t,\,\xi_*,\,\eta_*) := \partial_\xi x(t,\,\xi_*,\,\eta_*) \in \mathbb{M}^{n\times n}\,, \qquad W(t,\,\xi_*,\,\eta_*) := \partial_\eta x(t,\,\xi_*,\,\eta_*) \in \mathbb{M}^{n\times m}$$

sind wohldefiniert und erfüllen die *linearen*, aber im Allgemeinen *nicht-autonomen* Anfangswertprobleme

$$\partial_t V(t,\,\xi_*,\,\eta_*) = A(t,\,\xi_*,\,\eta_*)\,V(t,\,\xi_*,\,\eta_*)\,, \qquad V(0,\,\xi_*,\,\eta_*) = \mathsf{1}$$

und

$$\partial_t W(t,\,\xi_*,\,\eta_*) = A(t,\,\xi_*,\,\eta_*)\,W(t,\,\xi_*,\,\eta_*) + B(t,\,\xi_*,\,\eta_*)\,, \qquad W(0,\,\xi_*,\,\eta_*) = 0\,.$$

Hierbei können

$$A(t,\,\xi_*,\,\eta_*) = \partial_x f\big(t,\,x(t,\,\xi_*,\,\eta_*),\,\eta_*\big) \in \mathbb{M}^{n\times n}$$

und

$$B(t,\,\xi_*,\,\eta_*) = \partial_\eta f\big(t,\,x(t,\,\xi_*,\,\eta_*),\,\eta_*\big) \in \mathbb{M}^{n\times m}$$

aus den verschiedenen partiellen Ableitungen von f durch Einsetzen der festgehaltenen Lösung $x(t,\,\xi_*,\,\eta_*)$ berechnet werden.

Bemerkungen

1. Die Anfangswertprobleme für die Sensitivitäten sehen auf den ersten Blick sehr kompliziert aus, entstehen aber auf ganz einfache und natürliche Weise: Man differenziere die ursprüngliche Differentialgleichung sowie die Anfangsbedingung nach ξ bzw. η und wende dabei die höherdimensionale Kettenregel an. Mit dieser Erkenntnis kann man sich die Gleichungen für die Sensitivitäten auch jederzeit wieder herleiten.

2. Der Beweis des Theorems beruht (etwas vereinfacht gesprochen) auf folgender Idee: Die Sensitivitäten existieren, *weil* die angebenen Anfangswertprobleme eine eindeutige Lösung besitzen.

3. Eine direkte Folgerung ist:

   Die Lösungen einer Differentialgleichung hängen in stetig differenzierbarer Weise von den Anfangsdaten und den Parametern in der Gleichung ab, sofern die rechte Seite f hinreichend gut ist.

   Üblicherweise wird dieses wichtige Resultat direkt nach dem Satz von Picard-Lindelöf und ohne expliziten Bezug zu Sensitivitäten abgeleitet.

4. Die Sensitivitäten hängen von der Zeit t ab, d.h. sie unterliegen selbst einer Dynamik oder Evolution und für ihre Berechnung muss man lineare Differentialgleichungen lösen. Das kann sehr aufwendig sein und erfolgt in der Praxis meist numerisch mit dem Computer.

Beispiele

1. Im ersten der obigen eingeführten Beispiele können wir die Sensitivitäten direkt durch Differentiation der Lösungsformel berechnen. Wir erhalten

   $$v(t,\,\xi,\,\eta) = \partial_\xi x(t,\,\xi,\,\eta) = \exp(\eta\,t)\,, \quad w(t,\,\xi,\,\eta) = \partial_\eta x(t,\,\xi,\,\eta) = t\,\xi\exp(\eta\,t)$$

   und sind damit eigentlich schon fertig. Wir können aber zusätzlich durch Nachrechnen — oder alternativ durch Differentiation der Ursprungsgleichungen nach ξ und η — verifizieren, dass die beiden Anfangswertprobleme

