Analysis 2

Gero Friesecke, Department of Mathematics, TUM

25. 7. 2025

Zusammenfassung

Dieses Skript deckt den Standardstoff einer Analysis 2–Vorlesung für Mathematik-Studierende im 2. Bachelor-Semester ab, und richtet sich auch an mathematisch interessierte Studierende anderer Fächer. Wir geben eine Einführung in die mehrdimensionale Analysis, d.h. das Studium vektorwertiger Funktionen mehrerer Veränderlicher. Das zentrale Konzept ist das Ableiten solcher Funktionen. (Integrieren im Mehrdimensionalen lernen Sie in "Analsis 3".) Weitere wichtige Themen sind die durch solche Funktionen definierten nichtlinearen Gleichungs- und Differentialgleichungs-Systeme.

Das letzte Kapitel ist eine Einführung in die Topologie, also das systematische Studium stetiger Abbildungen sowie von Eigenschaften, die unter stetigen Deformationen erhalten bleiben.

Ich bemühe mich, den Stoff aus seinen anschaulichen Quellen heraus zu entwickeln und die Verbindung zu alten und neuen Anwendungen aufzuzeigen.

Inhaltsverzeichnis

Einleitung **1**

1 Konvergenz im $\mathbb{R}^n$ und daraus abgeleitete Konzepte **4**
- 1.1 Der n-dimensionale Raum $\mathbb{R}^n$ . . . 4
- 1.2 Euklidischer Abstand . . . 6
- 1.3 Konvergenz . . . 9
- 1.4 Satz von Bolzano-Weierstrass . . . 11
- 1.5 Offene und abgeschlossene Mengen; Abschluss, Inneres, Rand . . . 11
- 1.6 Funktionen im Mehrdimensionalen: Beispielklassen, Visualisierung . . 14
  - 1.6.1 Skalare Funktionen . . . 14
  - 1.6.2 Kurven . . . 17
  - 1.6.3 Vektorfelder . . . 19
- 1.7 Stetigkeit . . . 20
- 1.8 Satz vom Maximum und Minimum . . . 22

2 Ableitung im Mehrdimensionalen **25**
- 2.1 Partielle Ableitung . . . 25
- 2.2 Totale Ableitung . . . 29
- 2.3 Gradient; Richtungsableitung . . . 35
- 2.4 Höhere partielle Ableitungen . . . 44
- 2.5 Divergenz, Rotation, Laplaceoperator . . . 49
- 2.6 Taylorentwicklung im Mehrdimensionalen . . . 51
- 2.7 Anwendung: Maximieren/Minimieren . . . 58
  - 2.7.1 Optimalitätsbedingungen . . . 59
  - 2.7.2 Drei Beispiele . . . 62
  - 2.7.3 Gradientenverfahren . . . 67
- 2.8 Anwendung: partielle Differentialgleichungen . . . 71
  - 2.8.1 Beispiele partieller Differentialgleichungen . . . 71
  - 2.8.2 Herleitung der Wellengleichung für die schwingende Saite . . . 74
  - 2.8.3 Lösen der Wellengleichung via Separation der Variablen . . . 75
- 2.9 Anwendung: Maschinelles Lernen . . . 77
  - 2.9.1 Neuronale Netze vom Feedforward-Typ . . . 77
  - 2.9.2 Herleitung des Backpropagation-Algorithmus via Kettenregel . 80
  - 2.9.3 Training via Gradientenverfahren . . . 84

3 Normierte Räume; Banach'scher Fixpunktsatz **87**
3.1 Andere Normen auf dem $\mathbb{R}^n$ . . . 88
3.2 Äquivalenz aller Normen auf dem $\mathbb{R}^n$ . . . 90
3.3 Allgemeine normierte Räume . . . 91
3.4 Der Raum $C(\Omega;\mathbb{R}^m)$; gleichmässige Konvergenz . . . 94
3.5 Banach'scher Fixpunktsatz . . . 96

4 Inverse und implizite Funktionen **100**
4.1 Inverse Funktionen . . . 100
4.2 Implizite Funktionen . . . 108
4.3 Singuläre Punkte; Untermannigfaltigkeiten . . . 114
4.4 Lagrange'sche Multiplikatorregel . . . 121

5 Systeme gewöhnlicher Differentialgleichungen **127**
5.1 Einleitung . . . 127
5.2 Existenz- und Eindeutigkeitssatz . . . 133
5.3 Lineare Systeme . . . 139
5.4 Stabilität . . . 151

6 Einführung in die Topologie **160**
6.1 Der Begriff des topologischen Raumes . . . 161
6.2 Abgeschlossene Mengen, Abschluss, Inneres, Rand . . . 163
6.3 Konvergenz . . . 165
6.4 Kompaktheit . . . 166
6.5 Stetigkeit . . . 168
6.6 Zusammenhängend . . . 169
6.7 Was ist eine topologische Eigenschaft (oder topologische Invariante)? 174

7 Kurvenintegral **176**

CC BY-NC-SA 4.0
Attribution – Non-Commercial – ShareAlike 4.0 International
Nicht-kommerzielle Nutzung und Verbreitung unter den Lizenzbedingungen gestattet
https://creativecommons.org/licenses/by-sa/4.0/

Einleitung

Das Vorlesungsskript "Analysis 2" gibt eine Einführung in die mehrdimensionale Analysis, d.h. das Studium vektorwertiger Funktionen mehrerer Variablen. Ein zentrales Konzept ist das Ableiten solcher Funktionen. (Integrieren im Mehrdimensionalen lernen Sie in "Analysis 3".) Weitere wichtige Themen sind die durch solche Funktionen definierten nichtlinearen Gleichungs- und Differentialgleichungs-Systeme. Das letzte Kapitel ist eine Einführung in die Topologie. Wir bauen auf unserem Studium skalarer Funktionen einer reellen Variablen in "Analysis 1" auf, und nutzen zudem einige Ergebnisse aus der Linearen Algebra. Fast alle hier dargestellten Resultate und Methoden sind seit langem bekannt und finden sich in zahlreichen exzellenten Lehrbüchern.

Inhalt und Aufbau. In Kapitel 1 gebe ich eine kurze, im Gegensatz zu anderen Skripten und Büchern bewusst elementar gehaltene, Einführung zum Grundbegriff der Konvergenz und daraus abgeleiteter Konzepte (offene, abgeschlossene und kompakte Mengen; Stetigkeit) in der Hauptarena der Analysis 2, dem n-dimensionalen Raum $\mathbb{R}^n$. Durch diese Herangehensweise kommen wir schneller und einfacher an das nötige Hintergrundwissen für das zentrale Thema der Ableitung (Kapitel 2), ohne uns den Weg dahin durch die Subtilitäten undendlichdimensionaler Vektorräume oder topologischer Räume zu erschweren.[1] Als Hauptanwendung verallgemeinern wir den Satz vom Maximum und Minimum ins Mehrdimensionale.

In Kapitel 2 lernen wir, wie man Funktionen mehrerer Veränderlicher ableitet und wozu das gut ist. Ableiten ist jetzt weniger simpel als im Analysis-1-Fall eindimensionaler Funktionen, beruht aber in weiten Teilen auf denselben Ideen und Prinzipien, und enthält zusätzlich interessante geometrische Aspekte. So ist die Verallgemeinerung der Ableitung auf skalare Funktionen mehrerer Veränderlicher ein Vektor, der in die Richtung des steilsten Anstiegs der Funktion zeigt. Dieser Vektor heisst *Gradient*. Der Gradient spielt in der gesamten Mathematik und ihren Anwendungen eine grundlgenden Rolle. Zum Beispiel: Ohne Gradient kein Gradientenverfahren; ohne Gradientenverfahren kein maschinelles Lernen und keine Künstliche Intelligenz. Dieses prototypische numerische Minimierungsverfahren analysieren wir in §2.7.3, und benötigen dazu einen Großteil der bis dahin entwickelten Theorie.

In Kapitel 4 und 5 beschäftigen wir uns mit nichtlinearen Gleichungssystemen bzw. nichtlinearen Systemen gewöhnlicher Differentialgleichungen. Wir gehen insbesondere den folgenden grundlegenden Fragen nach: Wie kann man die Existenz von Lösungen solcher Systeme in der (typischen) Situation beweisen, wenn keine expliziten Lösungsformeln möglich sind? Was kann man über das qualitative Verhalten solcher Systeme sagen? Hierzu sind substantielle neue Methoden erforderlich: man

[1] Solche Räume – die den $\mathbb{R}^n$ weitreichend verallgemeinern – besprechen wir in späteren Kapiteln des Skripts (Kapitel 3 bzw. 6).

muss Analysis in allgemeinen unendlichdimensionalen Vektorräumen (z.B. Räumen von Funktionen) betreiben. Dies tun wir in Kapitel 3. Dort werden wir insbesondere die Begriffe der Konvergenz und der Vollständigkeit sinnvoll auf unendlichdimensionale Vektorräume verallgemeinern. Hauptresultat des Kapitels ist der Banach'sche Fixpunktsatz. In Kapiteln 4 und 5 setzen wir den Satz gewinnbringend ein, indem wir dessen Voraussetzungen in wichtigen konkreten Situationen (nichtlineare Gleichungs- bzw. Differentialgleichungs-Systeme) nachweisen.

Zum Schluss der Vorlesung, in Kapitel 6, beleuchten wir einige Grundbegriffe der Analysis (konvergent, abgeschlossen, kompakt, stetig) nochmal aus einer anderen Perspektive, derjenigen der *Topologie*. In der Topologie definiert man diese Begriffe nämlich nicht nur im $\mathbb{R}^n$ oder in normierten Vektorräumen, sondern viel allgemeiner und abstrakter, und beschäftigt sich mit Eigenschaften von Objekten, die unter stetigen Verformungen erhalten bleiben, wie z.B. "zusammenhängend" (was das ist, kann man interessanterweise präzise definieren). Die abstrakten Definitionen der Grundbegriffe sind zwar viel unanschaulicher als die "Fussgängerdefinitionen" aus Kapitel 1 für den $\mathbb{R}^n$. Aber dafür sind viele Beweise – selbst wenn man nur an Analysis im $\mathbb{R}^n$ interessiert ist – erstaunlich kurz und effizient. *Abstraktion macht manche Dinge einfacher!* Im Gegensatz zu vielen Skripten und Büchern fangen wir aber nicht mit abstrakter Topologie an, bevor wir mehrdimensionale Funktionen maximieren/minimieren, ableiten, taylorentwickeln, die brauchen Sie nämlich für dieses Material nicht. Nach meiner Erfahrung ist es für Sie leichter, Topologie am Ende des Semesters zu verstehen, wenn Sie sich im $\mathbb{R}^n$ schon auskennen.

Im optionalen kurzen Kapitel über Kurvenintegrale (Kapitel 7) lernen wir – als Vorschau auf Analysis 3 – die einfachste Verallgemeinerung des Integrals einer 1D Funktion $f : [a, b] \to \mathbb{R}$ kennen, nämlich das Integral eines Vektorfeldes über eine Kurve. Definition, wichtige Eigenschaften und schöne Anwendungen in Geometrie und Physik ergeben sich ohne viel Mühe durch Verbinden des Riemann-Integrals aus Analysis 1 mit unseren Kenntnissen über Vektorfelder und Kurven aus Analysis 2.

Darstellung. Meiner Präsentation des Stoffes liegen folgende Prinzipien zugrunde.

1. *Abstrakte Konzepte, Rechentechnik, Intuition und Motivation gehören zusammen.* Ich habe mich bemüht, neben dem präzisen Formulieren und Beweisen mathematischer Sachverhalte und dem Durchrechnen von Beispielen auch die zugrundeliegende Intuition und Motivation sorgfältig zu erklären: wie kommt man darauf?[2] wozu ist das gut?[3] wieso geht man dieses oder jenes Thema so abstrakt an?[4]

2. *Numerische Plots sind nicht nötig für einfache Beispiele (eine gute manuelle Skizze an der Tafel erledigt den Job), aber nützlich für fortgeschrittene Theorie.* Die übliche Stoffaufteilung und deren Begründung "Existenzsätze und qualitatives Ver-

[2] siehe z.B. die Diskussion des Gegenbeispiels in §2.4 oder des Beweises in §4.1

[3] für inner- und außermathematische Anwendungen siehe z.B. §2.8, §2.9, §4.4, §5.4

[4] siehe z.B. die Einleitungen zu Kapiteln 3, 5, 6

halten → Analysis (hierfür braucht man keine Numerik); Näherungsverfahren zur quantitativen Berechnung von Lösungen → Numerik (lernt man später, denn man braucht Analysis)" ist nämlich eine Übervereinfachung. Z.B. basiert der Beweis des Satzes über inverse Funktionen (eines typischen Existenzsatzes) auf einer iterativen Methode, in der man Näherungen für die Umkehrfunktion sukzessive verbessert (also einem typischen numerischen Verfahren); indem man das Verfahren kurz bespricht und exemplarisch solche Verbesserungen numerisch berechnet und plottet (siehe §4.1), wird der Beweis transparenter. Oder: Der Satz über implizite Funktionen enthält eine Rangbedingung an die Ableitung; numerische Plots interessanter impliziter Flächen (z.B. der Lösungsmengen kubischer Gleichungen mit drei Unbekannten) beleuchten den Satz und die Rangbedingung als spannende Methode zur Suche nach den singulären Punkten, die in den Plots zu sehen sind (siehe §4.3).

3. *Die in dieser Vorlesung erarbeitete Mathematik ist nicht nur wichtig in klassischen Anwendungsgebieten (Physik; Natur- und Ingenieurwissenschaften), sondern auch in neuen (maschinelles Lernen).* Das ein oder andere substantielle Beispiel aus den Naturwissenschaften, passend für Mathematik-Studierende aufbereitet, hat in Analysis-Vorlesungen Tradition. Ich behalte diese Tradition bei, und bespreche z.B. die Wellengleichung der schwingenden Saite. Wir leiten diese prototypische partielle Differentialgleichung her, lösen sie (in einer einfachen Situation), und interpretieren die Lösung physikalisch und musikalisch. Siehe §2.8. Ich ergänze diese Tradition um ein substantielles Beispiel zum maschinellen Lernen. Das scheint in der Skripten- und Lehrbuchliteratur neu zu sein. Ich präsentiere neuronale Netze und deren Training durch den Backpropagation-Algorithmus. Die Bedeutung dieser Lernmethode ist erst durch aktuelle Entwicklungen im Bereich der Künstlichen Intelligenz klar geworden, denen sie – im Gegensatz zu einer Vielzahl anderer Ansätze – zugrundeliegt. Gleichzeitig passt sie gut in eine "Analysis 2"-Vorlesung, denn sie ist gradientenbasiert und nutzt damit die Tatsache aus, dass es im allgemeinen viel einfacher ist, eine eingermaßen glatte stetige Funktion über viele Parameter zu minimieren als eine diskrete kombinatorische Funktion; darüber hinaus ist die Herleitung des Algorithmus eine schöne Anwendung eines klassischen Resultats aus Analysis 2, nämlich der mehrdimensionalen Kettenregel. Siehe §2.9.

Ich bedanke mich bei: Michiel Renger, auf dessen Notizen die Abschnitte 2.9.2–2.9.3 basieren; Christan Kuehn und Oliver Junge, für hilfreiche Kommentare zu Kapitel 5; Claudia Scheimbauer und Ulrich Bauer, für kurzes aber nützliches Feedback zu Kapitel 6; Hans-Peter Kruse, für fruchtbare Diskussionen über Stoffauswahl und Präsentation; und Generationen von Studierenden früherer Jahrgänge, insbesondere Samuel Belko, Oliver Kasper, Dennis Sander und Johanna Schneeberger, für zahlreiche Korrekturen und Verbesserungen.

G.F., München, 24.4.2025

1 Konvergenz im $\mathbb{R}^n$ und daraus abgeleitete Konzepte

Die Betrachtung von Grenzwerten ist für das Verständnis kontinuierlicher Funktionen und Prozesse unverzichtbar. Dieses Kapitel ist eine bewusst elementar gehaltene Einführung in den Grundbegriff des Grenzwertes und damit zusammenhängende Konzepte wie "stetig" im n-dimensionalen Raum $\mathbb{R}^n$; unser Ziel ist die Bereitstellung des nötigen Hintergrundwissens für das zentrale Thema der Ableitung im folgenden Kapitel.

1.1 Der n-dimensionale Raum $\mathbb{R}^n$

Die Haupt-Arena der Analysis 2 ist der n-dimensionale Raum $\mathbb{R}^n$.

Def. 1.1 Für $n \in \mathbb{N}$ ist

$$\mathbb{R}^n = \left\{ x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \,:\, x_1, ..., x_n \in \mathbb{R} \right\},$$

d.h. $\mathbb{R}^n$ ist die Menge der Spaltenvektoren mit n reellen Komponenten. Die Zahl x_k heisst k-te Komponente von x.

In der Sprache der Mengenlehre ist der $\mathbb{R}^n$ also das kartesische Produkt $\mathbb{R} \times \cdots \times \mathbb{R}$ von n Kopien der Menge der reellen Zahlen; seine Elemente sind geordnete n-Tupel reeller Zahlen.

Algebraische Eigenschaften. $\mathbb{R}^n$ ist $\mathbb{R}$–Vektorraum unter komponentenweiser Addition und komponentenweiser Multiplikation mit $\lambda \in \mathbb{R}$, d.h. unter den Operationen

$$x + y := \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix}, \quad \lambda x := \begin{pmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{pmatrix}.$$

Der Nullvektor

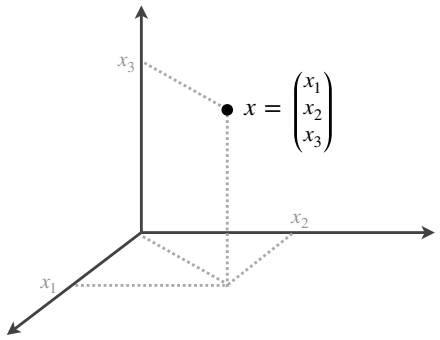
$$0_{\mathbb{R}^n} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

wird (sofern aus dem Kontext klar ist, dass nicht die Zahl 0 gemeint ist) ebenfalls mit 0 bezeichnet.

Visualisierung. Man kann sich Elemente des $\mathbb{R}^n$ auf verschiedene Weise vorstellen.

- **n=2:** $\mathbb{R}^2$ als Ebene; $x \in \mathbb{R}^2$ als Punkt oder Vektor in der Ebene

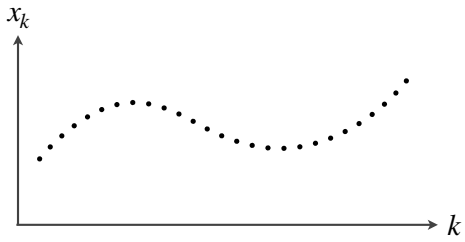
- **n=3:** $\mathbb{R}^3$ als 3D Raum; $x \in \mathbb{R}^3$ als Punkt oder Vektor im 3D Raum



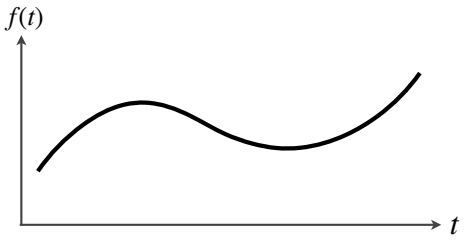
Punkt im $\mathbb{R}^3$

- **n groß:** $x \in \mathbb{R}^n$ als Punktgraph der Komponenten x_k als Funktion des Index k. Dies entspricht der Bedeutung vieler hochdimensionaler (Daten-)vektoren als Diskretisierung einer kontinuierlichen reellwertigen Funktion $f : [0, L] \to \mathbb{R}$,

$$x_k := f(k\,h) \;\; (k = 1, ..., n), \quad h = \frac{L}{n} \text{ (Gitterweite)}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$



Punkt im $\mathbb{R}^{27}$



Kontinuierliche Funktion

Verschiedene Zugänge zu Konvergenz im $\mathbb{R}^n$ in der Literatur. Im $\mathbb{R}^n$ sind nicht nur die beiden algebraischen Operationen Addition und Multiplikation mit Skalaren wichtig, sondern wir brauchen auch die analytische Operation der *Grenzwertbildung*. Dazu benötigen wir eine Verallgemeinerung des (auf dem Absolutbetrag beruhenden) *Abstandsbegriffes* reeller (bzw. komplexer) Zahlen auf den $\mathbb{R}^n$. Hierzu gibt es drei Zugänge steigender Allgemeinheit:

(A) Nimm den euklidischen (elementargeometrischen) Abstand und gehe dann wie in $\mathbb{R}$ oder $\mathbb{R}^2 \tilde{=} \mathbb{C}$ vor.

(B) Betrachte den $\mathbb{R}^n$ als Spezialfall eines Vektorraums; starte von einer beliebigen "Norm" auf diesem Vektorraum.

(C) Betrachte den $\mathbb{R}^n$ als Spezialfall einer Menge; starte von einer beliebigen "Metrik" auf dieser Menge.

Wir folgen in diesem Skript dem elementaren Zugang (A), weisen aber nach, dass er ein Spezialfall der Zugänge (B) und (C) ist.

1.2 Euklidischer Abstand

Ein elementargeomerisches Analogon des Absolutbetrages in $\mathbb{R}$ bzw. $\mathbb{R}^2 \tilde{=} \mathbb{C}$ ist die *euklidische Norm* eines Punktes $x \in \mathbb{R}^n$:

$$|x| := \sqrt{x_1^2 + ... + x_n^2} = \sqrt{\sum_{k=1}^{n} x_k^2}$$

Anschauliche Bedeutung von $|x|$ in n=2 und n=3 (siehe die Skizze auf der vorigen Seite): elementargeometrischer Abstand des Punktes x von 0 (= Länge des Vektors x). Dies lässt sich aus dem Satz des Pythagoras herleiten. Dementsprechend ist der *euklidische Abstand*

$$|x - y| = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2} \tag{1}$$

der elementargeometrische Abstand der Punkte x und y (=Länge des Vektors von x nach y).

Die euklidische Norm ist ein Spezialfall des folgenden allgemeineren Konzeptes (siehe (B)):

Def 1.2 (Norm) Sei X ein $\mathbb{K}$-Vektorraum, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$. Eine **Norm** auf X ist eine Abbildung $X \to \mathbb{R}$, $x \mapsto ||x||$, sodass gilt:
(i) Positivität: $||x|| \geq 0$, "=" $\Longleftrightarrow x = 0$
(ii) Homogenität: $||\lambda x|| = |\lambda|\,||x||$ für alle $\lambda \in \mathbb{K}$, $x \in X$
(iii) Dreiecksungleichung: $||x + y|| \leq ||x|| + ||y||$ für alle $x, y \in X$

In Bedingung (ii) bezeichnet $|\lambda|$ den Absolutbetrag der reellen oder komplexen Zahl λ.

Lemma 1.1 Die euklidische Norm $x \mapsto |x|$ ist eine Norm auf dem $\mathbb{R}$-Vektorraum $\mathbb{R}^n$ (d.h. sie besitzt die Eigenschaften (i), (ii), (iii)).

Der Beweis von (i) und (ii) ist trivial. Der Beweis von (iii) benutzt die folgende Definition und das nachfolgende Lemma.

Def. 1.4 (Skalarprodukt) Das *euklidische Skalarprodukt* zweier Vektoren $x, y \in \mathbb{R}^n$ ist

$$\langle x, y \rangle = \sum_{k=1}^{n} x_k y_k.$$

Die euklidische Norm kann mithilfe des Skalarproduktes ausgedrückt werden:

$$|x|^2 = \sum_{k=1}^{n} x_k^2 = \langle x, x\rangle, \quad \text{und folglich } |x| = \sqrt{\langle x, x\rangle}.$$

Lemma 1.2 (Cauchy-Schwarz-Ungleichung) Für $x, y \in \mathbb{R}^n$ gilt

$$|\langle x, y\rangle| \le |x|\,|y|.$$

Beweis der Cauchy-Schwarz-Ungleichung: O.B.d.A. sei $y \neq 0$, sonst ist die Behauptung trivial. Für beliebiges $\lambda \in \mathbb{R}$ gilt

$$\begin{aligned} 0 &\le |x - \lambda y|^2 = \langle x - \lambda y, x - \lambda y\rangle \\ &= \langle x, x\rangle - \lambda\langle x, y\rangle - \lambda\langle y, x\rangle + \lambda^2\langle y, y\rangle \\ &= |x|^2 - 2\lambda\langle x, y\rangle + \lambda^2|y|^2 =: f(\lambda). \end{aligned}$$

Um eine besonders "gute" Ungleichung zu erhalten, minimieren wir die rechte Seite über $\lambda \in \mathbb{R}$. An der Minimumsstelle muss $0 = f'(\lambda) = -2\langle x, y\rangle + 2\lambda|y|^2$ gelten, d.h. $\lambda = \frac{\langle x,y\rangle}{|y|^2}$. Einsetzen liefert

$$0 \le |x|^2 - \frac{2\langle x, y\rangle^2}{|y|^2} + \frac{\langle x, y\rangle^2}{|y|^4}|y|^2 = |x|^2 - \frac{\langle x, y\rangle^2}{|y|^2}.$$

Indem wir den rechten Term auf die linke Seite bringen und beide Seiten mit $|y|^2$ multiplizieren, folgt die Behauptung.

Beweis der Dreiecksungleichung:

$$|x + y|^2 = |x|^2 + 2\langle x, y\rangle + |y|^2 \underset{\text{Cauchy-Schwarz}}{\le} |x|^2 + 2|x|\,|y| + |y|^2 = \big(|x| + |y|\big)^2.$$

Die euklidische Norm ist nicht die einzige Norm auf dem $\mathbb{R}^n$, aber die wichtigste. Andere Normen – sowie die Tatsache, dass es für viele Zwecke egal ist, welche Norm man verwendet – lernen wir in Kapitel 3 kennen.

Analog zum euklidischen Abstand (1) liefert jede Norm einen Abstand $\|x - y\|$ zweier Punkte. Dies ist ein Spezialfall des folgenden allgemeineren Konzeptes (siehe (C)).

Def. 1.3 (Metrik) Sei X eine Menge. Eine **Metrik** auf X ist eine Abbildung $d : X \times X \to \mathbb{R}$ sodass gilt:
(i) Positivität: $d(x, y) \ge 0$, "=" $\Longleftrightarrow x = y$
(ii) Symmetrie: $d(x, y) = d(y, x)$
(iii) Dreiecksungleichung: $d(x, y) \le d(x, z) + d(z, y)$ für alle $x, y, z \in X$.

Lemma 1.3 Sei X ein $\mathbb{K}$-Vektorraum, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$, und sei $\|\cdot\|$ eine Norm auf X. Dann ist die Abbildung $d : X \times X \to \mathbb{R}$,

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

eine Metrik auf X.

Insbesondere ist also der euklidische Abstand (1) eine Metrik auf dem $\mathbb{R}^n$.

Beweis Die Positivität folgt sofort aus der Positivität der Norm. Die Symmetrie folgt aus der Homogenität der Norm, denn

$$d(y, x) = \|y - x\| = \|(-1)(x - y)\| = |-1|\,\|x - y\| = \|x - y\| = d(x, y).$$

Die Dreiecksungleichung folgt aus $x-y = (x-z)+(z-y)$ und der Dreiecksungleichung für die Norm.

In Analysis 2 sind die uns interessierenden Objekte in natürlicher Weise Elemente von Vektorräumen; deshalb benötigen wir den Begriff der Metrik nicht.

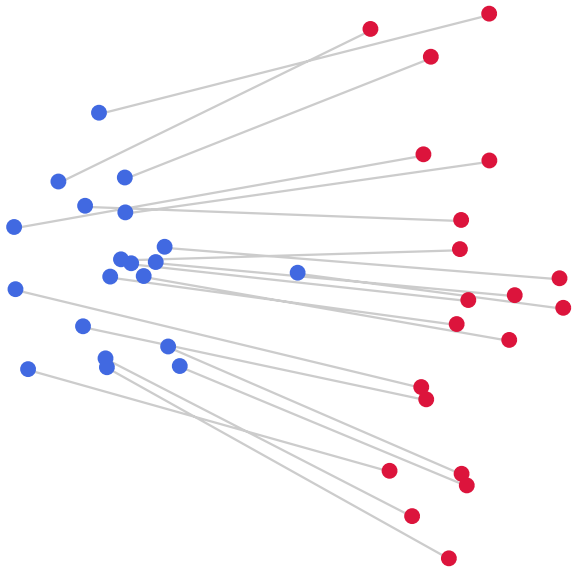
Beispiel einer Metrik: Wasserstein-Distanz zwischen Teilmengen des $\mathbb{R}^n$ Seien A und B zwei N-elementige Teilmengen des $\mathbb{R}^n$,

$$A = \{x_1, ..., x_N\}, \quad B = \{y_1, ..., y_N\},$$

jeweils bestehend aus N voneinander verschiedenen Punkten $x_1, ..., x_N \in \mathbb{R}^n$ und $y_1, ..., y_N \in \mathbb{R}^n$. Mengentheoretisch betrachten wir also – statt *geordenten* N-tupeln $(x_1, ..., x_N) \in \mathbb{R}^{nN}$ – *ungeordnete* N-tupel. Wie können wir einen sinnvollen Abstand definieren? Das geht z.B. mithilfe des euklidischen Abstandes zweier Punke x_i und y_j und Optimierung über die Zuordnung (siehe Schaubild):

$$d(A, B) = \min_{\sigma:\{1,...,N\}\to\{1,...,N\} \text{ Permutation}} \Big(\frac{1}{N} \sum_{i=1}^{N} |x_i - y_{\sigma(i)}|^2 \Big)^{1/2}.$$

Dies ist eine Metrik auf der Menge der N-elementigen Teilmengen (Beweis: siehe Übungen), kommt aber nicht von einer Norm, denn unsere Menge ist kein Vektorraum: um die Differenz von A und B zu definieren, bräuchte man eine *geordnete, d.h. nummerierte* statt einer ungeordneten Liste von Punkten. Hätte man die Punkte B schon (bzgl. A) korrekt nummeriert, entspräche d (bis auf den unwesentlichen Normierungsfaktor $\frac{1}{N}$ vor der Summe) dem euklidischen Abstand zwischen $(x_1, ..., x_N) \in \mathbb{R}^{n\cdot N}$ und $(y_1, ..., y_N) \in \mathbb{R}^{n\cdot N}$; aber die korrekte Nummerierung hängt von der Menge A ab. Die Metrik heisst *Wasserstein-Metrik*, und kann auf endliche Punktmengen verschiedener Grösse und sogar beliebige Wahrscheinlichkeitsverteilungen auf dem $\mathbb{R}^n$ verallgemeinert werden. Sie besitzt übrigens nützliche Anwendungen (z.B. in den Gebieten partielle Differentialgleichungen, Wahrscheinlichkeitstheorie, maschinelles Lernen). Siehe die regelmässig angebotene Spezialvorlesung *Optimal transport*.



Wasserstein-Distanz zwischen Punktmengen im $\mathbb{R}^2$. Die grauen Linien zeigen die optimale Zuordnung. Die Wasserstein-Distenz zum Quadrat ist der Mittelwert der Quadrate der Längen der grauen Linien.

1.3 Konvergenz

Wir kommen nun zum Grenzwertbegriff im $\mathbb{R}^n$. Eine **Folge** im $\mathbb{R}^n$ ist eine Abbildung $\mathbb{N} \to \mathbb{R}^n$, Schreibweise: $(a^{(j)})_{j\in\mathbb{N}}$, $a^{(j)} \in \mathbb{R}^n$ j-tes Folgenglied. Wir schreiben den Folgenindex oben statt unten, um ihn vom Komponentenindex (x_k = k^{te} Komponente des Vektors x) zu unterscheiden, und in Klammern eingehüllt, um ihn von der j-ten Potenz zu unterscheiden. D.h. $a_k^{(j)}$ bedeutet die k-te Komponente des j-ten Folgengliedes.

Def. 1.5 (Konvergenz) Eine Folge $(a^{(j)})_{j\in\mathbb{N}}$ im $\mathbb{R}^n$ heißt **konvergent** gegen $a \in \mathbb{R}^n$, Schreibweise: $a^{(j)} \to a$ oder $\lim_{j\to\infty} a^{(j)} = a$, wenn zu jedem $\epsilon > 0$ ein $N \in \mathbb{N}$ existiert mit $|a^{(j)} - a| < \epsilon$ für alle $j \geq N$. Der Punkt a heisst **Grenzwert** der Folge.

Beachte: in obiger Definition bezeichnet $|\cdot|$ die euklidische Norm. Ersetzt man in der Definition $\mathbb{R}^n$ durch einen beliebigen normierten Vektorraum und die euklidische Norm durch eine beliebige Norm, erhält man die Definition von "konvergent" und "Grenzwert" in normierten Vektorräumen. Dies sowie interessante Beispiele hierzu (z.B. Konvergenz im unendlichdimensionalen Vektorraum der stetigen Funktionen auf einem Intervall) werden ausführlich in Kapitel 3 besprochen.

Triviale aber nützliche Folgerung aus der Definition: $a^{(j)} \to a \iff |a^{(j)} - a| \to 0$. Auf der rechten Seite tritt nur noch Konvergenz in $\mathbb{R}$ auf.

Lemma 1.4 (Konvergenz = komponentenweise Konvergenz)
Sei $(a^{(j)})_{j\in\mathbb{N}}$ Folge in $\mathbb{R}^n$. Dann sind äquivalent:
(i) $(a^{(j)})$ konvergent (im $\mathbb{R}^n$)
(ii) Die Komponentenfolgen $(a_k^{(j)})_{j\in\mathbb{N}}$, $k = 1, ..., n$, sind konvergent (in $\mathbb{R}$).
Falls (i) oder (ii) gelten, folgt

$$\lim_{j\to\infty} \underbrace{\begin{pmatrix} a_1^{(j)} \\ \vdots \\ a_n^{(j)} \end{pmatrix}}_{=a^{(j)}} = \begin{pmatrix} \lim_{j\to\infty} a_1^{(j)} \\ \vdots \\ \lim_{j\to\infty} a_n^{(j)} \end{pmatrix}.$$

Die analytische Operation der Grenzwertbildung verhält sich also analog zu den algebraischen Operationen der Addition und der Multiplikation mit Skalaren: alle drei Operationen können komponentenweise ausgeführt werden.

Beweis Konvergenz ist äquivalent zu $|a^{(j)} - a| \to 0$ und komponentenweise Konvergenz zu $\max_{k\in\{1,...,n\}} |a_k^{(j)} - a_k| \to 0$. Es gilt aber

$$\max_{k\in\{1,...,n\}} |x_k| \;\le\; \underbrace{\sqrt{\sum_{k=1}^{n} x_k^2}}_{=|x|} \;\le\; \sqrt{n} \max_{k\in\{1,...,n\}} |x_k|. \qquad (*)$$

Die erste Ungleichung angewandt auf $x = a^{(j)} - a$ liefert (i)$\Longrightarrow$(ii), und die zweite die umgekehrte Implikation.

Beispiel

$$\lim_{j\to\infty} \begin{pmatrix} \frac{1}{j} \\ 1 + \frac{1}{j^2} \end{pmatrix} = \begin{pmatrix} \lim_{j\to\infty} \frac{1}{j} \\ \lim_{j\to\infty} (1 + \frac{1}{j^2}) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Die Äquivalenz von Konvergenz und komponentenweiser Konvergenz erlaubt die sofortige Übertragung grundlegender Eigenschaften von Grenzwerten in $\mathbb{R}$ auf Grenzwerte in $\mathbb{R}^n$, durch komponentenweise Anwendung der entsprechenden Sätze aus Analysis 1:

Korollar 1.1 Grenzwerte konvergenter Folgen im $\mathbb{R}^n$ sind eindeutig. Folgen im $\mathbb{R}^n$ sind Cauchyfolgen genau dann, wenn sie konvergent sind. Es gilt der "Kalkül der Grenzwerte" für die algebraischen Operationen im $\mathbb{R}^n$, d.h.

$$x^{(j)} \to x,\ y^{(j)} \to y \Longrightarrow x^{(j)} \pm y^{(j)} \to x \pm y;$$
$$\lambda^{(j)} \to \lambda \text{ (in } \mathbb{R}),\ x^{(j)} \to x \text{ (in } \mathbb{R}^n) \Longrightarrow \lambda^{(j)} x^{(j)} \to \lambda x \text{ (in } \mathbb{R}^n).$$

Def. 1.6 x_* heisst **Häufungspunkt** einer Folge $(x^{(j)})_{j\in\mathbb{N}}$ im $\mathbb{R}^n$, wenn es eine Teilfolge $(x^{(j_\ell)})_{\ell\in\mathbb{N}}$, $j_1 < j_2 < ...$, $j_\ell \in \mathbb{N}$, gibt sodass $x^{(j_\ell)} \to x_*$. Eine Folge $(x^{(j)})_{j\in\mathbb{N}}$ im $\mathbb{R}^n$ heisst **beschränkt**, wenn $C > 0$ existiert sodass $|x^{(j)}| \le C$ für alle j.

1.4 Satz von Bolzano-Weierstrass

Eine zentrale Eigenschaft des $\mathbb{R}^n$ ist:

Satz 1.1 (Satz von Bolzano-Weierstrass) *Jede beschränkte Folge im $\mathbb{R}^n$ besitzt einen Häufungspunkt.*

Diese Eigenschaft beruht sowohl auf der Vollständigkeit von $\mathbb{R}$ (die entsprechende Aussage in $\mathbb{Q}$ ist falsch) als auch der Endlichdimensionalität des Vektorraums $\mathbb{R}^n$ (siehe Kapitel 6 für ein Gegenbeispiel im Unendlichdimensionalen).

Beweis: Durch wiederholte Anwendung des Satzes von Bolzano-Weierstrass in $\mathbb{R}$ (Analysis 1 Satz 2.6).

Genauer: Sei $(x^{(j)})$ beschränkte Folge im $\mathbb{R}^n$. Wegen der ersten Ungleichung in (*) sind alle Komponentenfolgen $(x^{(j)}{}_k)_{j\in\mathbb{N}}$ beschänkte Folgen in $\mathbb{R}$. Nach Satz von Bolzano-Weierstrass in $\mathbb{R}$ gibt es eine Teilfolge $(x^{(j_\ell)})_{\ell\in\mathbb{N}}$ und ein $x_1^* \in \mathbb{R}$ sodass

$$x_1^{(j_\ell)} \to x_1^* \text{ in } \mathbb{R}.$$

Durch nochmalige Anwendung dieses Satzes finden wir eine Teilfolge $(x^{(j'_\ell)})_{\ell\in\mathbb{N}}$ dieser Teilfolge und ein $x_2^* \in \mathbb{R}$ sodass zusätzlich

$$x_2^{(j'_\ell)} \to x_2^* \text{ in } \mathbb{R}.$$

Nach n Schritten erhalten wir eine Teilfolge $(x^{(\tilde{j}_\ell)})_{\ell\in\mathbb{N}}$ sodass $x_k^{(\tilde{j}_\ell)} \to x_k^*$ in $\mathbb{R}$ für alle k.

1.5 Offene und abgeschlossene Mengen; Abschluss, Inneres, Rand

In Analysis 2 werden gewisse Teilmengen des $\mathbb{R}^n$ eine wichtige Rolle spielen: "abgeschlossene Mengen", auf denen wir Maximums- und Minimumsprobleme untersuchen werden, und "offene Mengen", auf denen man Differentialrechnung betreiben kann. Die präzisen Definitionen lauten wie folgt.

Def. 1.7 (Offene und abgeschlossene Mengen)
a) $\Omega \subseteq \mathbb{R}^n$ heisst **offen**, wenn zu jedem $x_0 \in \Omega$ ein $\varepsilon > 0$ existiert, sodaß $B_\varepsilon(x_0) \subseteq \Omega$. Hierbei ist

$$B_\varepsilon(x_0) = \{y \in \mathbb{R}^n : |y - x_0| < \varepsilon\}.$$

b) $A \subseteq \mathbb{R}^n$ heisst **abgeschlossen**, wenn für jede Folge $(x^{(j)})_{j\in\mathbb{N}}$ mit $x^{(j)} \in A$ für alle j und $x^{(j)} \to x \in \mathbb{R}^n$ gilt: $x \in A$.

Abgeschlossen bedeutet also *abgeschlossen unter der Operation der Grenzwertbildung*. Der folgende nichttriviale Zusammenhang besteht zwischen diesen beiden Begriffen:

Satz 1.2 *Eine Menge $A \subseteq \mathbb{R}^n$ ist genau dann abgeschlossen, wenn ihr Komplement $\mathbb{R}^n \backslash A$ offen ist.*

Die zweite Eigenschaft kann als alternative Definition von abgeschlossen benutzt werden.

Beweis "abgeschlossen $\Longrightarrow$ Komplement offen": Sei A abgeschlossen. Angenommen das Komplement $\mathbb{R}^n \backslash A$ wäre nicht offen. Dann gibt es einen Punkt $y \in \mathbb{R}^n \backslash A$ sodass keine Kugel um y in $\mathbb{R}^n \backslash A$ liegt. Also enthält jede Kugel $B_{1/j}(y)$, $j = 1, 2, ...$, einen Punkt $x^{(j)} \in A$. Wegen $|y - x^{(j)}| \leq 1/j$ konvergiert $x^{(j)}$ gegen y. Widerspruch zur Abgeschlossenheit von A.

"Komplement offen $\Longrightarrow$ abgeschlossen": Sei $\mathbb{R}^n \backslash A$ offen, $x^{(j)} \in A$, $x^{(j)} \to y$. Wir müssen zeigen: $y \in A$. Angenommen y läge in $\mathbb{R}^n \backslash A$. Da $\mathbb{R}^n \backslash A$ offen, existiert eine Kugel $B_r(y)$, $r > 0$, sodass $B_r(y) \subseteq \mathbb{R}^n \backslash A$. Also ist $|x^{(j)} - y| \geq r$ für alle j. Widerspruch zur Konvergenz $x^{(j)} \to y$.

Beispiele: 1) $B_r(x_0) := \{x \in \mathbb{R}^n : |x - x_0| < r\}$ ist offen, und heisst **offene Kugel um x_0 vom Radius r.**

Beweis der Offenheit: Sei $x \in B_r(x_0)$. Dann ist $\varepsilon := r - |x - x_0| > 0$, und jedes $y \in B_\varepsilon(x)$ liegt wegen Dreiecksungleichung

$$|y - x_0| = |(y - x) + (x - x_0)| \leq |y - x| + |x - x_0| < \varepsilon + |x - x_0| = r$$

auch in $B_r(x_0)$.

2) $\overline{B_r(x_0)} := \{x \in \mathbb{R}^n : |x - x_0| \leq r\}$ ist abgeschlossen, und heisst **abgeschlossene Kugel um x_0 vom Radius r.**

"Fussgänger"-Beweis (für einen eleganteren Beweis auf der Basis von mehr Theorie siehe Abschnitt 1.8): Sei $x^{(j)} \to x$, $x^{(j)} \in \overline{B_r(x_0)}$ für alle j. Nach Lemma 1.3 konvergiert jede Komponente von $x^{(j)}$ gegen die entsprechende Komponente von x und folglich (wegen der aus Analysis 1 bekannten Stetigkeit der eindimensionalen Funktionen $t \mapsto t^2$ und $t \mapsto \sqrt{t}$ $(t \geq 0)$)

$$|x^{(j)} - x_0| = \sqrt{(x^{(j)} - x_0)_1^2 + ... + (x^{(j)} - x_0)_n^2} \longrightarrow \sqrt{(x - x_0)_1^2 + ... + (x - x_0)_n^2} = |x - x_0|.$$

Die linke Seite ist $\leq r$, also wegen Kalkül der Grenzwerte (Analysis 1 Satz 2.2) auch die rechte Seite. Folglich $x \in \overline{B_r(x_0)}$.