   $$\partial_t v(t,\,\xi,\,\eta) = a(t,\,\xi,\,\eta)\,v(t,\,\xi,\,\eta)\,, \qquad v(0,\,\xi,\,\eta) = 1$$

   und

   $$\partial_t w(t,\,\xi,\,\eta) = a(t,\,\xi,\,\eta)\,w(t,\,\xi,\,\eta) + b(t,\,\xi,\,\eta)\,, \qquad w(0,\,\xi,\,\eta) = 0$$

in der Tat erfüllt sind, wobei sich die Koeffizientenfunktionen

$$a(t,\,\xi,\,\eta) = \partial_x f(t,\,x(t,\,\xi,\,\eta),\,\eta) = \eta$$

und

$$b(t,\,\xi,\,\eta) = \partial_\eta f(t,\,x(t,\,\xi,\,\eta),\,\eta) = \xi \exp(\eta)$$

direkt aus $f(t,\,x,\,\eta) = \eta\,x$ ergeben.

Bemerkung: In allen Formeln hätten wir natürlich auch ξ_* statt ξ und η_* statt η schreiben können.

2. Im Projektilproblem (siehe wieder oben) gibt es keine allgemeine Lösungsformel, aus der wir die Sensitivität durch direktes Ableiten gewinnen können. Wir wollen aber für die Lösung

$$\xi_* = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \eta_* = 0\,, \qquad x(t,\,\xi_*,\,\eta_*) = \begin{pmatrix} t - \frac{1}{2}t^2 \\ 1 - t \end{pmatrix}$$

ihre Sensitivitäten

$$V(t,\,\xi_*,\,\eta_*) = \begin{pmatrix} \partial_{\xi_1} x_1(t,\,\xi_*,\,\eta_*) & \partial_{\xi_2} x_1(t,\,\xi_*,\,\eta_*) \\ \partial_{\xi_1} x_2(t,\,\xi_*,\,\eta_*) & \partial_{\xi_2} x_2(t,\,\xi_*,\,\eta_*) \end{pmatrix} \in \mathbb{M}^{2\times 2}$$

und

$$W(t,\,\xi_*,\,\eta_*) = \begin{pmatrix} \partial_\eta x_1(t,\,\xi_*,\,\eta_*) \\ \partial_\eta x_2(t,\,\xi_*,\,\eta_*) \end{pmatrix} \in \mathbb{M}^{2\times 1}$$

mithilfe der angegebenen Anfangswertprobleme berechnen. Mit

$$f(t,\,x,\,\eta) = \begin{pmatrix} x_2 \\ -\dfrac{1}{(1+\eta\,x_1)^2} \end{pmatrix}$$

berechnen wir

$$\partial_x f(t,\,x,\,\eta) = \begin{pmatrix} 0 & +1 \\ \dfrac{2\,\eta}{(1+\eta\,x_1)^3} & 0 \end{pmatrix}, \qquad \partial_\eta f(t,\,x,\,\eta) = \begin{pmatrix} 0 \\ \dfrac{2\,x_1}{(1+\eta\,x_1)^3} \end{pmatrix}$$

und erhalten

$$A(t,\,\xi_*,\,\eta_*) = \begin{pmatrix} 0 & +1 \\ 0 & 0 \end{pmatrix}, \qquad B(t,\,\xi_*,\,\eta_*) = \begin{pmatrix} 0 \\ 2\,t - t^2 \end{pmatrix}$$

nach Einsetzen von η_* und der Lösung $x(t,\,\xi_*,\,\eta_*)$. Wir müssen nun die beiden Anfangswertprobleme

$$\dot V(t) = A(t)\,V(t)\,, \qquad V(0) = 1$$

und

$$\dot W(t) = A(t)\,W(t) + B(t)\,, \qquad W(0) = 0$$

lösen, wobei wir zur Vereinfachung der Notation die Abhängigkeit von ξ_* und η_* nicht mehr explizit geschrieben haben, da diese ja oben fixiert wurden. Da A hier nicht von t abhängt (das ist aber meist nicht so), erhalten wir