3) $(0, 1) \times [0, 1]$ ist weder offen noch abgeschlossen.

Weitere Beispiele siehe Übungen und Abschnitt 1.8.

Jeder beliebigen Teilmenge $A \subseteq \mathbb{R}^n$ kann man auf natürliche Weise eine "etwas kleinere" offene Menge und eine "etwas grössere" abgeschlossene Menge zuordnen. Die Differenzmenge heisst Rand von A. Der Rand von Mengen im $\mathbb{R}^n$ spielt dieselbe Rolle wie die Intervallgrenzen bei Intervallen. Zum Beispiel:
– bei Maximierungs- und Minimierungsproblemen muss man das Verhalten einer Funktion auf dem Rand getrennt untersuchen (siehe §2.7)

– bei Differentialgleichungen für Funktionen mehrerer Veränderlicher muss man typischerweise die Funktionswerte auf dem Rand vorgeben (siehe §2.8.3)
– Integrale etlicher Funktionen über mehrdimensonale Mengen lassen sich mithilfe der Randwerte ausdrücken (dies ist Gegenstand des Gauss'schen Satzes, der den Hauptsatz verallgemeinert und den Sie in Analysis 3 kennenlernen).

Def. 1.8 (Abschluss, Inneres, Rand) Sei $A \subseteq \mathbb{R}^n$ beliebig. Wir definieren:

$$\begin{aligned} int\, A &:= \{x \in A : \exists \varepsilon > 0 : B_\varepsilon(x) \subseteq A\} \text{ (**Inneres** von } A) \\ \overline{A} &:= \{x \in \mathbb{R}^n : \exists \text{ Folge } (x^{(j)})_{j=1}^{\infty} : x^{(j)} \in A,\, x^{(j)} \to x\} \text{ (**Abschluss** von } A) \\ \partial A &:= \overline{A} \setminus int\, A \text{ (**Rand** von } A). \end{aligned}$$

Offensichtlich gilt

$$int A \subseteq A \subseteq \overline{A}$$

sowie

$$A = int\, A \Longleftrightarrow A \text{ offen}, \quad A = \overline{A} \Longleftrightarrow A \text{ abgeschlossen}.$$

Des weiteren gilt $int\, A$ offen, $\overline{A}$ abgeschlossen. Ersteres folgt aus der Offenheit von $B_\varepsilon(x)$ (siehe Beispiel 1)) und letzteres aus der Äquivalenz A abgeschlossen $\Longleftrightarrow$ $\mathbb{R}^n \setminus A$ offen.

Beispiele:

1) Sei $A = (a, b)$ (offenes Intervall) oder $[a, b]$ (abgeschlossenes Intervall). In beiden Fällen ist $int\, A = (a, b)$, $\overline{A} = [a, b]$, $\partial A = \{a, b\}$. Der Rand entspricht also den Intervallgrenzen.

2) Sei $A = (0, 1) \times (0, 1)$. Dann folgt $int\, A = (0, 1) \times (0, 1)$, $\overline{A} = [0, 1] \times [0, 1]$, ∂A =Vereinigungsmenge der vier Strecken

$$[\binom{0}{0}, \binom{1}{0}] \text{ (unterer Rand)}, \; [\binom{1}{0}, \binom{1}{1}] \text{ (rechter Rand)}, \; [\binom{0}{1}, \binom{1}{1}] \text{ (oberer Rand)}, \; [\binom{0}{0}, \binom{0}{1}] \text{ (linker Rand)},$$

wobei $[x, y] = \{(1 - t)x + ty : t \in [0, 1]\}$ die Strecke von x nach y bezeichnet.

3) $A' = [0, 1] \times [0, 1]$: Inneres, Abschluss und Rand von A' sind identisch mit denen von A, obwohl $A' \neq A$.

4) Der Abschluss der offenen Kugel um x_0 vom Radius r, $B_r(x_0) = \{x \in \mathbb{R}^n : |x - x_0| < r\}$, ist die weiter oben elementar definierte abgeschlossene Kugel um x_0 vom Radius r, $\overline{B_r(x_0)} = \{x \in \mathbb{R}^n : |x - x_0| \leq r\}$. Umgekehrt ist die offene Kugel das Innere der abgeschlossenen Kugel. Der Rand beider Kugeln ist die **Sphäre um x_0 vom Radius** r, $S_r(x_0) := \{x \in \mathbb{R}^n : |x - x_0| = r\}$.

5) Für nicht so anschauliche Mengen sind Abschluss, Inneres und Rand ebenfalls nicht so anschaulich, z.B. gilt $\overline{\mathbb{Q}} = \mathbb{R}$, $int\, \mathbb{Q} = \varnothing$, $\partial \mathbb{Q} = \mathbb{R}$.

1.6 Funktionen im Mehrdimensionalen: Beispielklassen, Visualisierung

In Analysis 2 gilt unser Hauptinteresse Funktionen, die auf einer Teilmenge des $\mathbb{R}^n$ definiert sind und deren Werte im $\mathbb{R}^m$ liegen. Solche Funktionen bilden also Vektoren auf Vektoren ab. Die Dimensionen von Definitions- und Wertebereich können unterschiedlich sein.

Der *Graph* einer solchen Funktion $f : M \subseteq \mathbb{R}^n \to \mathbb{R}^m$,

$$\operatorname{graph} f = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m : y = f(x)\},$$

ist eine Teilmenge des (ziemlich hochdimensionalen) Raumes $\mathbb{R}^n \times \mathbb{R}^m$. Das sieht zunächst einmal unanschaulich aus – besonders im Vergleich zur aus Analysis 1 gewohnten Situation $f : M \subseteq \mathbb{R} \to \mathbb{R}$, bei der der Graph eine Teilmenge der Ebene $\mathbb{R}^2$ ist.

Bevor wir in die Theorieentwicklung einsteigen, verschaffen wir uns einen Überblick, wie man sich solche Funktionen – zumindest für wichtige Beispielklassen – anschaulich vorstellen kann.

1.6.1 Skalare Funktionen

Def. 1.9 Eine **skalare Funktion** ist eine Funktion mit 1D Wertebereich, d.h. $f : M \subseteq \mathbb{R}^n \to \mathbb{R}$.

n=2: Die grundlegenden Visualisierungsmöglichkeiten für $f : M \subseteq \mathbb{R}^2 \to \mathbb{R}$ sind:

a) Niveaulinien (oder Höhenlinien) im $\mathbb{R}^2$ skizzieren
b) Graph als Fläche im $\mathbb{R}^3$ zeichnen
c) Heatmap

Hierbei ist

$$\operatorname{graph} f = \left\{ \begin{pmatrix} x \\ y \\ f(x,y) \end{pmatrix} : \begin{pmatrix} x \\ y \end{pmatrix} \in M \right\},$$

$$\text{Niveaulinie zum Wert } c \in \mathbb{R} \ = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in M : f(x,y) = c \right\}.$$

Die Heatmap ist nur per Computer geeignet, und entspricht einem lückenlosen Kontinuum farb-codierter Niveaulinien.

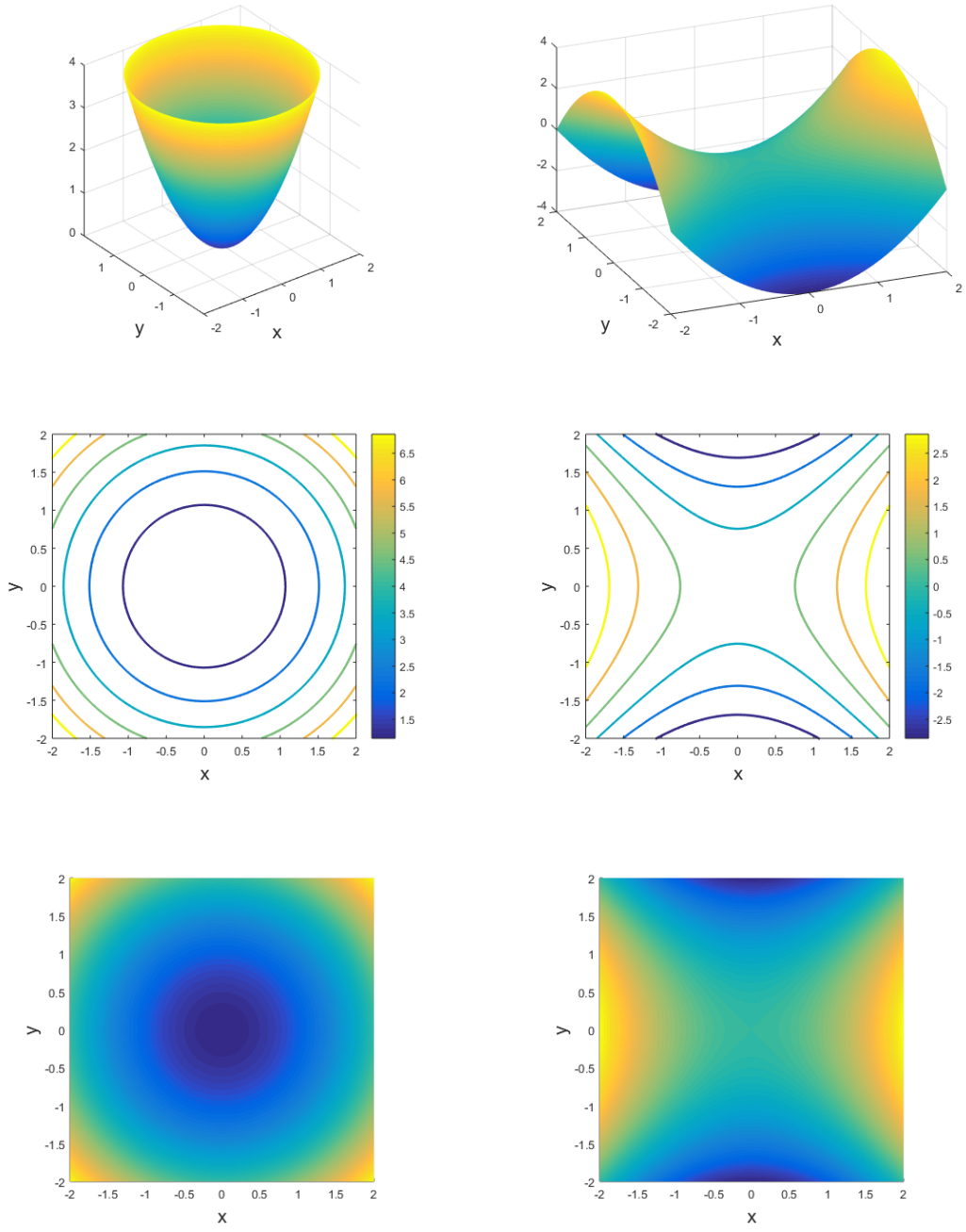
Beispiele (Skizzieren von Graph und Niveaulinien per Hand ist eine gute Übung):
1) $f(x, y) = x^2 + y^2$. Die Niveaulinie zum Wert 0 ist ein Punkt. Die Niveaulinien zu positiven Werten c sind Kreislinien vom Radius $\sqrt{c}$. Niveaulinien zu äquidistant wachsenden Werten, z.B. $c = 1$, $c = 2$, $c = 3$, ..., liegen näher und näher beieinander;

dies zeigt an, dass die Funktion nach aussen steiler und steiler wird. Der Graph entspricht einer "Schüssel".

2) $f(x,y) = x^2 - y^2$. Die Niveaulinie zum Wert 0 ist ein Kreuz bestehend aus den 2 sich im Ursprung schneidenden Geraden $y = \pm x$. Die Niveaulinie zum Wert 1 besteht aus den beiden (als Graphen über der y-Achse darstellbaren) Hyperbelästen $x = \pm\sqrt{1+y^2}$. Die Niveaulinie zum Wert -1 besteht aus den beiden (als Graphen über der x-Achse darstellenbaren) Hyperbelästen $y = \pm\sqrt{1+x^2}$. Der Graph entspricht einem "Sattel", oder einem Kartoffelchip der Marke "Pringles".

n=3: Visualisierungsmöglichkeit für $f : M \subseteq \mathbb{R}^3 \to \mathbb{R}$:
Niveauflächen $\{x \in \mathbb{R}^3 : f(x) = c\}$ als Flächen im $\mathbb{R}^3$ skizzieren.

Es lohnt sich einzuüben, solche Skizzen grob aber qualitativ korrekt per Hand zu zeichnen. Das hilft Ihnen, sich mehrdimensionale Funktionen im Kopf anschaulich vorzustellen, und ist eine wertvolle Quelle mathematischer Erkenntnisse.



Die Funktionen $f(x, y) = x^2 + y^2$ (links) und $f(x, y) = x^2 - y^2$ (rechts).
Oben: *Graph*. Mitte: *Niveaulinien*. Unten: *Heatmap*.

1.6.2 Kurven

Def. 1.10 Eine **Kurve** ist eine Abbildung mit 1D Definitionsbereich, d.h. $f : M \subseteq \mathbb{R} \to \mathbb{R}^n$.

Typischerweise interessiert man sich für den Fall M Intervall, f stetig (was stetig bedeutet, besprechen wir im nächsten Abschnitt). Manche Autoren fordern diese Zusatzeigenschaften bereits in der Definition; das ist Geschmackssache.

n=2: Visualisierungsmöglichkeiten für $f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} : M \subseteq \mathbb{R} \to \mathbb{R}^2$ sind:

a) Graph als Kurve im $\mathbb{R}^3$ skizzieren
b) Bild als Menge im $\mathbb{R}^2$ zeichnen.
Hierbei ist

$$\operatorname{graph} f = \left\{ \begin{pmatrix} t \\ f_1(t) \\ f_2(t) \end{pmatrix} : t \in M \right\},$$

$$\operatorname{Bild} f = \left\{ \begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix} : t \in M \right\}.$$

Für das Bild einer Kurve ist auch die alternative Bezeichnung *Spur* der Kurve üblich.

n=3: Visualisierungsmöglichkeit für $f : M \subseteq \mathbb{R} \to \mathbb{R}^3$:
Bild als Menge im $\mathbb{R}^3$ zeichnen.

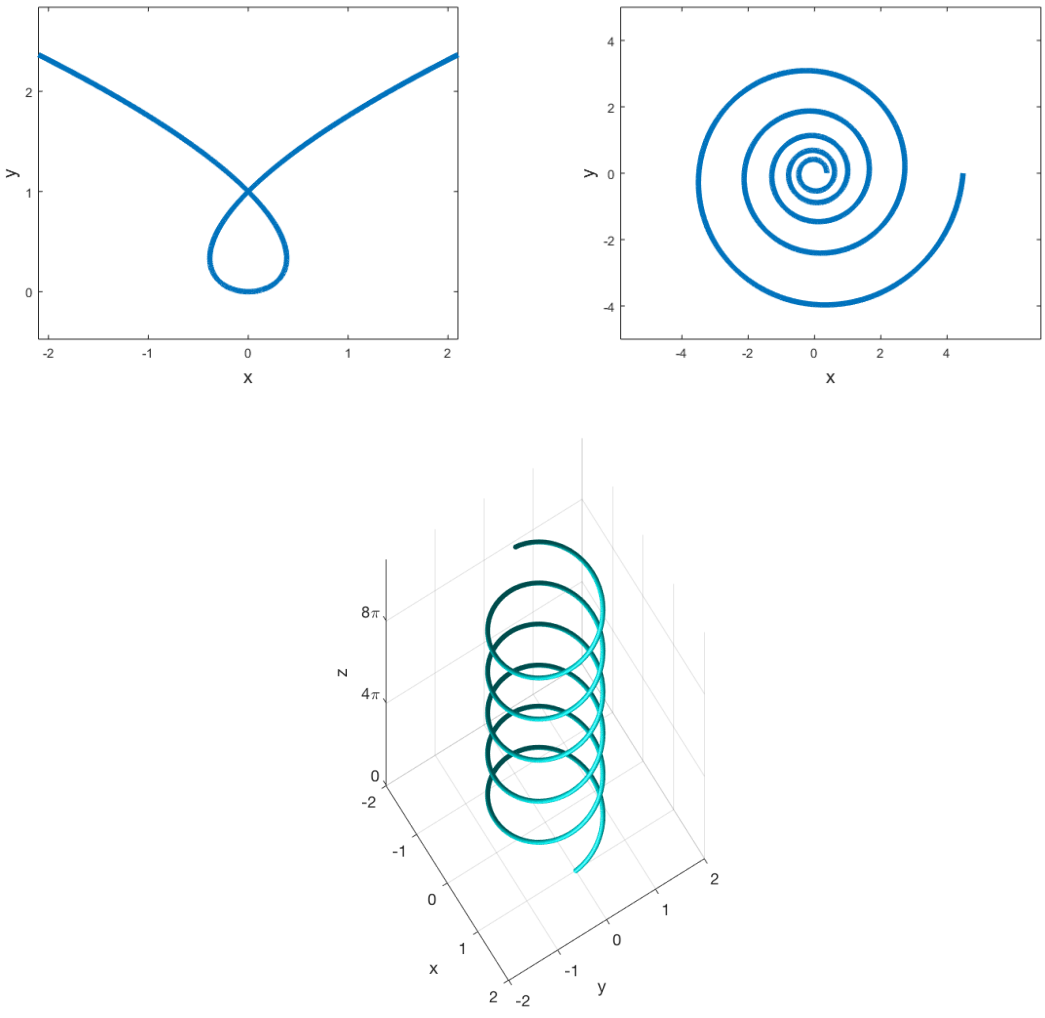
Beispiele (Skizzen per Hand anfertigen ist eine gute Übung):

1) $f(t) = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$, $t \in [0, \infty)$: der Graph ist eine Helix, und das Bild eine Kreislinie.
Beim Bild geht die Information verloren, für welches t der jeweilige Wert im $\mathbb{R}^2$ angenommen wird. Partielle Abhilfe: bei geeigneten Punkten das zugehörige t dazuschreiben; "Laufrichtung" durch Pfeil anzeigen.

2) $f(t) = \begin{pmatrix} t^3 - t \\ t^2 \end{pmatrix}$, $t \in \mathbb{R}$: das Bild ist eine "Schlaufe".

3) $f(t) = \begin{pmatrix} e^{t/2} \cos(2\pi t) \\ e^{t/2} \sin(2\pi t) \end{pmatrix}$, $t \in [-3, 3]$: das Bild ist eine Spirale.

4) $f(t) = \begin{pmatrix} \cos t \\ \sin t \\ t \end{pmatrix}$, $t \in \mathbb{R}$: das Bild ist eine Helix.
Schränken wir den Definitionsbereich auf $[0, N\pi]$, $N \in \mathbb{N}$, ein, erhalten wir $N/2$ Umdrehungen. Das Schaubild auf der nächsten Seite entspricht $N = 11$, d.h. wir sehen $5\frac{1}{2}$ Umdrehungen.



Die Kurven aus Beispiel 2 (links), Beispiel 3 (rechts), und Beispiel 4 (unten).

1.6.3 Vektorfelder

Def. 1.11 Ein **Vektorfeld** ist eine Abbildung, deren Definitions- und Wertebereich dieselbe Dimension haben, d.h. $f : M \subseteq \mathbb{R}^n \to \mathbb{R}^n$.

Visualisierung in den Dimensionen **n=2** und **n=3**: als "Vektorfeld" zeichnen.

D.h. für eine geeignete Menge von Punkten $x \in \mathbb{R}^n$ (z.B. ein Gitter $h\mathbb{Z}^n \cap M$, $h > 0$ geeignet) zeichnet man jeweils den zugehörigen Wert $f(x) \in \mathbb{R}^n$ als Richtungsvektor (Pfeil) mit Fusspunkt x.

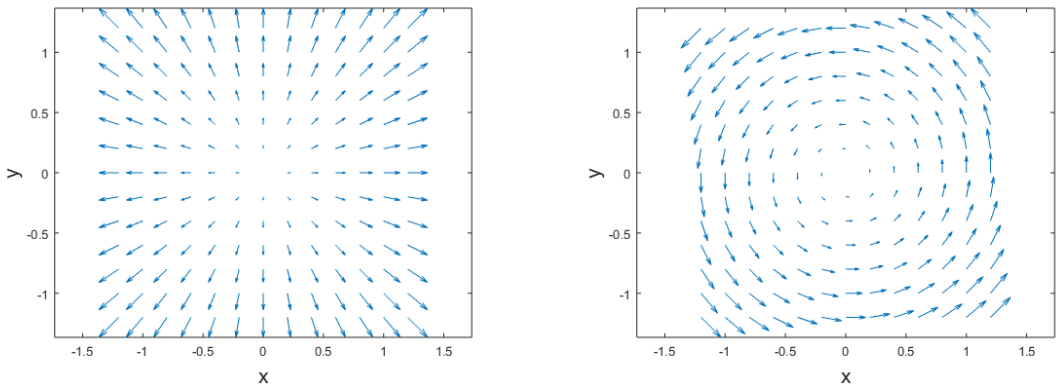
Physikalische Beispiele: Strömungsgeschwindigkeit eines Flusses; elektrische Felder und Magnetfelder; Windgeschwindigkeit auf einer Wetterkarte.

Mathematische Beispiele (Skizzen per Hand anfertigen ist eine gute Übung):

1) $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x,y) = \begin{pmatrix} x \\ y \end{pmatrix}$: das Vektorfeld ist eine "Quelle".

2) $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x,y) = \begin{pmatrix} -y \\ x \end{pmatrix}$: das Vektorfeld ist ein "Wirbel".

Die physikalischen Beispiele legen nahe, als Längeneinheit für die Richtungsvektoren nicht dieselbe Einheit wie für die Fusspunkte zu verwenden (Geschwindigkeiten kann man nicht in derselben Einheit messen wie Positionen). Wählen Sie auch bei mathematischen Beispielen die Längeneinheit so, dass die Vektoren kürzer sind als der Abstand zum nächsten Gitterpunkt, sonst wird's unübersichtlich!



Die Vektorfelder aus Beispiel 1 (links) und Beispiel 2 (rechts).

1.7 Stetigkeit

Informell heisst eine Funktion stetig, wenn hinreichend kleine Änderungen des Arguments nur kleine Änderungen des Funktionswerts bewirken. Dies kann man – wie im Eindimensionalen – entweder durch die Vertauschbarkeit von Funktionsanwendung und Grenzwertbildung oder durch das ε-δ-Kriterium präzise machen. Wir beginnen den Theorieaufbau für Funktionen im Mehrdimensionalen, indem wir den Begriff der Stetigkeit und einige Sätze aus Analysis 1 über stetige Funktionen (Verkettung stetiger Funktionen ist stetig; Satz vom Maximum und Minimum) auf den mehrdimensionalen Fall verallgemeinern.

Def. 1.12 Sei $\Omega \subseteq \mathbb{R}^n$, $f : \Omega \to \mathbb{R}^m$. f heißt **stetig** im Punkt $x \in \Omega$, wenn für jede Folge $(x^{(j)})_{j\in\mathbb{N}}$ mit $x^{(j)} \in \Omega$ gilt:

$$x^{(j)} \to x \implies f(x^{(j)}) \to f(x).$$

f heisst **stetig** (auf Ω), wenn f stetig in jedem Punkt $x \in \Omega$ ist.

Lemma 1.5 Sei $\Omega \subseteq \mathbb{R}^n$. Dann gilt:

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix} : \Omega \to \mathbb{R}^m \text{ ist stetig} \iff \begin{array}{c}\text{alle Komponentenfunktionen} \\ f_k : \Omega \to \mathbb{R} \text{ sind stetig.}\end{array}$$

Beweis: Nach Lemma 1.4 gilt $f(x^{(j)}) \to f(x)$ genau dann, wenn $f_k(x^{(j)}) \to f_k(x)$ für alle k.

Beispiele: 1) Konstante Funktionen $f : \mathbb{R}^n \to \mathbb{R}^m$ ($f(x) = y_0$ für ein $y_0 \in \mathbb{R}^m$ und alle x) sind stetig.

2) Lineare Abbildungen $L : \mathbb{R}^n \to \mathbb{R}^m$: diese sind gemäss Linearer Algebra von der Form

$$L(x) = Ax, \quad A = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \cdots & A_{mn} \end{pmatrix}, \quad Ax \text{ Matrix-Vektor-Produkt, d.h. } (Ax)_k = \sum_{\ell=1}^{n} A_{k\ell} x_\ell,$$

und daher wegen Lemma 1.4 stetig.

3) Monome $p : \mathbb{R}^n \to \mathbb{R}$, $p(x) = a_{i_1\ldots i_n} x_1^{i_1} \cdot \ldots \cdot x_n^{i_n}$, $(i_1, \ldots, i_n) \in (\mathbb{N} \cup \{0\})^n$, $a_{i_1\ldots i_n} \in \mathbb{R}$, sind stetig. Z.B. ist also $p(x, y) = 2\,x^2 y^3$ stetig.

4) Polynome $p : \mathbb{R}^n \to \mathbb{R}$, $p(x) = \sum_{(i_1,\ldots,i_n)\in\mathcal{S}} a_{i_1\ldots i_n} x_1^{i_1} \cdot \ldots \cdot x_n^{i_n}$, $\mathcal{S} \subset (\mathbb{N} \cup \{0\})^n$ endlich, $a_{i_1\ldots i_n} \in \mathbb{R}$, sind stetig. Z.B. ist also $p(x, y, z) = x^2 y^3 - 2xyz + 3$ stetig.

5) Sind $h : A \subseteq \mathbb{R}^n \to B \subseteq \mathbb{R}^m$ und $g : B \subseteq \mathbb{R}^m \to \mathbb{R}^k$ stetig, so ist auch deren Verkettung $f = g \circ h : A \to \mathbb{R}^k$, $f(x) = g(h(x))$, stetig.

Satz 1.3 (ε-δ-Kriterium) *Sei $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$, $x_0 \in \Omega$. Die folgenden Aussagen sind äquivalent:*
(i) f ist stetig im Punkt x_0
(ii) Zu jedem $\varepsilon > 0$ existiert ein $\delta > 0$ sodass gilt: $x \in \Omega$, $|x - x_0| < \delta \Longrightarrow |f(x) - f(x_0)| < \varepsilon$.

Die zweite Eigenschaft kann als alternative Definition von stetig benutzt werden.

Beweis: Wie im Eindimensionalen (siehe Analysis 1 Satz 10.1).

Beispiele (Fortsetzung): 6) Die euklidische Norm $x \mapsto |x|$ auf $\mathbb{R}^n$ ist stetig.
Beweis: es reicht zu zeigen, dass

$$\Big| |x| - |y| \Big| \leq |x - y| \text{ für alle } x, y \in \mathbb{R}^n \qquad (*)$$

(denn dann gilt (ii) mit $\delta = \varepsilon$). (*) ist äquivalent zu $(|x| - |y|)^2 \leq |x - y|^2$ und durch Ausmultiplizieren sieht man schnell ein, dass letztere Ungleichung äquivalent zur Cauchy-Schwarz-Ungleichung ist.

7) Sei $M \subseteq \mathbb{R}^n$ eine beliebige nichtleere Menge. Dann ist die Abstandsfunktion $d(x, M) := \inf\{|x - y| : y \in M\}$ stetig.
Beweis: es reicht zu zeigen, dass

$$|d(x, M) - d(x', M)| \leq |x - x'| \text{ für alle } x, x' \in \mathbb{R}^n \qquad (**)$$

(denn dann gilt (ii) mit $\delta = \varepsilon$). Um (**) einzusehen, argumentieren wir wie folgt: nach Dreiecksungleichung gilt

$$|x' - y| = |(x - y) + (x' - x)| \leq |x - y| + |x' - x| \text{ und analog } |x - y| \leq |x' - y| + |x - x'|.$$

Indem wir jeweils auf beiden Seiten das Infimum über $y \in M$ nehmen, erhalten wir

$$d(x', M) \leq d(x, M) + |x - x'| \text{ und } d(x, M) \leq d(x', M) + |x - x'|.$$

Diese beiden Ungleichungen zusammengenommen entsprechen (**).

Eleganterweise lassen sich ε und δ aus dem ε-δ-Kriterium eliminieren. Man benötigt stattdessen nur noch den Begriff der "offenen Menge" (siehe Def. 1.7 a)).

Satz 1.4 (Abstraktes ε-δ-Kriterium) *Für eine Funktion $f : \mathbb{R}^n \to \mathbb{R}^m$ sind äquivalent:*
(i) f ist stetig
(ii) Die Urbilder $f^{-1}(V)$ offener Mengen $V \subseteq \mathbb{R}^m$ sind offen.

Bemerkenswert! Nochmal durchlesen und staunen!

In obigem Kriterium kann "Urbilder offener Mengen" mitnichten durch "Bilder offener Mengen" ersetzt werden. Warum? Antwort 1: Die Bilder offener Mengen, also

$f(U)$ für offene $U \subseteq \mathbb{R}^n$, sind im allgemeinen nicht offen. Z.B. bildet die stetige Funktion $f(x) = x^2$, $f : \mathbb{R} \to \mathbb{R}$, die offene Menge $\mathbb{R}$ auf die nicht-offene Menge $[0, \infty)$ ab. Antwort 2: Im ε-δ-Kriterium heisst es "zu jedem ε existiert ein δ sodass..." und nicht "zu jedem δ existiert ein ε sodass...". Das ε hat aber mit dem Wertebereich der Funktion zu tun, und das δ mit dem Definitionsbereich. Das ε-δ-Kriterium ist also eine Aussage der Form "Zu gewissen Wertebereichs-Objekten existieren zugehörige Definitionsbereichs-Objekte sodass...", genau wie auch (ii).

Beweis "(i) $\Longrightarrow$ (ii)": Sei $V \subseteq \mathbb{R}^m$ offen. O.B.d.A. $f^{-1}(V)$ nicht leer (sonst ist die Behauptung trivial). Sei $x_0 \in f^{-1}(V)$ beliebig. Da $f(x_0) \in V$ und V offen, enthält V eine Kugel $B_\varepsilon(f(x_0))$ von positivem Radius $\varepsilon > 0$. Da f nach Voraussetzung stetig, existiert nach ε-δ-Kriterium ein $\delta > 0$ sodass gilt: $|x - x_0| < \delta \Longrightarrow |f(x) - f(x_0)| < \varepsilon$. Anders formuliert heisst das: es existiert eine Kugel $B_\delta(x_0)$ von positivem Radius $\delta > 0$ sodass $f(B_\delta(x_0)) \subseteq B_\varepsilon(f(x_0))$. Letztere Inklusion ist aber dasselbe wie $B_\delta(x_0) \subseteq f^{-1}(B_\varepsilon(f(x_0)))$. Insbesondere gilt also $B_\delta(x_0) \subseteq f^{-1}(V)$.

"(ii) $\Longrightarrow$ (i)": Wir zeigen, dass f in jedem Punkt x_0 das ε-δ-Kriterium erfüllt. Seien also $x_0 \in \mathbb{R}^n$ und $\varepsilon > 0$ beliebig. Wir betrachten die folgende Kugel im Wertebereich: $B_\varepsilon(f(x_0))$. Diese Kugel ist gemäss Abschnitt 1.3 offen, folglich nach Voraussetzung auch deren Urbild $f^{-1}(B_\varepsilon(f(x_0))$. Somit existiert um den Punkt x_0 im Urbild eine ebenfalls im Urbild enthaltene Kugel $B_\delta(x_0)$ von positivem Radius $\delta > 0$, d.h. $B_\delta(x_0) \subseteq f^{-1}(B_\varepsilon(f(x_0)))$. Letztere Inklusion ist aber dasselbe wie $f(B_\delta(x_0)) \subseteq B_\varepsilon(f(x_0))$, oder, elementarer formuliert, $|x - x_0| < \delta \Longrightarrow |f(x) - f(x_0)| < \varepsilon$.

1.8 Satz vom Maximum und Minimum

Nachdem wir uns ausführlich aus theoretischer Sicht mit dem Begriff der Stetigkeit beschäftigt haben, benutzen wir ihn nun als Voraussetzung in einem für Anwendungen grundlegenden Satz, der die Lösbarkeit von Extremwertaufgaben unter sehr allgemeinen (und realistischen) Voraussetzungen garantiert. Das eindimensionale Analogon haben Sie schon in Analysis 1 kennengelernt (Satz 6.5). Dieser Satz spielt in verschiedenen Gebieten (z.B. Analysis, Geometrie, Optimierung, theoretische Ökonomie) eine wichtige Rolle und wird Ihnen im Laufe Ihres Studiums wiederholt begegnen.

Def. 1.13 Eine Teilmenge $K \subseteq \mathbb{R}^n$ heisst **kompakt**, wenn jede Folge in K einen Häufungspunkt $x_* \in K$ besitzt.

Manche Autoren nennen diese Eigenschaft *folgenkompakt*, um sie von alternativen – im $\mathbb{R}^n$ äquivalenten – Definitionen von kompakt abzugrenzen. Siehe Kapitel 6.

Lemma 1.6 $K \subseteq \mathbb{R}^n$ kompakt $\Longleftrightarrow$ K abgeschlossen und beschränkt.

Beweis Wichtige Richtung "$\Longleftarrow$": Jede Folge in K ist eine beschränkte Folge im $\mathbb{R}^n$ und besitzt folgtlich nach einer zentralen Eigenschaft des $\mathbb{R}^n$, nämlich dem Satz

von Bolzano-Weierstrass (Satz 1.1), einen Häufungspunkt $x_* \in \mathbb{R}^n$. Wegen K abgeschlossen gilt $x_* \in K$.

Andere Richtung "$\Longrightarrow$": einfache Übungsaufgabe.

Satz 1.5 (Satz vom Maximum und Minimum) *Sei $K \subset \mathbb{R}^n$ kompakt (d.h. abgeschlossen und beschränkt), $f : K \to \mathbb{R}$ stetig. Dann besitzt f eine Maximumsstelle $x^* \in K$ (d.h. $f(x) \le f(x^*)$ für alle $x \in K$) und eine Minimumsstelle $x_* \in K$ (d.h. $f(x) \ge f(x_*)$ für alle $x \in K$).*

Beweis Analog zum eindimensionalen Fall (Analysis 1 Satz 3.5): 1. Wähle eine Minimalfolge $(x^{(j)})_{j\in\mathbb{N}}$, d.h. $x^{(j)} \in K$, $f(x^{(j)}) \to \inf_K f \in \mathbb{R} \cup \{-\infty\}$. 2. Nach Satz von Bolzano-Weierstrass im $\mathbb{R}^n$ (Satz 1.1) besitzt die Minimalfolge einen Häufungspunkt $x_* \in \mathbb{R}^n$. 3. Da K abgeschlossen, gilt $x_* \in K$. 4. Da f stetig, folgt für jede gegen den Häufungspunkt konvergierende Teilfolge $(x^{(j_\ell)})_{\ell\in\mathbb{N}}$: $f(x_*) = \lim_{\ell\to\infty} f(x^{(j_\ell)})$. Wegen Schritt 1. ist aber die rechte Seite gleich $\inf_K f$, d.h. x_* ist Minimumsstelle (und insbesondere $\inf_K f > -\infty$).

Der obige Satz und die nachfolgenden Beispiele klären nur die Existenz von Maximums- und Minimumsstellen. Wie man diese konkret bestimmt, besprechen wir in §2.7. Wir betonen aber, dass bereits die blosse Existenz relevante Konsequenzen hat; z.B. benötigen wir Satz 1.5 beim Beweis der Taylorentwicklung (Korollar 2.2), der Äquivalenz aller Normen auf dem $\mathbb{R}^n$ (Satz 3.1), oder des Spektralsatzes für symmetrische Matrizen (Korollar 4.4).

Beispiele 1) (Minimaler Abstand) Sei $M \subseteq \mathbb{R}^n$ abgeschlossen und nicht leer, $x \in \mathbb{R}^n$. Dann existiert ein Punkt $y_* \in M$ mit minimalem Abstand von x, d.h.

$$|x - y_*| \le |x - y| \text{ für alle } y \in M. \tag{1}$$

Begründung: Sei y_0 ein beliebiger Punkt in M, und sei $r = |x - y_0|$. Dann gilt

$$\inf\{|x-y| : y \in M\} = \inf\{|x-y| : y \in \underbrace{M \cap \overline{B_r(x)}}_{=:K}\}. \tag{2}$$

Die Menge K ist abgeschlossen und beschränkt, also kompakt. Nach Satz 1.5 existiert eine Minimumsstelle y_* der Funktion $y \mapsto |x - y|$ auf K, und wegen (2) ist dieses y_* auch Minimumsstelle der Funktion auf ganz M.

Als Korollar von Beispiel 1) erhalten wir einen anschaulichen Beweis der Tatsache "A abgeschlossen $\Longrightarrow \mathbb{R}^n \backslash A$ offen". Sei $x \notin A$. Da A abgeschlossen, gibt es einen Punkt y_* in A mit minimalem Abstand von x. Da $x \notin A$, gilt $|x - y_*| > 0$. Die Minimalitätseigenschaft (1) von y_* besagt $B_{|x-y_*|}(x) \subseteq \mathbb{R}^n \backslash A$.

2) (Extremwertprobleme mit Nebenbedingungen) Seien $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$

stetig, $|g(x)| \to \infty$ für $|x| \to \infty$, $c \in g(\mathbb{R}^n)$. Dann besitzt die Extremwertaufgabe mit Nebenbedingungen

$$\text{Minimiere } f \text{ bezüglich der Nebenbedingung } g(x) = c$$

eine Lösung.

Begründung: Obiges Problem ist äquivalent zur Minimierung von f über die Menge $K = \{x \in \mathbb{R}^n : g(x) = c\}$. Diese ist wegen der Stetigkeit von g abgeschlossen (denn $g(x^{(j)}) = c$, $x^{(j)} \to x \Longrightarrow g(x) = c$) und wegen der Wachstumsbedingung an g beschränkt. Die Behauptung folgt aus Satz 1.5.

Korollar 1.2 $f : \mathbb{R}^n \to \mathbb{R}^m$ stetig $\Longleftrightarrow$ die Urbilder abgeschlossener Mengen sind abgeschlossen.

Beweis: siehe Übungen.

Korollar 1.3 (Riesige Beispielklasse abgeschlossener und offener Mengen)

a) Niveaumengen $\{x \in \mathbb{R}^n : f(x) = c\}$ stetiger Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ sind abgeschlossen.

b) Subniveaumengen $\{x \in \mathbb{R}^n : f(x) \leq c\}$ stetiger Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ sind abgeschlossen.

c) Strikte Subniveaumengen $\{x \in \mathbb{R}^n : f(x) < c\}$ stetiger Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ sind offen.

Beweis: Die Mengen in a), b), c) sind genau die Urbilder der Mengen $\{c\}$, $(-\infty, c]$, $(-\infty, c)$. Die Behauptung folgt aus Satz 1.4 und Korollar 1.2 wegen $\{c\}$ abgeschlossen, $(-\infty, c]$ abgeschlossen, $(-\infty, c)$ offen.

Beispiel: Aus b) und c) gewinnen wir die bereits bekannten Tatsache zurück, dass die Kugel $B_r(x_0)$ offen und die Kugel $\overline{B_r(x_0)}$ abgeschlossen ist, indem wir $f(x) = |x - x_0|$ und $c = r$ wählen. a) liefert die neue Tatsache, dass die **Sphäre** mit Mittelpunkt x_0 und Radius $r > 0$,

$$S_r(x_0) := \{x \in \mathbb{R}^n : |x - x_0| = r\},$$

abgeschlossen ist.

2 Ableitung im Mehrdimensionalen

Die Ableitung einer Funktion gehört auch im Mehrdimensionalen zu den wichtigsten und meistverwendeten Konzepten der Analysis. Einerseits gibt es – wie im eindimensionalen Fall – einen mächtigen Kalkül, der erlaubt, die Ableitung nahezu aller handelsüblichen Funktionen mit Papier und Bleistift auszurechen. Andererseits gibt es ein riesiges Spektrum von Anwendungen sowohl in der Mathematik als auch ausserhalb, etwa
- Approximation von Funktionen durch Polynome (Taylor-Entwicklung)
- Maximieren/Minimieren
- Modellierung (Aufstellen partieller Differentialgleichungen).
Siehe Abschnitte 2.5–2.9.

Nun zum "Kleingedruckten". Entsprechend der Komplexität mehrdimensionaler Funktionen (Definitions- und Wertebereich sind mehrdimensional) gibt es verschiedene, allesamt nützliche Varianten der Ableitung: partielle Ableitung, totale Ableitung, Gradient, Richtungsableitung. Wir führen sie der Reihe nach ein und besprechen, wozu sie gut sind.

2.1 Partielle Ableitung

Im folgenden bezeichnen $e^{(1)}, ..., e^{(n)}$ die Standard-Basisvektoren des $\mathbb{R}^n$, d.h. $(e^{(j)})_k = 1$ für $k = j$ und 0 sonst. Die j-te partielle Ableitung einer Funktion beschreibt, wie sich die Funktion in Richtung des j-ten Standard-Basisvektors ändert. Genauer:

Def. 2.1 Sei $\Omega \subseteq \mathbb{R}^n$ offen. Eine Funktion

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix} : \Omega \longrightarrow \mathbb{R}^m, \quad f(x) = \begin{pmatrix} f_1(x_1, ..., x_n) \\ \vdots \\ f_m(x_1, ..., x_n) \end{pmatrix}$$

heisst partiell nach x_j differenzierbar im Punkt $x \in \Omega$, wenn der Grenzwert

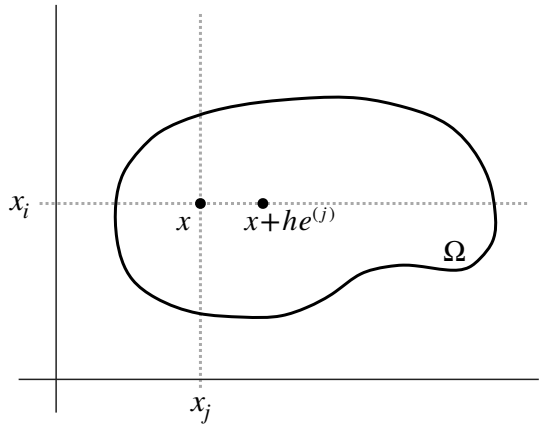
$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{1}{h}\big[f(x+he^{(j)}) - f(x)\big] = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{1}{h}\big[f(x_1, ..., x_j+h, ..., x_n) - f(x_1, ..., x_j, ..., x_n)\big] \in \mathbb{R}^m \tag{PA}$$

existiert. Falls ja, heisst der Grenzwert **partielle Ableitung von f nach x_j** im Punkt x, Schreibweise:

$$\frac{\partial f}{\partial x_j}(x) \text{ oder } \frac{\partial}{\partial x_j} f(x) \text{ oder } \partial_j f(x) \in \mathbb{R}^m.$$

Ist die Funktion an jedem Punkt $x \in \Omega$ partiell nach x_j differenzierbar [bzw. an jedem Punkt partiell nach jeder Komponente $x_1, ..., x_n$ differenzierbar], so heisst sie

partiell nach x_j differenzierbar (in Ω) [bzw. partiell differenzierbar (in Ω)].[5]



Wegen Äquivalenz von Konvergenz und komponentenweiser Konvergenz ist f partiell nach x_j differenzierbar genau dann, wenn jede Komponente f_k partiell nach x_j differenzierbar ist, und falls ja, gilt

$$\frac{\partial f}{\partial x_j}(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_j}(x) \\ \vdots \\ \frac{\partial f_m}{\partial x_j}(x) \end{pmatrix}.$$

Geometrisch ist $\frac{\partial f_i}{\partial x_j}$ die Steigung der i-ten Komponentenfunktion in der Koordinatenrichtung x_j.