$$V(t) = \exp\left(t \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$$

und das Duhamel-Prinzip liefert wegen der Anfangsdaten

$$\begin{aligned} W(t) &= \int_0^t \exp\left((t-\sigma) \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right) B(\sigma)\,\mathrm{d}\sigma = \int_0^t \begin{pmatrix} (t-\sigma)\,(2\sigma-\sigma^2) \\ (2\sigma-\sigma^2) \end{pmatrix} \mathrm{d}\sigma \\ &= \begin{pmatrix} \frac{1}{3}\,t^3 - \frac{1}{12}\,t^4 \\ t^2 - \frac{1}{3}\,t^3 \end{pmatrix}. \end{aligned}$$

Wir können diese Ergebnisse nun in die approximative Taylor-Formel einsetzen. Wenn wir zum Beispiel zunächst η als variabel, die Anfangsdaten aber als fixiert, betrachten, erhalten wir

$$\begin{aligned} x_1(t,\,\xi_*,\,\eta) &= x_1(t,\,\xi_*,\,\eta_*) + W_{11}(t,\,\xi_*,\,\eta_*)\,(\eta-\eta_*) + O\big((\eta-\eta_*)^2\big) \\ &= \big(t - \tfrac{1}{2}\,t^2\big) + \big(\tfrac{1}{3}\,t^3 - \tfrac{1}{12}\,t^4\big)\eta + O\big(\eta^2\big) \end{aligned}$$

als Näherungsformel für die Flughöhe, wobei die rechte Seite nur von t und η abhängt. Insbesondere sehen wir, dass der Einfluss des Parameters η für kleine Zeiten t in der Tat sehr klein ist, aber dann relativ schnell anwächst. Die Formel mit Störung der Anfangsdaten beinhaltet auch die Beiträge von den Sensitivitäten $V_{11}(t,\,\xi_*,\,\eta_*)$, $V_{12}(t,\,\xi_*,\,\eta_*)$ und lautet

$$x_1(t,\,\xi,\,\eta) = \big(t - \tfrac{1}{2}\,t^2\big) + \xi_1 + t\,(\xi_2 - 1) + \big(\tfrac{1}{3}\,t^3 - \tfrac{1}{12}\,t^4\big)\eta + O\big(\eta^2 + \xi_1^2 + (\xi_2-1)^2\big)\,,$$

wobei für hinreichend kleine Zeiten Störungen von ξ_1 und ξ_2 deutlich größere Auswirkungen haben als eine Änderung von η. Für große Zeiten ist das aber nicht mehr so.

Bemerkung: Wir hätten die Formel für $V(t,\,\xi_*,\,0)$ auch aus der oben angegeben Lösungsformel für $\eta = 0$ durch Differentiation nach ξ_1 und ξ_2 direkt ableiten können. Die Sensitivität $W(t,\,\xi_*,\,0)$ kann aber so nicht berechnet werden, eben weil die Lösungsformel nur für $\eta = 0$ gilt.

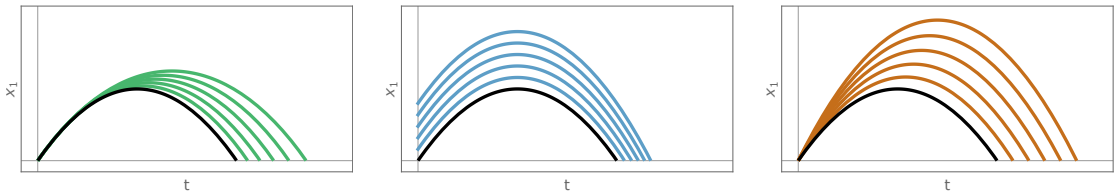

Abbildung Verschiedene Lösungen des Projektilproblems, wobei $x_1(t,\,\xi_*,\,\eta_*)$ schwarz dargestellt ist und die grünen bzw. blauen bzw. braunen Kurven Störungen in η bzw. ξ_1 bzw. ξ_2 entsprechen.