Praktisch bestimmt man partielle Ableitungen wie gewöhnliche Ableitungen, wobei man alle übrigen Variablen, von denen die Funktion abhängt, als Konstanten betrachtet. Analog zum Ableitungskalkül im Eindimensionalen gelten folgende Rechenregeln: für $f, g : \Omega \to \mathbb{R}^m$ und $\lambda, \mu \in \mathbb{R}$ ist

$$\frac{\partial}{\partial x_i}(\lambda f + \mu g) = \lambda \frac{\partial}{\partial x_i} f + \mu \frac{\partial}{\partial x_i} g \quad \text{(Linearität)}.$$

Für $f : \Omega \to \mathbb{R}^m$, $g : \Omega \to \mathbb{R}$ ist

$$\frac{\partial}{\partial x_i}(f\, g) = (\frac{\partial}{\partial x_i} f) g \; + \; f \, \frac{\partial}{\partial x_i} g \quad \text{(Produktregel)}.$$

[5]Um die partielle Ableitung nach x_j im Punkt x definieren zu können, ist nicht unbedingt notwendig, dass Ω offen; es würde reichen, dass $x + h_k e_j \in \Omega$ für eine Nullfolge (h_k) mit $h_k \neq 0$ für alle k. Das reicht – ausser im eindimensonalen Fall – aber nicht, um partielle Ableitungen auch auf dem Rand einer offenen Menge zu definieren. Ist z.B. Ω die "Sichel" $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 > 0, \sqrt{x_1} < x_2 < 2\sqrt{x_1}\}$, so existieren solche Folgen $x + h_k e$ im Randpunkt $x = (0,0)$ für keine Richtung $e \in \mathbb{R}^2$, $|e| = 1$.

Insgesamt gibt es m Komponentenfunktionen, die wir ableiten können, und n Komponenten von x, nach denen wir ableiten können. Es ist sinnvoll (mehr dazu später), die resultierenden $m \cdot n$ partiellen Ableitungen im Punkt x in eine $m \times n$ Matrix anzuordnen. Die Matrix der Spaltenvektoren $\frac{\partial f}{\partial x_j}$,

$$J_f(x) = \begin{pmatrix} | & & | \\ \frac{\partial f}{\partial x_1}(x) & \cdots & \frac{\partial f}{\partial x_n}(x) \\ | & & | \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{pmatrix}$$

heisst **Matrix der partiellen Ableitungen** (oder Ableitungsmatrix oder Jacobi-Matrix) von f im Punkt x.

Beispiele: 1) $f : \mathbb{R}^2 \to \mathbb{R}^3$ (d.h. $n = 2$, $m = 3$), $f(x,y) = \begin{pmatrix} x^3 + y^3 \\ 7xy \\ 2x + 5 \end{pmatrix}$. Dann ist

$$\frac{\partial f}{\partial x}(x,y) = \begin{pmatrix} 3x^2 \\ 7y \\ 2 \end{pmatrix}, \quad \frac{\partial f}{\partial y}(x,y) = \begin{pmatrix} 3y^2 \\ 7x \\ 0 \end{pmatrix}, \quad J_f(x,y) = \begin{pmatrix} 3x^2 & 3y^2 \\ 7y & 7x \\ 2 & 0 \end{pmatrix}.$$

Die Matrix der partiellen Ableitungen im Punkt (x,y) ist eine 3×2 Matrix. Auswerten an verschiedenen Punkten liefert verschiedene Matrizen, z.B.

$$J_f(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 2 & 0 \end{pmatrix}, \quad J_f(1,0) = \begin{pmatrix} 3 & 0 \\ 0 & 7 \\ 2 & 0 \end{pmatrix}, \quad J_f(0,1) = \begin{pmatrix} 0 & 3 \\ 7 & 0 \\ 2 & 0 \end{pmatrix}.$$

2) lineare Abbildungen $f : \mathbb{R}^n \to \mathbb{R}^m$: Nach Ergebnissen der Linearen Algebra ist f von der Form

$$f(x) = Ax, \quad A = \begin{pmatrix} A_{11} & \dots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \dots & A_{mn} \end{pmatrix}$$

für eine $m \times n$ Matrix A, wobei $Ax =$ Matrix-Vektor-Produkt, d.h. $(Ax)_i = \sum_{j=1}^n A_{ij} x_j$. Die Matrix der partiellen Ableitungen ist gleich A, d.h. $J_f(x) = A$.

3) Partiell differenzierbare Funktionen müssen nicht stetig sein! Betrachte z.B. $f : \mathbb{R}^2 \to \mathbb{R}$, $f = 0$ auf den Koordinatenachsen und $f = 1$ sonst. Offenbar ist f im Nullpunkt partiell differenzierbar mit $\frac{\partial f}{\partial x}(0,0) = \frac{\partial f}{\partial y}(0,0) = 0$, aber dort unstetig. Dieses Beispiel ist nicht überall partiell differenzierbar, denn an Punkten auf der x-Achse ausserhalb des Nullpunkts ist f nur in x- aber nicht in y-Richtung partiell differenzierbar. Eine in $(0,0)$ unstetige, aber sogar überall partiell diff'bare Funktion erhalten wir, indem wir den Wert unserer Funktion auf den Koordinatenachsen (Null) und der Diagonalen (Eins) beibehalten, aber dazwischen interpolieren:

$$f(x,y) = \begin{cases} \frac{2xy}{x^2+y^2} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0). \end{cases}$$

Nach wie vor ist f partiell differenzierbar im Punkt $(0,0)$ mit $\frac{\partial f}{\partial x}(0,0) = \frac{\partial f}{\partial y}(0,0) = 0$, und unstetig im Punkt $(0,0)$ wegen $f(t,0)) = f(0,t) = 0$ und $f(t,t) = 1$ $(t \neq 0)$, aber jetzt auch an allen Punkten $\neq (0,0)$ partiell diff'bar, da dort der Nenner $\neq 0$ ist.

Interpretation: Partielle Diff'barkeit untersucht nur das Verhalten in Richtung der Koordinatenachsen, und liefert daher zu wenig Information über eine Funktion. Darüber hinaus ist unbefriedigend, dass partielle Diff'barkeit von der Wahl der Koordinaten abhängt. Abhilfe: Koordinateninvariante Definition der Ableitung, siehe nächster Abschnitt.

Der Fall von Kurven. Ist der Definitionsbereich eindimensional, d.h. $f : I \subseteq \mathbb{R} \to \mathbb{R}^m$ mit I Intervall (d.h., in der Terminologie von Def. 1.10, f *Kurve*), reduziert sich der Grenzwert (PA) zu

$$\lim_{h\to 0} \frac{1}{h}\big(f(t+h) - f(t)\big) \in \mathbb{R}^m.$$

Falls er existiert, heisst er statt "partielle Ableitung" einfach *Ableitung von f im Punkt t*, Schreibweise:

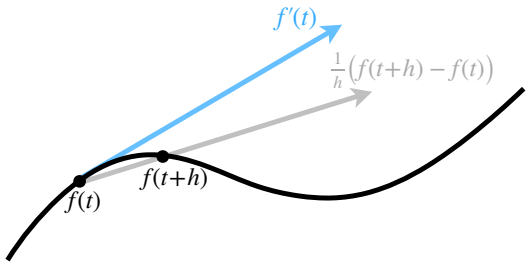
$$f'(t) \text{ oder } \frac{df}{dt}(t) \text{ oder } \frac{d}{dt}f(t),$$

und f heisst *differenzierbar im Punkt $t \in I$*. Existiert er für alle $t \in I$, so heisst die Kurve f *differenzierbar*. Wie bereits oben erläutert, ist f diff'bar genau dann wenn jede Komponente f_k von f diff'bar, und falls ja, gilt

$$f'(t) = \begin{pmatrix} f_1'(t) \\ \vdots \\ f_m'(t) \end{pmatrix} = J_f(t) \in \mathbb{R}^m;$$

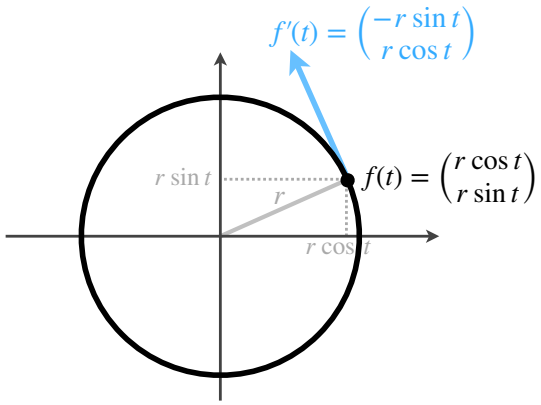
hier ist $f_k'(t)$ die aus Analysis 1 bekannte Ableitung von f_k im Punkt t. Die Jacobi-Matrix reduziert sich also auf den Spaltenvektor bestehend aus den gewöhnlichen Ableitungen der Komponenten.

Aus geometrischen Gründen (siehe Schaubild) wird die Ableitung $f'(t)$ auch *Tangentialvektor der Kurve im Punkt $f(t)$* genannt; zeichnen wir ihn als Vektor mit Fusspunkt $f(t)$, liegt er "tangential" an der Kurve.



Differenzenquotient $\frac{1}{h}\big(f(t+h) - f(t)\big)$ (grauer Vektor)
und sein Grenzwert $f'(t)$ (blauer Vektor)

Beispiel: Kreislinie $f : \mathbb{R} \to \mathbb{R}^2$, $f(t) = \begin{pmatrix} r\cos t \\ r\sin t \end{pmatrix}$ $(r > 0)$. Dann ist $f'(t) = \begin{pmatrix} -r\sin t \\ r\cos t \end{pmatrix}$.



Wie man sieht, liegt die berechnete Ableitung (blauer Vektor) tangential an der Kreislinie.

2.2 Totale Ableitung

Motiviert durch den Satz von Taylor im Eindimensionalen, d.h. der Näherungsformel

$$f(x+h) = f(x) + f'(x)h + \eta(h), \quad \eta(h) = \text{ kleiner Rest d.h. } \frac{\eta(h)}{h} \to 0 \ (h \to 0),$$

fassen wir die Ableitung einer eindimensionalen Funktion $f : \mathbb{R} \to \mathbb{R}$ an der Stelle x nicht als eine Zahl, sondern als eine lineare Abbildung $h \mapsto f'(x)h$ auf. Nämlich diejenige lineare Abbildung, die die nichtlineare Abbildung $h \mapsto f(x+h) - f(x)$ für kleine h "besonders gut approximiert". Dieser Standpunkt lässt sich auf mehrdimensionale Funktionen verallgemeinern.

Def. 2.2 (Totale Ableitung als bestapproximierende lineare Abbildung) Sei $\Omega \subseteq \mathbb{R}^n$ offen. Eine Funktion $f : \Omega \to \mathbb{R}^m$ heisst **(total) differenzierbar** im Punkt $x \in \Omega$, wenn eine lineare Abbildung $L : \mathbb{R}^n \to \mathbb{R}^m$ existiert sodass

$$\frac{f(x+h) - f(x) - L(h)}{|h|} \to 0 \quad (h \neq 0,\ h \text{ sodass } x + h \in \Omega,\ h \to 0). \tag{TA}$$

Falls ja, heisst die lineare Abbildung L **(totale) Ableitung** von f im Punkt x, Schreibweise: $L = Df(x)$. f heisst (total) differenzierbar in Ω, wenn f (total) differenzierbar in jedem Punkt $x \in \Omega$.

Die (totale) Ableitung $Df(x)$ ist also eine lineare Abbildung $\mathbb{R}^n \to \mathbb{R}^m$. Aus elementaren Überlegungen oder Satz 2.1 folgt, dass sie – sofern existent – eindeutig ist. Ein

attraktiver Aspekt der obigen Definition ist ihre Koordinatenunabhängigkeit. Insbesondere spielen keine bestimmten Richtungen im $\mathbb{R}^n$ (wie etwa, in der Definition der Matrix der partiellen Ableitungen, die Richtungen der Koordinatenachsen) eine Extra-Rolle.

Die aus Analysis 1 bekannten Landau'schen Symbole klein-o und Gross-O, die sich sofort ins Mehrdimensionale übertragen lassen (siehe die Definition unten), erlauben eine nützliche Kurzschreibweise der Bedingung (TA):

$$f(x+h) = f(x) + L(h) + o(|h|) \quad (h \to 0). \tag{TA'}$$

Die Definition von $o(|h|)$ sieht genauso aus wie im Eindimensionalen, wir ersetzen nur den Absolutbetrag auf $\mathbb{R}$ durch die euklidische Norm $|\cdot|$ auf dem $\mathbb{R}^n$:

Def. (Landau'sche Symbole klein-o und Gross-O) Für $f : \Omega \to \mathbb{R}^m$, $g : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$ offen, $0 \in \Omega$ ist $f(h) = O(g(h))$ für $h \to 0$ ("f ist Groß-O von g für h gegen 0") wenn $\delta > 0$ und $C > 0$ existieren sodass

$$|f(h)| \leq C|g(h)| \ \forall h \in \Omega \text{ mit } |h| < \delta,$$

und $f(h) = o(g(h))$ für $h \to 0$ ("f ist klein-o von g für h gegen 0") wenn $g(h) \neq 0$ für alle $h \neq 0$ und

$$\frac{f(h)}{g(h)} \to 0 \text{ für } h \to 0,\ h \neq 0.$$

Der folgende grundlegende Satz klärt den Zusammenhang zwischen totaler und partieller Ableitung.

Satz 2.1 *Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}^m$.*

(i) f (total) diff'bar im Punkt $x \in \Omega \Longrightarrow$ f partiell diff'bar im Punkt x, und $Df(x)(h) = J_f(x)h$ für alle $h \in \mathbb{R}^n$.

(ii) f partiell diff'bar in Ω, alle partiellen Ableitungen $\frac{\partial f}{\partial x_i}$ stetig $\Longrightarrow$ f (total) diff'bar in Ω.

Die Formel $Df(x)(h) = J_f(x)h$ in (i) bedeutet in Worten: die totale Ableitung angewendet auf den Vektor h ist gleich der Matrix der partiellen Ableitungen multipliziert mit h. *Die Matrix $J_f(x)$ der partiellen Ableitungen ist also nichts anderes als die darstellende Matrix der linearen Abbildung $Df(x)$.*

Die Voraussetzung der Stetigkeit der partiellen Ableitungen in (ii) kann nicht weggelassen werden (siehe etwa Beispiel 3) im vorherigen Abschnitt).

Aussage (i) zeigt insbesondere: die totale Ableitung ist, falls sie existiert, eindeutig.

Beweis von (i): Anwenden der Definition (TA) auf Vektoren der Form $h = te^{(j)}$

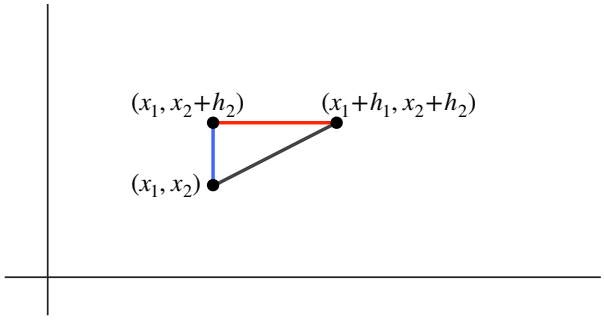
($t \in \mathbb{R}$) liefert (mit der Notation $L = Df(x)$)

$$\frac{|f(x+h) - f(x) - L(h)|}{|h|} = \left| \frac{f(x + te^{(j)}) - f(x)}{t} - L(e^{(j)}) \right| \to 0,$$

d.h. f partiell nach x_j diff'bar im Punkt x, $L(e^{(j)}) = \frac{\partial f}{\partial x_j}(x)$. Wegen Linearität von L folgt somit für beliebiges $h = \sum_j h_j e^{(j)} \in \mathbb{R}^n$:

$$L(h) = \sum_j h_j L(e^{(j)}) = \sum_j h_j \frac{\partial f}{\partial x_j}(x) = J_f(x)h.$$

Beweis von (ii): Zur Maximierung der Verständlichkeit beschränken wir uns auf $n = 2$. Beweisidee: spalte die zu untersuchende Differenz $f(x_1+h_1, x_2+h_2) - f(x_1, x_2)$ der Funktionswerte in der "schrägen" Richtung (h_1, h_2) in eine Differenz in horizontaler Richtung und eine Differenz in vertikaler Richtung auf, um die partiellen Ableitungen ins Spiel zu bringen.



Details:

$$f(x_1+h_1, x_2+h_2) - f(x_1, x_2) - \overbrace{\begin{pmatrix} | & | \\ \frac{\partial f}{\partial x_1}(x) & \frac{\partial f}{\partial x_2}(x) \\ | & | \end{pmatrix}}^{=J_f(x)} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

$$= \Big(f(x_1+h_1, x_2+h_2) - f(x_1, x_2+h_2) - \frac{\partial f}{\partial x_1}(x)h_1 \Big) + \Big(f(x_1, x_2+h_2) - f(x_1, x_2) - \frac{\partial f}{\partial x_2}(x)h_2 \Big).$$

Mit Mittelwertsatz schreiben wir die i-te Komponente der ersten Klammer als

$$\Big(\frac{\partial f_i}{\partial x_1}(\xi_1, x_2+h_2) - \frac{\partial f_i}{\partial x_1}(x) \Big) h_1 \quad \text{für ein } \xi_1 \in [x_1, x_1 + h_1]$$

und die i-te Komponente der zweiten Klammer als

$$\Big(\frac{\partial f_i}{\partial x_2}(x_1, \xi_2) - \frac{\partial f_i}{\partial x_2}(x) \Big) h_2 \quad \text{für ein } \xi_2 \in [x_2, x_2 + h_2].$$

Also folgt

$$\Big|(f(x_1{+}h_1, x_2{+}h_2) - f(x_1,x_2) - \begin{pmatrix} | & | \\ \frac{\partial f}{\partial x_1}(x) & \frac{\partial f}{\partial x_2}(x) \\ | & | \end{pmatrix}\begin{pmatrix} h_1 \\ h_2 \end{pmatrix})_i\Big|$$
$$\leq |h_1| \sup_{|\xi|\leq|h|} \Big|\frac{\partial f_i}{\partial x_1}(x{+}\xi) - \frac{\partial f_i}{\partial x_1}(x)\Big| + |h_2| \sup_{|\xi|\leq|h|} \Big|\frac{\partial f_i}{\partial x_2}(x{+}\xi) - \frac{\partial f_i}{\partial x_2}(x)\Big|.$$

Wegen Stetigkeit der partiellen Ableitungen gehen die Suprema für $|h| \to 0$ gegen 0, und es folgt *linke Seite*$/|h| \to 0$ ($|h| \to 0$), d.h. f total diff'bar mit Ableitung $Df(x)(h) = J_f(x)h$.

Der Fall von Kurven. Ist der Definitionsbereich eindimensional, d.h. $f : I \subseteq \mathbb{R} \to \mathbb{R}^m$ mit I Intervall, gilt in (i) Äquivalenz, d.h. die verschiedenen Differenzierbarkeitsbegriffe fallen zusammen (total diff'bar = partiell diff'bar). In diesem Fall bedeutet nämlich partielle Diff'barkeit von f im Punkt t Existenz des Grenzwertes $\lim_{h\to 0}\big(f(t+h) - f(t)\big)/h$, also wie in Abschnitt 2.1 besprochen gewöhnliche Diff'barkeit der Komponenten, und diese impliziert nach Satz von Taylor, dass f total diff'bar im Punkt t mit $Df(t)(h) = J_f(t)h$. Die Kurve heisst dann einfach *differenzierbar im Punkt* t.

Das merkwürdige Phänomen aus Abschnitt 2.1, dass partiell differenzierbare Funktionen unstetig sein können, wird durch den Begriff der totalen Ableitung behoben:

Korollar 2.1 Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$, $x \in \Omega$. Dann gilt:

$$f \text{ (total) differenzierbar im Punkt } x \implies f \text{ stetig im Punkt } x.$$

Beweis: In der Darstellung (TA') für $f(x+h)$ geht der letzte Term auf der rechten Seite gegen 0 für $|h| \to 0$, und ebenso der zweite Term wegen der Stetigkeit linearer Abbildungen (siehe §1.7).

Wir fassen zusammen: für Funktionen $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$ gelten die Implikationen

$$\begin{matrix} f \text{ partiell differenzierbar,} \\ \text{partielle Ableitungen stetig} \end{matrix} \implies f \text{ (total) differenzierbar} \implies f \text{ stetig.}$$

Notation: Funktionen mit der linken Eigenschaft heissen *stetig differenzierbar*, oder $\mathcal{C}^1$*-Funktionen*.

Kettenregel. Als nächstes besprechen wir die Kettenregel, die auf die Mathematik-Pioniere Isaac Newton und Gottfried Wilhelm Leibniz zurückgeht. Erinnerung: im Eindimensionalen gilt: falls $f, g : \mathbb{R} \to \mathbb{R}$ diff'bar, ist die Verkettung $u(x) = f(g(x))$ diff'bar, mit Ableitung

$$u'(x) = f'(g(x))\, g'(x).$$

Im Mehrdimensionalen haben wir folgende Situation: die "innere" Funktion g bildet eine Teilmenge $\Omega' \subseteq \mathbb{R}^k$ in eine Teilmenge $\Omega \subseteq \mathbb{R}^n$ ab; die "äussere" Funktion f bildet den Wertebereich $\Omega \subseteq \mathbb{R}^n$ von g in den $\mathbb{R}^m$ ab; die Verkettung $u = f \circ g$, $u(x) = f(g(x))$, ist dann wohldefiniert und bildet $\Omega' \subseteq \mathbb{R}^k$ in den $\mathbb{R}^m$ ab. *Was ist die Ableitung von u?*

Es ist hilfreich, sich das mithilfe des Begriffs der totalen Ableitung zu überlegen. Die lineare Bestapproximation der Verkettung der nichtlinearen Abbildungen f und g im Punkt x sollte die Verkettung der bestapproximierenden linearen Abbildungen (von f im Punkt $g(x)$, und g im Punkt x) sein. Wenn dem aber so ist, muss sich – nach der zentralen Einsicht der linearen Algebra "Verkettung linearer Abbildungen entspricht Multiplikation der darstellenden Matrizen" – die Jacobi-Matrix der Verkettung als Matrizenprodukt der Jacobi-Matrizen von f und g ergeben.

Satz 2.2 (Mehrdimensionale Kettenregel) *Seien $\Omega \subseteq \mathbb{R}^n$, $\Omega' \subseteq \mathbb{R}^k$ offen, $f : \Omega \to \mathbb{R}^m$, $g : \Omega' \to \Omega$, $u = f \circ g$ (Verkettung von f und g), d.h. $u(x) = f(g(x))$. Falls g diff'bar im Punkt $x \in \Omega'$ und f diff'bar im Punkt $g(x) \in \Omega$, ist u diff'bar im Punkt x, und es gilt:*

(i) Die Ableitung der Verkettung ist die Verkettung der Ableitungen, d.h.

$$Du(x)(h) = Df(g(x))\Big(Dg(x)(h)\Big) \text{ für alle } h \in \mathbb{R}^k.$$

(ii) Die Jacobi-Matrix der Verkettung ist das Produkt der Jacobi-Matrizen, d.h.

$$J_u(x) = J_f(g(x))\, J_g(x).$$

Wichtiger als ein Verständnis des Beweises ist es, **sich klarzumachen, was die Ausdrücke in (i) bedeuten**. Schauen wir uns die rechte Seite an, indem wir von innen anfangen und systematisch Schritt für Schritt vorgehen. h ist ein Vektor im $\mathbb{R}^k$. $Dg(x)$ ist eine lineare Abbildung vom $\mathbb{R}^k$ in den $\mathbb{R}^n$. Folglich können wir $Dg(x)$ auf h anwenden; das Bild von h unter dieser linearen Abbildung, also $Dg(x)(h)$, ist ein Vektor im $\mathbb{R}^n$. Auf diesen können wir die lineare Abbildung $Df(g(x))$ anwenden, denn letztere ist eine lineare Abbildung vom $\mathbb{R}^n$ in den $\mathbb{R}^m$. Das Bild des Vektors $Dg(x)(h)$ unter dieser Abbildung, also die rechte Seite von (i), ist dementsprechend ein Vektor im $\mathbb{R}^m$.

Beispiel $k = n = m = 2$, $f, g : \mathbb{R}^2 \to \mathbb{R}^2$,

$$g(x_1, x_2) = \begin{pmatrix} x_1^2 - x_2^2 \\ 2x_1x_2 \end{pmatrix}, \quad f(y_1, y_2) = \begin{pmatrix} e^{y_1}\cos y_2 \\ e^{y_1}\sin y_2 \end{pmatrix},$$

$u(x) = f(g(x))$. (Aufgefaßt als Abbildung von $\mathbb{C}$ nach $\mathbb{C}$ ist $u(z) = \exp(z^2)$.) Dann ist

$$u(x_1, x_2) = \begin{pmatrix} e^{x_1^2 - x_2^2}\cos(2x_1x_2) \\ e^{x_1^2 - x_2^2}\sin(2x_1x_2) \end{pmatrix}.$$

Wir berechnen

$$J_f(y) = \begin{pmatrix} e^{y_1}\cos y_2 & -e^{y_1}\sin y_2 \\ e^{y_1}\sin y_2 & e^{y_1}\cos y_2 \end{pmatrix}, \quad J_g(x) = \begin{pmatrix} 2x_1 & -2x_2 \\ 2x_2 & 2x_1 \end{pmatrix}$$

und folglich nach (mehrdimensionaler) Kettenregel

$$J_u(x) = e^{x_1^2 - x_2^2} \begin{pmatrix} 2x_1\cos(2x_1x_2) - 2x_2\sin(2x_1x_2) & -2x_2\cos(2x_1x_2) - 2x_1\sin(2x_1x_2) \\ 2x_2\cos(2x_1x_2) + 2x_1\sin(2x_1x_2) & 2x_1\cos(2x_1x_2) - 2x_2\sin(2x_1x_2) \end{pmatrix}.$$

Probe: direktes Berechnen der partiellen Ableitungen von u liefert dieselbe Matrix, wobei sich die zwei Summanden jeder Matrixkomponente aus der Produktregel ergeben. Weitere Beispiele zur Kettenregel: siehe Übungen und Abschnitt 2.3.

Beweis der Kettenregel: Wir setzen

$$\begin{aligned} \eta(h) &:= g(x+h) - g(x) - \underbrace{Dg(x)(h)}_{=J_g(x)h}, \\ \xi(k) &:= f(y+k) - f(y) - \underbrace{Df(y)(k)}_{=J_f(y)k}, \quad y := g(x). \end{aligned}$$

Dann gilt $\eta(0) = 0$, $\xi(0) = 0$, und die vorausgesetzte Differenzierbarkeit von f im Punkt x und g im Punkt y bedeutet: $|\eta(h)|/|h| \to 0$ ($|h| \to 0$), $|\xi(k)|/|k| \to 0$ ($|k| \to 0$). Wir berechnen

$$\begin{aligned} f(g(x+h)) - f(g(x)) &= f\Big(g(x) + Dg(x)(h) + \eta(h)\Big) - f(g(x)) \\ &= Df(g(x))\Big(Dg(x)(h) + \eta(h)\Big) + \xi\Big(Dg(x)(h) + \eta(h)\Big) \\ &\underset{Df(g(x)) \text{ linear}}{=} Df(g(x))\Big(Dg(x)(h)\Big) + \underbrace{Df(g(x))\Big(\eta(h)\Big)}_{=:u(h)} + \underbrace{\xi\Big(Dg(x)(h) + \eta(h)\Big)}_{=:v(h)}. \end{aligned}$$

Wir müssen zeigen: die Summe der beiden Terme $u(h)$ und $v(h)$ dividiert durch $|h|$ geht gegen Null für $|h| \to 0$. Wir zeigen (a) $|u(h)|/|h| \to 0$ ($|h| \to 0$), (b) $|v(h)|/|h| \to 0$ ($|h| \to 0$). Dazu benutzen wir die euklidische Norm für Matrizen, definiert für $A \in \mathbb{R}^{m\times n}$ als

$$|A| := \Big(\sum_{i,j} A_{ij}^2\Big)^{1/2} = \Big(\operatorname{tr} A^T A\Big)^{1/2},$$

und die Cauchy-Schwarz-Ungleichung für Matrix-Vektor-Produkte,

$$|Av| \le |A|\,|v| \quad \text{für alle } A \in \mathbb{R}^{m\times n} \text{ und alle } v \in \mathbb{R}^n.$$

Beweis dieser Ungleichung: Wegen der üblichen Cauchy-Schwarz-Ungleichung (Lemma 1.2) ist

$$|(Ax)_i| = \sum_j A_{ij}x_j \le \Big(\sum_j A_{ij}^2\Big)^{1/2}\Big(\sum_j x_j^2\Big)^{1/2};$$

Quadrieren und Summieren über i liefert die Behauptung. Somit können wir $u(h)$ wie folgt abschätzen:

$$|u(h)| = |J_f(g(x))\eta(h)| \le |J_f(g(x))||\eta(h)|$$

und (a) folgt aus $|\eta(h)|/|h| \to 0$. Setze $k := Dg(x)(h) + \eta(h) = J_g(x)h + \eta(h)$, dann gilt

$$|k| \le |J_g(x)||h| + |\eta(h)|$$

und folglich

$$\frac{|v(h)|}{|h|} = \begin{cases} 0, & k = 0 \\ \frac{|\xi(k)|}{|k|}\frac{|k|}{|h|}, & k \neq 0 \end{cases} \leq \begin{cases} 0, & k = 0 \\ \frac{|\xi(k)|}{|k|}\left(|J_g(x)| + \frac{|\eta(h)|}{|h|}\right), & k \neq 0. \end{cases}$$

Wegen $|\eta(h)|/|h| \to 0$ ($|h| \to 0$), $|\xi(k)|/|k| \to 0$ ($|k| \to 0$), und $|k| \to 0$ für $|h| \to 0$ (siehe obige Abschätzung für $|k|$) folgt (b).

Zum Schluss dieses Abschnitts leiten wir eine aus der Linearen Algebra bekannte Funktion, die Determinante $A \mapsto \det A$, ab. Am einfachsten geht das, indem man die totale Ableitung ausrechnet, nicht indem man einzeln nach den Komponenten der Matrix ableitet. Die totale Ableitung einer Funktion an einem Punkt x_0 ist eine lineare Abbildung L; um sie auszurechnen, muss man ihren Wert $L(h)$ für alle h angeben.

Beispiel Betrachte die Funktion

$$\det : \mathbb{R}^{n\times n} \to \mathbb{R}.$$

Behauptung 1: $D\det(I)(H) = \operatorname{tr} H$ für alle $H \in \mathbb{R}^{n\times n}$. In Worten: *Die Ableitung der Determinante an der Einheitsmatrix ist die Spur.* Zum Beweis berechnen wir

$$\begin{aligned} \det(I+H) &= \det\begin{pmatrix} 1+H_{11} & H_{12} & \cdots & H_{1n} \\ H_{21} & 1+H_{22} & \cdots & H_{2n} \\ \vdots & \vdots & & \vdots \\ H_{n1} & H_{n2} & \cdots & 1+H_{nn} \end{pmatrix} \\ &= (1+H_{11})\cdot(1+H_{22})\cdot\ldots\cdot(1+H_{nn}) + O(|H|^2) \\ &= \underbrace{1}_{=\det I} + \underbrace{(H_{11}+\ldots+H_{nn})}_{=\operatorname{tr} H} + O(|H|^2) \end{aligned}$$

(wobei $|H|$ die oben im Beweis der Kettenregel eingeführte euklidische Norm der Matrix H bezeichnet); also folgt die Behauptung direkt aus Def. 2.2.

Behauptung 2: Falls A invertierbar, ist $D\det(A)(H) = \det A \cdot \operatorname{tr}(A^{-1}H)$ für alle H. Dies folgt via Einschieben einer "multiplikativen Eins" und Behauptung 1:

$$\det(A+H) = \det(A + AA^{-1}H) = \det(A(I+A^{-1}H)) = \det A \cdot \underbrace{\det(I+A^{-1}H)}_{\underset{\text{Beh. 1}}{=} 1+\operatorname{tr}(A^{-1}H)+O(|H|^2)}.$$

2.3 Gradient; Richtungsableitung

Wir haben bisher zwei Sichtweisen der Ableitung einer mehrdimensionalen Funktion ($f : \mathbb{R}^n \to \mathbb{R}^m$) in einem Punkt x kennengelernt:

1) als $m \times n$ Matrix (nämlich der Jacobi-Matrix $J_f(x)$)
2) als lineare Abbildung $\mathbb{R}^n \to \mathbb{R}^m$ (nämlich der totalen Ableitung $Df(x)$).

Diese beiden, eng miteinander zusammenhängenden Begriffe (die Jacobi-Matrix ist die darstellende Matrix der totalen Ableitung) haben einen Vorteil und einen Nachteil.

Vorteil: allgemein (Dimensionen n, m von Def.- u. Wertebereich beliebig)
Nachteil: unanschaulich (nicht überzeugend visualisierbar)

Wir besprechen nun zwei weitere Varianten der Ableitung:

3) Gradient
4) Richtungsableitung.

Vorteil: anschaulich (visualisierbar, geometrisch)
Nachteil: "Nur" für skalare Funktionen ($m = 1$)

Def. 2.3 (Gradient) Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$ partiell differenzierbar, $x \in \Omega$. Der Vektor

$$\operatorname{grad} f(x) = J_f(x)^T = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \in \mathbb{R}^n$$

heißt **Gradient** von f am Punkt x. Das Vektorfeld $\operatorname{grad} f : \Omega \to \mathbb{R}^n$ heißt Gradient von f.

Alternative Schreibweise: $\nabla f(x)$ statt $\operatorname{grad} f(x)$. Das Symbol

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix}$$

heisst *Nabla*.

Obige Definition ist auf den ersten Blick koordinatenabhängig. Alternativ können wir den Gradienten auch ohne Benutzung spezieller Koordinatenrichtungen definieren, und zwar mithilfe der totalen Ableitung:

Def. 2.3′ Sei $f : \Omega \to \mathbb{R}$ total differenzierbar. Der Gradient $\operatorname{grad} f(x)$ von f am Punkt x ist derjenige Vektor im $\mathbb{R}^n$, sodass

$$Df(x)(h) = \langle \operatorname{grad} f(x), h \rangle \quad \text{für alle } h \in \mathbb{R}^n.$$

Hierbei ist $\langle a, b \rangle$ das euklidische Skalarprodukt $a_1 b_1 + \ldots + a_n b_n$ zweier Vektoren im $\mathbb{R}^n$.[6]

[6] Für jede lineare Abbildung $L : \mathbb{R}^n \to \mathbb{R}$ existiert genau ein Vektor $a \in \mathbb{R}^n$ sodass $L(h) = \langle a, h \rangle$ für alle $h \in \mathbb{R}^n$ ('Riesz'scher Darstellungssatz' aus der Linearen Algebra). Die Existenz folgt, indem man $a_i := L(e^{(i)})$ mit $e^{(i)} = i$-ter Einheitsvektor setzt; die Eindeutigkeit folgt aus der positiven Definitheit des Skalarproduktes.

Die Äquivalenz beider Definitionen folgt wegen

$$Df(x)(h) \underset{\text{Satz 2.1}}{=} J_f(x)h = \left(\tfrac{\partial f}{\partial x_1}(x) \cdots \tfrac{\partial f}{\partial x_n}(x)\right)\begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} = \left\langle \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}, \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} \right\rangle.$$

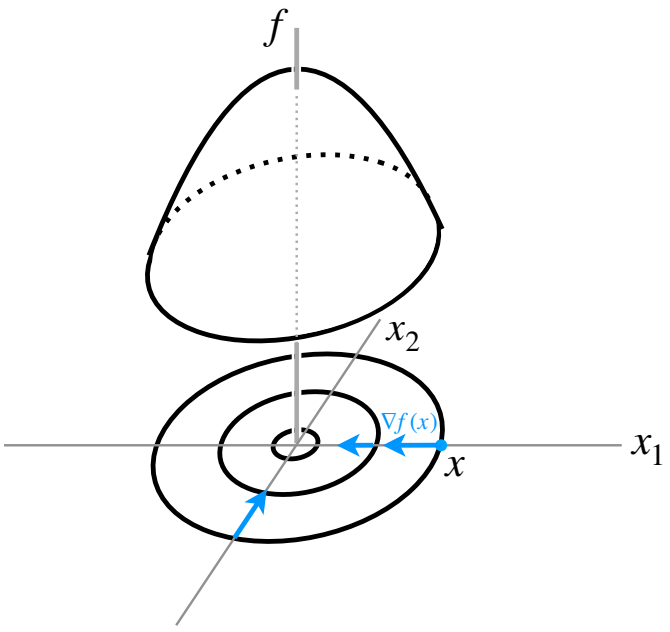
Der Gradient an einem Punkt x ist – vom Informationsgehalt her – dasselbe wie die Jacobi-Matrix am Punkt x (nämlich eine Liste der partiellen Ableitungen nach den Koordinatenrichtungen); der auf den ersten Blick irrelevant scheinende aber fruchtbare Unterschied ist, dass wir ihn nicht in einem abstrakten Raum von $1 \times n$ Matrizen verorten, sondern ihn als Vektor im $\mathbb{R}^n$ auffassen. Dementsprechend ist die Abbildung $x \mapsto \operatorname{grad} f(x)$ ein Vektorfeld.

Beispiel: $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = 2 - x_1^2 - x_2^2$ (siehe Skizze). Dann ist

$$\operatorname{grad} f(x) = \begin{pmatrix} -2x_1 \\ -2x_2 \end{pmatrix}$$

und z.B.

$$x = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \Longrightarrow \operatorname{grad} f(x) = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \quad x = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \Longrightarrow \operatorname{grad} f(x) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \quad x = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix} \Longrightarrow \operatorname{grad} f(x) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$



Der Vektor $\operatorname{grad} f(x)$ mit Fusspunkt x (blauer Vektor) zeigt immer in Richtung des Ursprungs! *Wieso?* Allgemeiner gefragt:

In welche Richtung zeigt der Gradient?

Die Richtung hat eine bemerkenswert einfache geometrische Bedeutung, die durch Satz 2.3 unten offengelegt wird. Zur Vorbereitung definieren wir:

Def. 2.4 (Richtungsableitung) Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$, $x \in \Omega$, $e \in \mathbb{R}^n$, $|e| = 1$, $t \in \mathbb{R}$, $|t|$ hinreichend klein sodass $x + te \in \Omega$. Die reelle Zahl

$$\partial_e f(x) := \frac{d}{dt} f(x + te)\Big|_{t=0} = \lim_{t \to 0} \frac{f(x + te) - f(x)}{t}$$

heißt – falls der Grenzwert auf der rechten Seite existiert – **Richtungsableitung** von f am Punkt x in Richtung e.

Wenn f diff'bar im Punkt x ist, existiert die obige Richtungsableitung, denn dann ist $t \mapsto f(x + te)$ als Verkettung der diff'baren Funktionen f und $t \mapsto x + te$ diff'bar.

Anschauliche Bedeutung der Richtungsableitung: Wir schränken die Funktion f auf das durch x hindurchgehende Geradenstück $\{x+te : t \in \mathbb{R}, |t| \text{ hinreichend klein}\}$ ein. Der Graph von f über diesem Geradenstück ist der Graph einer eindimensionalen Funktion, $u(t) := f(x + te)$, und die Richtungsableitung ist nichts als die Ableitung dieser eindimensionalen Funktion an der Stelle $t = 0$, d.h.

$$\frac{d}{dt} f(x + te)\Big|_{t=0} = u'(0).$$

Die Richtungsableitung ist also die Steigung von f in Richtung e.

Im Spezialfall von Koordinatenrichtungen, d.h.

$$e = e^{(i)} = \mathrm{i}^{ter} \text{ Einheitsvektor } = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow \mathrm{i}^{te} \text{ Komponente},$$

reduziert sich die Richtungsableitung auf die partielle Ableitung, denn dann ist

$$u(t) = f(x + te^{(i)}) = f(x_1, ..., x_{i-1}, x_i + t, x_{i+1}, ..., x_n)$$

und somit

$$\begin{aligned} u'(0) &= \frac{d}{dt} f(x + te^{(i)})\Big|_{t=0} \\ &= \lim_{t \to 0} \frac{f(x_1, ..., x_{i-1}, x_i + t, x_{i+1}, ..., x_n) - f(x_1, ..., x_n)}{t} = \frac{\partial f}{\partial x_i}(x_1, ..., x_n). \end{aligned}$$

Wie berechnet man Richtungsableitungen für allgemeine Richtungen? Hier hilft die Kettenregel weiter. Obige Funktion $u(t) = f(x + te)$ ist eine Verkettung, nämlich $u = f \circ g$, $g(t) = x + te$ (insbesondere ist $g(0) = x$; Definitions- und Wertebereiche:

$g : (-\varepsilon, \varepsilon) \to \Omega$ ($\varepsilon > 0$ hinreichend klein), $f : \Omega \to \mathbb{R}$, $u : (-\varepsilon, \varepsilon) \to \mathbb{R}$). Nach Kettenregel (Satz 2.2) folgt:

$$\frac{d}{dt} f(x + te)\Big|_{t=0} = u'(0) = J_f(g(0))\, J_g(0) = J_f(x)e = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x)\, e_i = \langle \operatorname{grad} f(x),\, e \rangle.$$

Die Richtungsableitung lässt sich also auf zweierlei Weise auffassen und berechnen: als Produkt der $1 \times n$ Matrix $J_f(x)$ mit der $n \times 1$ Matrix e; oder alternativ als Skalarprodukt zwischen Gradient von f an der Stelle x (einem Vektor im $\mathbb{R}^n$) und Richtungsvektor $e \in \mathbb{R}^n$. Wir halten dieses Ergebnis als Lemma fest:

Lemma 2.1 Ist f diff'bar im Punkt x, so gilt für alle Richtungen $e \in \mathbb{R}^n$, $|e| = 1$

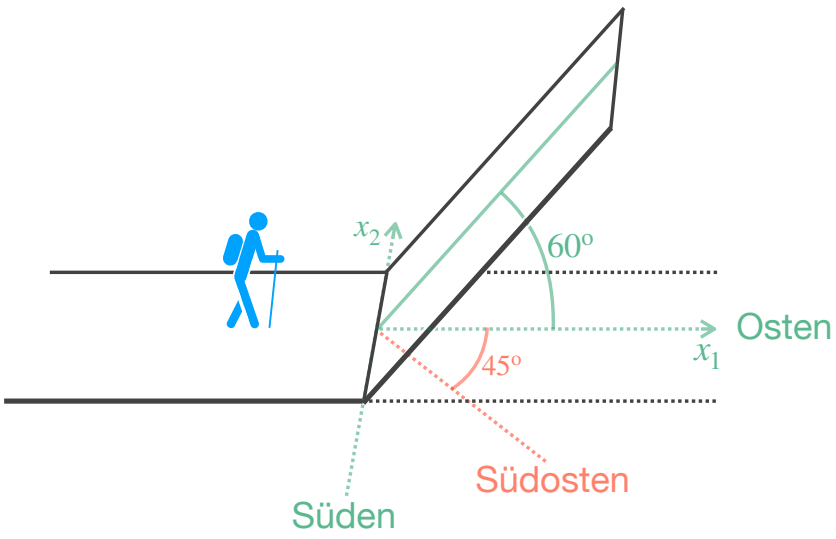
$$\partial_e f(x) = J_f(x)\, e = \langle \operatorname{grad} f(x),\, e \rangle.$$

Beispiel zur Richtungsableitung: *Bergsteigerin an einem* 60^o *steilen Hang.* Siehe Schaubild.

Liegt der Hang Richtung Osten und identifizieren wir wie üblich Osten mit der x_1-Achse und Norden mit der x_2-Achse, so ist der Hang der Graph der Höhenfunktion

$$h(x_1, x_2) = \sqrt{3} x_1, \quad h : \mathbb{R}^2 \to \mathbb{R}.$$

(Wieso $\sqrt{3}$? Weil $\cos 60^o = \frac{1}{2}$, $\sin 60^o = \frac{\sqrt{3}}{2}$, also $\tan 60^o = \sqrt{3}$, und so-und-so-viel Grad steil bedeutet Steigung gleich tan(so-und-so-viel Grad).)



Sieht steil aus. Bergsteigerin entscheidet, statt nach Osten schräg in Richtung Südosten zu gehen. *Was ist die Steigung? Vielleicht um die* 30^o*?* Überraschende Antwort: Viel

steiler als man denkt! Genauer:
Osten, d.h. $e = \begin{pmatrix}1\\0\end{pmatrix}$: die Richtungsableitung ist

$$\frac{d}{dt}h(0+te)\Big|_{t=0} = \langle \operatorname{grad} h(x),\, e\rangle = \left\langle \begin{pmatrix}\sqrt{3}\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}\right\rangle = \sqrt{3};$$

dies entspricht $\arctan(\sqrt{3}) = 60^o$.
Süden, d.h. $e = \begin{pmatrix}0\\-1\end{pmatrix}$: die Richtungsableitung ist

$$\frac{d}{dt}h(0+te)\Big|_{t=0} = \left\langle \begin{pmatrix}\sqrt{3}\\0\end{pmatrix}, \begin{pmatrix}0\\-1\end{pmatrix}\right\rangle = 0;$$

dies entspricht $\arctan(0) = 0^o$.
Südosten, d.h. $e = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}$: die Richtungsableitung ist

$$\frac{d}{dt}h(0+te)\Big|_{t=0} = \left\langle \begin{pmatrix}\sqrt{3}\\0\end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}\right\rangle = \sqrt{\tfrac{3}{2}};$$

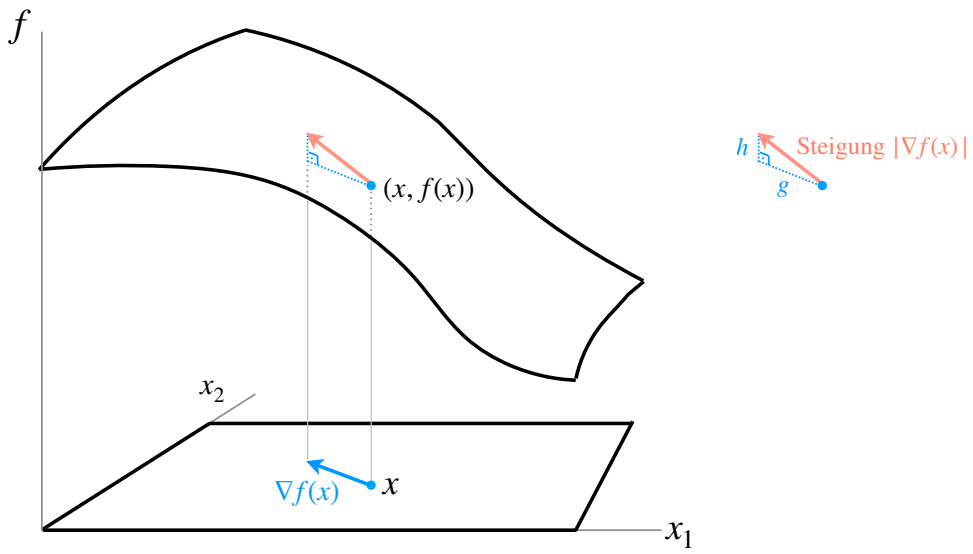
dies entspricht $\arctan(\sqrt{\frac{3}{2}}) \approx 51^o$. *Immer noch sehr steil. Wieso?*
Einer von vielen Sachverhalten, die durch folgenden Satz aufgehellt werden (siehe die Diskussion am Ende des Beweises).

Satz 2.3 (Geometrische Bedeutung des Gradienten) *Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$ differenzierbar, $x \in \Omega$.*
a) Der Gradient $\operatorname{grad} f(x)$ zeigt in Richtung des stärksten Anstiegs von f; die Norm des Gradienten gibt die Steigung von f in dieser Richtung an. D.h.

$$\frac{d}{dt}f(x+te)\Big|_{t=0} \le |\operatorname{grad} f(x)| \ \forall e \in \mathbb{R}^n \text{ mit } |e| = 1, \ \text{“=”} \Longleftrightarrow \operatorname{grad} f(x) = \lambda e \text{ für ein } \lambda \ge 0.$$

b) Minus der Gradient, $-\operatorname{grad} f(x)$, zeigt in Richtung des steilsten Abfalls von f; minus die Norm des Gradienten gibt die Steigung von f in dieser Richtung an. D.h.

$$\frac{d}{dt}f(x+te)\Big|_{t=0} \ge -|\operatorname{grad} f(x)| \ \forall e \in \mathbb{R}^n \text{ mit } |e| = 1, \ \text{“=”} \Longleftrightarrow \operatorname{grad} f(x) = \lambda e \text{ für ein } \lambda \le 0.$$



Der Gradient (blauer Vektor) liegt in der $(x_1, ..., x_N)$-Ebene, und zeigt in Richtung des stärksten Anstiegs von f. Seine Länge ist die Steigung, oder Richtungsableitung, in dieser Richtung.

Der Tangentialvektor an den Funktionsgraphen (roter Vektor), dessen Projektion auf die $(x_1, ..., x_N)$-Ebene den Gradienten ergibt, ist gleich $(\nabla f(x), |\nabla f(x)|^2)$. Das eingezeichnete (rechtwinklige) Steigungsdreieck muss nämlich wegen obiger Eigenschaft des Gradienten die Steigung $h/g = |\nabla f(x)|$ besitzten, aber nach Konstruktion ist die Grundlinie $g = |\nabla f(x)|$, und somit $h = |\nabla f(x)|^2$.

Beweis: Die beiden Ungleichungen folgen sofort aus Lemma 2.1 und der Cauchy-Schwarz-Ungleichung. Dass Gleichheit genau in den behaupteten Situationen eintritt, ist eine Konsequenz des nachfolgenden Lemmas.

Lemma 2.2 (Gleichheit in der Cauchy-Schwarz-Ungleichung) Seien $v, w \in \mathbb{R}^n$, $w \neq 0$.

(i) In der Cauchy-Schwarz-Ungleichung "$|\langle v, w\rangle| \leq |v||w|$" gilt Gleichheit genau dann, wenn v ein skalares Vielfaches von w ist, d.h. wenn $v = \lambda w$ für ein $\lambda \in \mathbb{R}$.

(ii) $\langle v, w\rangle = |v||w|$ (bzw. $= -|v||w|$) g.d.w. zusätzlich $\lambda \geq 0$ (bzw. ≤ 0).

Beweis: "wenn" ist elementar. Wir zeigen "nur dann wenn". Im Beweis der Cauchy-Schwarz-Ungleichung hatten wir durch Ausmultiplizieren des Ausdrucks $|v-\lambda w|^2$ und Einsetzen von $\lambda = \langle v, w\rangle/|w|^2$ folgende Identität erhalten:

$$|v - \lambda w|^2 = |v|^2 - \frac{|\langle v, w\rangle|^2}{|w|^2}.$$

"Nur dann wenn" in (i): Es gelte Gleichheit in der Cauchy-Schwarz-Ungleichung. Dies ist offenbar äquivalent zu "rechte Seite gleich Null", und folglich zu "linke Seite

gleich Null", also zu $v = \lambda w$ mit obigem speziellen λ; insbesondere folgt die Existenz eines solchen λ. Dies etabliert die "nur dann wenn" Aussage in (i). Die "nur dann wenn" Aussagen in (ii) folgen ebenfalls, denn wenn z.B. $\langle v, w\rangle = |v|\,|w|$, ist $\langle v, w\rangle \geq 0$, und somit auch das spezielle λ.

Bergsteigerin, revisited: Wir kommen nun auf unser merkwürdiges Ergebnis zurück: *Obwohl die Bergsteigerin ihre Gangrichtung um 45^o gedreht hat, ist die Steigung des Hanges fast gleich geblieben.* Wieso?

Die Antwort lautet: Die ursprüngliche Gangrichtung zeigt in Richtung des steilsten Anstieges des Hanges (dies sieht man durch Konsultieren des Schaubildes oder durch Berechnen des Gradienten $\nabla h(x)$ und Satz 2.3). D.h. in diese Richtung ist die Richtungsableitung von h maximal. Also ist (nach einem grundlegenden Sachverhalt aus Analysis 1) die Ableitung der Richtungsableitung nach dem Winkel φ der Gangrichtung zur Ost-Achse gleich Null an der Stelle $\varphi = 0$. Also ändert sich die Steigung des Hanges bei kleiner Änderung der Gangrichtung zunächst fast überhaupt nicht.

Unser Beispiel erklärt, warum Serpentinen typischerweise fast waagerecht zum Hang verlaufen; der Winkel zur Richtung des steilsten Anstiegs liegt weit weg von 0^o und viel näher an 90^o. Illustrative Bilder sind leicht im Netz zu finden.

Nachdem wir erarbeitet haben, in welche Richtung der Gradient einer Funktion f zeigt (nämlich in Richtung des steilsten Anstiegs), untersuchen wir die Beziehung zwischen Gradient und Niveaumengen von f. Zur Vorbereitung erinnern wir an den Begriff und die geometrische Bedeutung des Tangentialvektors einer Kurve (siehe die Diskussion am Ende von §2.1): Ist $I \subseteq \mathbb{R}$ Intervall und $\gamma : I \to \mathbb{R}^n$ differenzierbare Kurve, so heisst die Ableitung

$$\lim_{h\to 0} \frac{1}{h}\big(\gamma(t+h) - \gamma(t)\big) = \gamma'(t) \in \mathbb{R}^n$$

Tangentialvektor der Kurve im Punkt $\gamma(t)$; zeichnen wir ihn als Vektor mit Fusspunkt $\gamma(t)$, liegt er tangential an die Kurve.

Satz 2.4 (Beziehung zwischen Gradient und Niveaumengen) *Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$ differenzierbar, sei N_d die Niveaumenge von f zum Wert d, d.h.*

$$N_d = \{x \in \Omega : f(x) = d\},$$

und sei $x^* \in N_d$. Dann gilt: Der Gradient von f im Punkt x^* steht senkrecht auf allen Tangentialvektoren differenzierbarer Kurven in der Niveaumenge im Punkt x^*. D.h. wenn $\gamma : I \to N_d$ differenzierbare Kurve in N_d (d.h. $I \subseteq \mathbb{R}$ offen, γ differenzierbar, $\gamma(I) \subseteq N_d$) mit $\gamma(t_*) = x^*$ für ein $t_* \in I$, so gilt

$$\langle grad\, f(x^*), \gamma'(t_*)\rangle = 0.$$

Saloppe Sprechweise: "Der Gradient steht senkrecht auf der Niveaumenge". (Damit ist natürlich nicht die offensichtlich falsche Aussage gemeint, der Gradient stünde senkrecht auf allen Elementen der Niveaumenge. Man kann zeigen: die Gesamtheit aller Tangentialvektoren differenzierbarer Kurven in der Niveaumenge im Punkt x^* bildet – z.B. unter der typischerweise erfüllten Voraussetzung $\nabla f(x) \neq 0$ – einen $\mathbb{R}$-Vektorraum, den sogenannten *Tangentialraum* der Niveaumenge im Punkt x^*. Besser wäre also die Sprechweise "Der Gradient steht senkrecht auf dem Tangentialraum der Niveaumenge".

Beispiel: $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = 2 - x_1^2 - x_2^2$, $d > 0$. Siehe die Skizze zu Beginn dieses Abschnitts. Die Niveaumenge N_d ist die Kreislinie um 0 vom Radius $r = \sqrt{d}$. Zu festem aber beliebigem $x^* \in N_d$ wählen wir Polarkoordinaten r und t_* sodass $x_1^* = r\cos t_*$, $x_2^* = r\sin t_*$ (zum Beweis der – anschaulich offensichtlichen – Existenz von Polarkoordinaten siehe Analysis 1). Die Kurve $\gamma : \mathbb{R} \to \mathbb{R}^2$,

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

ist also differenzierbare Kurve in N_d mit $\gamma(t_*) = x^*$. Der Tangentialvektor der Kurve im Punkt $\gamma(t_*)$ ist

$$\gamma'(t_*) = \begin{pmatrix} -r\sin t_* \\ r\cos t_* \end{pmatrix} = \begin{pmatrix} -x_2^* \\ x_1^* \end{pmatrix}.$$

Laut Satz 2.4 muss also gelten:

$$\left\langle \operatorname{grad} f(x^*), \begin{pmatrix} -x_2^* \\ x_1^* \end{pmatrix} \right\rangle = 0.$$

Dies sieht man in der Skizze. Zur Probe berechnen wir:

$$\operatorname{grad} f(x^*) = -\begin{pmatrix} 2x_1^* \\ 2x_2^* \end{pmatrix}$$

und somit $\langle \operatorname{grad} f(x^*), \gamma'(t_*) \rangle = 2x_1^* x_2^* - 2x_2^* x_1^* = 0$.

Beweis von Satz 2.4 Die Verkettung $t \mapsto f(\gamma(t))$ ist wegen der Voraussetzung $\gamma(I) \subseteq N_d$ konstant und daher

$$0 = \frac{d}{dt} f(\gamma(t)) \underset{\text{Kettenregel}}{=} J_f(\gamma(t))\,\gamma'(t) = \langle \operatorname{grad} f(\gamma(t)), \gamma'(t) \rangle \quad \text{für alle } t.$$

Auswerten an der Stelle $t = t_*$ liefert die Behauptung.

Weitere Beispiele siehe Übungen.

2.4 Höhere partielle Ableitungen

Def. 2.5 Sei $\Omega \subseteq \mathbb{R}^n$ offen,

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix} : \Omega \to \mathbb{R}^m$$

partiell differenzierbar. Falls die partielle Ableitung $\frac{\partial f}{\partial x_j} = \partial_j f : \Omega \to \mathbb{R}^m$ partiell differenzierbar nach x_i im Punkt $x \in \Omega$, heisst

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) := \frac{\partial}{\partial x_i}\Big(\frac{\partial f}{\partial x_j}\Big)(x)$$

zweite partielle Ableitung von f nach x_i, x_j im Punkt x. Alternative Schreibweise: $\partial_{ij} f(x)$.

Zweite partielle Ableitungen sind also durch hintereinander Ausführen der Ableitungen von rechts nach links definiert. Typischerweise kommt es nicht auf die Reihenfolge an, siehe Satz 2.5 unten. Falls $i = j$, schreibt man

$$\frac{\partial^2 f}{\partial x_i^2}(x) := \frac{\partial^2 f}{\partial x_i \partial x_i}(x).$$

Alternative Schreibweise: $\partial_{ii} f(x)$. Sind alle (ersten) partiellen Ableitungen $\frac{\partial f}{\partial x_j}$ $(j = 1, ..., n)$ an allen Punkten $x \in \Omega$ partiell differenzierbar nach allen Koordinaten x_i $(i = 1, ..., n)$, heisst f 2mal partiell differenzierbar.
Analog sind höhere partielle Ableitungen

$$\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}(x), \quad i_1, ..., i_k \in \{1, ..., n\},$$

definiert.

Satz 2.5 (Satz von Schwarz) *Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$ 2mal partiell diff'bar, und seien die zweiten partiellen Ableitungen von f stetig im Punkt x. Dann gilt*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x).$$

In manchen Anwendungen ist es nützlich, die zweiten partiellen Ableitungen als Matrix zusammenzufassen. Die *Matrix der zweiten partiellen Ableitungen* (oder *Hesse-Matrix*) von f im Punkt x ist definiert als

$$H_f(x) := \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{pmatrix}.$$

Der Satz von Schwarz besagt dann: die Hesse-Matrix ist symmetrisch.

Beweis Es reicht, den Fall $n = 2$ zu betrachten, denn für gegebenes x hängen obige partielle Ableitungen am Punkt x nur von der Funktion zweier Variablen $(\tilde{x_i}, \tilde{x_j}) \mapsto f(x_1, ..., x_{i-1}, \tilde{x_i}, x_{i+1}, ..., x_{j-1}, \tilde{x_j}, x_{j+1}, ..., x_n)$ ab. Sei also $f = f(x, y)$. Dann ist

$$\begin{aligned}\frac{\partial^2 f}{\partial x \partial y}(x,y) &= \lim_{h\to 0} \frac{\frac{\partial f}{\partial y}(x+h,y) - \frac{\partial f}{\partial y}(x,y)}{h} \\ &= \lim_{h\to 0}\Big(\lim_{k\to 0} \underbrace{\frac{f(x+h,y+k) - f(x+h,y) - f(x,y+k) + f(x,y)}{h\,k}}_{=:A(h,k)}\Big)\end{aligned}$$

und analog

$$\frac{\partial^2 f}{\partial y \partial x}(x,y) = \lim_{k\to 0}\Big(\lim_{h\to 0} A(h,k)\Big).$$

Wir wissen aus Analysis 1: oft, aber selbst in anschaulichen Alltagsbeispielen nicht immer ist die Reihenfolge zweier Grenzwerte vertauschbar. Was tun?

Um die Voraussetzung der Stetigkeit der zweiten partiellen Ableitungen ins Spiel zu bringen, schreiben wir $A(h,k)$ via Hauptsatz als Integral:

$$A(h,k) = \frac{1}{hk}\int_x^{x+h}\Big(\frac{\partial f}{\partial x}(s,y+k) - \frac{\partial f}{\partial x}(s,y)\Big)\,ds = \frac{1}{hk}\int_x^{x+h}\Big(\int_y^{y+k}\frac{\partial^2 f}{\partial y \partial x}(s,t)\,dt\Big)\,ds.$$

Wegen Stetigkeit der zweiten partiellen Ableitung im Punkt (x, y) existiert zu gegebenem $\varepsilon > 0$ ein $\delta > 0$ sodass

$$\Big|\frac{\partial^2 f}{\partial y \partial x}(s,t) - \frac{\partial^2 f}{\partial y \partial x}(x,y)\Big| \le \varepsilon \quad \forall |s-x|,\, |t-y| \le \delta.$$

Mithilfe der Standardabschätzung *Absolutbetrag eines Integrals ≤ Länge des Integrationsgebietes mal Supremum des Integranden* (Analysis 1 Lemma 11.1) folgt

$$\Big|A(h,k) - \frac{\partial^2}{\partial y \partial x} f(x,y)\Big| = \Big|\frac{1}{hk}\int_x^{x+h}\Big[\int_y^{y+k}\Big(\frac{\partial^2 f}{\partial y \partial x}(s,t) - \frac{\partial^2 f}{\partial y \partial x}(x,y)\Big)\,dt\Big]\,ds\Big| \le \frac{1}{hk}\cdot hk\varepsilon = \varepsilon$$

für alle $|h|, |k| \le \delta$. In dieser Ungleichung können wir nun h, k in derjenigen Reihenfolge gegen Null gehen lassen, die die andere gemischte partielle Ableitung ergibt:

$$\varepsilon \ge \lim_{h\to 0}\Big(\lim_{k\to 0}\Big|A(h,k) - \frac{\partial^2 f}{\partial y \partial x}(x,y)\Big|\Big) \underset{|\cdot| \text{ stetig}}{=} \Big|\underbrace{\lim_{h\to 0}\Big(\lim_{k\to 0} A(h,k)\Big)}_{=\frac{\partial^2 f}{\partial x \partial y}(x,y)} - \frac{\partial^2 f}{\partial y \partial x}(x,y)\Big|.$$

Da $\varepsilon > 0$ beliebig, folgt die Behauptung.

Im Beweis haben wir die Stetigkeit der anderen zweiten partiellen Ableitung gar nicht benutzt. Die Voraussetzung des Satzes von Schwarz kann also zu f stetig

diff'bar, $\frac{\partial f}{\partial x_i}$ partiell nach x_j diff'bar, $\frac{\partial^2 f}{\partial x_j \partial x_i}$ stetig abgeschwächt werden, dann folgt sowohl die Existenz der zweiten partiellen Ableitung in anderer Reihenfolge als auch die behauptete Gleichheit.

Die Voraussetzung der Stetigkeit mindestens einer der beiden zweiten partiellen Ableitungen ist in der Praxis typischerweise erfüllt, kann aber nicht weggelassen werden. Wir konstruieren ein Gegenbeispiel.[7] Es reicht aus, Funktionen auf dem $\mathbb{R}^2$ zu betrachten; wir suchen also eine Funktion mit

$$\frac{\partial^2 f}{\partial x \partial y}(0,0) \neq \frac{\partial^2 f}{\partial y \partial x}(0,0). \tag{*}$$

Sagen wir, die linke Seite soll 1 und die rechte Seite -1 sein. Die linke Seite hängt nur von den Werten von $\frac{\partial f}{\partial y}$ auf der x-Achse ab, d.h. von $\frac{\partial f}{\partial y}(x,0)$. Setzen wir also

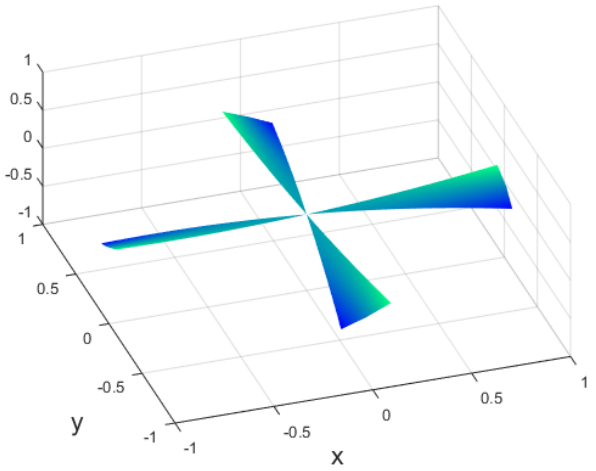
$$\frac{\partial f}{\partial y}(x,0) = x \quad \text{für alle } x \in \mathbb{R}, \tag{1}$$

so ist die linke Seite von (*) gleich 1. Analog hängt die rechte Seite nur von den Werten von $\frac{\partial f}{\partial x}$ auf der y-Achse ab, d.h. von $\frac{\partial f}{\partial x}(0,y)$. Setzen wir

$$\frac{\partial f}{\partial x}(0,y) = -y \quad \text{für alle } y \in \mathbb{R}, \tag{2}$$

ist die rechte Seite von (*) gleich -1. Die beiden Bedingungen (1) und (2) implizieren also (*). Wie basteln wir eine Funktion, die (1) und (2) erfüllt? Für (1) reicht aus, dass $f \approx xy$ nahe der x-Achse. Für (2) reicht aus: $f \approx -xy$ nahe der y-Achse. Anschaulich bedeutet dies: bewegt man sich im Gegenuhrzeigersinn auf dem Einheitskreis in der (x,y)-Ebene, so sollte beim Durchqueren der positiven x-Achse, der positiven y-Achse, der negativen x-Achse und der negativen y-Achse der Wert von f immer Null und die Steigung von f immer $+1$ sein. Anders gesagt: der Funktionsgraph in der Nähe der beiden Achsen ist ein "Hubschrauberpropeller".

[7] In vielen exzellenten Lehrbüchern und Skripten fällt ein solches Gegenbeispiel vom Himmel; Studierende sollen nur durch Ausrechnen der partiellen Ableitungen die Nichtvertauschbarkeit nachprüfen. Auf diese Weise versteht man nicht, was die beiden gemischten partiellen Ableitungen $\frac{\partial^2 f}{\partial x \partial y}(x,y)$ und $\frac{\partial^2 f}{\partial y \partial x}(x,y)$ eigentlich geometrisch über den Graphen einer Funktion zweier Veränderlicher aussagen, und erst recht nicht, warum sie das Verhalten des Graphen in völlig verschiedenen Regionen betreffen und deshalb beliebige Wertepaare annehmen können.



Die Funktion $f(x,y) = \begin{cases} xy & \text{nahe der } x\text{-Achse} \\ -xy & \text{nahe der } y\text{-Achse.} \end{cases}$

Fehlt nur noch ein nahtloser Übergang, der f "zwischendurch" von positiven auf negative Werte absenkt, damit f die nächste Achse von unten kommend mit Steigung 1 erreichen kann. Diesen Job erledigten wir, indem wir den Vorfaktor von xy nahtlos zwischen 1 auf der x-Achse und -1 auf der y-Achse variieren. Dies tut man am einfachsten durch Benutzung von Polarkoordinaten (siehe Analysis 1 Satz 7.7) $x = r\cos\varphi$, $y = r\sin\varphi$: wir wollen zwischen 1 und -1 interpolieren, also bietet sich eine Sinus- oder Cosinusfunktion des Winkels φ an, und die gewünschten Werte 1 auf der x-Achse (also bei $\varphi = 0, \pi$) und -1 auf der y-Achse (also bei $\varphi = \pi/2, 3\pi/2$) ergeben sich, wenn man $\cos(2\varphi)$ nimmt. Wir setzen also

$$f(x,y) = \cos(2\varphi) \cdot xy \tag{3}$$

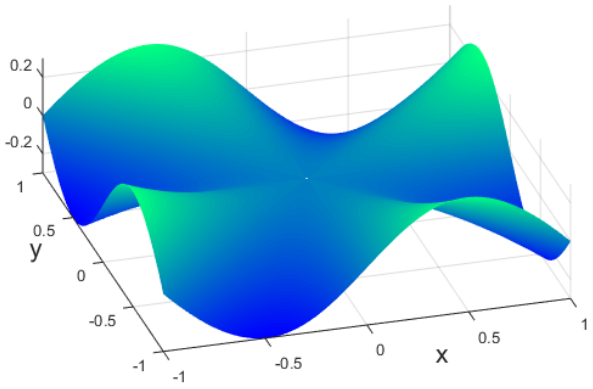
(dies ist wohldefiniert für $(x,y) \neq (0,0)$; im Nullpunkt setzen wir $f(0,0) = 0$). Siehe Schaubild. Zum Schluss drücken wir den Cosinus-Faktor noch mithilfe von x und y aus. Nach Doppelwinkelformel gilt

$$\cos(2\varphi) = \cos^2\varphi - \sin^2\varphi = (\tfrac{x}{r})^2 - (\tfrac{y}{r})^2 = \frac{x^2-y^2}{x^2+y^2},$$

d.h. unsere Funktion ist

$$f(x,y) = \frac{x^2-y^2}{x^2+y^2} \cdot xy. \tag{3}$$

Wie man leicht nachrechnet, erfüllt obiges f die Bedingungen (1) und (2) und somit folgt (*).



Die Funktion $f(x,y) = \frac{x^2-y^2}{x^2+y^2} \cdot xy$.

Ausblick: *Höhere totale Ableitungen.* Dieses Konzept benötigen wir in dieser Vorlesung nicht; für interessierte Studierende sei es aber kurz erklärt. Das Konzept ist einerseits rechentechnisch unhandlich, aber andererseits konzeptuell elegant: *die zweite totale Ableitung ist die totale Ableitung der ersten totalen Ableitung.* Um dies präzise zu machen, betrachten wir – etwas allgemeiner als in Abschnitt 2.2 – Abbildungen

$$f : \Omega \subseteq \mathbb{R}^n \to V$$

von einer offenen Teilmenge des $\mathbb{R}^n$ in einen beliebigen endlichdimensionalen Vektorraum V versehen mit einer Norm $\|\cdot\|$. Die totale Ableitung im Punkt $x \in \Omega$ ist dann diejenige lineare Abbildung

$$Df(x) : \mathbb{R}^n \to V$$

mit der Bestapproximationseigenschaft $\frac{\|f(x+h)-f(x)-Df(x)(h)\|}{|h|} \to 0$ $(|h| \to 0)$. Wir betrachten nun eine Abbildung

$$f : \Omega \to V_0 = \mathbb{R}^m.$$

Die totale Ableitung im Punkt $x \in \Omega$ ist eine lineare Abbildung vom $\mathbb{R}^n$ in den Bildraum von f, $Df(x) : \mathbb{R}^n \to V_0$, d.h.

$$Df(x) \in V_1 = \mathcal{L}(\mathbb{R}^n, V_0) = \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m),$$

wobei $\mathcal{L}(\mathbb{R}^n, V_0)$ den Vektorraum der linearen Abbildungen von $\mathbb{R}^n$ nach V_0 bezeichnet. Die totale Ableitung als Funktion von x ist somit eine Abbildung mit *selbem Definitionsbereich wie f aber neuem Bildraum*,

$$Df : \Omega \to V_1.$$

Versehen wir den Vektorraum V_1 mit einer Norm, z.B. indem wir einer linearen Abbildung $L \in V_1$ ihre darstellende Matrix $A \in \mathbb{R}^{m\times n}$ bezüglich der Standardbasen von $\mathbb{R}^n$ und $\mathbb{R}^m$ zuordnen und die euklidische Norm $\|L\| = (\sum_{i,j} A_{ij}^2)^{1/2}$ nehmen, können wir das Konzept der totalen Ableitung auf Df anwenden:

$$D^2 f(x) := D\,Df\,(x) \in V_2 = \mathcal{L}(\mathbb{R}^n, V_1) = \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)).$$

Die zweite totale Ableitung von f im Punkt x ist also eine lineare Abbildung vom $\mathbb{R}^n$ in den Bildraum von Df, d.h. den Raum der linearen Abbildungen vom $\mathbb{R}^n$ in den $\mathbb{R}^m$. Folglich ist $D^2 f$ als Funktion von x eine Abbildung

$$D^2 f : \Omega \to V_2.$$

Analog ist dann $D^3 f(x) := D\,D^2 f(x) \in V_3 = \mathcal{L}(\mathbb{R}^n, V_2) = \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)))$ usw. Zusammenfassung: $D^k f : \Omega \to V_k$ ist die totale Ableitung von $D^{k-1} f$, mit Bildraum $V_k = \mathcal{L}(\mathbb{R}^n, V_{k-1})$.

Die obige Konstruktion ist recht abstrakt. Eine Koordinatendarstellung von $D^2f(x)$ sieht wie folgt aus: $D^2f(x)$ ist eine lineare Abbildung vom $\mathbb{R}^n$ in den Vektorraum V_1 der linearen Abbildungen vom $\mathbb{R}^n$ in den $\mathbb{R}^m$; $D^2f(x)(h)$ ist für gegebenes $h \in \mathbb{R}^n$ eine lineare Abbildung vom $\mathbb{R}^n$ in den $\mathbb{R}^m$; $D^2f(x)(h)(k)$ ist für gegebenes $k \in \mathbb{R}^n$ ein Vektor im $\mathbb{R}^m$. Diesen Vektor können wir wie folgt mithilfe der zweiten partiellen Ableitungen von f ausdrücken:

$$D^2f(x)(h)(k) = \begin{pmatrix} h^t H_{f_1}(x)k \\ \vdots \\ h^t H_{f_m}(x)k \end{pmatrix} \in \mathbb{R}^m \quad \text{für alle } h, k \in \mathbb{R}^n.$$

Hierbei ist $H_{f_i}(x)$ die Hesse-Matrix der i-ten Komponentenfunktion $f_i : \Omega \to \mathbb{R}$ im Punkt x. (Diese Formel ist das Analogon der Formel "$Df(x)(h) = J_f(x)h$ für alle $h \in \mathbb{R}^n$" aus Satz 2.1 für die erste Ableitung.) Insbesondere ist im skalaren Fall ($m = 1$) $D^2f(x)(h)(k) = h^t H_f(x)k$ für alle $h, k \in \mathbb{R}^n$.

Wir merken noch folgendes an. Der Raum $\mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m))$ ist isomorph zum Raum der bilinearen Abbildungen $B : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$, indem wir $L \in \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m))$ als Abbildung $(h, k) \mapsto L(h)(k)$ auffassen. (Bilinear bedeutet linear in jedem der beiden Argumente.) Analog ist der Raum $\mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)))$ isomorph zum Raum der trilinearen Abbildungen $T : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$, usw. Manche Autoren fassen deshalb die k-te totale Ableitung einer skalaren Funktion im Punkt x als k-lineare Abbildung $\mathbb{R}^n \times \ldots \times \mathbb{R}^n \to \mathbb{R}^m$ auf.

2.5 Divergenz, Rotation, Laplaceoperator

In vielen Anwendungen treten die folgenden universellen Kombinationen partieller Ableitungen auf.

Def. 2.6 a) Sei $\Omega \subseteq \mathbb{R}^n$ offen, $v : \Omega \to \mathbb{R}^n$ differenzierbares Vektorfeld. Die Zahl

$$\nabla \cdot v(x) = \operatorname{div} v(x) = \sum_{j=1}^{n} \partial_j v_j(x) = \sum_{j=1}^{n} \frac{\partial v_j}{\partial x_j}(x)$$

heisst *Divergenz von v im Punkt x*. Die skalare Funktion $\nabla \cdot v = \operatorname{div} v : \Omega \to \mathbb{R}$ heisst *Divergenz von v*.

b) Sei $\Omega \subseteq \mathbb{R}^3$ offen, $v : \Omega \to \mathbb{R}^3$ differenzierbares Vektorfeld im $\mathbb{R}^3$. Der Vektor

$$\nabla \times v(x) = \operatorname{curl} v(x) = \operatorname{rot} v(x) = \begin{pmatrix} \partial_2 v_3(x) - \partial_3 v_2(x) \\ \partial_3 v_1(x) - \partial_1 v_3(x) \\ \partial_1 v_2(x) - \partial_2 v_1(x) \end{pmatrix}$$

heisst *Rotation von v im Punkt x*. Das Vektorfeld $\nabla \times v = \operatorname{curl} v = \operatorname{rot} v : \Omega \to \mathbb{R}^3$ heisst *Rotation von v*.

c) Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$ 2mal differenzierbare skalare Funktion. Die skalare Funktion

$$\Delta f = \sum_{j=1}^{n} \partial_{jj} f = \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j^2}$$

heisst *Laplace von f*. Der Operator $\Delta : f \mapsto \Delta f$ heisst *Laplace-Operator*.

Zum bequemen Merken und Ausrechnen kann man die Rotation symbolisch als 3×3 "Determinante" schreiben[8]

$$\operatorname{rot} v = \begin{vmatrix} \vec{e_1} & \vec{e_2} & \vec{e_3} \\ \partial_1 & \partial_2 & \partial_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$$

[8] sorry an die Lineare Algebra, ich weiss, da stehen verbotene Objekte in der Determinante

(hierbei sind die $\overrightarrow{e_i}$ die üblichen Einheitsvektoren im $\mathbb{R}^3$) und diese nach der 1. Zeile entwickeln

$$= \overrightarrow{e_1}\begin{vmatrix}\partial_2 & \partial_3\\ v_2 & v_3\end{vmatrix} - \overrightarrow{e_2}\begin{vmatrix}\partial_1 & \partial_3\\ v_1 & v_3\end{vmatrix} + \overrightarrow{e_3}\begin{vmatrix}\partial_1 & \partial_2\\ v_1 & v_2\end{vmatrix},$$

der Koeffizienten von $\overrightarrow{e_i}$ ist dann die i-te Komponente der Rotation.

Beispiele 1) "Quelle": Das Vektorfeld $v : \mathbb{R}^3 \to \mathbb{R}^3$, $v(x) = x$ (siehe §1.6.3, linkes Bild) erfüllt

$$\operatorname{div} v = 3, \quad \operatorname{rot} v = 0.$$

2) "Wirbel": Das Vektorfeld $v : \mathbb{R}^3 \to \mathbb{R}^3$, $v(x) = \begin{pmatrix}-x_2\\ x_1\\ 0\end{pmatrix}$ (§1.6.3, rechtes Bild) erfüllt

$$\operatorname{rot} v = \begin{pmatrix}0\\ 0\\ 2\end{pmatrix}, \quad \operatorname{div} v = 0.$$

Anschauung: Die Divergenz misst das Vorhandensein und die Stärke von "Quellen". Die Rotation misst das Vorhandensein und die Stärke von "Wirbeln". Diese Anschauung wird in Analysis 3 mithilfe mehrdimensionaler Integration (Sätze von Gauss und Stokes) untermauert.

3) "Newton-Potential": Die Funktion $f : \mathbb{R}^3\backslash\{0\} \to \mathbb{R}$, $f(x) = \frac{1}{|x|}$, erfüllt

$$\Delta f = 0.$$

Grundlegende Beziehungen zwischen Gradient, Divergenz und Rotation sind:

Lemma 2.3 Sei $\Omega \subseteq \mathbb{R}^n$ offen, und seien $f : \Omega \to \mathbb{R}$, $v : \Omega \to \mathbb{R}^n$ 2mal stetig diff'bar.
1) $\operatorname{div}(\nabla f) = \Delta f$
2) $\operatorname{rot}(\nabla f) = 0$ $(n = 3)$
3) $\operatorname{div}(\operatorname{rot} v) = 0$ $(n = 3)$.

Beweis 1) folgt aus der Definition von Divergenz und Gradient und 2), 3) aus dem Satz von Schwarz.

Dieses Lemma erklärt die Tatschen $\operatorname{rot} v = 0$ bzw. $\operatorname{div} v = 0$ in Beispielen 1) und 2), denn

$$x = \nabla\frac{|x|^2}{2}, \qquad \begin{pmatrix}-x_2\\ x_1\\ 0\end{pmatrix} = \operatorname{rot}\begin{pmatrix}0\\ 0\\ -\frac{x_1^2+x_2^2}{2}\end{pmatrix}.$$

Für mehr zu Aussage 2) des Lemmas siehe §7.

2.6 Taylorentwicklung im Mehrdimensionalen

Auch im Mehrdimensionalen gilt: jede hinreichend oft differenzierbare Funktion kann in der Nähe eines gegebenen Punktes in ein Polynom und einen kleinen Rest aufgespalten werden.

In 1D findet man Polynom und Rest wie folgt (Analysis 1 Satz 9.5): falls $f : \mathbb{R} \to \mathbb{R}$ (k+1)-mal stetig differenzierbar, ist

$$f(x_0+h) = \underbrace{f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2!}h^2 + \ldots + \frac{f^{(k)}(x_0)}{k!}h^k}_{=:T_k(x_0;h)} + \underbrace{\frac{f^{(k+1)}(\xi)}{(k+1)!}h^{k+1}}_{=:R_k(x_0;h)},$$

für ein ξ zwischen x_0 und x_0+h, wobei $\frac{R_k(x_0;h)}{h^k} \to 0$ $(h \to 0)$. Die Koeffizienten des Taylorpolynoms sind also durch Ableitungen der Funktion f am Entwicklungspunkt x_0 gegeben.

Im Mehrdimensionalen ist die analytische Form, und die praktische Berechnung, von Taylorpolynomen komplizierter als in 1D, da Funktionen auf dem $\mathbb{R}^n$ von vielen Koordinaten abhängen und höhere partielle Ableitungen nach allen möglichen Kombinationen von Koordinaten, also z.B. $\frac{\partial^3 f}{\partial x_1^2 \partial x_3}$, auftreten. Diese Schwierigkeit ist aber "nur" notationell bzw. rechenaufwandstechnisch; prinzipiell kann die mehrdimensionale Taylorentwicklung direkt aus der 1D Taylorentwicklung (Analysis 1 Satz 9.5) gefolgert werden, ohne deren Beweis anfassen oder kennen zu müssen.

Dafür ist die geometrische Bedeutung der Taylorentwicklung interessanter als in 1D: der Graph einer allgemeinen 2D Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ ist eine *2D Fläche im 3D Raum*; die Taylorentwicklung sagt uns, dass und wie wir eine solche Fläche durch *einfache spezielle Flächen* (nämlich Graphen von Polynomen) annähern können. Graphische Beispiele siehe unten.

Notation: Eine Funktion $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$ heisst *k-mal stetig diff'bar* oder *$\mathcal{C}^k$-Funktion*, wenn f k-mal partiell diff'bar und alle k-ten Ableitungen stetig.

Wir benutzen folgende übliche Schreibweise für die Verbindungsstrecke zwischen zwei Punkten $x, y \in \mathbb{R}^n$: $[x,y] := \{(1-t)x + ty : t \in [0,1]\}$. In 1D ist $[x,y]$ gleich dem Intervall $[\min\{x,y\}, \max\{x,y\}]$.

Satz 2.6 (Taylorentwicklung) *Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$ (k+1)-mal stetig diff'bar, $k \in \mathbb{N} \cup \{0\}$, $x_0 \in \Omega$.*
a) Für jedes $h \in \mathbb{R}^n$ sodass $[x_0, x_0+h] \subset \Omega$ existiert ein $\xi \in [x_0, x_0+h]$ sodass

$$f(x_0+h) = T_k(x_0;h) + R_k(x_0;h),$$

wobei

$$\begin{aligned}
T_0(x_0;h) &= f(x_0),\\
T_1(x_0;h) &= T_0(x_0;h) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0)\, h_i,\\
T_2(x_0;h) &= T_1(x_0;h) + \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(x_0)\, h_i h_j,\\
T_k(x_0;h) &= T_{k-1}(x_0;h) + \frac{1}{k!}\sum_{i_1=1}^{n} \ldots \sum_{i_k=1}^{n} \frac{\partial^k f}{\partial x_{i_1}\ldots\partial x_{i_k}}(x_0)\, h_{i_1}\ldots h_{i_k}
\end{aligned}$$

und

$$R_k(x_0;h) = \frac{1}{(k+1)!}\sum_{i_1=1}^{n} \ldots \sum_{i_{k+1}=1}^{n} \frac{\partial^{k+1} f}{\partial x_{i_1}\ldots\partial x_{i_{k+1}}}(\xi)\, h_{i_1}\ldots h_{i_{k+1}}.$$

b) Es gilt

$$\frac{R_k(x_0;h)}{|h|^k} \to 0 \quad (h \to 0).$$

Die Abbildung $h \mapsto T_k(x_0;h)$ heisst *Taylorpolynom k-ter Ordnung* von f an der Stelle x_0. Die Abbildung $h \mapsto R_k(x_0;h)$ heisst Restglied k-ter Ordnung.

Die Funktion $x \mapsto T_k(x_0;x-x_0)$ (d.h. das Taylorpolynom an der Stelle x_0 ausgewertet auf dem Inkrement $h = x - x_0$, aufgefasst als Funktion von x statt h) nennen wir in dieser Vorlesung *Taylorapproximation k-ter Ordnung* von f an der Stelle x_0, denn gemäss b) gilt

$$f(x) \approx T_k(x_0;x-x_0) \quad \text{falls } |x-x_0| \text{ klein.}$$

In der Literatur wird typischerweise notationell nicht zwischen den beiden Abbildungen $h \mapsto T_k(x_0;h)$ und $x \mapsto T_k(x_0;x-x_0)$ unterschieden, da aus dem Kontext klar ist, als Abbildung von was T_k jeweils aufgefasst wird.

Beweis b): Wir schätzen das Restglied mithilfe der Dreiecksungleichung sowie der trivialen Tatsache $|h_{i_j}| \le |h|$ für alle $j = 1,\ldots,k+1$ nach oben ab:

$$\frac{|R_k(x_0;h)|}{|h|^k} \le \frac{1}{(k+1)!}\sum_{i_1=1}^{n} \ldots \sum_{i_{k+1}=1}^{n} \Big|\frac{\partial^{k+1} f}{\partial x_{i_1}\ldots\partial x_{i_{k+1}}}(\xi)\Big| \underbrace{\frac{|h|^{k+1}}{|h|^k}}_{=|h|}.$$

Für $h \to 0$ konvergiert die oben auftretende partielle Ableitung wegen der Stetigkeitsvoraussetzung gegen ihren Wert an der Stelle x_0 und der Faktor $|h|$ konvergiert gegen 0; folglich strebt die rechte Seite gegen 0.

a): Betrachte die Funktion $g(t) := f(\varphi(t)) = f(x_0 + th)$ mit $\varphi(t) = x_0 + th$, $t \in [0,1]$. Offenbar gilt $g(0) = f(x_0)$, $g(1) = f(x_0+h)$. Anwenden der 1D Taylorformel auf g

mit $t_0 = 0$ und dem eindimensionalen (nicht mit dem Vektor h zu verwechselnden) Inkrement $h = 1$ liefert

$$\underbrace{g(1)}_{=f(x_0+h)} = \underbrace{g(0)}_{=f(x_0)} + g'(0)1 + \frac{g''(0)}{2!}1^2 + \ldots + \frac{g^{(k)}(0)}{k!}1^k + \frac{g^{(k+1)}(\tau)}{(k+1)!}1^{k+1} \text{ für ein } \tau \in [0,1]. \quad (*)$$

Ableiten von g durch wiederholte Anwendung der Kettenregel liefert

$$\begin{aligned}
g'(t) &= \underbrace{J_f(x_0+th)}_{=\left(\frac{\partial f}{\partial x_1}(x_0+th) \cdots \frac{\partial f}{\partial x_n}(x_0+th)\right)} \underbrace{\frac{d}{dt}(x_0+th)}_{=\begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}} = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x_0+th)\, h_i, \\
g''(t) &= \sum_{j=1}^n \sum_{i=1}^n \frac{\partial^2 f}{\partial x_j \partial x_i}(x_0+th)\, h_j h_i, \\
g^{(k)}(t) &= \sum_{i_1=1}^n \ldots \sum_{i_k=1}^n \frac{\partial^k f}{\partial x_{i_1} \ldots \partial x_{i_k}}(x_0+th)\, h_{i_1} \ldots h_{i_k}.
\end{aligned}$$

Auswerten der obigen Ableitungen an der Stelle $t = 0$ sowie der $(k+1)$-ten Ableitung $g^{(k+1)}(t)$ an der Stelle $t = \tau$ und Einsetzen in (*) liefert a), mit $\xi = x_0 + \tau h$.

Index-freie Darstellung des linearen und quadratischen Terms: die Summe im linearen Term kann als Skalarprodukt aufgefasst werden; im zweiten Term können wir die Summe über i als Skalarprodukt und diejenige über j als Matrix-Vektor-Produkt darstellen. Genauer:

$$T_1(x_0;h) - T_0(x_0;h) = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x_0)\, h_i = \langle \nabla f(x_0),\, h \rangle$$

und

$$T_2(x_0;h) - T_1(x_0;h) = \frac{1}{2} \sum_{i=1}^n \underbrace{\left(\sum_{j=1}^n \frac{\partial^2 f}{\partial x_j \partial x_i}(x_0) h_j \right)}_{=(H_f(x_0)h)_i} h_i = \frac{1}{2} \langle h,\, H_f(x_0)h \rangle,$$

wobei $H_f(x)$ die Hesse-Matrix von f im Punkt x ist, d.h. die Matrix mit Komponenten $(H_f(x))_{ij} = \frac{\partial^2 f}{\partial x_j \partial x_i}(x)$. Offenbar ist $H_f(x) = J_{\nabla f}(x)$, d.h. die Hesse-Matrix ist die Jacobi-Matrix des Gradienten. Nach Satz von Schwarz ist $H_f(x)$ symmetrisch, wenn die zweiten partiellen Ableitungen stetig sind.

Beispiel $f(x_1, x_2) = \log x_1 \cdot e^{x_2}$, $f : (0,\infty) \times \mathbb{R} \to \mathbb{R}$. Wir bestimmen die Taylorpolynome erster und zweiter Ordnung an der Stelle $x_0 = (2,0)$. Wir berechnen zunächst Gradient und Hesse-Matrix:

$$\nabla f(x) = \begin{pmatrix} \frac{1}{x_1} e^{x_2} \\ \log x_1\, e^{x_2} \end{pmatrix}, \quad H_f(x) = \begin{pmatrix} -\frac{1}{x_1^2} e^{x_2} & \frac{1}{x_1} e^{x_2} \\ \frac{1}{x_1} e^{x_2} & \log x_1\, e^{x_2} \end{pmatrix}.$$

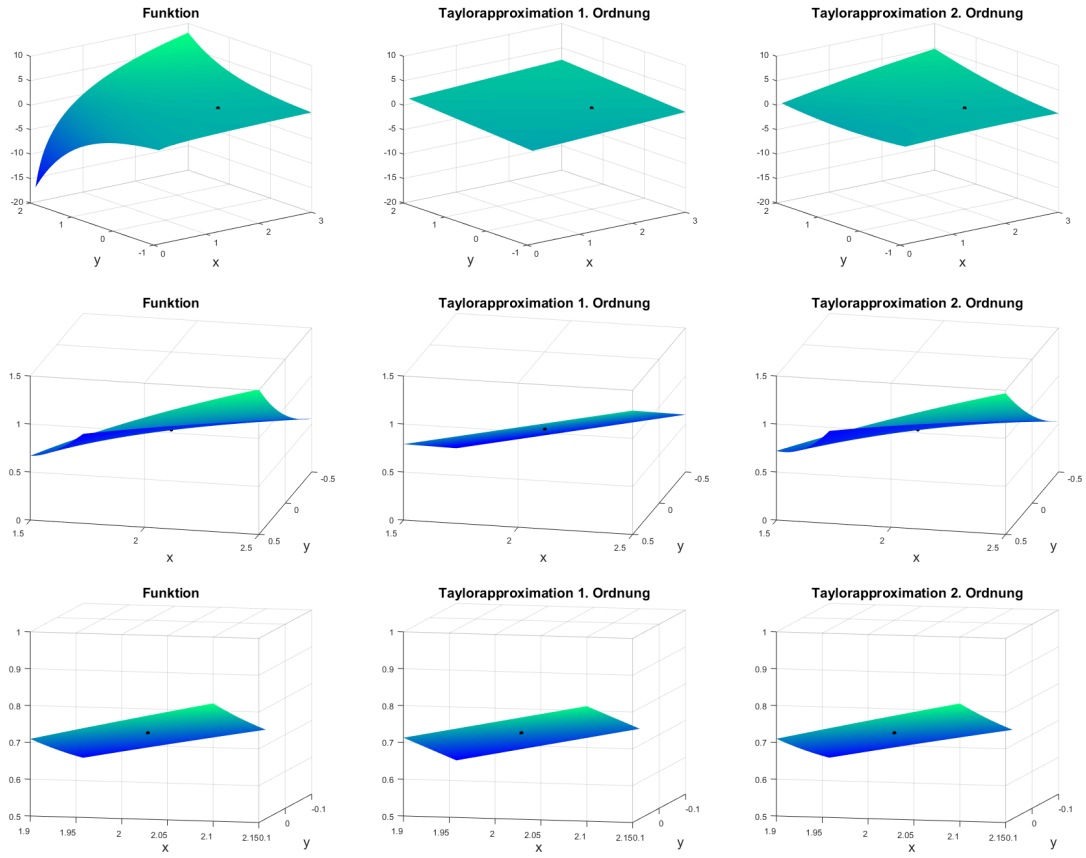
Auswerten an der Stelle $x = (2,0)$ liefert

$$\nabla f(2,0) = \begin{pmatrix} \frac{1}{2} \\ \log 2 \end{pmatrix}, \quad H_f(2,0) = \begin{pmatrix} -\frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & \log 2 \end{pmatrix}.$$

Folglich gilt

$$\begin{aligned} T_1(x_0;h) &= T_1(2,0;h) = \log 2 + \frac{1}{2}h_1 + \log 2 \cdot h_2, \\ T_2(x_0;h) &= T_2(2,0;h) = \log 2 + \frac{1}{2}h_1 + \log 2 \cdot h_2 - \frac{1}{8}h_1^2 + \frac{1}{2}h_1h_2 + \frac{\log 2}{2}h_2^2. \end{aligned}$$

Wie sehen die Graphen der entsprechenden Taylorapproximationen (also $x \mapsto T_1(2,0;x_1-2,x_2-0)$ *und* $x \mapsto T_2(2,0;x_1-2,x_2-0)$*) im Vergleich zum Graph von* f *aus?*



Tayorapproximationen der Funktion $f(x, y) = \log x \cdot e^y$

Oben: Auf einer grossen Längenskala schaffen die Taylorapproximationen nur unzulänglich, die Funktion zu approximieren, da sie versuchen, das globale Verhalten aus dem Verhalten nahe dem Entwicklungspunkt (2,0) zu "extrapolieren". Z.B. ist am linken Rand die Taylorapproximation 2. Ordnung nach oben gekrümmt, die Funktion aber nach unten.

Mitte: Auf einer mittleren Längenskala ist zumindest die Taylorapproximation 2. Ordnung bereits eine gute Näherung. Sie erfasst nicht nur die Position, die Steigung und die Krümmung der Fläche, sondern – beeindruckenderweise – auch die "Verwindung".

Unten: Auf kleinen Längenskalen sehen glatte Funktionen nach Vergrösserung linear aus! Selbst die Taylorapproximation 1. Ordnung gibt die Funktion korrekt wieder. Bei genauem Hinschauen sieht man, dass die Vorder- und Hinterkante des Funktionsgraphen nicht ganz parallel sind, sondern nach hinten "auseinanderlaufen"; dies kann die Approximation 1. Ordnung nicht erfassen, denn affin lineare Abbildungen bilden parallele Geraden auf parallele Geraden ab. Die Taylorapproximation 2. Ordnung hingegen erfasst auch dieses Detail korrekt, und ist optisch nicht von der Funktion zu unterscheiden.

Um die Terme höherer Ordnung in der Taylorentwicklung zu vereinfachen, benutzen wir Multi-indices.

Def. 2.7 (Multi-index) Ein $\alpha = (\alpha_1, ..., \alpha_n) \in (\mathbb{N} \cup \{0\})^n$ heisst Multi-index. Für einen Multi-index definieren wir:

$$|\alpha| = \sum_{i=1}^{n} \alpha_i \text{ (Ordnung)}, \quad \alpha! = \prod_{i=1}^{n} \alpha_i! \text{ (Fakultät)}.$$

Für ein $|\alpha|$-mal stetig differenzierbares $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ ist

$$\partial^\alpha f(x) = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} ... \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}} f(x) \text{ (α-te partielle Ableitung von } f).$$

(Hierbei benutzen wir die Konvention $\frac{\partial^0 f}{\partial x_i^0}(x) = f(x)$.) Für $h \in \mathbb{R}^n$ ist

$$h^\alpha = h_1^{\alpha_1} \cdot ... \cdot h_n^{\alpha_n}.$$

Beispiel $n = 3$, $\alpha = (2, 0, 1)$:
$|\alpha| = 3$, $\alpha! = 2!1! = 2$, $\partial^\alpha f(x) = \frac{\partial^2}{\partial x_1^2} \frac{\partial}{\partial x_3} f(x)$, $h^\alpha = h_1^2 h_3$.

Wir betrachten zunächst den Beitrag 3. Ordnung, also die Dreifachsumme

$$T_3(x_0; h) - T_2(x_0; h) = \frac{1}{3!} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} \frac{\partial^3 f}{\partial x_{i_1} \partial x_{i_2} \partial x_{i_3}}(x_0) h_{i_1} h_{i_2} h_{i_3}. \qquad (*)$$

Für $n = 2$, also Funktionen $f = f(x_1, x_2)$, gibt es 8 Terme, entsprechend den möglichen Kombinationen von Werten der drei Indizes i_1, i_2, i_3:

$$\begin{aligned} (i_1, i_2, i_3) = \quad & (1,1,1) \\ & (1,1,2) \\ & (1,2,1) \\ & (1,2,2) \\ & (2,1,1) \\ & (2,1,2) \\ & (2,2,1) \\ & (2,2,2) \end{aligned}$$

Der erste Term ist gleich $\frac{\partial^3 f}{\partial x_1^3}(x_0) h_1^3$. Zweiter, dritter und fünfter Term sind wegen Schwarz identisch und liefern zusammengenommen den Beitrag $3 \frac{\partial^3 f}{\partial x_1^2 \partial x_2}(x_0) h_1^2 h_2$. Entsprechend liefern vierter, sechster und siebenter Term zusammen den Beitrag

$3\,\frac{\partial^3 f}{\partial x_1 \partial x_2^2}(x_0)h_1h_2^2$, und der letzte Term ist gleich $\frac{\partial^3 f}{\partial x_2^3}(x_0)h_2^3$. In Multiindex-Schreibweise gilt also: (*) ist gleich

$$\frac{1}{3!}\Big(\partial^{(3,0)}f(x_0)h^{(3,0)} + 3\partial^{(2,1)}f(x_0)h^{(2,1)} + 3\partial^{(1,2)}f(x_0)h^{(1,2)} + \partial^{(0,3)}f(x_0)h^{(3,0)}\Big).$$

Die hier vorkommenden Multi-indices sind genau die $\alpha \in (\mathbb{N}\cup\{0\})^2$ mit $|\alpha| = 3$, und wegen $(3,0)! = 3!0! = 3!$, $(2,1)! = 2!1! = 2$, $(1,2)! = 2$, $(0,3)! = 3!$ lässt sich obige Summe in der eleganten Form

$$\sum_{\substack{\alpha\in(\mathbb{N}\cup\{0\})^2\\ |\alpha|=3}} \frac{1}{\alpha!}\partial^\alpha f(x_0)h^\alpha$$

schreiben.

Im allgemeinen Fall (n, k beliebig) sieht der entsprechende Term in der Taylorentwicklung genauso aus.

Lemma 2.4 (Multiindex-Darstellung der Taylorentwicklung) Für $n, k \in \mathbb{N}$ gilt unter den Voraussetzungen von Satz 2.6

$$\begin{aligned} T_k(x_0;h) - T_{k-1}(x_0) &= \sum_{|\alpha|=k} \frac{1}{\alpha!}\partial^\alpha f(x_0)h^\alpha \\ T_k(x_0;h) &= \sum_{|\alpha|\le k} \frac{1}{\alpha!}\partial^\alpha f(x_0)h^\alpha, \\ R_k(x_0;h) &= \sum_{|\alpha|=k+1} \frac{1}{\alpha!}\partial^\alpha f(\xi)h^\alpha. \end{aligned}$$

Zum Abschluss dieses Abschnitts geben wir eine nützliche Variante des Satzes von Taylor an, bei der die Differenzierbarkeitsvoraussetzungen minimal sind, wir dafür aber auf eine explizite Darstellung des Restgliedes mithilfe höherer Ableitungen verzichten. Die entsprechende Aussage im Eindimensionalen wurde bereits in Analysis 1 bewiesen (Korollar 9.2).

Korollar 2.2 (Qualitative Taylorentwicklung) Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$, $x_0 \in \Omega$.
a) f $(k+1)$-mal stetig diff'bar $\Longrightarrow$ $f(x_0+h) = T_k(x_0;h) + O(|h|^{k+1})$ für $h \to 0$.
b) f k-mal stetig diff'bar $\Longrightarrow$ $f(x_0+h) = T_k(x_0;h) + o(|h|^k)$ für $h \to 0$.
Hierbei bezeichnet T_k das Taylorpolynom k-ter Ordnung im Punkt x_0, d.h. $T_k(x_0;h) = \sum_{|\alpha|\le k} \frac{1}{\alpha!}\partial^\alpha f(x_0)h^\alpha$.

Beweis a): Wir stellen die Differenz $f(x_0+h) - T_k(x_0;h)$ gemäss Satz von Taylor (Satz 2.6) dar, wählen $\delta > 0$ sodass die Kugel $\overline{B_\delta(x_0)} \subset \Omega$, und setzen

$$C := \max_{\xi\in\overline{B_\delta(x_0)}} \frac{1}{(k+1)!} \sum_{|\alpha|=k+1} \Big|\frac{\partial^\alpha f}{\partial x^\alpha}(\xi)\Big|.$$

(Beachte: die Funktionen $\partial^\alpha f/\partial x^\alpha$, $|\alpha| = k+1$, sind nach Voraussetzung stetig und somit wird das obige Maximum nach Satz 1.5 angenommen.) Indem wir noch die elementare Abschätzung $|h^\alpha| \le |h|^{k+1}$ benutzen, folgt für alle $|h| \le \delta$

$$|f(x_0+h) - T_k(x_0;h)| = \Big|\frac{1}{(k+1)!}\sum_{|\alpha|=k+1}\frac{\partial^\alpha f}{\partial x^\alpha}(\xi)h^\alpha\Big| \le C|h|^{k+1}.$$

b): Wir benutzen den Satz von Taylor (Satz 2.6) für $k-1$ statt k. Dies liefert die folgende Darstellung (mit geeignetem $\xi \in [x_0, x_0+h]$):

$$\begin{aligned} f(x_0+h) &= \sum_{|\alpha|\le k-1}\frac{1}{\alpha!}\partial^\alpha f(x_0)h^\alpha + \sum_{|\alpha|=k}\frac{1}{\alpha!}\underbrace{\partial^\alpha f(\xi)}_{=\partial^\alpha f(x_0)+(\partial^\alpha f(\xi)-\partial^\alpha f(x_0))}h^\alpha \\ &= \underbrace{\sum_{|\alpha|\le k}\frac{1}{\alpha!}\partial^\alpha f(x_0)h^\alpha}_{=T_k(x_0;h)} + \underbrace{\sum_{|\alpha|=k}\frac{1}{\alpha!}\Big(\partial^\alpha f(\xi)-\partial^\alpha f(x_0)\Big)h^\alpha}_{=:R_k(x_0;h)}. \end{aligned}$$

Es bleibt zu zeigen, dass das so definierte Restglied von der Ordnung $o(|h|^k)$ ist. Dies folgt aber aus der Stetigkeit der k-ten partiellen Ableitungen sowie der elementaren Abschätzung $|h^\alpha| \le |h|^k$ für alle $|\alpha| = k$:

$$\frac{|R_k(x_0;h)|}{|h|^k} \le \sum_{|\alpha|=k}\frac{1}{\alpha!}\Big|\underbrace{\partial^\alpha f(\xi)-\partial^\alpha f(x_0)}_{\to 0}\Big| \to 0 \quad (|h|\to 0).$$

Literatur: Ein Standardlehrbuch für das Material in §§2.1–2.6 ist *Konrad Königsberger, Analysis II, Springer-Verlag.* Eine frei im Netz verfügbare, schlank und zugänglich geschriebene Quelle ist das Skript meines Kollegen Martin Brokate (*https://mediatum.ub.tum.de/doc/1701089/1701089.pdf*).

2.7 Anwendung: Maximieren/Minimieren

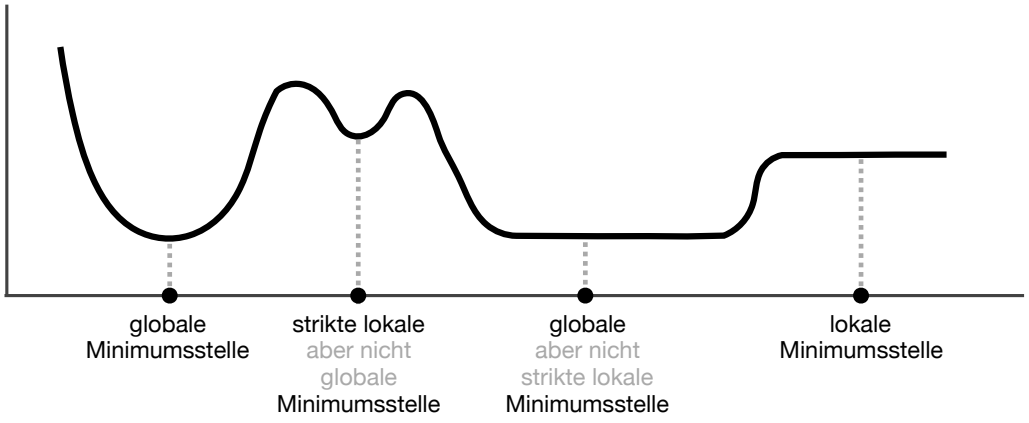
In Abschnitt 1.8 hatten wir bewiesen, dass stetige Funktionen mehrerer Veränderlicher unter allgemeinen und realistischen Voraussetzungen Maximums- und Minimumsstellen besitzen. Die Ableitungsbegriffe liefern Methoden, um diese zu bestimmen – bei einfachen, niedrig-dimensionalen Problemen per Hand und bei komplizierten oder hochdimensionalen Problemen numerisch per Computer.

Wir beginnen mit einer langen Vokabel-Liste.

Def. 2.8 Sei $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$. Ein Punkt $x_0 \in \Omega$ heisst

$$\begin{cases} \text{Maximumsstelle} \\ \text{Minimumsstelle} \\ \text{lokale Maximumsstelle} \\ \text{lokale Minimumsstelle} \\ \text{Extrempunkt oder Extremstelle} \\ \text{lokaler Extrempunkt oder lokale Extremstelle} \end{cases} \text{von } f\text{, wenn gilt:} \begin{cases} f(x_0) \geq f(x)\ \forall x \in \Omega \\ \leq \\ \exists \delta > 0 :\ f(x_0) \geq f(x)\ \forall x \in \Omega \text{ mit } |x - x_0| < \delta \\ \leq \\ x_0 \text{ Maximums- oder Minimumsstelle} \\ x_0 \text{ lokale Max.- oder lokale Min.stelle.} \end{cases}$$

Eine lokale Extremstelle heisst *strikt*, wenn zusätzlich $f(x) \neq f(x_0)$ für alle $x \in \Omega \backslash \{x_0\}$ mit $|x - x_0| < \delta$.



2.7.1 Optimalitätsbedingungen

Das folgende grundlegende Resultat geht (im Spezialfall von Funktionen einer Variablen) auf den Mathematik-Pionier Pierre de Fermat zurück.

Satz 2.7 *[Optimalitätsbedingungen erster Ordnung] Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$, $x_0 \in \Omega$ lokale Extremstelle von f, f partiell diff'bar im Punkt x_0. Dann gilt*

$$\nabla f(x_0) = 0.$$

Die Umkehrung gilt nicht. Z.B. gilt für $f(x, y) = x^2 - y^2$, $x_0 = (0, 0)$: $\nabla f(x_0) = 0$, aber x_0 ist keine lokale Extremstelle. Solche Punkte heissen *Sattelpunkte*.

Beweisidee: Betrachte f entlang der Koordinatenachsen durch x_0 und argumentiere wie beim Beweis des analogen Resultates in Analysis 1.

Details: Sei z.B. x_0 lokale Minimumsstelle. Sei e_i der i^{te} Einheitsvektor im $\mathbb{R}^n$. Da x_0 lokale Minimumsstelle, folgt insbesondere für alle $h \in \mathbb{R}$ mit hinreichend kleinem Absolutbetrag und $h \neq 0$:

$$0 \leq f(x_0 + he_i) - f(x_0). \qquad (**)$$

Multiplikation mit $1/h > 0$ und Grenzübergang $h \to 0$ liefert

$$0 \le \frac{f(x_0 + he_i) - f(x_0)}{h} \to \frac{\partial f}{\partial x_i}(x_0).$$

Multiplikation von (**) mit $1/h < 0$ (beachte: bei Multiplikation mit negativen Zahlen kehren sich Ungleichheitszeichen um) und Grenzübergang liefert andererseits

$$0 \ge \frac{f(x_0 + he_i) - f(x_0)}{h} \to \frac{\partial f}{\partial x_i}(x_0).$$

Beide Ungleichungen zusammengenommen ergeben $\frac{\partial f}{\partial x_i}(x_0) = 0$. Da i beliebig, folgt die Behauptung.

Der folgende Satz benutzt "Definitheits"-Eigenschaften quadratischer Matrizen; wir formulieren zuerst den Satz und definieren anschliessend diese Begriffe.

Satz 2.8 *[Optimalitätsbedingungen zweiter Ordnung] Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$ 2mal stetig diff'bar, $x_0 \in \Omega$. Dann gilt:*

a) Notwendige Bedingungen:

$$x_0 \text{ lokale } \begin{cases} \text{Minimumsstelle} \\ \text{Maximumsstelle} \end{cases} \implies \text{die Hesse-Matrix } H_f(x_0) \text{ ist } \begin{cases} \text{positiv semidefinit} \\ \text{negativ semidefinit} \end{cases}$$

b) Lokal hinreichende Bedingungen:

$$x_0 \text{ strikte lokale } \begin{cases} \text{Minimumsstelle} \\ \text{Maximumsstelle} \end{cases} \impliedby \nabla f(x_0) = 0 \text{ und } H_f(x_0) \begin{cases} \text{positiv definit} \\ \text{negativ definit} \end{cases}$$

Von besonderem Interesse ist Aussage b): während a), genau wie Satz 2.7, die Menge der möglichen Extrempunkte nur *einschränkt*, erlaubt uns b), lokale Extremalität allein mithilfe des Verhaltens der ersten und zweiten Ableitung von f an der Stelle x_0 *nachzuweisen*.

Def. 2.9 Eine reelle $n \times n$ Matrix A heisst
positiv definit, wenn $\langle h, Ah \rangle > 0$ für alle $h \in \mathbb{R}^n$, $h \neq 0$
negativ definit .....<...
positiv semidefinit, wenn $\langle h, Ah \rangle \ge 0$ für alle $h \in \mathbb{R}^n$
negativ semidefinit .....≤...
indefinit wenn A weder positiv semidefinit noch negativ semidefinit.

Geometrische Bedeutung der positiven Definitheit der Hesse-Matrix. Sei $\Omega \subseteq \mathbb{R}^n$ offen und konvex (d.h. für $x, y \in \Omega$ liegt die gerade Verbindungsstrecke $[x, y] = \{(1-t)x + ty : t \in [0,1]\}$ in Ω). Eine Funktion $f : \Omega \to \mathbb{R}$ heisst *konvex*, wenn $f((1-t)x + ty) \le (1-t)f(x) + tf(y)$ für alle x, $y \in \Omega$ und alle $t \in (0,1)$. Geometrisch bedeutet das: die gerade Verbindungsstrecke zwischen den Punkten $(x, f(x))$ und

$(y, f(y))$ liegt oberhalb des Graphen von f. Ist f 2mal stetig partiell differenzierbar, so gilt:

$$f \text{ konvex} \iff H_f(z) \geq 0 \text{ für alle } z \in \Omega.$$

Anschaulich ist das klar, denn Positivität der zweiten Ableitung auf der Verbindungsstrecke von x nach y heisst, dass die Steigung von f entlang der Strecke zunimmt und daher der Graph von f unterhalb der Verbindungsstrecke von $(x, f(x))$ nach $(y, f(y))$ liegen muss. Für einen strengen Beweis siehe Abschnitt 2.7.2.

Beweisidee, Satz 2.8 Differenz $f(x_0 + h) - f(x)$ bis zur zweiten Ordnung Taylorentwickeln.

Details: a): Sei x_0 lokale Minimumsstelle. Dann ist für hinreichend kleines $|h|$

$$0 \leq f(x_0 + h) - f(x_0) \underset{\text{Taylor}}{=} \underbrace{\langle \nabla f(x_0), h \rangle}_{=0 \text{ (Satz2.7)}} + \langle h, H_f(x_0) h \rangle + o(|h|^2).$$

Division durch $|h|^2$ und Grenzübergang $|h| \to 0$ liefert $\langle \frac{h}{|h|}, H_f(x_0) \frac{h}{|h|} \rangle \geq 0 \; \forall h \neq 0$, d.h. $H_f(x_0) \geq 0$.

b): Sei $\nabla f(x_0) = 0$, $H_f(x_0)$ pos. definit. Nach Satz vom Maximum und Minimum nimmt $g(e) := \langle e, H_f(x_0) e \rangle > 0$ sein Minimum auf der (kompakten) Sphäre $S^{n-1} = \{e \in \mathbb{R}^n : |e| = 1\}$ an, d.h. $g(e) \geq m > 0 \; \forall e \in S^{n-1}$. Folglich

$$f(x_0 + h) - f(x_0) \underset{\text{Taylor}}{=} \langle h, H_f(x_0) h \rangle + o(|h|^2) \geq m|h|^2 + o(|h|^2) > 0 \text{ für } |h| \text{ hinreichend klein.}$$

Das folgende Beispiel zeigt: die Voraussetzung der Definitheit von $H_f(x_0)$ in Satz 2.8 b) kann nicht zu semidefinit abgeschwächt werden. Ist $H_f(x_0)$ nur semidefinit, kann man keine Aussage darüber machen, ob x_0 lokale Extremstelle ist oder nicht.

Beispiel: $f(x, y) = x^2 + ay^4$, $x_0 = (0, 0)$, $a \in \mathbb{R}$. Die Hessematrix ist positiv semidefinit und unabhängig vom Vorfaktor a des y^4 Terms, aber für $a > 0$ ist x_0 strikte lokale Minimumsstelle, für $a = 0$ nicht-strikte lokale Minimumsstelle, und für $a < 0$ keine lokale Extremstelle. Im semidefiniten Fall wird das lokale Verhalten einer Funktion also durch Terme höherer Ordnung bestimmt, die die Hessematrix nicht "sieht".

Wie stellt man fest, ob eine Matrix positiv bzw. negativ definit ist?

Lemma 2.5 Sei A eine reelle symmetrische $n \times n$ Matrix.

a) A positiv (bzw. negativ) definit $\iff$ alle Eigenwerte von A sind > 0 (bzw. < 0)

b) Falls $n = 2$: A positiv (bzw. negativ) definit $\iff$ mindestens ein Diagonalelement ist > 0 (bzw. < 0) und $\det A > 0$.

Analoge Aussagen gelten, wenn man "definit" durch "semidefinit", starke durch schwache Ungleichungen und "mindestens ein Diagonalelement" durch "beide Diagonalargumente" ersetzt.

Beweis für Diagonalmatrizen: Sei

$$A = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}:$$

In diesem Fall ist $\langle h, Ah\rangle = \sum_{i=1}^{n} \lambda_i h_i^2$. Dies impliziert sofort a), und auch b) wegen $\det A = \lambda_1\lambda_2$. Der allgemeine Fall folgt z.B. aus einem Resultat der Linearen Algebra, nach dem jede symmetrische reelle $n \times n$ Matrix in einer geeigneten Orthonormalbasis Diagonalform hat. Einen analytischen Beweis dieses Resultates lernen wir in Abschnitt 4.4 kennen.

2.7.2 Drei Beispiele

Beispiel 1) Wir bestimmen und klassifizieren alle lokalen Extremstellen der Funktion $f(x,y) = x^2 - xy + y^4$, $f : \mathbb{R}^2 \to \mathbb{R}$.

Wir gehen nach folgendem Verfahren vor
a) Gradient bestimmen
b) Nullstellen des Gradienten bestimmen
c) an den Nullstellen des Gradienten die Hessematrix bestimmen
d) Hessematrix auf Definitheit untersuchen.

a) Wir berechnen: $\nabla f(x,y) = \begin{pmatrix} 2x - y \\ -x + 4y^3 \end{pmatrix}$.

b) Die erste Komponente der Gleichung $\nabla f(x,y) = 0$ lautet

$$2x - y = 0,$$

folglich $y = 2x$. Einsetzen in die zweite Komponente der Gleichung,

$$-x + 4y^3 = 0,$$

liefert die Gleichung

$$-x + 32x^3 = 0.$$

Dies ist eine kubische Gleichung für nur noch eine Unbekannte, x. Die linke Seite schreiben wir mithilfe von Ausklammern des Faktors x als $x(-1 + 32x^2)$, folglich verschwindet sie, wenn entweder $x = 0$ oder $-1 + 32x^2 = 0$. Letztere Gleichung hat die beiden Lösungen $x = \pm\frac{1}{\sqrt{32}} = \pm\frac{1}{2\sqrt{8}}$.

Insgesamt hat die Gleichung $\nabla f(x,y) = 0$ also die folgenden drei Lösungen (und nur diese):

$$(x,y) = (0,0), \quad (x,y) = (\tfrac{1}{2\sqrt{8}}, \tfrac{1}{\sqrt{8}}), \quad (x,y) = (-\tfrac{1}{2\sqrt{8}}, -\tfrac{1}{\sqrt{8}}).$$

c) Die Hessematrix ist $H_f(x,y) = \begin{pmatrix} 2 & -1 \\ -1 & 12y^2 \end{pmatrix}$ und folglich

$$H_f(0,0) = \begin{pmatrix} 2 & -1 \\ -1 & 0 \end{pmatrix}, \quad H_f(\pm\tfrac{1}{2\sqrt{8}}, \pm\tfrac{1}{\sqrt{8}}) = \begin{pmatrix} 2 & -1 \\ -1 & \frac{3}{2} \end{pmatrix}.$$

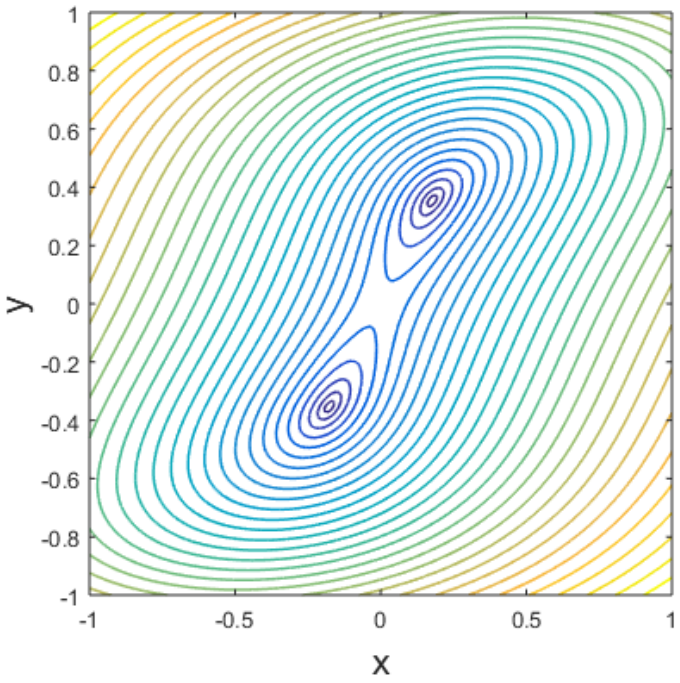
d) Die erste Matrix hat Determinante $2 \cdot 0 - (-1) \cdot (-1) = -1 < 0$ und ist folglich wegen Lemma 2.5 b) weder positiv noch negativ semidefinit. Wegen Satz 2.8 a) ist somit $(0,0)$ kein lokales Extremum, d.h. ein Sattelpunkt. Die zweite Matrix hat positive Diagonalelemente und Determinante $2 \cdot \frac{3}{2} - (-1) \cdot (-1) = 2 > 0$ und ist folglich wegen Lemma 2.5 b) positiv definit. Somit sind die Punkte $\pm(\frac{1}{2\sqrt{8}}, \frac{1}{\sqrt{8}})$ lokale Minimumsstellen.

Zum Abschluss bestimmen wir noch den Wert von f an den lokalen Minimumsstellen:

$$f(\pm\tfrac{1}{2\sqrt{8}}, \pm\tfrac{1}{\sqrt{8}}) = \frac{1}{32} - \frac{1}{16} + \frac{1}{64} = -\frac{1}{64}.$$

Insbesondere stimmt der Wert an beiden lokalen Minimumsstellen überein. Da ausserdem $f(x,y) \to \infty$ für $|(x,y)| \to \infty$, sind diese Punkte sogar Minimumsstellen.

Wie sieht die untersuchte Funktion aus? Siehe Schaubild.



Niveaulinien der Funktion $f(x,y) = x^2 - xy + y^4$

Beispiel 2) *Lineare Regression.* Reale Daten liegen (nach geeigneter Skalierung, z.B. Log-Plot, siehe Analysis 1 Abschnitt 10) bestenfalls näherungsweise auf einer Geraden. Lineare Regression ist eine grundlegende und vielverwendete Methode aus der Statistik, zu gegebenen Daten eine "bestapproximierende" Gerade zu finden. Sie ist auch eine Urversion für maschinelles Lernen: statt *Funktion* → *Wertetabelle* möchte man das umgekehrte Problem *Wertetabelle* → *Funktion* lösen, also aus gegebenen Daten eine funktionale Abhängigkeit extrahieren, um anschliessend nicht verfügbare Daten vorhersagen zu können.

Genauere Formulierung der Problemstellung: Finde zu gegebenen Daten

x	x_1	x_2	...	x_n
y	y_1	y_2	...	y_n

(mit x_i, $y_i \in \mathbb{R}$) eine Gerade $y(x) = ax + b$ sodass (ein geeignetes Fehlermaß für) die "durchschnittliche Abweichung" zwischen realen Daten y_i und theoretischen Werten $y(x_i)$ minimal wird.

Methode der linearen Regression (die auf C.F.Gauss zurückgeht): definiere den Fehler im i-ten Datenpunkt, $r_i = y_i - y(x_i) = y_i - (ax_i + b)$, und wähle als Fehlermaß den *mittleren quadratischen Fehler*

$$R(a,b) = \frac{1}{n}\Big(r_1^2 + ... + r_n^2\Big).$$

Dieser ist eine Funktion, die von den Parametern a (Steigung) und b (Achsenabschnitt) der Geraden abhängt. Wähle nun a und b so, dass R minimal wird.

(Auch andere Fehlermaße sind möglich und sinnvoll, z.B. $R(a,b) = \frac{1}{n}\sum_i |r_i|$, aber dann wird die Theorie komplizierter.)

Die Minimierung von R können wir mit der in dieser Vorlesung erarbeiteten Methode durchführen, d.h. wir führen wie in Beispiel 1) folgende Schritte aus:

a) Gradient bestimmen
b) Nullstellen des Gradienten bestimmen
c) an den Nullstellen des Gradienten die Hessematrix bestimmen
d) Hessematrix auf Definitheit untersuchen.

Unsere Fehlerfunktion lautet $R(a,b) = \frac{1}{n}\sum_{i=1}^n (y_i - ax_i - b)^2$. Der Gradient ist

$$\nabla R(a,b) = \begin{pmatrix} -\frac{2}{n}\sum_{i=1}^n (y_i - ax_i - b)x_i \\ -\frac{2}{n}\sum_{i=1}^n (y_i - ax_i - b) \end{pmatrix}.$$

Die zweite Komponente von ∇R ist Null genau dann wenn

$$0 = \frac{1}{n}\sum_{i=1}^n (y_i - ax_i - b) = \overline{y} - a\overline{x} - b,$$

wobei $\overline{z} := \frac{z_1+\ldots+z_n}{n}$ den Mittelwert von n Zahlen $z_1, \ldots, z_n \in \mathbb{R}$ bezeichnet. Somit ist also

$$b = \overline{y} - a\overline{x}. \tag{1}$$

Die erste Komponente von ∇R ist Null genau dann wenn

$$0 = \sum_{i=1}^{n} \Big(y_i - ax_i - b \Big) x_i.$$

Einsetzen der Gleichung für b und Verwenden der Identität $\sum_{i=1}^{n}(z_i - \overline{z}) = 0$ liefert

$$0 = \sum_{i=1}^{n} \Big((y_i - \overline{y}) - a(x_i - \overline{x}) \Big) x_i = \sum_{i=1}^{n} \Big((y_i - \overline{y}) - a(x_i - \overline{x}) \Big) (x_i - \overline{x}).$$

Im ersten Ausdruck ist nicht klar, ob der Vorfaktor von a ungleich Null ist, was aber zwecks Auflösbarkeit nach a erforderlich ist. Der zweite Ausdruck ist besser, denn wir sehen ihm an, dass - sofern $n \geq 2$ und die x_i alle verschieden - der Vorfaktor von a ungleich Null ist. Unter dieser natürlichen und minimalen Annahme (sonst macht lineare Regression offensichtlich keinen Sinn) folgt also

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{2}$$

Der Nenner ist übrigens gerade die *Varianz* der x_i; diese ist ein Maß dafür, wie breit die x_i gestreut sind. (Auch der Zähler hat eine interessante statistische Bedeutung, er ist die sogenannte *Covarianz* der x_i und y_i; diesen Begriff benötigen wir in unserer Diskussion der linearen Regression nicht, Sie werden ihn im 2. Studienjahr kennenlernen.)

Die Hessematrix ist unabhängig von a und b,

$$H_R(a,b) = \begin{pmatrix} \frac{2}{n}\sum_{i=1}^{n} x_i^2 & 2\overline{x} \\ 2\overline{x} & 2 \end{pmatrix} = 2 \begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}$$

(mit der Notation $\overline{x^2}$ = Mittelwert der $x_i^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$). Das rechte untere Diagonalelement ist offensichtlich positiv und ebenso die Determinante

$$\det H_R(a,b) = 4(\overline{x^2} - \overline{x}^2) = 4 \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}_{=\text{Varianz der } x_i} > 0.$$

Nach Lemma 2.5 ist somit der durch Gleichungen (1) und (2) gegebene Punkt (a,b) eine lokale Minimumsstelle. Da ausserdem $R(a,b) \to \infty$ ($|(a,b)| \to \infty$), ist (a,b) sogar die eindeutige Minimumsstelle. Die zugehörige Gerade $y(x) = ax + b$ heisst *Regressionsgerade*.

Wir fassen zusammen: *Für beliebige Daten* $y_1, ..., y_n \in \mathbb{R}$ *und* $x_1, .., x_n \in \mathbb{R}$ *mit* $n \geq 2$ *und paarweise verschiedenen* x_i *existiert eine eindeutige Gerade* $y : \mathbb{R} \to \mathbb{R}$, $y(x) = ax + b$, *sodass der mittlere quadratische Fehler* R *minimal wird. Steigung* a *und Achsenabschnitt* b *der Geraden sind durch die Formeln (1) und (2) gegeben.*

Illustrative Beispiele experimenteller Datensätze und deren Interpretation nebst Vorhersage fehlender Werte (z.B. aus der Biologie oder der Halbleiterindustrie, Stichwort "Moore'sches Gesetz") mithilfe von linearer Regression sind leicht im Netz zu finden.

Ausblick: Analog lassen sich bestapproximierende Funktionen aus anderen Funktionenklassen finden, z.B. den kubischen Polynomen $y(x) = a_3x^3 + a_2x^2 + a_1x + a_0$. Wichtig für obiges Prinzip sind nur die *Vektorraumeigenschaft* der Funktionenklasse und die *Quadratizität* der Fehlerfunktion; dann erhält man ein lineares Gleichungssystem für die Koeffizienten. Siehe Übungen.

Beispiel 3) *Konvexe Funktionen.* Hier ist unser Ziel nicht die explizite Optimierung einer konkreten Funktion; vielmehr benutzen wir die Optimalitätsbedingungen aus Satz 2.7 und Satz 2.8, um Konvexität einer Funktion mithilfe von Eigenschaften ihrer Ableitungen zu charakterisieren.

Satz 2.9 *Sei* $\Omega \subseteq \mathbb{R}^n$ *konvex,* $f : \Omega \to \mathbb{R}$ *differenzierbar. Dann sind äquivalent:*
(a) f *ist konvex*
(b) $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ *für alle* $x, y \in \Omega$.
Ist f *2mal differenzierbar, so ist ausserdem äquivalent:*
(c) $H_f(x) \geq 0$ *für alle* $x \in \Omega$.

Beweis (a)$\Longrightarrow$(b): Sei f konvex. Dann hat für $x, y \in \Omega$ die Funktion $\varphi : [0,1] \to \mathbb{R}$, $\varphi(t) = (1-t)f(x) + tf(y) - f((1-t)x + ty)$ eine Minimumsstelle bei $t = 0$ und somit

$$0 \leq \varphi'(0) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

(b)$\Longrightarrow$(c): Nach Voraussetzung hat die Funktion $u(y) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle$ eine Minimumsstelle bei $y = x$ und somit ist nach Satz 2.8

$$0 \leq H_u(x) = H_f(x).$$

(c)$\Longrightarrow$(a): Es gelte (c). Wir argumentieren indirekt. Ist f nicht konvex, existieren $x, y \in \Omega$ sodass die im ersten Teil des Beweises eingeführte Funktion φ nicht ≥ 0 auf ganz $[0,1]$ ist. Also hat φ eine Minimumsstelle $t_0 \in (0,1)$ mit $\varphi(t_0) < 0 = \varphi(0) = \varphi(1)$. Nach Satz 2.7 ist $\varphi'(t_0) = 0$ und wegen Hauptsatz und (c) folgt für beliebiges $t \in [t_0, 1]$

$$\varphi'(t) = \underbrace{\varphi'(t_0)}_{=0} + \int_{t_0}^{t} \varphi''(s)\, ds = \int_{t_0}^{t} -\big\langle y - x,\, H_f((1-s)x + sy)(y - x) \big\rangle\, ds \leq 0.$$

Dies impliziert aber $\varphi(t_0) \geq \varphi(1)$, Widerspruch.

2.7.3 Gradientenverfahren

Wie gehen wir vor, wenn wir komplizierte oder hochdimensionale Funktionen maximieren/minimieren wollen?

Mithilfe geeigneter numerischer Verfahren, die auf den in dieser Vorlesung entwickelten theoretischen Konzepten aufbauen. Genau wie das – in Analysis 1 besprochene – Newtonverfahren das grundlegende numerische Verfahren zum Gleichungslösen ist, ist das Gradientenverfahren das grundlegende numerische Verfahren zum Minimieren von Funktionen. Die Idee besteht darin, den (durch Satz 2.3 offengelegten) Sachverhalt auszunutzen, dass Minus der Gradient in Richtung des stärksten Abfalls der Funktion zeigt.

Gradientenverfahren (informell): Starte an irgendeinem Punkt. Gehe eine geeignete Distanz in Richtung von Minus Gradient. Iteriere.

Gradientenverfahren (mathematisch): Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine gegebene stetig differenzierbare Funktion, und sei $x^{(0)} \in \mathbb{R}^n$ ein gegebener Startpunkt. Definiere rekursiv:

$$x^{(i+1)} = x^{(i)} - \tau \nabla f(x^{(i)}) \quad (i = 0, 1, 2, ...),$$

wobei $\tau > 0$ eine geeignet zu wählende Zahl (Schrittweite) ist.

Hoffnung: unter geeigneten Voraussetzungen konvergiert die Folge $(x^{(i)})$ gegen eine Minimumsstelle, oder zumindest eine lokale Minimumsstelle, von f. Ein Folgenglied mit hinreichend großem Index sollte eine gute Näherung liefern.

Die richtige Wahl der Schrittweite ist nicht offensichtlich. Anhand der Graphen eindimensionaler Beispiele kann man sich zumindest klarmachen, dass bei zu grosser Schrittweite die Funktionswerte nicht "bergab" gehen, und man andererseits bei zu kleiner Schrittweite zu sehr "auf der Stelle tritt".

Um das Verhalten des Verfahrens analytisch zu untersuchen, sind einige Vorüberlegungen notwendig.

Def. 2.10 Sei $M \subseteq \mathbb{R}^n$. Eine Funktion $f : M \to \mathbb{R}^m$ heisst *L-Lipschitzstetig* (wobei $L > 0$), wenn

$$|f(x) - f(y)| \leq L|x - y| \quad \text{für alle } x, y \in M.$$

Eine Funktion heisst *Lipschitzstetig*, wenn ein L existiert sodass sie L-Lipschitzstetig ist.

Es gelten die Implikationen

$$\text{stetig} \quad \begin{array}{c} \Longleftarrow \\ \not\Longrightarrow \end{array} \quad \text{Lipschitzstetig} \quad \overset{\text{z.B. falls } M \text{ kompaktes Intervall}}{\begin{array}{c} \Longleftarrow \\ \not\Longrightarrow \end{array}} \quad \text{stetig differenzierbar}$$

(wobei für die Wohldefiniertheit von "stetig diff'bar" notwendig ist, dass M Intervall (in 1D) oder offen (in nD)). Die linke Implikation folgt aus dem ε-δ-Kriterium durch

Wahl von $\delta = \varepsilon/L$; dies zeigt, dass Lipschitzstetigkeit eine quantitative Version der Stetigkeit ist, in der das δ linear im ε gewählt werden kann. Beweis der rechten Implikation in Dimension $n = 1$: gute Übung. Gegenbeispiele für die ungültigen Implikationen: $f(x) = \sqrt{x}$ auf $[0,1]$ ist stetig (und sogar gleichmässig stetig), aber nicht Lipschitzstetig; $f(x) = |x|$ auf $[-1,1]$ ist 1-Lipschitzstetig, aber am Punkt 0 nicht differenzierbar.

Def. 2.11 Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$ diff'bar. Ein Punkt $x_* \in \Omega$ heisst *kritischer Punkt* von f, wenn $\nabla f(x_*) = 0$.

Satz 2.10 (Verhalten des Gradientenverfahrens) *Sei $f : \mathbb{R}^n \to \mathbb{R}$ diff'bar, $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ L-Lipschitzstetig, $f(x) \to \infty$ ($|x| \to \infty$). Die Schrittweite $\tau \in (0, \infty)$ genüge der Bedingung*

$$\tau < \tfrac{1}{L}.$$

Dann gilt für die durch das Gradientenverfahren mit beliebigem Startwert $x^{(0)} \in \mathbb{R}^n$ definierte Folge $(x^{(j)})$:

1. *Die Folge der Funktionswerte, $(f(x^{(j)}))$, ist monoton fallend.*
2. *Die Menge der Häufungspunkte von $(x^{(j)})$ ist eine nichtleere Teilmenge der Menge der kritischen Punkte von f.*

Wichtigster Aspekt der Voraussetzungen: die Schrittweite muss kleiner als der Kehrwert der Lipschitzkonstante des Gradienten sein, $\tau < \frac{1}{L}$. Dies bedeutet: falls sich der Gradient der Funktion schnell ändert (grosses L), muss die Schrittweite klein gewählt werden (kleines τ).

Wichtigster Aspekt der Konklusion: es existiert mindestens ein Häufungspunkt, und dieser löst die gewünschte (für lokale oder globale Minimumsstellen notwendige, siehe Satz 2.7) Gleichung.

Beispiel, dass die Folge mehr als einen Häufungspunkt haben kann: siehe Übungen. Beispiel, dass Häufungspunkte nicht notwendigerweise lokale Minimumsstellen sein müssen: falls der Startwert $x^{(0)}$ ein beliebiger kritischer Punkt ist, gilt $x^{(j)} = x^{(0)}$ für alle j.

Darauf, dass wir mit den bisher entwickelten Methoden eine derartige interessante und nichttriviale Aussage beweisen können, können wir stolz sein; hierbei kommt sowohl der *Ableitungskalkül* als auch die *konzeptionelle Seite* der Analysis (also Begriffe wie abgeschlossen, kompakt, Subniveaumenge, stetig, ... und deren Zusammenspiel) zum Einsatz; siehe unten.

Den ersten Schritt des Beweises von Satz 2.10 formulieren wir wie folgt.

Lemma 2.6 Ist $f : \mathbb{R}^n \to \mathbb{R}$ diff'bar und ∇f L-Lipschitzstetig, so gilt $\forall x^+, x \in \mathbb{R}^n$

$$f(x^+) \leq f(x) + \langle \nabla f(x),\, x^+ - x \rangle + L|x^+ - x|^2.$$

Die ersten beiden Terme auf der rechten Seite sind nichts als die Taylorapproximation erster Ordnung an der Stelle x für $f(x^+)$, aber die rechte Seite hängt – im Gegensatz zum Restglied im Satz von Taylor – nicht von höheren Ableitungen ausserhalb des Entwicklungspunktes, sondern nur von der ersten Ableitung am Entwicklungspunkt ab.

Beweis Nach Satz von Taylor sowie anschliessendem Einschieben einer "additiven Null" gilt mit $\xi \in [x, x^+]$

$$\begin{aligned} f(x^+) &= f(x) + \langle \nabla f(\xi),\, x^+ - x\rangle \\ &= f(x) + \langle \nabla f(x),\, x^+ - x\rangle + \langle \big(\nabla f(\xi) - \nabla f(x)\big), x^+ - x\rangle. \end{aligned}$$

Die ersten beiden Terme entsprechen denjenigen im Lemma, und den letzten Term können wir nach oben abschätzen, indem wir nacheinander Cauchy-Schwarz, die Lipschitzstetigkeit von ∇f und die triviale Ungleichung $|\xi - x| \le |x^+ - x|$ benutzen:

$$\langle \big(\nabla f(\xi) - \nabla f(x)\big), x^+ - x\rangle \le |\nabla f(\xi) - \nabla f(x)|\,|x^+ - x| \le L|\xi - x|\,|x^+ - x| \le L|x^+ - x|^2.$$

Damit ist die behauptete Ungleichung bewiesen. Wir merken noch an, dass die Ungleichung gültig bleibt, wenn man L durch $L/2$ ersetzt; der Beweis ist dann aber deutlich schwieriger.

Beweis von Satz 2.10 Zuerst zeigen wir 1). Anwenden von Lemma 2.6 mit $x = x^{(j)}$, $x^+ = x^{(j+1)}$ liefert wegen $x^+ - x = -\tau\nabla f(x)$

$$f(x^+) \le f(x) + \langle \nabla f(x), -\tau\nabla f(x)\rangle + L|\tau\nabla f(x)|^2 = f(x) - \tau(1 - L\tau)|\nabla f(x)|^2. \quad (*)$$

Der Faktor $(1 - L\tau)$ ist genau dann positiv, wenn $\tau < \frac{1}{L}$. Dies erklärt die Voraussetzung an τ, und etabliert 1). Nun zu 2). Zunächst zeigen wir: die Menge der Häufungspunkte ist nichtleer. Betrachte dazu die Subniveaumenge $N := \{x \in \mathbb{R}^n : f(x) \le f(x^{(0)})\}$. Die Menge N ist wegen $f(x) \to \infty$ ($|x| \to \infty$) beschränkt, und wegen der Stetigkeit von f abgeschlossen, folglich (siehe Lemma 1.6) kompakt. Wegen 1) liegt die Folge in N; somit besitzt sie einen Häufungspunkt. Es bleibt zu zeigen: falls x_* Häufungspunkt, gilt $\nabla f(x_*) = 0$. Dies zeigen wir indirekt. Angenommen es gelte $\nabla f(x_*) \ne 0$. Dann gilt wegen (*) mit $\varepsilon = \tau(1 - L\tau)|\nabla f(x_*)|^2 > 0$

$$f(x_* - \tau\nabla f(x^*)) \le f(x_*) - \varepsilon.$$

Wegen Stetigkeit von f und ∇f existiert ein $\delta > 0$ sodass

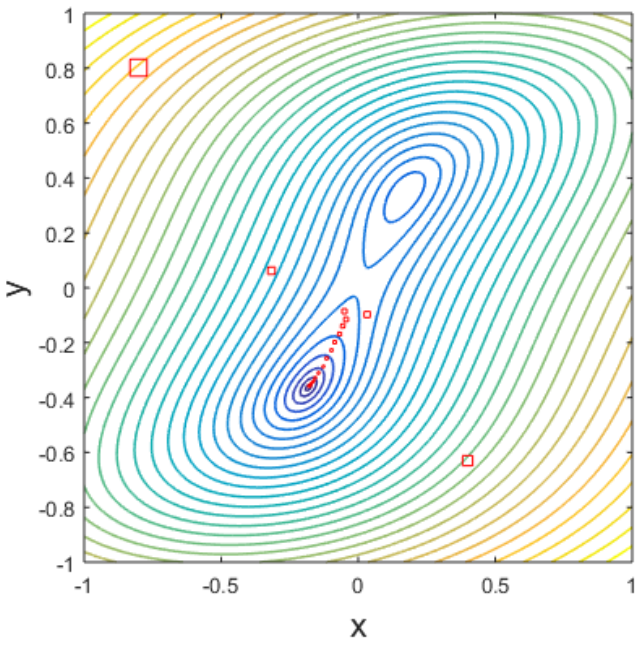
$$f(x - \tau\nabla f(x)) \le f(x_*) - \tfrac{\varepsilon}{2} \quad \text{für alle } x \in B_\delta(x_*).$$

Da x_* Häufungspunkt, existiert ein $x^{(m)} \in B_\delta(x_*)$. Somit folgt $f(x^{(m+1)}) \le f(x_*) - \frac{\varepsilon}{2}$ und – wegen 1) –

$$f(x^{(j)}) \le f(x_*) - \frac{\varepsilon}{2} \text{ für alle } j \ge m + 1.$$

Indem wir eine gegen x_* konvergente Teilfolge $(x^{(j_k)})$ wählen und k gegen Unendlich gehen lassen, erhalten wir einen Widerspruch, denn die linke Seite konvergiert gegen $f(x_*)$. Damit ist der Beweis beendet.

Wie sieht die durch das Gradientenverfahren definierte Folge in einem typischen numerischen Beispiel aus?



Erste 20 Folgenglieder des Gradientenverfahrens für die Funktion $f(x,y) = x^2 - xy + \varepsilon y + y^4$, $\varepsilon = 0.01$. Als Schrittweite haben wir $\tau = 0.5$ und als Startwert $(-0.8, 0.8)$ gewählt. Frühere Folgenglieder entsprechen grösseren Markern.

Im Fall $\varepsilon = 0$ besitzt die Funktion zwei gleich tiefe Minimimumsstellen $\pm(\frac{1}{2\sqrt{8}}, \frac{1}{\sqrt{8}})$ (siehe unsere Analyse in Abschnitt 2.7); durch Addition des Terms εy haben wir die Funktion in der oberen Halbebene etwas angehoben und in der unteren Halbebene etwas abgesenkt, sodass die Minimumsstelle in der oberen Halbebene nur noch lokal ist.

Die Folge springt vom Startwert (links oben) zunächst auf die gegenüberliegende Seite, dann ein paarmal hin und her, bis sie knapp unterhalb des Sattelpunkts zwischen lokaler und globaler Minimumsstelle landet; von dort geht's dann schnurstracks bergab Richtung globale Minimumsstelle. Das asymptotische Verhalten entspricht also in Satz 2.10 der Option ein Häufungspunkt, Häufungspunkt ist Minimumsstelle.

Wie bei jeder durch das Gradientenverfahren definierten Folge steht die Richtung vom vorherigen zum nächsten Folgenglied, also Minus Gradient am vorherigen Folgenglied, senkrecht zur dortigen Niveaulinie (siehe Satz 2.4).

2.8 Anwendung: partielle Differentialgleichungen

Ableitungen sind nicht nur ein zentrales Hilfsmittel in der Analyse gegebener Funktionen (z.B. der Bestimmung von Extremstellen), sondern auch in der Modellierung.

Bisher haben wir nur "gegebene" Funktionen betrachtet, und deren Ableitungen untersucht. In Anwendungen ist die Problemstellung oft genau andersherum: viele Gesetzmässigkeiten und Modelle in Natur- und Ingenieurwissenschaften lassen sich als Beziehungen zwischen einer *unbekannten* Funktion und ihren Ableitungen formulieren. Gesucht ist dann eine Funktion, die diesen Beziehungen genügt.

Eine solche Beziehung zwischen einer Funktion und ihren Ableitungen, in der partielle Ableitungen nach mindestens zwei verschiedenen Koordinaten auftreten, heisst *partielle Differentialgleichung*. (Falls nur partielle Ableitungen nach einer Koordinate auftreten – z.B. weil die Funktion nur von einer Variablen abhängt – spricht man von einer *gewöhnlichen Differentialgleichung*, siehe Analysis 1 Abschnitt 12.)

2.8.1 Beispiele partieller Differentialgleichungen

Wir geben einige grundlegende Beispiele partieller Differentialgleichungen an. Den Fragen,
– wie man solche Gleichungen aus Modellierungsannahmen herleitet
– wie man sie (in einfachen Situationen) löst
– was die Lösungen "machen"
gehen wir exemplarisch in Abschnitten 2.8.2 und 2.8.3 nach.

Beispiele 1) **Poissongleichung** Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}^n$ gegeben. Die partielle Differentialgleichung

$$-\Delta u = f$$

oder, in Langversion,

$$-\sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} u(x_1, ..., x_n) = f(x_1, ..., x_n) \text{ für alle } x = (x_1, ..., x_n) \in \Omega,$$

für $u : \Omega \to \mathbb{R}$ heisst *Poissongleichung*. Hierbei ist Δ der Laplaceoperator (siehe Abschnitt 2.5), und u eine unbekannte Funktion, die es zu bestimmen gilt. Diese Gleichung tritt für $\Omega = \mathbb{R}^3$ z.B. in der Elektrostatik und der Astrophysik auf (f Ladungsverteilung, u elektrostatisches Potential; bzw. f Massenverteilung, u Gravitationspotential). Der Spezialfall $f = 0$, d.h.

$$-\Delta u = 0,$$

heisst *Laplacegleichung*. Lösungen dieser Gleichung heissen *harmonische Funktionen*.

2) **Wellengleichung** Die partielle Differentialgleichung

$$\frac{\partial^2}{\partial t^2} u = \Delta u$$

für $u : \Omega \times [0,\infty) \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$, $u = u(x,t)$, heisst *Wellengleichung*. Hierbei modelliert $u(x,t)$ z.B. die transversale Auslenkung einer schwingenden Saite $\Omega = [0,L]$ oder Membran $\Omega \subset \mathbb{R}^2$ am Punkt $x \in \Omega$ zur Zeit t, oder die Amplitude zur Zeit t einer Schall- oder elektromagnetischen Welle an einem Punkt x des dreidimensionalen Raumes. Wir benutzen hier die übliche Konvention, dass sich der Laplace-Operator nur auf die Ortskoordinaten $x = (x_1, ..., x_n) \in \Omega$ bezieht, d.h. $\Delta u = \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2}$.

3) **Wärmeleitungsgleichung** Die partielle Differentialgleichung

$$\frac{\partial}{\partial t} u = \Delta u$$

für $u : \Omega \times [0,\infty) \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$, $u = u(x,t)$, heisst *Wärmeleitungsgleichung*. Wie bei der Wellengleichung bezieht sich der Laplace-Operator nur auf die Ortskoordinaten $x = (x_1, ..., x_n) \in \Omega$. Diese Gleichung tritt z.B. in der Physik ($\Omega = [0,L]$, $u(x,t)$=Temperatur eines Drahtes $[0,L]$ am Punkt x zur Zeit t) und in der Stochastik ($\Omega = \mathbb{R}^n$, $u(x,t)$=Wahrscheinlichkeitsdichte, dass sich ein sich zufällig bewegendes Teilchen zur Zeit t am Ort x aufhält) auf.

4) **Schrödingergleichung** Sei $V : \mathbb{R}^3 \to \mathbb{R}$ gegeben. Die partielle Differentialgleichung

$$-\tfrac{1}{2}\Delta u + Vu = \lambda u$$

für $u : \mathbb{R}^3 \to \mathbb{C}$ und $\lambda \in \mathbb{R}$ heisst (zeitunabhängige) *Schrödingergleichung* im Potential V. Prototypisch ist der Fall $V(x) = -1/|x|$ (Schrödingergleichung des Wasserstoffatoms). Diese Gleichung spielt eine grundlegende Rolle in der Quantenphysik und der Theoretischen Chemie. Physikalisch ist $|u(x)|^2$ die Wahrscheinlichkeitsdichte, dass sich das Elektron am Ort x aufhält (wobei aufgrund der Invarianz der Gleichung unter $u \mapsto \alpha u$ für $\alpha \in \mathbb{R}$ angenommen werden kann, dass $\int |u|^2 = 1$), $V(x)$ die potentielle Energie aufgrund der elektrostatischen Anziehung durch den (am Punkt 0 positionierten) Atomkern, und λ die Energie des Elektrons. Mathematisch ist λ ein Eigenwert der linearen Abbildung $u \mapsto -\frac{1}{2}\Delta u + Vu$; man kann zeigen, dass – wie beim Eigenwertproblem für Matrizen – λ, also die Energie, nur diskrete ("quantisierte") Werte annehmen kann.

Auf den ersten Blick scheint es vollkommen hoffnungslos, solche Gleichungen zu lösen, da die Beziehungen zwischen partiellen Ableitungen an *jedem* Punkt x bzw. *jedem* Punkt (x,t) gelten müssen. Trotzdem existiert oft unter sinnvollen Zusatzbedingungen (siehe exemplarisch Abschnitt 2.8.2) eine eindeutige Lösung. Eine systematische Einführung in das Gebiet "partielle Differentialgleichungen" bietet eine gleichnamige Vorlesung im 3. Studienjahr.

Beispiele

1) Das *Newton-Potential* $u : \mathbb{R}^n \backslash \{0\} \to \mathbb{R}$, $u(x) = \frac{1}{|x|^{n-2}}$ $(n \geq 3)$ löst die Laplacegleichung.

2) Die Funktion $u : (\mathbb{R}^3\backslash\{0\}) \times \mathbb{R} \to \mathbb{C}$, $u(x,t) = \frac{e^{ik(|x|-t)}}{|x|}$ $(k > 0)$ sowie ihr Realteil $\tilde{u}(x,t) = \frac{\cos(k(|x|-t))}{|x|}$ lösen die Wellengleichung.

3) Die zeitabhaengige Gauss-Funktion $u : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, $u(x) = \frac{1}{t^{n/2}} e^{-\frac{|x|^2}{4t}}$ löst die Wärmeleitungsgleichung.

4) Die Funktion $u(x) = e^{-|x|}/\sqrt{\pi}$ (*Grundzustand*) löst die Schrödingergleichung für das Wasserstoffatom mit $\lambda = -\frac{1}{2}$; die Funktion $u(x) = x_1 e^{-|x|/2}/\sqrt{32\pi}$ (*Beispiel eines angeregten Zustands*) löst sie mit $\lambda = -\frac{1}{8}$.

Es ist eine exzellente Übung, die jeweiligen Funktionsgraphen in Abhängigkeit von x_1, x_2 (bei 2) und 3) für verschiedene t) zu skizzieren und die Gültigkeit der jeweiligen partiellen Differentialgleichungen nachzuprüfen, mithilfe der folgenden

Regel: Für 2mal differenzierbare radialsymmetrische Funktionen, d.h. $v : \mathbb{R}^n\backslash\{0\} \to \mathbb{R}$ mit $v(x) = g(r)$, $r = |x|$, gilt

$$\Delta v(x) = g''(r) + \frac{n-1}{r} g'(r).$$

Herleitung der Regel: Wir berechnen zunächst gemäss Kettenregel

$$\frac{\partial r}{\partial x_j} = \frac{\partial}{\partial x_j}\sqrt{x_1^2 + \ldots + x_n^2} = \frac{1}{2\sqrt{x_1^2 + \ldots + x_n^2}} 2x_j = \frac{x_j}{r} \quad \text{und somit } \nabla r = \frac{x}{r}.$$

Nach Kettenregel folgt

$$\nabla v(x) = g'(r)\frac{x}{r}.$$

Die Produktregel für die Divergenz, d.h. $\operatorname{div}(\varphi u)(x) = \langle \nabla\varphi(x), u(x)\rangle + \varphi(x)\operatorname{div} u(x)$ für eine skalare Funktion φ und ein Vektorfeld u, liefert

$$\Delta v(x) = \operatorname{div}\nabla v(x) = \operatorname{div}\Big(g'(r)\frac{x}{r}\Big) = \langle \nabla g'(r), \frac{x}{r}\rangle + g'(r)\operatorname{div}\frac{x}{r} = \langle g''(r)\frac{x}{r}, \frac{x}{r}\rangle + g'(r)\Big[\langle \nabla\frac{1}{r}, x\rangle + \frac{1}{r}\operatorname{div} x\Big].$$

Wegen $\nabla\frac{1}{r} = -\frac{1}{r^2}\nabla r = -\frac{x}{r^3}$ und $\operatorname{div} x = \sum_{j=1}^n \frac{\partial}{\partial x_j} x_j = n$ sowie $\langle x, x\rangle = r^2$ folgt

$$\Delta v(x) = g''(r) + g'(r)\Big[-\frac{1}{r} + \frac{n}{r}\Big].$$

5) Wir bestimmen alle radialsymmetrischen harmonischen Funktionen. Sei $n \geq 2$, $v : \mathbb{R}^n\backslash\{0\} \to \mathbb{R}$ mit $v(x) = g(r)$, $r = |x|$. Sei $\Delta v = 0$. Nach obiger Regel bedeutet dies

$$0 = g''(r) + \frac{n-1}{r} g'(r) = \frac{1}{r^{n-1}}\frac{d}{dr}\Big(r^{n-1} g'(r)\Big).$$

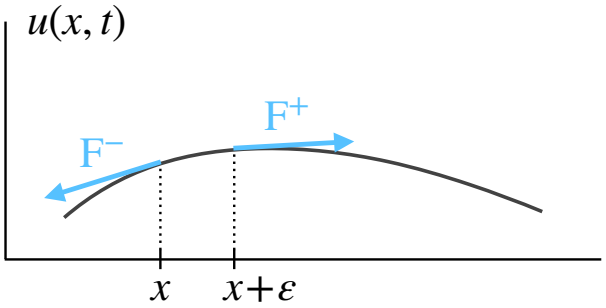
Somit $r^{n-1}g'(r) = a$ für eine Konstante $a \in \mathbb{R}$. Folglich $g'(r) = \frac{a}{r^{n-1}}$ und somit

$$g(r) = \begin{cases} a\log r + b & \text{für } n = 2 \\ \frac{a'}{r^{n-2}} + b & \text{für } n \geq 3 \end{cases}$$

(mit $a' = -\frac{a}{n-2}$). Insbesondere ist das Newtonpotential $\frac{1}{r}$ bis auf eine multiplikative und eine additive Konstante die einzige radialsymmetrische harmonische Funktion auf $\mathbb{R}^3\backslash\{0\}$.

2.8.2 Herleitung der Wellengleichung für die schwingende Saite

Wir betrachten eine elastische Saite der Länge L, hergestellt aus einem homogenen Material. Die transversale Auslenkung am Punkt x zur Zeit t beschreiben wir durch eine skalare Funktion $u = u(x,t)$, $u : [0,L] \times \mathbb{R} \to \mathbb{R}$; siehe Skizze. Die Position der gesamten Saite zur Zeit t entspricht dem Graph der Funktion $u(\cdot,t)$.



Die auf das Teilstück zwischen x und $x+\varepsilon$ durch den Rest der Saite ausgeübte Kraft wirkt tangential an beiden Enden des Teilstücks und hat (wegen Homogenität) den selben Betrag, d.h. die Netto-Kraft ist

$$F = F^+ + F^- \quad \text{wobei } |F^+| = |F^-|$$

(siehe Skizze). Der Tangentialvektor der Länge 1 an den Graphen von $u(\cdot,t)$ in Richtung von F^+ bzw. F^- ist

$$\begin{pmatrix} 1 \\ u'(x+\varepsilon) \end{pmatrix} \frac{1}{\sqrt{1+u'(x+\varepsilon)^2}} \quad \text{bzw.} \quad -\begin{pmatrix} 1 \\ u'(x) \end{pmatrix} \frac{1}{\sqrt{1+u'(x)^2}}.$$

Folglich gilt

$$F \;=\; |F^+| \Big[\begin{pmatrix} 1 \\ u'(x+\varepsilon) \end{pmatrix} \frac{1}{\sqrt{1+u'(x+\varepsilon)^2}} \;-\; \begin{pmatrix} 1 \\ u'(x) \end{pmatrix} \frac{1}{\sqrt{1+u'(x)^2}} \Big].$$

Wir sind an der Nettokraft F_{vert} in vertikaler Richtung interessiert. Unter der Annahme, dass die Saite fast horizontal bleibt (d.h. u' klein) ergibt sich wegen Taylor (beachte $1/\sqrt{1+u'^2} = 1 + O(u'^2)$)

$$F_{vert} \;\approx\; |F^+| \Big[u'(x+\varepsilon) - u'(x) \Big] \;\approx\; \varepsilon |F^+|\, u''(x).$$

(Analog sehen wir: Nettokraft in horizontaler Richtung ≈ 0, d.h. die Gesamtkraft ist nahezu vertikal.) Das Newton'sche Gesetz (Masse $\times$ Beschleunigung = Kraft) liefert wegen Masse = $\varepsilon \times$ Massendichte

$$\varepsilon \times \text{Massendichte} \times u_{tt}(x,t) \;\approx\; \varepsilon |F^+|\, u''(x).$$

Im Limes $\varepsilon \to 0$ erhalten wir mit $c^2 := |F^+|/\text{Massendichte}$ die Wellengleichung

$$u_{tt}(x,t) = c^2 u_{xx}(x,t).$$

2.8.3 Lösen der Wellengleichung via Separation der Variablen

Ein vollständiges Modell der schwingenden Saite ist gegeben durch die Wellengleichung zusammen mit Randbedingungen und Anfangsbedingungen:

$$u_{tt}(x,t) = c^2 u_{xx}(x,t) \tag{W}$$
$$u(0,t) = 0,\ u(L,t) = 0 \text{ für alle } t \tag{R}$$
$$u(x,0) = u_0(x),\ u_t(x,0) = 0 \text{ für alle } x \tag{A}$$

(Randbedingung: die Saite ist an den Enden fest eingespannt; Anfangsbedingung: die Anfangsauslenkung ist vorgegeben, und die Anfangsgeschwindigkeit ist Null.) Eine typische Anfangsbedingung ist

$$u_0(x) = \sin(\tfrac{\pi x}{L}) + \gamma \sin(\tfrac{3\pi x}{L}).$$

Schritt 1 Separation der Variablen: Zunächst ignorieren wir die erste Anfangsbedingung, und suchen Lösungen von (W), (R), und der zweiten Anfangsbedingung der Form $u(x,t) = X(x)T(t)$. Einsetzen in (W) liefert

$$XT'' = c^2 X''T$$

und somit, indem wir (wie in Analysis 1 bei der Behandlung gewöhnlicher Differentialgleichungen erster Ordnung) alle x-abhängigen Terme auf die linke Seite und alle t-abhängigen Terme auf die rechte Seite bringen,

$$\frac{X''(x)}{X(x)} = \frac{1}{c^2}\frac{T''(t)}{T(t)}.$$

Somit linke Seite = rechte Seite = k für eine Konstante $k \in \mathbb{R}$. Unser System (W), (R), (A2) reduziert sich also auf

$$\begin{cases} X'' = kX \\ X(0) = X(L) = 0, \end{cases} \qquad \begin{cases} T'' = c^2 kT \\ T'(0) = 0. \end{cases}$$

Schritt 2 Die Gleichungen für X und T sind gewöhnliche Differentialgleichungen. (Wir haben also eine partielle Differentialgleichung auf zwei gewöhnliche Differentialgleichungen reduziert!) Wie man sie löst, ist uns bereits aus Analysis 1 bekannt (siehe Abschnitt 12.3 "Schwingungsgleichungen"). Die allgemeine Lösung von $X'' = kX$, $X(0) = 0$ (mit $k < 0$) ist

$$X(x) = A\sin(\sqrt{|k|}x),$$

und wegen $X(L) = 0$ muss $\sqrt{|k|}L = n\pi$ für ein $n \in \mathbb{N}$ gelten, d.h.

$$k = -\Big(\frac{n\pi}{L}\Big)^2 \ (n = 1, 2, 3, ...). \tag{*}$$

Die allgemeine Lösung von $T'' = c^2 k T$ ist

$$\begin{aligned} T(t) &= B_1 \cos(\sqrt{c^2|k|}t) + B_2 \sin(\sqrt{c^2|k|}t) \\ &\underset{(*)}{=} B_1 \cos(c\tfrac{n\pi}{L}t) + B_2 \sin(c\tfrac{n\pi}{L}t), \end{aligned}$$

und wegen $T'(0) = 0$ muss $B_2 = 0$ sein. Insgesamt erhalten wir also

$$u(x,t) = \alpha_n \sin(\tfrac{n\pi x}{L}) \cos(c\tfrac{n\pi t}{L}) \text{ für ein } n \in \mathbb{N} \text{ und ein } \alpha_n \in \mathbb{R}.$$

Schritt 3 Superposition: Wegen Linearität des Systems (W), (R), (A2) sind Summen der in Schritt 2 gefunden Lösungen wiederum Lösungen (und unter geeigneten Abkling-Annahmen an die Koeffizienten α_n gilt dies sogar für unendliche Summen),

$$u(x,t) = \sum_{n=1}^{\infty} \alpha_n \sin(\tfrac{n\pi x}{L}) \cos(c\tfrac{n\pi t}{L}).$$

Schritt 4 Die Koeffizienten α_n können wir nun aus der bisher ignorierten Anfangsbedingung (A1) bestimmen. Für unsere Beispiel-Anfangsauslenkung muss gelten:

$$u(x,0) = \sum_{n=1}^{\infty} \alpha_n \sin(\tfrac{n\pi x}{L}) \overset{!}{=} \sin(\tfrac{\pi x}{L}) + \gamma \sin(\tfrac{3\pi x}{L}).$$

Folglich

$$\alpha_n = \begin{cases} 1 & n = 1 \\ \gamma & n = 3 \\ 0 & \text{sonst}, \end{cases}$$

und

$$u(x,t) = \sin(\tfrac{\pi x}{L}) \cos(c\tfrac{\pi t}{L}) + \gamma \sin(\tfrac{3\pi x}{L}) \cos(c\tfrac{3\pi t}{L}).$$

Man kann zeigen, dass dies die eindeutige Lösung ist. Wie sie aussieht, wie sie sich anhört, und wie die Wahl von γ den Sound verändert, wird in der Vorlesung demonstriert.

Ausblick: In Analysis 3 wird gezeigt (Theorie der "Fourier-Reihen"), dass jede hinreichend glatte Funktion u_0 auf dem Intervall $[0, L]$ mit $u_0(0) = u_0(L) = 0$ durch eine Reihe $\sum_{n=1}^{\infty} \alpha_n \sin(\tfrac{n\pi x}{L})$ dargestellt werden kann. Unsere Methode zur Lösung der Wellengleichung funktioniert somit für beliebige Anfangsauslenkungen u_0.

2.9 Anwendung: Maschinelles Lernen

Die bisher in dieser Vorlesung erarbeitete Mathematik spielt nicht nur in Natur- und Ingenieurwissenschaften, sondern auch im maschinellen Lernen eine wichtige Rolle. In diesem Abschnitt besprechen wir – aus mathematischer Sicht – eine grundlegende Methode, gradientenbasiertes Lernen neuronaler Netze, anhand einer typischen Aufgabe, Klassifikation von Inputs fester Grösse (z.B. Bilder) in verschiedene Klassen (z.B. Hund, Katze,...). Wir setzen keine Kenntnisse über maschinelles Lernen voraus.

Mathematisch gesehen ist ein neuronales Netz eine multivariate vektorwertige Funktion.

Beispiel (Neuronale Netze zur Klassifikation von 64×64 Schwarz-Weiss-Bildern): ein solches Netz bildet den Vektor $x \in \mathbb{R}^{4096}$, dessen i-te Komponente dem Grauwert des i-ten Pixels entspricht, auf den Vektor $\hat{y} \in \mathbb{R}^C$ ab, der eine Wahrscheinlichkeit für jede der C möglichen Klassen des Bildes angibt.

Neuronale Netze sind aber nicht beliebige Funktionen, sondern eine neuere Funktionenklasse spezieller Bauart, die in der Informatik enwickelt wurde und auf *vielfacher Verkettung* einfacher Funktionen beruht (für eine präzise Definition siehe Abschnitt 2.9.1). Dies steht im Gegensatz zu traditionellen Funktionenklassen wie etwa Polynomen oder trigonometrischen Polynomen, die auf *Linearkombination* einfacher Funktionen beruhen.[9]

Neuronale Netze hängen von a priori unbekannten Parametern ab, die nicht von Hand bestimmt werden. Stattdessen wählt man eine geeinete Anzahl von Testbeispielen, bei denen man die korrekte Klassifizierung kennt, und bestimmt die Parameter durch Minimierung der (geeignet zu messenden) Abweichung des für die Testbeispiele vorhergesagten Outputs vom korrekten Output. Diese Vorgehensweise trägt zwar den modernen Namen *Training*, ist aber in Wirklichkeit das Lösen eines mathematischen Optimierungsproblems, und man verwendet hierzu traditionelle mathematische Verfahren (typischerweise Varianten des Gradientenverfahrens aus Abschnitt 2.7.3). Das Bestimmen des Gradienten gelingt durch wiederholte Anwendung der *mehrdimensionalen Kettenregel* aus Abschnitt 2.2.

2.9.1 Neuronale Netze vom Feedforward-Typ

1. Netzwerk: Eine wichtige Klasse neuronaler Netze sind die *Feedforward-Netze*, die aus mehreren hintereinandergeschalteten Schichten von Neuronen bestehen.

[9] Der Erfolg der traditionellen Funktionenklassen bei Fragestellungen aus Natur- und Ingenieurwissenschaften hat damit zu tun, dass man dort häufig Funktionenklassen approximieren möchte, die selbst eine lineare (oder nur mild nichtlineare) Struktur haben, wie etwa die Lösungsmengen linearer (oder nur mild nichtlinearer) partieller Differentialgleichungen (siehe Abschnitt 2.8) in Abhängigkeit variierender Rand- und Anfangsbedingungen. Der Erfolg neuronaler Netze für maschinelles Lernen hat damit zu tun, dass die Abbildungen, die man dort approximieren möchte, hochgradig nichtlinear sind.

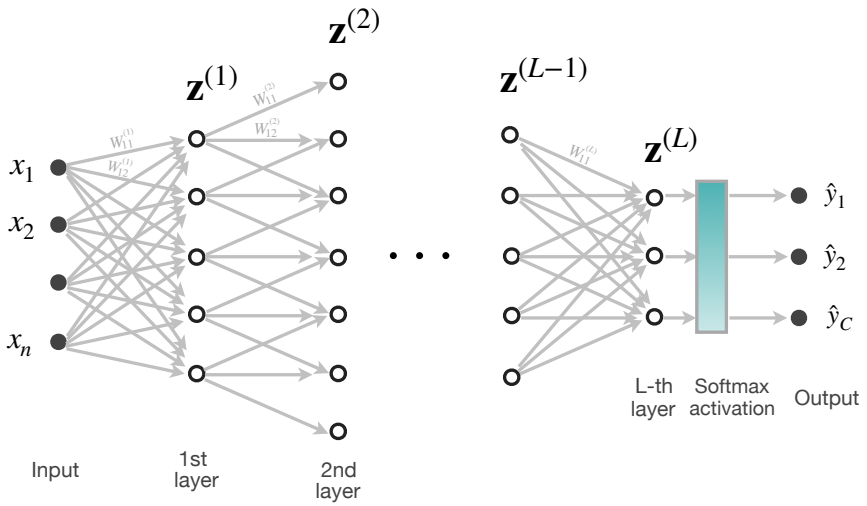
Def. 2.12 Ein Feedforward-Netz mit L Schichten ist eine Funktion $\hat{y} : \mathbb{R}^n \to \mathbb{R}^m$ der folgenden Form (siehe Skizze):

$$\hat{y} = f^{(L)} \circ A^{(L)} \circ f^{(L-1)} \circ A^{(L-1)} \circ \ldots \circ f^{(2)} \circ A^{(2)} \circ f^{(1)} \circ A^{(1)}$$

mit affin linearen Abbildungen $A^{(k)}$, d.h.

$$A^{(k)}(z^{(k)}) = (W^{(k)})^T z^{(k)} + b^{(k)},$$

und nichtlinearen Abbildungen $f^{(k)}$.



Die $f^{(k)}$ agieren (im Normalfall) nur komponentenweise; dies entspricht der Modellierungsannahme, dass Neuronen benachbarter Schichten nur linear (d.h. nur durch eine lineare Abbildung) gekoppelt sind. Ein Standardbeispiel ist die rectifying linear unit (ReLu)

$$\left(\left(f^{(k)}(z^{(k)})\right)_i = \mathrm{ReLU}(z_i^{(k)}), \quad \mathrm{ReLu}(t) = \max\{t, 0\}.\right.$$

Die $W^{(k)}$ und $b^{(k)}$ heissen *weights* und *biases*, und die nichtlinearen Abbildungen heissen *Aktivierungsfunktionen*. Den Wert der k-ten Schicht nach Anwenden von $A^{(k)}$ bezeichnen wir mit $z^{(k)}$, d.h.

$$z^{(1)} = (W^{(1)})^T x + b^{(1)}, \;\; z^{(2)} = (W^{(2)})^T f^{(1)}(z^{(1)}) + b^{(2)}, \;\; \text{usw.}$$

Siehe Skizze. A priori kann die Anzahl der Knoten der k-ten Schicht (alias Komponenten des Vektors $z^{(k)}$) beliebig sein, d.h. $z^{(k)} \in \mathbb{R}^{d_k}$. (Die Definitions- und Wertebereiche der Abbildungen sind also $A^{(k)} : \mathbb{R}^{d_{k-1}} \to \mathbb{R}^{d_k}$ und $f^{(k)} : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$ mit $d_0 = n$, $d_L = m$.) Die i-te Komponente $(z^{(k)})_i$ gibt den Wert des i-ten Knotens der k-ten Schicht an.

Der Koeffizient $W_{ij}^{(1)}$ bezeichnet den Faktor, mit dem der Wert des i-ten Knotens der 0. Schicht multipliziert wird, bevor er an den j-ten Knoten der 1. Schicht weitergeleitet wird (siehe Skizze); daher tritt bei Verwenden von Matrixschreibweise die Transponierte $(W^{(1)})^T$ auf, denn $(W^T x)_j = \sum_i W_{ji}^T x_i = \sum_i W_{ij} x_i$.

2. Softmax: Für Klassifizierungsaufgaben benutzt man als finale Aktivierungsfunktion meist die Softmax-Funktion, d.h. $f^{(L)} = \sigma$ mit

$$\sigma(z_1, ..., z_C) = \begin{pmatrix} \frac{e^{z_1}}{\sum_j e^{z_j}} \\ \vdots \\ \frac{e^{z_C}}{\sum_j e^{z_j}} \end{pmatrix}.$$

(Insbesondere muss die Anzahl d_L der Knoten der letzten Schicht der Anzahl C der Klassen entsprechen.) Der Sinn dieser Funktion liegt darin, dass sie (i) einen beliebigen Input-Vektor $z \in \mathbb{R}^C$ in einen Wahrscheinlichkeitsvektor umwandelt, d.h. einen Vektor mit nichtnegativen Komponenten, die sich zu 1 summieren, (ii) grösseren Input-Komponenten grössere Wahrscheinlichkeiten zuteilt.

3. Parameter des Netzwerks: Die zunächst unbestimmten Parameter des Netzwerkes sind die affin linearen Abbildungen, also die weights und biases. (Demgegenüber sind die nichtlinearen Aktivierungsfunktionen von Anfang an festgelegt.) Die Parameter der k-ten Schicht fassen wir als Vektor auf, indem wir die Spalten der Matrix $W^{(k)}$ untereinanderschreiben und unten noch den Vektor $b^{(k)}$ dazuschreiben, d.h.

$$\theta^{(k)} = \begin{pmatrix} W_{\cdot 1}^{(k)} \\ W_{\cdot 2}^{(k)} \\ \vdots \\ W_{\cdot d_k}^{(k)} \\ b^{(k)} \end{pmatrix} \in \mathbb{R}^{d_{k-1} \cdot d_k + d_k} =: \Omega_k.$$

Insgesamt ist unser neuronales Netz mit *unbestimmten Parametern* also eine Abbildung

$$\begin{aligned} \hat{y}: \quad & \mathbb{R}^n \times \Omega_1 \times ... \times \Omega_L \quad \to \mathbb{R}^C \\ \hat{y}: \quad & (x, \theta^{(1)}, ..., \theta^{(k)}) \quad \mapsto \left(\sigma \circ A_{\theta^{(L)}}^{(L)} \circ f^{(L-1)} \circ A_{\theta^{(L)}}^{(L-1)} \circ ... \circ A_{\theta^{(2)}}^{(2)} \circ f^{(1)} \circ A_{\theta^{(1)}}^{(1)}\right)(x). \end{aligned}$$

4. Verlustfunktion: Um das Netzwerk zu trainieren, d.h. die Parameter zu bestimmen, benutzt man Datenpunkte $x^{(\nu)}$, $\nu = 1, ..., N$, für die die zugehörigen Klassen $c_\nu \in \{1, ..., C\}$ bekannt sind. Bei korrekter Klassifizierung sollte das Netzwerk für diese Datenpunkte folgende Output-Vektoren $y^{(\nu)}$ liefern:

$$y_i^{(\nu)} = \begin{cases} 1 & i = c_\nu \\ 0 & \text{sonst.} \end{cases}$$

Den Fehler misst man mithilfe einer *Verlustfunktion* (englisch: loss function) $\ell : \mathbb{R}^C \times \mathbb{R}^C \to \mathbb{R}$. Deren Wert $\ell(y, \hat{y})$ sollte umso niedriger sein, je näher $\hat{y}$ an y liegt. Eine oft verwendete Verlustfunktion für Klassifizierung ist die *Kreuzentropie* (englisch: cross entropy)

$$\mathrm{CE}(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log \hat{y}_i.$$

Diese Funktion möchte, dass $\hat{y}_i$ gross ist, wenn $y_i = 1$; sie belohnt also das Modell, wenn es dem korrekten Label eine hohe Wahrscheinlichkeit zuteilt. Genauer:

Lemma 2.7 *Sei $y_i = 1$ für $i = c$ und 0 sonst. Dann ist die Kreuzentropie $\mathrm{CE}(y, p)$, aufgefasst als Funktion auf der Menge der Wahrscheinlichkeitsvektoren $\{p \in \mathbb{R}^C : p_i \geq 0, \sum_i p_i = 1\}$ (mit der Konvention $a \log b = 0$ wenn $a = b = 0$ und $-\infty$ wenn $a > 0$ und $b = 0$), genau dann minimal, wenn $p = y$.*

Beweis Für das gegebene y ist $\mathrm{CE}(y, p) = -1 \cdot \log p_c$. Da $p_c \in [0, 1]$ und $-\log$ streng monoton fallend, ist die eindeutige Minimumsstelle $p_c = 1$. Da p Wahrscheinlichkeitsvektor, folgt $p_i = 0$ für $i \neq c$, d.h. $p = y$.

5. Training: Nun kann man die Parameter des Netzwerks durch (näherungweises numerisches) Lösen des folgenden Optimierungsproblems bestimmen:

$$\min_{\theta^{(1)},\ldots,\theta^{(L)}} \sum_{\nu=1}^{N} \mathrm{CE}\Big(y^{(\nu)},\, \hat{y}(x^{(\nu)}, \theta^{(1)}, \ldots, \theta^{(L)})\Big). \tag{T}$$

2.9.2 Herleitung des Backpropagation-Algorithmus via Kettenregel

Um das Trainingsproblem (T) näherungsweise numerisch zu lösen, benutzt man das Gradientenverfahren; man muss also die rechte Seite von (T) nach den Netzwerkparametern ableiten. Für Studierende, die sich mit mehrdimensionalem Ableiten auskennen, ist – aufgrund der Verkettungsstruktur des Netzes – offensichlich, dass man die mehrdimensionale Kettenregel anwenden kann, und dass somit die Jacobimatrix ein Matrizenprodukt sein muss. Dies wurde in der machine learning community von verschiedenen Autoren unabhängig voneinander bemerkt. Numerisch hat das die nützliche Konsequenz, dass sich der Gradient durch eine Hintereinanderausführung von Matrix-Vektor-Multiplikationen effizient berechen lässt (selbst dann, wenn das Netz viele Schichten und Millionen von Parametern enthält); diese Methode heisst Backpropagation-Algorithmus.

Um den Algorithmus herzuleiten, bestimmen wir für einen gegebenen Input x mit gegebenem korrektem Output y den Gradienten bezüglich der Parameter, z.B. denjenigen der ersten Schicht,

$$\nabla_{\theta^{(1)}} \mathrm{CE}\Big(y, \hat{y}(x, \theta^{(1)})\Big).$$

Dies bereitet zwar keine "prinzipiellen" Schwierigkeiten, ist aber natürlich aufgrund der grossen Anzahl verschiedener Variablen und Funktionen und der vielen Verkettungen eine Menge Arbeit.

Schritt 1: Kreuzentropie und und Softmax. Wir berechnen

$$\begin{aligned}\frac{\partial}{\partial z_k}\mathrm{CE}(y,\sigma(z)) &= -\sum_{i=1}^{c} y_i \frac{\partial}{\partial z_k}\Big[\log e^{z_i} - \log\sum_j e^{z_j}\Big] \\ &= -y_k \underbrace{\frac{1}{e^{z_k}} e^{z_k}}_{=1} + \underbrace{\sum_i y_i}_{=1} \underbrace{\frac{1}{\sum_j e^{z_j}} e^{z_k}}_{=\sigma(z)_k} \\ &= -y_k + \sigma(z)_k.\end{aligned}$$

Somit gilt (wir bezeichnen im folgenden die Jacobimatrix, d.h. die Matrix der partiellen Ableitungen, von $\mathrm{CE}(y,\sigma(z))$ bzgl. z mit $J_z\mathrm{CE}(y,\sigma(z))$ und erinnern an die Beziehung $\nabla_z = (J_z)^T$)

$$J_z\mathrm{CE}(y,\sigma(z)) = -(y-\sigma(z))^T, \qquad \nabla_z\mathrm{CE}(y,\sigma(z)) = -(y-\sigma(z)).$$

Schritt 2: Erste Schicht. Wir schreiben $\theta^{(1)} =: \theta$, $W^{(1)} =: W$, $b^{(1)} =: b$, $d_1 =: d$, und bezeichnen die Jacobimatrix (alias Matrix der partiellen Ableitungen) der Abbildung $x \mapsto z^{(1)}(x,\theta)$ bezüglich θ mit $J_\theta z^{(1)}(x,\theta)$. Komponentenweise ist die Abbildung gegeben durch

$$x \mapsto z^{(1)} = W^T x + b = \begin{pmatrix} \langle W_{\cdot 1}, x\rangle + b_1 \\ \vdots \\ \langle W_{\cdot d}, x\rangle + b_d \end{pmatrix}$$

wobei $W_{\cdot i}$ die i-te Spalte der Matrix W ist, d.h.

$$W = \begin{pmatrix} | & & | \\ W_{\cdot 1} & \cdots & W_{\cdot d} \\ | & & | \end{pmatrix} \in R^{n\times d}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_d \end{pmatrix} \in \mathbb{R}^d.$$

Der Vektor θ ist

$$\theta = \begin{pmatrix} W_{\cdot 1} \\ \vdots \\ W_{\cdot d} \\ b \end{pmatrix} \in \mathbb{R}^{d\cdot n+d}.$$

Folglich ist (mit $\mathbf{0}$ =Nullvektor im $\mathbb{R}^n$)

$$J_\theta z^{(1)}(x,\theta) = \left(\begin{array}{cccc|cccc} x^T & \mathbf{0}^T & \cdots & \mathbf{0}^T & 1 & 0 & & 0 \\ \mathbf{0}^T & x^T & \cdots & \mathbf{0}^T & 0 & 1 & & 0 \\ & & \ddots & & & & \ddots & \\ \mathbf{0}^T & \mathbf{0}^T & \cdots & x^T & 0 & 0 & \cdots & 1 \end{array}\right).$$

Insbesondere hängt die Jacobimatrix gar nicht von θ ab, sondern nur von x, und ist nur dünn besetzt (d.h. enthält viele Nullen).

Die eigenartige Struktur dieser Matrix ist verursacht durch das – für numerische Implementierung nötige aber konzeptuell etwas künstliche – Umformen der Matrix $W \in \mathbb{R}^{n\times d}$ und des Vektors $b \in \mathbb{R}^n$ in einen langen Parametervektor. Alternativ können wir direkt (ohne dieses Umformen) die totale Ableitung (§2.2) bestimmen, wie bei unserer Diskussion der Determinante: Die Abbildung $z^{(1)} : (W,b) \mapsto W^T x + b$ ist offensichtlich linear; folglich ist die totale Ableitung $Dz^{(1)}(W,b) : \mathbb{R}^{n\times d}\times\mathbb{R}^n \to \mathbb{R}^d$ die lineare Abbildung $Dz^{(1)}(W,b)(H,h) = H^T x + h$. Es ist eine gute Übung nachzurechnen, dass die darstellende Matrix dieser linearen Abbildung gerade die oben hergeleitete Jacobimatrix ist, wenn wir H und h in einen langen Parametervektor umformen.

Schritt 3: Parametergradient für Netz mit einer Schicht. Sei $L = 1$. Da x und y fest, schreiben wir im folgenden $\hat{y}(\theta)$, $z^{(1)}(\theta)$, $\mathrm{CE}(\hat{y}(\theta))$ statt $\hat{y}(x,\theta)$, $z^{(1)}(x,\theta)$, $\mathrm{CE}(y,\hat{y}(x,\theta))$. Es gilt

$$\hat{y} = \sigma \circ z^{(1)}$$

und somit nach Kettenregel

$$J_\theta \mathrm{CE}(\hat{y}(\theta)) = J_{z^{(1)}}\mathrm{CE}(\sigma(z^{(1)}(\theta)))\ \ J_\theta z^{(1)}(\theta).$$

D.h. die Jacobimatrix ist ein Matrizenprodukt. Durch Transponieren erhalten wir den Gradienten, wobei das Matrizenprodukt in umgekehrter Reihenfolge auftritt:

$$\nabla_\theta \mathrm{CE}(\hat{y}(\theta)) = \Big(J_\theta z^{(1)}(\theta)\Big)^T\ \nabla_{z^{(1)}}\mathrm{CE}(\sigma(z^{(1)}(\theta))).$$

Einsetzen der Ergebnisse aus Schritt 1 und 2 liefert schliesslich folgende explizite Formel für den Gradienten:

$$\nabla_\theta \mathrm{CE}(\hat{y}(\theta)) = \left(\begin{array}{cccc} x & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & x & \cdots & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & \cdots & x \\ \hline 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 1 \end{array}\right) \Big[-(y-\sigma(z))\Big]. \tag{G}$$

Hierbei ist (da die erste=letzte Schicht aus C Knoten besteht)
– die linke Seite ein Vektor im $\mathbb{R}^{nC+C}$
– die Matrix eine $(nC+C)\times C$ Matrix
– der Vektor auf der rechten Seite ein Vektor im $\mathbb{R}^C$.

Schritt 4: Parametergradient für Netz mit vielen Schichten. Analog ergibt sich für ein Netz mit 2 Schichten (unter Benutzung von $J_z(Bz) = B$ für eine beliebige Matrix B)

$$J_{z^{(1)}} z^{(2)}(z^{(1)}) = W^{(2)T} Jf^{(1)}(z^{(1)})$$

und somit nach Kettenregel

$$J_\theta \mathrm{CE}(\hat{y}(\theta)) = J_{z^{(2)}} \mathrm{CE}(\sigma(z^{(2)}))\, W^{(2)T}\, Jf^{(1)}(z^{(1)})\, J_\theta z^{(1)}(\theta).$$

Analog erhalten wir für L Schichten

$$J_\theta \mathrm{CE}(\hat{y}(\theta)) = J_{z^{(L)}} \mathrm{CE}(\sigma(z^{(L)}))\, W^{(L)T} Jf^{(L-1)}(z^{(L-1)}) \cdots W^{(2)T}\, Jf^{(1)}(z^{(1)})\, J_\theta z^{(1)}(\theta),$$

d.h. ein $(L{+}1)$-faches Matrizenprodukt. Transponieren und Einsetzen der Ergebnisse aus Schritt 1 und 2 liefert das Endresultat

$$\nabla_\theta \mathrm{CE}(\hat{y}(\theta)) = \left(\begin{array}{cccc} x & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & x & \cdots & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & \cdots & x \\ \hline 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 1 \end{array}\right) (Jf^{(1)}(z^{(1)}))^T W^{(2)} \cdots (Jf^{(L-1)}(z^{(L-1)}))^T W^{(L)} \Big[-(y - \sigma(z^{(L)}))\Big].$$

Schritt 5: Backpropagation-Algorithmus. Eine rekursive Implementierung der hergeleiteten Formel, die auch für grosse Netze noch effizient ist, geht wie folgt:

grad = $-(y - \sigma(z^{(L)}))$; (gradient of loss w.r.to $z^{(L)}$)
for k from $L-1$ backwards to 1
    grad = $\left(Jf^{(k)}(z^{(k)})\right)^T W^{k+1}$grad; (gradient of loss w.r.to $z^{(k)}$)
end
grad = $\left(J_\theta z^{(1)}(\theta)\right)^T$grad. (gradient of loss w.r.to $\theta^{(1)}$)

Man arbeitet sich also sukzessive von hinten nach vorne durch die Schichten des Netzes (bzw. das hergeleitete Produkt) durch, um den Einfluss der Parameter der ersten Schicht auf den Output zu bestimmen; daher der Name des Algorithmus. Wir merken noch an, dass es wichtig ist, Aktivierungsfunktionen $f^{(k)}$ zu verwenden, deren Ableitung *exakt* bekannt ist; numerisches Differenzieren in der **for** Schleife würde zu sich aufschaukelnden Instabilitäten führen.

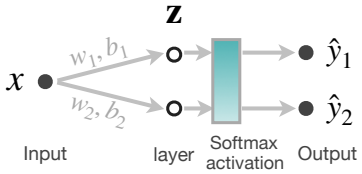
Literatur: §§2.9.1–2.9.2 basieren auf zwei Originalarbeiten (*LeCun, Bottou, Bengio, and Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11), 1998, 2278-2324; Rumelhart, Hinton, and Williams, Learning representations by back-propagating errors, Nature 323 (6088), 1986, 533-536*); meine Ausführungen sind eine "Übersetzung" und aktualisierte detaillierte Ausarbeitung einiger Haupteinsichten für Mathematik-Studierende.

2.9.3 Training via Gradientenverfahren

Wir trainieren mithilfe unserer allgemeinen Erkenntnisse aus dem vorherigen Abschnitt ein einfaches Netzwerk.

Netzwerk. Wir betrachten ein "kleinstmögliches" Modellnetzwerk, mit einer Schicht ($L = 1$), einem eindimensionalem Input-Vektor ($n = 1$), und zwei Klassen ($C = 2$). Das Netzwerk ist also die folgende vektorwertige Funktion $\hat{y} : \mathbb{R} \to \mathbb{R}^2$:

$$\hat{y}(x) = \begin{pmatrix} \sigma_1(z) \\ \sigma_2(z) \end{pmatrix}, \quad z = \begin{pmatrix} w_1 x + b_1 \\ w_2 x + b_2 \end{pmatrix}$$



mit Parametervektor

$$\theta = \begin{pmatrix} w_1 \\ w_2 \\ b_1 \\ b_2 \end{pmatrix} \in \mathbb{R}^4$$

(wobei σ die softmax-Funktion aus Abschnitt 2.9.1 bezeichnet).

Klassifizierungsproblem. Das Netzwerk soll folgendes Problem lösen: Bestimme für Datenpunkte auf der reellen Achse, die entweder zufällige Stichproben einer 1D Gauss-Verteilung mit Mittelwert 2 und Standardabweichung 1 sind (Klasse 1) oder zufällige Stichproben einer 1D Gauss-Verteilung mit Mittelwert 5 und Standardabweichung 1 (Klasse 2), die korrekte Klasse.

Parametergradient. Der Parametergradient der Verlustfunktion für einen gegebenen Input $x \in \mathbb{R}$ mit gegebenem korrektem Output $y \in \mathbb{R}^2$ ist, gemäss Formel (G) aus dem vorigen Abschnitt,

$$\nabla_\theta \mathrm{CE}(y, \hat{y}(x, \theta)) = -\begin{pmatrix} x & 0 \\ 0 & x \\ 1 & 0 \\ 0 & 1 \end{pmatrix} (y - \sigma(x)).$$

Trainingsdaten. Als Trainingsdaten nehmen wir 50 zufällige Stichproben der ersten Gauss-Verteilung und 50 zufällige Stichproben der zweiten Gauss-Verteilung. Wir haben also insgesamt 100 Trainingspaare $(x^{(\nu)}, y^{(\nu)})$ zur Verfügung.

Training. Um für unser Netzwerk das Trainingsproblem (T) aus Abschnitt 2.9.1 näherungsweise numerisch zu lösen, führen wir N Trainingsschritte mit der folgenden – im maschinellen Lernen üblichen – Variante des Gradientenverfahrens durch: in jedem Schritt verschieben wir den Parametervektor ein bisschen in Richtung Minus Parametergradient der Verlustfunktion bezüglich einem einzelnen Trainingspaar.

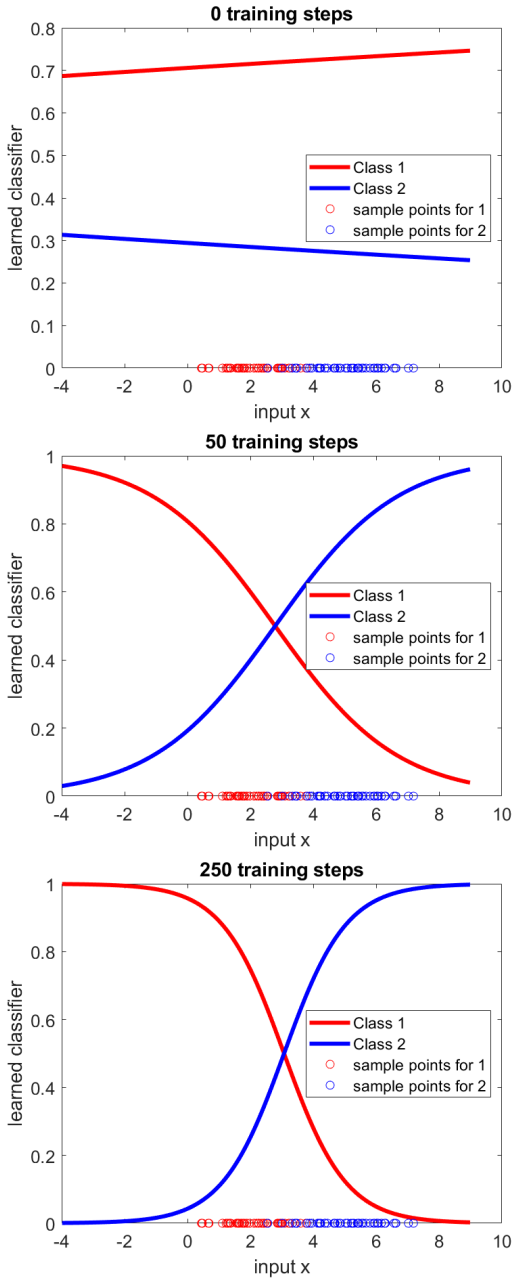
Genauer: wir initialisieren θ zufällig, ziehen in jedem Schritt ein zufälliges Paar $(x^{(k)}, y^{(k)})$ aus dem Trainingsdatensatz, und setzen

$$\theta^{k+1} \;=\; \theta^k \;-\; 0.05\ \nabla_\theta \mathrm{CE}(y^{(k)}, \hat{y}(x^{(k)}, \theta)).$$

Diese Variante des Gradientenverfahrens heisst *stochastic gradient descent*, da der Gradientenabstiegsschritt jeweils bezüglich eines einzelnen zufällig gewählten Trainingsdatenpunktes ausgeführt wird. Die empirische Rechtfertigng ist, dass man auf diese Weise nicht so leicht in lokalen Minima steckenbleibt. (In der Praxis werden für grosse Netze Verfeinerungen dieser Methode benutzt, die Grundidee bleibt aber dieselbe.)

Interessierte Studierende mit Programmierkenntnissen sind eingeladen, das Training unseres Modellnetzwerks schnell selbst in einer Programmiersprache ihrer Wahl zu implementieren; das in der Vorlesung erarbeitete kurze MATLAB-Programm liefert folgendes Ergebnis.

Ergebnis. Das Netzwerk – d.h. die beiden Funktionen $x \mapsto \hat{y}_1(x)$ (rote Kurve) und $x \mapsto \hat{y}_2(x)$ (blaue Kurve), die die Wahrschenlichkeit angeben, dass x zu Klasse 1 bzw. 2 gehört – sieht nach 0, 50, und 250 Trainingsschritten wie folgt aus. (Die Trainingsdaten sind ebenfalls geplottet.)



Das Netzwerk identifiziert also korrekt die Regionen, in denen die beiden Klassen typischerweise liegen, inklusive der Überlappregion, in der es richtigerweise beiden Klassen eine Wahrscheinlichkeit von ca. 0.5 zuweist.

3 Normierte Räume; Banach'scher Fixpunktsatz

Bisher haben wir so getan, als wären Punkte im $\mathbb{R}^n$ etwas vollkomen Verschiedenes von Funktionen $f : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$.

Diese Sichtweise wird fragwürdig, wenn man die Funktion "diskretisiert". Diskretisieren bedeutet: wir ersetzen den Definitionsbereich, sagen wir z.B. $[a, b]$, durch n Gitterpunkte $x_1, ..., x_n$, und approximieren die Funktion durch den Vektor $(f(x_1), ..., f(x_n))$ ihrer Werte an den Gitterpunkten. Dieser Vektor ist ein Element des Vektorraums $\mathbb{R}^n$. Anders gesagt: die Diskretisierung ist ein Punkt im $\mathbb{R}^n$. Je feiner die Diskretisierung, desto grösser wird die Dimension n des Vektorraums.

Wir gehen nun einen Schritt weiter. Eine seit dem Beginn des 20. Jahrhunderts übliche, fruchtbare alternative Sichtweise reellwertiger Funktionen ist:

Wir stellen uns Funktionen als Punkte in einem geeigneten (unendlich-dimensionalen) Vektorraum von Funktionen vor.

(Die Vektorraumeigenschaft ergibt sich durch punktweise Addition von Funktionen bzw. punktweise Multiplikation mit Skalaren $\lambda \in \mathbb{R}$.) Diese Sichtweise heisst

funktionalanalytischer Standpunkt.[10]

Der Vorteil ist, dass wir nun Ideen über Folgen von Punkten im $\mathbb{R}^n$ auf Folgen von Funktionen f_n (kurz: Funktionenfolgen) übertragen können, sofern uns auf dem Vektorraum eine Norm – d.h. ein Analogon der euklidischen Norm auf dem $\mathbb{R}^n$, mit dem wir den Abstand zwischen zwei Funktionen messen können – zur Verfügung steht. Uns interessiert insbesondere, ob, wann und wogegen Funktionenfolgen konvergieren und welche Eigenschaften sich auf den Grenzwert vererben. Funktionenfolgen treten in Anwendungen häufig auf.

- Aus Analysis 1 kennen wir bereits die Folge (T_n) der Taylorpolynome einer gegebenen Funktion.
- In Kapitel 5 untersuchen wir Systeme gewöhnlicher Differentialgleichungen, für die man keine explizite Lösung angeben kann. Stattdessen konstruieren wir zunächst nur Funktionen f_n, die die Dgl. nicht exakt erfüllen, aber näherungsweise, und immer genauer für $n \to \infty$. Wir zeigen dann, dass die Folge konvergiert und der Grenzwert die Dgl. löst. Hierbei werden sich die in diesem Kapitel erarbeiteten Ergebnisse als sehr hilfreich erweisen.

Wir beginnen damit, alternative Normen auf dem $\mathbb{R}^n$ zu besprechen.

[10] Funktioalanalysis ist ein Teilgebiet der Analysis, das sich systematisch mit dem Studium unendlichdimensionaler Vektorräume und Abbildungen zwischen diesen beschäftigt.

3.1 Andere Normen auf dem $\mathbb{R}^n$

Die euklidische Norm

$$|x| = \left(x_1^2 + ... + x_n^2\right)^{\frac{1}{2}} = \left(1 \cdot x_1^2 + ... + 1 \cdot x_n^2\right)^{\frac{1}{2}}$$

ist nicht die einzige Norm auf $\mathbb{R}^n$. Natürliche Verallgemeinerungen ergeben sich, indem man Exponenten $\neq 2$ oder Gewichtsfaktoren $\neq 1$ benutzt. Zunächst zu Ersterem. Der Ausdruck

$$|x|_p = \left(|x_1|^p + ... + |x_n|^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \quad (1 \leq p < \infty)$$

heisst p-Norm oder ℓ^p-Norm, und der Ausdruck

$$|x|_\infty = \max\Big\{|x_1|, ..., |x_n|\Big\} = \max\Big\{|x_i| \,:\, i = 1, ..., n\Big\}$$

heisst Maximumsnorm oder ℓ^∞-Norm. Offensichtlich reduziert sich die p-Norm für $p = 2$ auf die euklidische Norm, d.h. $|x|_2 = |x|$. Für grosse p nähert sich die p-Norm der Maximumsnorm: für alle $x \in \mathbb{R}^n$ gilt

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

denn sei x_{i_0} eine Komponente mit maximalem Absolutbetrag, d.h. $|x_{i_0}| \geq |x_i|$ für alle i, so folgt

$$|x|_\infty = |x_{i_0}| = \left(|x_{i_0}|^p\right)^{1/p} \leq \left(\sum_i |x_i|^p\right)^{1/p} \leq \left(n|x_{i_0}|^p\right)^{1/p} = \underbrace{n^{1/p}}_{\to 1} |x|_\infty.$$

Lemma 3.1 $x \mapsto |x|_p$ ist eine Norm auf $\mathbb{R}^n$ (für $1 \leq p \leq \infty$).

Beweis: Homogenität und Positivität sind offensichtlich, nicht aber die Dreiecksungleichung. Diese folgt aus einer geeigneten Verallgemeinerung der Cauchy-Schwarz'schen Ungleichung, der *Hölder'schen Ungleichung*:

$$|\langle x, y\rangle| \leq |x|_p |y|_{p'} \quad (x, y \in \mathbb{R}^n,\ p' \text{ Lösung der Gleichung } \frac{1}{p} + \frac{1}{p'} = 1)$$

(mit der Konvention $p' = \infty$ für $p = 1$, und $p' = 1$ für $p = \infty$). Details siehe Übungen.

Für $0 < p < 1$ ist die p-"Norm" immer noch wohldefiniert und erfüllt die zwei Norm-Axiome Homogenität und Positivität, verletzt aber die Dreiecksungleichung und ist deshalb keine Norm mehr, siehe Übungen.

Wie sieht die Einheitskugel $\{x \in \mathbb{R}^n : |x|_p \leq 1\}$ aus? Im $\mathbb{R}^2$ ergibt sich: für $p = 1$ ein "Diamant", d.h. ein Viereck mit Ecken

$$\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}-1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}0\\-1\end{pmatrix};$$

für $p = 2$ ein Kreis; für $p = \infty$ ein achsenparalleles Quadrat mit Ecken

$$\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1\\-1\end{pmatrix}, \begin{pmatrix}-1\\1\end{pmatrix}, \begin{pmatrix}-1\\-1\end{pmatrix};$$

und in den dazwischenliegenden Parameterbereichen $1 < p < 2$ bzw. $2 < p < \infty$ ein "abgerundeter Diamant" bzw. ein "abgerundetes achsenparalleles Quadrat".

Für grosse n verhalten sich die p-Normen sehr unterschiedlich. Um dies zu verstehen, betrachten wir Datenvektoren, die durch Diskretisierung einer Funktion entstehen (siehe auch Abschnitt 1.1): sei $f : [0,T] \to \mathbb{R}$ eine stetige Funktion und $v^{(h)}$ der durch Zerlegung des Intervalls $[0,T]$ in n gleich lange Teilintervalle erhaltene Datenvektor im $\mathbb{R}^n$, d.h. $h = \frac{T}{n}$ (Gitterweite),

$$v^{(h)} = \begin{pmatrix} v_1^{(h)} \\ \vdots \\ v_n^{(h)} \end{pmatrix}, \quad v_i^{(h)} = f(i\,h),\ i = 1, ..., n.$$

Zunächst betrachten wir das Verhalten der Maximumsnorm für grosses n:

$$|v^{(h)}|_\infty \to \sup\{|f(t)| : t \in [0,T]\} = \max\{|f(t)| : t \in [0,T]\} =: \|f\|_\infty \ \ (h \to 0)$$

(das Supremum ist ein Maximum wegen Satz 1.5). Die Abbildung $f \mapsto \|f\|_\infty$ ist eine Norm auf dem Vektorraum $C([0,T]) := \{f : [0,T] \to \mathbb{R} : f \text{ stetig}\}$, und heisst Supremumsnorm. *Im Grenzwert kleiner Gitterweite wird die Maximumsnorm also unabhängig von der Gitterweite; sie ist dementsprechend eine sinnvolle Norm für grosse Datenvektoren, die durch Diskretisierung einer kontinuierlichen Funktion entstehen.*

Was passiert mit der euklidischen Norm für grosses n?

Offenbar gilt

$$|v^{(h)}|_2 \to \infty \ \ (h \to 0) \ \text{ es sei denn } f \text{ ist die Nullfunktion,}$$

d.h. die Norm divergiert! *Die euklidische Norm ist also keine sinnvolle Norm für grosse Datenvektoren, die durch Diskretisierung einer kontinuierlichen Funktion entstehen.*

Dieses Fehlverhalten der euklidischen Norm lässt sich aber nicht nur durch Übergang zur Maximumsnorm, sondern alternativ durch einen geeigneten Gewichtsfaktor beheben. Die "richtig" gewichtete euklidische Norm auf dem $\mathbb{R}^n$

$$\|v\|_{2,h} := \sqrt{h}|v|_2, \quad h = \tfrac{T}{n}$$

erfüllt

$$\|v^{(h)}\|_{2,h} = \sqrt{\sum_{j=1}^{n} \big(f(j\,h)\big)^2 h} \to \sqrt{\int_0^T \big(f(t)\big)^2 dt} =: \|f\|_2 \;\; (h \to 0),$$

wobei wir den Term unter der ersten Wurzel als *Riemann-Summe* der Funktion $t \mapsto f(t)^2$ erkannt und interpretiert haben. Die rechte Seite ist eine Norm auf $C([0,T])$, und heisst L^2-Norm.

3.2 Äquivalenz aller Normen auf dem $\mathbb{R}^n$

Wie wir gesehen haben, verhalten sich Maximumsnorm, euklidische Norm, und gewichtete euklidische Norm im Limes grosser Datenvektoren sehr unterschiedlich. Im $\mathbb{R}^n$ für festes n ist es aber für viele Zwecke (siehe die Folgerungen nach Satz 3.1) egal, welche Norm wir verwenden.

Zur Vorbereitung definieren wir: Zwei Normen $x \mapsto \|x\|_A$ und $x \mapsto \|x\|_B$ auf einem $\mathbb{R}$-Vektorraum V heissen *äquivalent*, wenn Konstanten $c > 0$, $C > 0$ existieren mit

$$c\|x\|_B \le \|x\|_A \le C\|x\|_B \quad \text{für alle } x \in V. \tag{*}$$

Satz 3.1 *Auf dem $\mathbb{R}^n$ sind alle Normen äquivalent.*

Folgerung: Die Begriffe

- *konvergent* (für Folgen im $\mathbb{R}^n$)
- *offen, abgeschlossen, beschränkt, kompakt* (für Teilmengen des $\mathbb{R}^n$)
- *stetig* (für Funktionen $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$)

hängen nicht von der Wahl der Norm ab! Wir hätten also in Kapitel 1 mit einer beliebigen anderen Norm auf dem $\mathbb{R}^n$ starten können; die erhaltenen Ergebnisse bleiben gleich.

Um Satz 3.1 als bemerkenswert würdigen zu können, lohnt ein Blick in unendlichdimensionale Vektorräume wie den oben betrachteten Raum $C([0,T])$ der stetigen

Funktionen auf dem Intervall $[0,T]$. Die beiden oben durch Grenzübergang $h \to 0$ gewonnenen Normen, d.h. die Supremumsnorm und die L^2-Norm, sind nicht äquivalent. Dies sieht man am besten ohne gross herumzurechnen durch Skizzieren von Funktionen wie $f^{(j)}$ = Zacken der Höhe 1 und der Breite T/j auf dem Teilintervall $[0,T/j]$ und 0 sonst. Da die Zackenhöhe konstant bleibt, ist $\|f^{(j)}\|_\infty = 1$; da die Zackenhöhe konstant bleibt und die Zackenbreite gegen Null geht, gilt $\|f^{(j)}\|_2 \to 0$ $(j \to \infty)$.

Beweis von Satz 3.1 O.B.d.A. können wir annehmen: $\|x\|_A = |x|$ (euklidische Norm). Die erste Ungleichung in (*) folgt sofort aus der Dreiecksungleichung und der Homogenität der B-Norm, denn mit $e_i := i^{ter}$ Einheitsvektor im $\mathbb{R}^n$ (d.h. $(e_i)_k = 1$ für $k = i$ und 0 sonst) folgt

$$\|x\|_B = \Big\| \sum_{i=1}^n x_i e_i \Big\|_B \le \sum_{i=1}^n \|x_i e_i\|_B = \sum_{i=1}^n \underbrace{|x_i|}_{\le |x|} \|e_i\|_B \le \Big(\sum_{i=1}^n \|e_i\|_B \Big) |x|$$

d.h. die erste Ungleichung in (*) mit $c = \frac{1}{\sum_{i=1}^n \|e_i\|_B}$.

Die zweite Ungleichung ist nichttrivial. Zunächst behaupten wir: die Funktion $x \mapsto \|x\|_B$ ist stetig. Dies folgt aus der schon bewiesenen ersten Ungleichung, angewandt auf Differenzen $x - y$, und dem ε-δ-Kriterium. Nach Satz vom Maximum und Minimum (Satz 1.5) besitzt die Funktion auf der (abgeschlossenen und beschränkten, und somit wegen Lemma 1.6 kompakten) Menge $K = \{x \in \mathbb{R}^n : |x| = 1\}$ eine Minimumsstelle x_*. Wegen Positivität der B-Norm ist $m := \|x_*\|_B > 0$, und aus der Homogenität der B-Norm folgt für alle $x \neq 0$

$$\|x\|_B = |x| \, \|\tfrac{x}{|x|}\|_B \ge |x| \, m.$$

Diese beiden Tatsachen zusammengenommen liefern die zweite Ungleichung in (*), mit $C = \frac{1}{m}$.

3.3 Allgemeine normierte Räume

[Erinnerung: Was eine Norm ist, wurde in Abschnitt 1.2 Def. 1.2 definiert.]

Def. 3.1 (normierter Vektorraum) Sei $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$. Ein *normierter $\mathbb{K}$-Vektorraum* ist ein Paar $(V, \|\cdot\|)$ mit V $\mathbb{K}$-Vektorraum, $\|\cdot\|$ Norm auf V. (Wir benutzen hier die übliche Notation $\|\cdot\|$ für die Abbildung $v \mapsto \|v\|$.)

Als nächstes übertragen wir Grundbegriffe wie Cauchyfolge, konvergent, abgeschlossen vom $\mathbb{R}^n$ auf beliebige, möglicherweise unendlichdimensionale normierte Vektorräume.

Def. 3.2 (Grundbegriffe)
Eine Folge (v_n) in V heisst

- Cauchyfolge, wenn zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ existiert sodass $\|v_n - v_m\| < \varepsilon$ für alle $m, n \geq N$
- konvergent gegen $v \in V$, Schreibweise: $v_n \to v$, wenn zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ existiert sodass $\|v_n - v\| < \varepsilon$ für alle $n \geq N$
- konvergent, wenn ein $v \in V$ existiert sodass $v_n \to v$.

Eine Teilmenge $A \subseteq V$ heisst

- abgeschlossen, wenn für jede Folge (v_n) in A mit $v_n \to v \in V$ gilt: $v \in A$
- offen, wenn für jedes $v_0 \in A$ ein $\varepsilon > 0$ existiert sodass $B_\varepsilon(v_0) := \{v \in V : \|v - v_0\| < \varepsilon\} \subseteq A$.

Eine Abbildung $f : M \subseteq V \to V'$, $(V', \|\cdot\|')$ normierter Vektorraum, heisst

- stetig, wenn für jede Folge (v_n) in M mit $v_n \to v \in M$ gilt: $f(v_n) \to f(v)$.

Wir kommen nun zum Begriff der Vollständigkeit. Sie war in Analysis 1 eine unverzichtbare Eigenschaft des Vektorraumes $\mathbb{R}$: *ohne Vollständigkeit keine Wurzeln, keine Kreiszahl π, kein Zwischenwertsatz, keine Häufungspunkte und damit kein Satz vom Maximum und Minimum, kein Summieren von Reihen, kein Integral.* Das Vollständigkeitsaxiom beruhte aber auf der Anordnung von $\mathbb{R}$, die wir selbst im $\mathbb{R}^n$ nicht haben (dort konnten wir uns noch durch komponentenweises Vorgehen helfen, das uns z.B. in Abschnitten 1.4 und 1.7 Häufungspunkte und den Satz vom Maximum und Minimum geliefert hat).

Es gibt aber eine Aussage aus Analysis 1, die wir aus dem Vollständigkeitsaxiom gefolgert hatten und die sogar zu diesem äquivalent ist (was wir hier nicht beweisen), nämlich das *Cauchy-Kriterium* (Analysis 1 Satz 2.5). Dessen Formulierung benötigt nur den Absolutbetrag auf $\mathbb{R}$. Das Cauchy-Kriterium eignet sich somit als Definition von Vollständigkeit eines normierten Vektorraums.

Def. 3.3 (Vollständigkeit); Banachraum) Ein normierter $\mathbb{R}$-Vektorraum heisst **vollständig**, wenn jede Cauchyfolge konvergent ist. Ein vollständiger normierter $\mathbb{R}$-Vektorraum heisst *Banachraum*.

Beispiele 1) $(\mathbb{R}^n, |\cdot|)$, $|x| = \sqrt{x_1^2 + ... + x_n^2}$ euklidische Norm, ist vollständig. Siehe Abschnitt 1.3 Korollar 1.1.

2) $(\mathbb{R}^n, \|\cdot\|)$, $\|\cdot\|$ beliebige Norm auf $\mathbb{R}^n$, ist vollständig. Dies folgt aus 1) und der Äquivalenz aller Normen auf dem $\mathbb{R}^n$ (Satz 3.1).

3) Der normierte $\mathbb{Q}$-Vektorraum $(\mathbb{Q}^n, |\cdot|)$ ist **nicht** vollständig.

4) $(C([0,T], \|\cdot\|_\infty)$, also der in §3.1 eingeführte Raum der stetigen Funktionen von $[0,T]$ nach $\mathbb{R}$ versehen mit der Supremumsnorm, ist vollständig. Das beweisen wir im nächsten Abschnitt.

5) $(C([0,T],\|\cdot\|_2)$, also der in §3.1 eingeführte Raum der stetigen Funktionen von $[0,T]$ nach $\mathbb{R}$ versehen mit der L^2-Norm $\|f\|_2 = (\int_0^T f(t)^2 dt)^{1/2}$, ist **nicht** vollständig.

Aus den Gegenbeispielen lernen wir: Für Vollständigkeit ist sowohl die "Löcherfreiheit" des zugrundeliegenden Körpers als auch die Endlichdimensionalität wichtig. Aus Beispiel 4) lernen wir: auch im Unendlichdimensionalen ist Vollständigkeit möglich, allerdings ist hierfür eine "gute" Wahl der Norm notwendig.

Ein Spezialfall normierter Räume sind innere Produkträume. Das bedeutet, dass die Norm – wie die euklidische Norm im $\mathbb{R}^n$ – von einem inneren Produkt herkommt. Die genauen Definitionen sollten aus der Linearen Algebra bekannt sein, und lauten wie folgt:

Def. 3.4 Sei V ein $\mathbb{K}$-Vektorraum, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$. Ein **inneres Produkt** in V ist eine Abbildung $V \times V \to \mathbb{K}$, $(x,y) \mapsto \langle x,y\rangle$, sodass gilt:
(i) Positivität: $\langle x, x\rangle \geq 0$, " $=$"$\Longleftrightarrow x = 0$
(ii) Symmetrie: $\langle x, y\rangle = \overline{\langle y, x\rangle}$
(iii) Bilinearität: $\langle x, \lambda y + \mu z\rangle = \lambda\langle x, y\rangle + \mu\langle x, z\rangle$.
Hierbei sind $x, y, z \in V$ und $\lambda, \mu \in \mathbb{K}$.

Bedingung (iii) sagt, dass das innere Produkt linear im zweiten Argument ist. Wegen (ii) impliziert dies, dass es konjugiert-linear im ersten Argument ist, d.h. $\langle \lambda x + \mu z, y\rangle = \overline{\lambda}\langle x,y\rangle + \overline{\mu}\langle x,z\rangle$. Im reellen Fall ($\mathbb{K} = \mathbb{R}$) ist das innere Produkt also bilinear. Standardbeispiele innerer Produkte sind

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i \quad \text{auf } \mathbb{R}^n$$

und

$$\langle x, y\rangle = \sum_{i=1}^{n} \overline{x_i} y_i \quad \text{auf } \mathbb{C}^n.$$

Ein unendlichdimensionales Beispiel ist das innere Produkt

$$\langle f, g\rangle = \int_a^b \overline{f(x)}\, g(x)\, dx$$

auf dem Vektorraum der stetigen Funktionen von $[a,b]$ nach $\mathbb{C}$; es heisst L^2-inneres Produkt. Es spielt in vielen Anwendungen eine wichtige Rolle, z.B. in der Quantenphysik oder der Fourieranalysis. In Analysis 3 wird es zum Verständnis von Fourierreihen beitragen.

Ist $\langle\cdot,\cdot\rangle$ ein inneres Produkt auf V, so ist die Abbildung

$$\|\cdot\| : V \to \mathbb{R}, \quad \|x\| = \sqrt{\langle x, x\rangle} \tag{N}$$

eine Norm. Beweis: Analog zumBeweis für das euklidische Skalarprodukt im $\mathbb{R}^n$ in Abschnitt 1.2. Im Fall obiger drei Beispiele erhält man die euklidische Norm auf $\mathbb{R}^n$, die euklidische Norm auf $\mathbb{C}^n$, und die L^2-Norm auf $C([0,T])$,

$$\|f\|_2 = \Big(\int_a^b |f(x)|^2 dx\Big)^{1/2}.$$

Ein Vektorraum versehen mit einem inneren Produkt sowie der zugehörigen Norm (N) heisst *innerer Produktraum*. Ein (bezüglich der Norm (N)) vollständiger innerer Produktraum heisst *Hilbertraum*.

Ausblick: Wie sich herausstellt, ist beim Aufbau der analytischen Grundbegriffe sowie der Herleitung sinnvoller Folgerungen weder die Vektorraumstruktur des Raumes noch das Homogenitätsaxiom der Norm wesentlich. Dies führt auf den noch allgemeineren Begriff des *metrischen Raumes*. Wie schon in Abschnitt 1.3 besprochen, ist eine Metrik auf einer Menge X eine Abbildung $d : X \times X \to \mathbb{R}$ sodass
(i) $d(x,y) \geq 0$, "$=$" $\Longleftrightarrow x = y$ (Positivität),
(ii) $d(x,y) = d(y,x)$ (Symmetrie),
(iii) $d(x,z) \leq d(x,y) + d(y,z)$ (Dreiecksungleichung).
Die Zahl $d(x,y)$ interpretieren wir als Abstand zwischen x und y. Ein metrischer Raum ist ein Paar (X,d) mit X Menge, d Metrik auf X. Jeder Vektorraum $(V, \|\cdot\|)$ ist mittels $d(x,y) := \|x-y\|$ ein metrischer Raum. Aber auch auf Teilmengen von Vektorräumen gibt es natürliche Metriken, die nicht von dieser Form sind. Auf der Sphäre $\{x \in \mathbb{R}^3 : |x| = 1\}$ könnten wir z.B. statt der Länge $|x-y|$ der geraden Strecke "durch's Innere" zwischen zwei Punkten x und y die Länge der kürzesten Verbindungslinie auf der Sphäre wählen, d.h. das Bogenmass $d(x,y) = \arccos(\langle x,y\rangle) \in [0,\pi]$. Für die in diesem Abschnitt eingeführten qualitativen Begriffe wie konvergent, stetig, abgeschlossen etc. macht das aber keinen Unterschied, denn $|x-y| \leq d(x,y) \leq \frac{\pi}{2}|x-y|$.

3.4 Der Raum $C(\Omega;\mathbb{R}^m)$; gleichmässige Konvergenz

Wir lernen nun das Paradebeispiel eines unendlichdimensionalen aber trotzdem vollständigen Raumes, stetige Funktionen mit Supremumsnorm, genauer kennen.

Als Definitionsbereich lassen wir eine beliebige Menge $\Omega \subseteq \mathbb{R}^n$ zu, und betrachten den Vektorraum der stetigen beschränkten Funktionen von Ω nach $\mathbb{R}^m$,

$$C(\Omega;\mathbb{R}^m) := \{f : \Omega \to \mathbb{R}^m : f \text{ stetig, } f \text{ beschränkt}\},$$

versehen mit der Norm

$$\|f\|_\infty := \sup_{x\in\Omega} |f(x)|.$$

Diese Norm heisst Supremumsnorm. Hierbei bezeichnet $|f(x)|$ die euklidische Norm des Vektors $f(x) \in \mathbb{R}^m$ bzw. im skalaren Fall $m = 1$ den Absolutbetrag der Zahl $f(x)$. Ist Ω kompakt, z.B. ein Intervall $[a,b]$, kann die Bedingung "f beschränkt" in obiger Definition weggelassen werden, denn dann ist f nach Satz vom Maximum und Minimum automatisch beschränkt.

Was bedeutet Konvergenz im Raum $(C(\Omega;\mathbb{R}^m), \|\cdot\|)$? Nichts anderes als die zweite der folgenden beiden klassischen Arten der Konvergenz von Funktionenfolgen.

Def. 3.5 (Punktweise und gleichmässige Konvergenz) Sei $\Omega \subseteq \mathbb{R}^n$, $f^{(k)}$, $f : \Omega \to \mathbb{R}^m$. Die Folge $(f^{(k)})$ heisst

a) *punktweise konvergent* gegen f, wenn $f^{(k)}(x) \to f(x)$ für jedes $x \in \Omega$
b) *gleichmässig konvergent* gegen f, wenn $\sup_{x\in\Omega} |f^{(k)}(x) - f(x)| \to 0$ $(k \to \infty)$.

Offenbar gilt: punktweise Konvergenz $\not\Longleftarrow$ gleichmässige Konvergenz. Gegenbeispiel: Folge der Dreiecksfunktionen der Höhe 1 mit Grundlinie $[0, \frac{1}{k}]$, durch 0 fortgesetzt auf $[0, 1]$.

Konkret bedeutet b): zu jedem $\varepsilon > 0$ existiert ein $N \in \mathbb{N}$ sodass

$$|f^{(k)}(x) - f(x)| < \varepsilon \ \forall k \geq N \quad (*).$$

Im prototypischen Fall $\Omega = [a, b]$ besagt Bedingung (*) also anschaulich: der Graph von $f^{(k)}$ liegt in einem Schlauch vom Radius ε um den Graphen von f.

Wir kommen nun zum Hauptergebnis dieses Abschnittes.

Satz 3.2 *Der normierte Raum* $(C(\Omega; \mathbb{R}^m), \|\cdot\|_\infty)$ *ist vollständig.*

Beweisskizze: Man beschafft sich zunächst einen Kandidaten für den Grenzwert einer Cauchyfolge; dies tut man punktweise mithilfe des Cauchy'schen Konvergenzkriteriums in $\mathbb{R}$. Gleichmässigkeit der Konvergenz weist man durch Grenzübergang $\ell \to \infty$ in der Ungleichung $|f^{(k)}(x) - f^{(\ell)}(x)| < \varepsilon$ nach. Stetigkeit des Grenzwertes folgt aus Satz 3.3.

Satz 3.3 (Der gleichmässige Limes stetiger Funktionen ist stetig) *Sei* $\Omega \subseteq \mathbb{R}^n$, $f^{(k)} : \Omega \to \mathbb{R}^m$ *(*$k \in \mathbb{N}$*), und* $(f^{(k)})$ *konvergiere gleichmässig gegen* $f : \Omega \to \mathbb{R}^m$. *Sind alle* $f^{(k)}$ *stetig, so ist auch* f *stetig.*

Beweis Wir verifizieren das ε-δ-Kriterium für f im Punkt $x_0 \in \Omega$ mithilfe eines $\frac{\varepsilon}{3}$-Argumentes. Sei $\varepsilon > 0$. Wegen der gleichmässigen Konvergenz der $f^{(k)}$ existiert $N \in \mathbb{N}$ sodass $|f(x) - f^{(N)}(x)| < \varepsilon/3$ für alle $x \in \Omega$. Wegen der Stetigkeit von $f^{(N)}$ existiert $\delta > 0$ sodass $|f^{(N)}(x) - f^{(N)}(x_0)| < \varepsilon/3$ für alle $x \in \Omega$ mit $|x - x_0| < \delta$. Insgesamt folgt

$$\begin{aligned} |f(x) - f(x_0)| &= |f(x) - f^{(N)}(x) + f^{(N)}(x) - f^{(N)}(x_0) + f^{(N)}(x_0) - f(x_0)| \\ &\leq \underbrace{|f(x) - f^{(N)}(x)|}_{<\varepsilon/3} + \underbrace{|f^{(N)}(x) - f^{(N)}(x_0)|}_{<\varepsilon/3} + \underbrace{|f^{(N)}(x_0) - f(x_0)|}_{<\varepsilon/3} < \varepsilon \end{aligned}$$

für alle $x \in \Omega$ mit $|x - x_0| < \delta$.

Zum Abschluss dieses Abschnittes untersuchen wir, wie sich gleichmässige Konvergenz mit Ableiten und Integrieren verträgt.

Satz 3.4 (Vertauschbarkeit von Integration und gleichmässiger Konvergenz) *Sei* $y^{(k)} : [a, b] \to \mathbb{R}^n$ *stetig für alle* $k \in \mathbb{N}$, *und* $(y^{(k)})$ *konvergiere gleichmässig gegen* $y : [a, b] \to \mathbb{R}^n$. *Dann gilt*

$$\int_a^b y^{(k)}(t)\, dt \ \to \ \int_a^b y(t)\, dt.$$

Beachte: das Integral auf der rechten Seite ist wohldefiniert, da y nach Satz 3.3 stetig.

Die Voraussetzung der gleichmässigen Konvergenz kann nicht durch punktweise Konvergenz ersetzt werden. Gegenbeispiel: die Dreiecksfunktion $y^{(k)}$ mit Grundlinie $[0, 1/k]$ und Höhe k, durch Null fortgesetzt auf $[0, 1]$, konvergiert punktweise gegen die Nullfunktion, ihr Integral ist aber gleich $\frac{1}{2}$ für alle k und konvergiert somit nicht gegen das Integral der Nullfunktion.

Beweis Wir benutzen der Reihe nach die Linearität des Integrals, die Standardabschätzung *Absolutbetrag des Integrals kleiner gleich Supremum des Absolutbetrags des Integranden mal Länge des Integrationsgebiets* (Analysis 1 Lemma 11.1 b)), und die Definition der gleichmässigen Konvergenz. Dies liefert

$$\Big| \int_a^b y^{(k)} - \int_a^b y \Big| = \Big| \int_a^b \big(y^{(k)} - y\big) \Big| \le (b-a) \cdot \sup_{s \in [a,b]} |y^{(k)}(s) - y(s)| \to 0.$$

Satz 3.5 (Bedingung für Vertauschbarkeit von Ableiten und Konvergenz) *Sei $y^{(k)} : [a, b] \to \mathbb{R}^n$ stetig diff'bar mit*

$$\begin{aligned} (a) \qquad & y^{(k)} \;\to\; y \text{ punktweise}, \\ (b) \qquad & \big(y^{(k)}\big)' \;\to\; z \text{ gleichmässig}. \end{aligned}$$

Dann ist y stetig diff'bar mit $y' = z$.

Beweis Nach Voraussetzung ist $(y^{(k)})'$ stetig, folglich wegen Hauptsatz

$$y^{(k)}(t) = y^{(k)}(a) + \int_a^t (y^{(k)})'(\tau)\, d\tau \text{ für jedes } t \in [a, b].$$

Nach Voraussetzung (a) konvergiert die linke Seite gegen $y(t)$, und aufgrund der Voraussetzungen (a) und (b) sowie Satz 3.3 und Satz 3.4 ist z stetig und die rechte Seite strebt gegen $y(a) + \int_a^t z(\tau)\, d\tau$; folglich

$$y(t) = y(a) + \int_a^t z(\tau)\, d\tau.$$

Wegen Hauptsatz folgt die Behauptung.

3.5 Banach'scher Fixpunktsatz

Der folgende Satz wird uns sowohl beim Lösen von Gleichungssystemen als auch beim Lösen von Differentialgleichungen zunutze kommen. Deshalb ist ein hoher Grad an Abstraktion sinnvoll.

Def. 3.6 Sei M Menge, $f : M \to M$. Ein Punkt $x \in M$ heisst *Fixpunkt* von f, wenn

$$f(x) = x.$$

Ein "Fixpunktsatz" ist ein Satz, der unter geeigneten Annahmen an die Menge M und die Abbildung f die Existenz eines Fixpunktes garantiert. Die "Fixpunktgleichung" $f(x) = x$ sieht auf den ersten Blick speziell aus, ist in Wirklichkeit aber extrem allgemein, wie das folgende Beispiel zeigt.

Beispiel Betrachte n beliebige Gleichungen für n reelle Unbekannte,

$$\begin{array}{rcl} g_1(x_1,...,x_n) & = & c_1 \\ g_2(x_1,...,x_n) & = & c_2 \\ \vdots & & \\ g_n(x_1,...,x_n) & = & c_n \end{array} \qquad (*)$$

mit $g_1,...,g_n : \mathbb{R}^n \to \mathbb{R}$, $c_1,...,c_n \in \mathbb{R}$. Dieses System können wir eleganter schreiben als

$$g(x) = c$$

mit

$$g = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix} : \mathbb{R}^n \to \mathbb{R}^n, \quad c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} \in \mathbb{R}^n.$$

Durch Addition von x und Subtraktion von c auf beiden Seiten können wir es auf "Fixpunktform" bringen:

$$\underbrace{g(x) + x - c}_{=:f(x)} = x.$$

Die Lösungen von (*) sind also genau die Fixpunkte von f.

Satz 3.6 (Banach'scher Fixpunktsatz) *Sei $(V, \|\cdot\|)$ ein Banachraum. Sei A abgeschlossene Teilmenge von V. Sei $F : A \to V$ mit*
(i) $F(A) \subseteq A$
(ii) es gibt $\lambda \in (0,1)$ sodass $\|F(x) - F(y)\| \leq \lambda \|x - y\|$ für alle $x, y \in A$.
Dann existiert genau ein Fixpunkt $x_* \in A$ von F.

Sprechweise: Eine Abbildung, die die Eigenschaft (i) erfüllt, heisst *Selbstabbildung* (von A). Eine Abbildung, die die Eigenschaft (ii) erfüllt, heisst *Kontraktion* (auf A).

Beweis: 1. Betrachte zu beliebigem $x_0 \in A$ die rekursiv durch wiederholte Anwendung von F definierte Folge

$$x_k = F(x_{k-1}) \quad (k = 1, 2, ...)$$

Die Folge ist wegen (i) wohldefiniert.

2. Mithilfe von (ii) weisen wir nach, dass (x_k) Cauchyfolge. Zunächst gilt

$$\|x_{k+1} - x_k\| = \|F(x_k) - F(x_{k-1})\| \underset{\text{(ii)}}{\leq} \lambda \|x_k - x_{k-1}\|$$

und somit durch wiederholtes Anwenden $\|x_{k+1} - x_k\| \leq \lambda^k \|x_1 - x_0\|$. Damit folgt für $m > n \geq N$

$$\begin{aligned} \|x_m - x_n\| &\leq \|x_{n+1} - x_n\| + \|x_{n+2} - x_{n+1}\| + \ldots + \|x_m - x_{m-1}\| \\ &\leq \|x_1 - x_0\| \Big(\underbrace{\lambda^n + \lambda^{n+1} + \ldots + \lambda^{m-1}}_{\leq \lambda^n \sum_{k=0}^{\infty} \lambda^k = \frac{\lambda^n}{1-\lambda}} \Big) \leq \|x_1 - x_0\| \frac{\lambda^N}{1-\lambda}; \end{aligned}$$

da die rechte Seite für $N \to \infty$ gegen 0 konvergiert, folgt (x_k) Cauchy.

3. Da V nach Voraussetzung Banachraum, ist die Folge konvergent gegen ein $x_* \in V$. Da A nach Voraussetzung abgeschlossen, gilt $x_* \in A$. Indem man in der Rekursionsgleichung auf beiden Seiten zum Limes $k \to \infty$ übergeht und die, wegen (ii) vorliegende, Stetigkeit von F ausnutzt, erhält man $x_* = F(x_*)$, d.h. x_* ist Fixpunkt.

4. Eindeutigkeit folgt aus (ii), denn seien x, x' Fixpunkte, so gilt $\|x - x'\| = \|F(x) - F(x')\| \leq \lambda \|x - x'\|$ und damit $\|x - x'\| = 0$.

Korollar 3.1 Unter den Voraussetzungen von Satz 3.6 gilt für die im Beweis definierte Folge (x_k) die Fehlerabschätzung

$$\|x_k - x_*\| \leq \frac{\lambda}{1-\lambda} \|x_k - x_{k-1}\|.$$

Diese Fehlerabschätzung ist interessant, da man zum Auswerten der rechten Seite den Fixpunkt nicht zu kennen braucht; sie kann allein aus Kenntnis der – z.B. numerisch berechneten – ersten Folgenglieder x_0, x_1, ..., x_k bestimmt werden. Hat man diese Folgenglieder berechnet, weiss man, dass der – wegen Satz 3.6 existierende und eindeutige – wahre Fixpunkt von der berechneten Näherung x_k höchstens $\frac{\lambda}{1-\lambda} \|x_k - x_{k-1}\|$ entfernt ist.

Beispiel $V = \mathbb{R}$, $A = [0,1]$, $f(x) = \frac{1}{2} \cos x$, $f : [0,1] \to \mathbb{R}$. Da $\cos x \in [0,1]$ für $x \in [0, \frac{\pi}{2}]$, gilt insbesondere $\cos x \in [0,1]$ für $x \in [0,1]$, d.h.

$$f(A) \subseteq A.$$

Damit ist (i) erfüllt. Des weiteren gilt für $x, y \in [0,1]$ nach Mittelwertsatz

$$|f(x) - f(y)| = |f'(\xi)| |x - y|$$

mit ξ zwischen x und y, aber $|f'(\xi)| = |\frac{1}{2} \sin \xi| \leq \frac{1}{2}$ und somit ist f λ-Lipschitz mit $\lambda = \frac{1}{2}$. Also existiert nach Banach'schem Fixpunktsatz genau ein Fixpunkt $x_* \in [0,1]$ von f, d.h. eine genau eine Lösung $x_* \in [0,1]$ der Gleichung

$$\tfrac{1}{2} \cos x_* = x_*.$$

(Skizze der ersten Folgenglieder der Folge $x_0 = 0$, $x_k = f(x_{k-1})$ anhand des Graphen von f.) Numerische Werte (Matlab, auf 5 Nachkommastellen gerundet):

$$
\begin{aligned}
x_1 &= \tfrac{1}{2} = 0.5 \\
x_2 &= \tfrac{1}{2}\cos(\tfrac{1}{2}) = 0.43879 \\
x_3 &= \tfrac{1}{2}\cos(\tfrac{1}{2}\cos(\tfrac{1}{2})) = 0.45263 \\
x_4 &= \tfrac{1}{2}\cos(\tfrac{1}{2}\cos(\tfrac{1}{2}\cos(\tfrac{1}{2}))) = 0.44965 \\
x_5 &= \tfrac{1}{2}\cos(\tfrac{1}{2}\cos(\tfrac{1}{2}\cos(\tfrac{1}{2}\cos(\tfrac{1}{2})))) = 0.45030 \\
x_6 &= 0.45016, \;\; x_7 = 0.45019, \;\; x_8 = 0.45018, \;\; x_9 = 0.45018.
\end{aligned}
$$

Genauerer Wert (Matlab): $x_9 = 0.45018387....$ Nach Korollar 3.1 gilt $x_* \approx x_9$ mit maximalem Fehler $\frac{\lambda}{1-\lambda}(x_9 - x_8) = 1.4567... \cdot 10^{-6}$, d.h. die ersten 5 Nachkommastellen des exakten Fixpunktes x_* sind 0.45018.

4 Inverse und implizite Funktionen

Als weitere Anwendung der Ableitung studieren wir Systeme von k nichtlinearen Gleichungen für n Unbekannte. Unter geeigneten Voraussetzungen verhalten sich solche Systeme lokal ähnlich wie lineare Gleichungssysteme; z.B. kann man eine einzige Gleichung $f(x_1, x_2) = 0$ für zwei Unbekannte typischerweise lokal durch eine Funktion $x_2 = g(x_1)$ oder eine Funktion $x_1 = h(x_2)$ auflösen. In der Nähe sogenannter "Singularitäten" ist allerdings ein solches Auflösen nicht möglich.

4.1 Inverse Funktionen

Wir beginnen mit dem etwas einfacheren Fall von Gleichungssystemen mit n Gleichungen und n Unbekannten,

$$\begin{array}{rcl} f_1(x_1, ..., x_n) & = & y_1 \\ f_2(x_1, ..., x_n) & = & y_2 \\ \vdots & & \\ f_n(x_1, ..., x_n) & = & y_n \end{array} \qquad (*)$$

($y = (y_1, ..., y_n) \in \mathbb{R}^n$ gegeben, $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ gegeben, $x = (x_1, ..., x_n) \in \mathbb{R}^n$ gesucht). Für $n \geq 2$ steht uns, im Gegensatz zu einer einzigen Gleichung $f(x_1) = y_1$ für eine einzige Unbekannte, der Zwischenwertsatz aus Analysis 1 leider nicht zur Verfügung. Dieser Satz beruht nämlich entscheidend auf der Anordnung von $\mathbb{R}$, und besitzt *kein Analogon* im *nicht angeordneten* $\mathbb{R}^n$ ($n \geq 2$).

Ausgangspunkt unserer Betrachtungen ist die Situation im linearen Fall, $f(x) = Ax$, A reelle $n \times n$ Matrix. Unser Gleichungssystem (*) lautet dann: $Ax = y$, oder in Langform

$$\begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Dieses System wurde in der Linearen Algebra untersucht. Ein Hauptresultat lautet: *Falls die Matrix A invertierbar ist, besitzt das Gleichungssystem für jedes $y \in \mathbb{R}^n$ eine eindeutige Lösung $x \in \mathbb{R}^n$.* Darüber hinaus liefert die Lineare Algebra ein nützliches (für nicht zu grosse Matrizen leicht nachprüfbares) Kriterium für Invertierbarkeit: *A invertierbar* $\Longleftrightarrow \det A \neq 0$, wobei $\det A$ die Determinante von A bezeichnet.

Eine derartig einfache, globale Existenz- und Eindeutigkeitstheorie können wir für nichtlineare Systeme nicht erwarten.

Beispiel 1) $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x(1 - x)$. Phänomene:
A) Für $y > \frac{1}{4}$ besitzt die Gleichung $f(x) = y$ keine Lösung.
B) Für $y < \frac{1}{4}$ besitzt die Gleichung $f(x) = y$ zwei Lösungen.

Die Existenz- und Eindeutigkeitstheorie für lineare Systeme besitzt aber ein weitreichendes *lokales* Analogon für nichtlineare Systeme. Dies beruht letztendlich auf der *lokalen Approximierbarkeit* nichtlinearer durch lineare Abbildungen (siehe insbesondere Abschnitt 2.2).

Satz über inverse Funktionen, informell: Falls das nichtlineare System (*) für eine rechte Seite y_0 eine Lösung x_0 besitzt, und die Ableitung $Df(x_0)$ invertierbar ist, besitzt das System für alle rechten Seiten y nahe y_0 eine eindeutige Lösung x nahe x_0.

Die formalisierte Version dieser Aussage lautet wie folgt.

Satz 4.1 (Satz über inverse Funktionen) *Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}^n$ stetig differenzierbar, $f(x_0) = y_0$, $Df(x_0)$ invertierbar (oder, äquivalent dazu, $J_f(x_0)$ invertierbar). Dann existieren offene Mengen $U_0 \ni x_0$, $V_0 \ni y_0$ sodass:*

(i) Für alle $y \in V_0$ besitzt die Gleichung $f(x) = y$ eine eindeutige Lösung $x \in U_0$.

(ii) $f|_{U_0} : U_0 \to V_0$ ist bijektiv. Die Umkehrabbildung $x =: f^{-1}(y)$, $f^{-1} : V_0 \to U_0$, ist stetig differenzierbar und es gilt

$$Df^{-1}(y) = Df\big(f^{-1}(y)\big)^{-1}, \quad J_{f^{-1}}(y) = J_f\big(f^{-1}(y)\big)^{-1}.$$

Beispiel 1) (Fortsetzung): In diesem Fall ist $x_0 = 0$ offensichtlich Lösung der Gleichung mit rechter Seite $y_0 = 0$, d.h. $f(x_0) = y_0$. Die Jacobimatrix $J_f(x_0)$ ist in diesem Fall die 1×1 Matrix $f'(x_0)$ und diese ist invertierbar, denn $f(x) = x - x^2$, also $f'(x) = 1 - 2x$, also $f'(0) = 1$. Also ist der Satz über inverse Funktionen anwendbar, und für y nahe 0 besitzt die Gleichung $f(x) = y$ eine eindeutige Lösung x nahe 0. Die Voraussetzung "y nahe 0" schliesst Phänomen A) der Nichtexistenz von Lösungen aus. Die Voraussetzung "x nahe 0" schliesst Phänomen B) der Nichteindeutigkeit aus. (Skizze.)

Beweis, Teil 1 ((i) und Bijektivität von f) Aus Analysis 1 kennen wir zwecks Lösen einer nichtlinearen Gleichung $f(x) = y$, $f : \mathbb{R} \to \mathbb{R}$, das Newtonverfahren. Dieses konstruiert eine Folge (x_n) von Näherungslösungen, mit Updating-Schritt

$$x_{n+1} = y\text{-Stelle der Tangente an (den Graphen von) } f \text{ im Punkt } x_n$$

(Skizze), in Formelsprache:

$$x_{n+1} = \text{Lösung der Gleichung } \underbrace{f(x_n) + f'(x_n)(x - x_n)}_{\text{Tangente}} = y,$$

also – indem wir diese Gleichung nach x auflösen –

$$x_{n+1} = x_n - f'(x_n)^{-1}(f(x_n) - y).$$

Diese Iteration mathematisch auf Konvergenz zu untersuchen ist möglich, aber aufwändig. Leichter zu untersuchen ist folgendes *vereinfachte Newtonverfahren*, in dem man – motiviert durch die Tatsache x nahe x_0 – $f'(x_n)$ durch $f'(x_0)$ ersetzt:

$$x_{n+1} = x_n - f'(x_0)^{-1}(f(x_n) - y).$$

Graphisch entspricht dies folgendem Updating-Schritt:

x_{n+1} = y-Stelle der Parallele durch $(x_n, f(x_n))$ zur Tangente an Graph f im Punkt x_0

(Skizze). Dieser Schritt lässt sich sofort ins Mehrdimensionale verallgemeinern:

$$x_{n+1} = \underbrace{x_n - J_f(x_0)^{-1}(f(x_n) - y)}_{=:F_y(x_n)}.$$

Offensichtlich ist x genau dann Fixpunkt von F_y, d.h. Lösung der Fixpunktgleichung

$$x - J_f(x_0)^{-1}\Big(f(x) - y\Big) = x,$$

wenn x Lösung der Gleichung (*) $f(x) = y$. Jetzt kommt die Magie des Banach'schen Fixpunktsatzes ins Spiel. Mithilfe dieses Satzes lässt sich folgende Behauptung zeigen:

Beh.: Für geeignet gewähltes $r > 0$ und $r' > 0$ besitzt die Abbildung F_y für jedes $y \in \overline{B_{r'}(y_0)}$ genau einen Fixpunkt $x \in \overline{B_r(x_0)}$.

Beweis *1. Kontraktion?* Es gilt

$$F_y(x) - F_y(x') = (x - x') - J_f(x_0)^{-1}(f(x) - f(x')).$$

Nach Taylor ist $f_i(x) - f_i(x') = \langle \nabla f_i(\xi_i), x - x' \rangle$ für ein ξ_i zwischen x und x', und somit

$$f(x) - f(x') = \underbrace{\begin{pmatrix} - & \nabla f_1(\xi_1) & - \\ & \vdots & \\ - & \nabla f_n(\xi_n) & - \end{pmatrix}}_{=:M} (x - x').$$

Folglich

$$F_y(x) - F_y(x') = J_f(x_0)^{-1}(J_f(x_0) - M)\,(x - x').$$

Um die rechte Seite abzuschätzen, benutzen wir – wie schon im Beweis der Kettenregel – die euklidische Norm für Matrizen $A \in \mathbb{R}^{m\times n}$, $|A| = (\sum_{i,j} A_{ij}^2)^{1/2}$, und die Cauchy-Schwarz-Ungleichung für Matrix-Vektor-Produkte, $|Av| \le |A|\,|v|$ für alle $A \in \mathbb{R}^{m\times n}$ und alle $v \in \mathbb{R}^n$.

Wähle nun $r > 0$ so klein, dass $\overline{B_r(x_0)} \subseteq \Omega$ und $|J_f(x_0) - M| \le 1/(2|J_f(x_0)^{-1}|)$ für alle $\xi_1, ..., \xi_n \in \overline{B_r(x_0)}$ (das ist möglich, da J_f stetig im Punkt x_0). Somit

$$|F_y(x) - F_y(x')| \underset{\text{Cauchy-Schwarz}}{\le} |J_f(x_0)^{-1}|\,|J_f(x_0) - M|\,|x - x'| \underset{\text{Wahl von } r}{\le} \tfrac{1}{2}|x - x'|.$$

2. Selbstabbildung? Wir müssen zeigen: wenn $y \in \overline{B_{r'}(y_0)}$, gilt $F_y(x) \in \overline{B_r(x_0)}$ für alle $x \in \overline{B_r(x_0)}$. Da x Fixpunkt von F_y genau dann wenn $f(x) = y$, und nach Voraussetzung $f(x_0) = y_0$, ist x_0 Fixpunkt von F_{y_0}, d.h. $F_{y_0}(x_0) = x_0$. Deshalb ist

$$\begin{aligned} F_y(x) - x_0 &= F_y(x) - F_{y_0}(x_0) \\ &= \Big(F_y(x) - F_y(x_0)\Big) + \Big(F_y(x_0) - F_{y_0}(x_0)\Big) \end{aligned}$$

und somit

$$|F_y(x) - x_0| \le \tfrac{1}{2}|x - x_0| + \underbrace{|J_f(x_0)^{-1}(y - y_0)|}_{\le |J_f(x_0)^{-1}|\,|y - y_0|}.$$

Wählen wir $r' \le r/(2|J_f(x_0)^{-1}|)$, dann folgt für $y \in \overline{B_{r'}(y_0)}$

$$\text{linke Seite} \;\le\; \tfrac{r}{2} + |J_f(x_0)^{-1}|r' \underset{\text{Wahl von } r}{\le} \tfrac{r}{2} + \tfrac{r}{2} \;=\; r.$$

3. Anwenden des Banach'schen Fixpunktsatzes: Der Satz ist wegen 1. und 2. anwendbar, und liefert die Behauptung.
Also besitzt für jedes y in $\overline{B_{r'}(y_0)}$ die Gleichung (*) genau eine Lösung x in $\overline{B_r(x_0)}$. Dies impliziert i), und auch die Bijektivität von f mit

$$\begin{aligned} V_0 &:= B_{r'}(y_0), \\ U_0 &:= \{x \in B_r(x_0) : f(x) \in V_0\} = f^{-1}(B_{r'}(y_0)) \cap B_r(x_0) \end{aligned}$$

(beachte: das Urbild der offenen Menge $B_{r'}(y_0)$ unter f ist nach abstraktem ε-δ-Kriterium offen, und somit auch U_0).

Bevor wir zur stetigen Diff'barkeit der Umkehrabbildung kommen, reflektieren wir noch einmal über den obigen, höchst interessanten, Beweis von deren Existenz.

In unserem Existenzbeweis der Lösung x der Gleichung $f(x) = y$ haben wir x als Fixpunkt des vereinfachten Newtonverfahrens (also der Abbildung $F_y(x) = x - J_f(x_0)^{-1}(f(x) - y)$) konstruiert. Im Beweis des Banach'schen Fixpunktsatzes haben wir den Fixpunkt von F_y als Grenzwert der Folge $\alpha_0(y) = x_0$, $\alpha_k(y) = F_y(\alpha_{k-1}(y))$ konstruiert. Insgesamt haben wir also gezeigt: für y hinreichend nahe y_0 konvergiert die Folge $(\alpha_k(y))$ gegen $f^{-1}(y)$.

Wie sieht die Folge $(\alpha_k(y))$ von Näherungslösungen in einem typischen Beispiel aus?

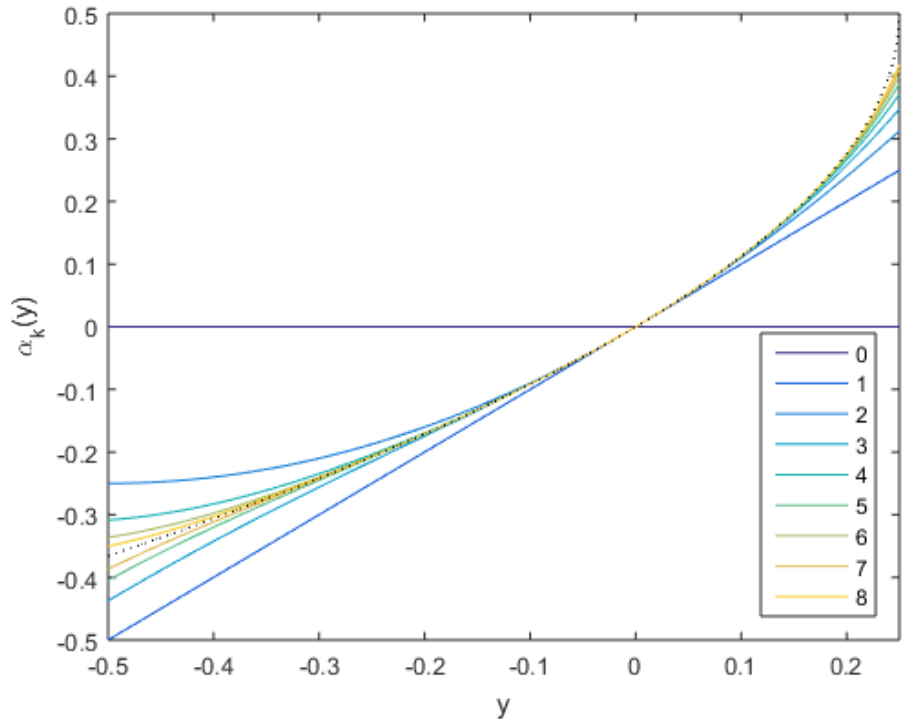
Beispiel 1), Fortsetzung d.h. $f(x) = x(1-x)$, $x_0 = y_0 = 0$.
Da $x_0 = 0$, ist $\alpha_0(y) = 0$. Die Iterationsvorschrift lautet wegen $f'(x_0) = f'(0) = 1$

$$\begin{aligned}\alpha_k(y) &= F_y(\alpha_{k-1}(y)) \\ &= \alpha_{k-1}(y) - 1^{-1}\big(f(\alpha_{k-1}(y)) - y\big) \\ &= \alpha_{k-1}(y) - \big(\alpha_{k-1}(y) - \alpha_{k-1}(y)^2\big) + y \\ &= \alpha_{k-1}(y)^2 + y.\end{aligned}$$

D.h. die nächste Iteration erhalten wir durch Quadrieren der vorherigen und Addition von y. Da $y_0 = 0$, besagt Aussage (i) von Satz 4.1: für y hinreichend nah bei 0 konvergiert die Folge gegen $f^{-1}(y)$. In unserem Fall ist $f^{-1}(y)$ für $y \le \frac{1}{4}$ definiert, als die eindeutige nahe x_0 liegende Lösung x der quadratischen Gleichung $x(1-x) = y$, d.h. $x^2 - x + y = 0$. Die p-q-Formel liefert $x_{1/2} = \frac{1}{2} \pm \sqrt{(1/2)^2 - y}$ und wegen x nahe $x_0 = 0$ ist das x aus Satz 4.1 (i) die '-' Lösung, also

$$f^{-1}(y) = \tfrac{1}{2} - \sqrt{\tfrac{1}{4} - y}, \quad f^{-1} : (-\infty, \tfrac{1}{4}] \to \mathbb{R}.$$

Wie sich die α_k dieser Umkehrfunktion annähern, zeigt der folgende (in der Vorlesung mit Matlab erstellte) Plot.



Die Folge von Näherungen α_k der Umkehrfunktion f^{-1} (schwarze gepunktete Linie) aus dem Beweis des Satzes über inverse Funktionen im Fall $f(x) = x(1-x)$, $x_0 = y_0 = 0$.

Beweis von Satz 4.1, Teil 2 (stetige Differenzierbarkeit der Umkehrfunktion) Wir gehen in 4 Schritten vor:
1. Stetigkeit von $y \mapsto f^{-1}(y) =: \alpha(y)$
2. Herleitung der Formel für die Ableitung $J_f^{-1}(y)$ unter der Annahme α diff'bar
3. Differenzierbarkeit von α
4. Stetigkeit der Ableitung.

1. Stetigkeit von α: Seien y, $y' \in V_0$. Wir schreiben $x := \alpha(y)$, $x' := \alpha(y')$. Indem wir zunächst die Fixpunkteigenschaft $x = F_y(x)$, $x' = F_{y'}(x')$ benutzen und anschliessend die "additive Null" $-F_{y'}(x) + F_{y'}(x)$ einschieben, folgt

$$|\alpha(y) - \alpha(y')| = |x - x'| = |F_y(x) - F_{y'}(x')| \leq \underbrace{|F_y(x) - F_{y'}(x)|}_{\leq |J_f(x_0)^{-1}|\,|y-y'|} + \underbrace{|F_{y'}(x) - F_{y'}(x')|}_{\leq \frac{1}{2}|x-x'|}.$$

(Die Abschätzung für den ersten Term auf der rechten Seite folgt aus $F_y(x) - F_{y'}(x) = J_f(x_0)^{-1}(y - y')$, und diejenige für den zweiten Term aus der bereits gezeigten λ-Lipschitz-Eigenschaft von F_y mit $\lambda = \frac{1}{2}$.) Indem wir auf beiden Seiten $\frac{1}{2}|x - x'|$ abziehen, folgt

$$\tfrac{1}{2}|x - x'| \leq |J_f(x_0)^{-1}|\,|y - y'|$$

oder, nach Multiplikation beider Seiten mit 2, wegen $x = \alpha(y)$, $x' = \alpha(y')$

$$|\alpha(y) - \alpha(y')| \leq 2|J_f(x_0)^{-1}|\,|y - y'|. \qquad (*)$$

Die Umkehrabbildung α ist also nicht nur stetig, sondern sogar Lipschitzstetig. Dies werden wir im Beweis der Differenzierbarkeit ausnutzen.

2. Formel für die Ableitung: diese folgt sofort aus der Kettenregel, denn indem wir die Identität

$$f(\alpha(y)) = y = id(y)$$

(mit id=Identitätsabbildung) nach y ableiten, erhalten wir

$$J_f(\alpha(y))J_\alpha(y) = I$$

(mit I=Einheitsmatrix) und somit $J_\alpha(y) = J_f(\alpha(y))^{-1}$.

3. Differenzierbarkeit: Um zu zeigen, dass α total differenzierbar im Punkt y_0 mit Ableitung $D\alpha(y_0) = Df(x_0)^{-1}$, müssen wir zeigen (siehe Definition der totalen Diff'barkeit): für jede Folge $y_k \to y_0$, $y_k \neq y_0$ gilt

$$\frac{|\alpha(y_k) - \alpha(y_0) - J_f(x_0)^{-1}(y_k - y_0)|}{|y_k - y_0|} \to 0. \qquad (**)$$

Sei $x_k = \alpha(y_k)$. Nach Erweitern mit $|x_k - x_0|$ ist die linke Seite gleich

$$\underbrace{\frac{|x_k - x_0 - J_f(x_0)^{-1}(f(x_k) - f(x_0))|}{|x_k - x_0|}}_{=:A} \underbrace{\frac{|x_k - x_0|}{|y_k - y_0|}}_{=:B}.$$

Der zweite Bruch, B, ist wegen (*) $\leq 2|J_f(x_0)^{-1}|$. Den ersten Bruch, A, bearbeiten wir, indem wir den Vorfaktor $J_f(x_0)^{-1}$ ausklammern:

$$\begin{aligned} A &= \frac{|J_f(x_0)^{-1}\Big(f(x_k) - f(x_0) - J_f(x_0)(x_k - x_0)\Big)|}{|x_k - x_0|} \\ &\leq |J_f(x_0)^{-1}| \underbrace{\frac{|f(x_k) - f(x_0) - J_f(x_0)(x_k - x_0)|}{|x_k - x_0|}}_{\to 0 \text{ da } f \text{ diff'bar}}. \end{aligned}$$

Damit ist (**) bewiesen.

4. Stetigkeit der Ableitung: Wir schreiben die Formel aus 2. für die Ableitung $J_\alpha(y)$ als Dreifach-Verkettung,

$$J_\alpha(y) = J_f(\alpha(y))^{-1} = \Phi\Big(J_f\big(\alpha(y)\big)\Big)$$

mit der Matrixinversionsabbildung

$$\Phi(A) = A^{-1}, \quad \Phi : \{A \in \mathbb{R}^{n\times n} : \det A \neq 0\} \to M^{n\times n}.$$

Die drei auftretenden Abbildungen sind allesamt stetig (α wie in 1. gezeigt, J_f nach Voraussetzung, und Φ aufgrund des nachfolgenden Lemmas), und somit auch deren Verkettung J_α. Es bleibt noch zu zeigen:

Lemma 4.1 Die Matrixinversionsabbildung $\Phi : A \mapsto A^{-1}$ ist stetig.

Beweis Dies folgt z.B. aus der Formel

$$A^{-1} = \frac{1}{\det A}(\operatorname{cof} A)^T,$$

wobei $(\operatorname{cof} A)_{ij} = (-1)^{i+j} \det \widehat{A_{ij}}$ und $\widehat{A_{ij}}$ die (n-1)×(n-1) Matrix bezeichnet, die man aus A durch Streichen der i^{ten} Zeile und j^{ten} Spalte erhält. Gemäss dieser Formel sind die Komponenten der Matrix A^{-1} rationale Funktionen (d.h. Quotienten von Polynomen) in den Komponenten der Matrix A, und somit stetig.

Zum Abschluss dieses Abschnittes schauen wir uns noch ein Beispiel an, das zeigt, dass an Punkten mit nichtinvertierbarer Ableitung oft auch die Funktion selbst nicht

lokal invertierbar ist. Weitere Beispiele zum inversen Funktionensatz siehe Übungen.

Beispiel 2) 2D Polarkoordinaten. Sei f die Abbildung von Polar- auf kartesische Koordinaten, d.h.

$$f(r,\varphi) = \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix}, \quad f : \mathbb{R}^2 \to \mathbb{R}^2.$$

Wir berechnen

$$J_f(r,\varphi) = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix}.$$

Es gilt $\det J_f(r,\varphi) = r\cos^2\varphi + r\sin^2\varphi = r$ und folglich

$$J_f(r,\varphi) \text{ nicht invertierbar } \iff r = 0,$$

d.h. J_f ist genau auf der vertikalen Achse der (r,φ)-Ebene nicht invertierbar. In der Tat bildet f die gesamte vertikale Achse der (r,φ)-Ebene auf den Punkt $(x,y) = (0,0)$ der kartesischen Ebene ab, und folglich existiert für jeden Punkt $(0,\varphi)$ auf der vertikalen Achse *keine* offene Menge $U_0 \ni (0,\varphi)$ sodass $f|_{U_0}$ injektiv.

Vergleich mit Analysis 1. Es ist legitim und üblich, aber in gewisser Weise willkürlich, die Polarkoordinatenabbildung auf einen festen Bereich einzuschränken, auf dem sie bijektiv ist (Satz 10.7, Analysis 1): $f : U \to V$ bijektiv mit $U = (0,\infty) \times (-\pi,\pi]$, $V = \mathbb{R}^2\backslash\{0\}$. Dann ist aber die zugehörige Umkehrabbildung g auf der negativen x-Achse in der kartesischen Ebene (d.h. an den Punkten $(x_0,0)$, $x_0 < 0$) unstetig, denn sie ist für kleine $y \geq 0$ ungefähr gleich $(|x_0|,\pi)$ und für kleine $y < 0$ ungefähr gleich $(|x_0|,-\pi)$. Andererseits sind – aus Sicht des Satzes über inverse Funktionen – die Punkte $(r_0,\varphi_0) = (|x_0|,\pi)$, $x_0 < 0$, regulär: es gilt $f(r_0,\varphi_0) = (x_0,0)$ und die Jacobimatrix $J_f(r_0,\varphi_0)$ ist invertierbar. Also existiert gemäss des Satzes über inverse Funktionen auf einer offenen Menge $U_0 \ni (x_0,0)$ eine *stetige* lokale Umkehrabbildung f^{-1}. Diese stimmt für kleine $y < 0$ nicht mit der globalen Umkehrabbildung g überein, genauer:

$$f^{-1}(x,y) = \begin{cases} g(x,y) \approx \begin{pmatrix} |x| \\ \pi \end{pmatrix} & \text{für } (x,y) \in U_0,\ y \geq 0 \\ g(x,y) + \begin{pmatrix} 0 \\ 2\pi \end{pmatrix} \approx \begin{pmatrix} |x| \\ \pi \end{pmatrix} & \text{für } (x,y) \in U_0,\ y < 0. \end{cases}$$

Zusammengefasst: die lokale Umkehrabbildung f^{-1} vermeidet die künstliche Winkelsprungstelle und liefert nahe der negativen x-Achse der kartesischen Ebene die eindeutige Lösung φ nahe π des Gleichungssystems $x = r\cos\varphi$, $y = r\sin\varphi$.

4.2 Implizite Funktionen

Im vorangegangenen Abschnitt über inverse Funktionen haben wir uns mit dem Lösen von *gleich vielen* Gleichungen und Unbekannten beschäftigt. Nun interessieren wir uns für das Lösen von *weniger* Gleichungen als Unbekannten.

Wir wollen also die Lösungsmenge von k nichtlinearen Gleichungen mit n Unbekannten beschreiben, wenn $k < n$:

$$\begin{aligned} f_1(z_1, ..., z_n) &= 0 \\ &\vdots \\ f_k(z_1, ..., z_n) &= 0 \end{aligned} \qquad (*)$$

für eine gegebene Funktion $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^k$. (Dass die rechten Seiten Null sind, ist keine Beschränkung der Allgemeinheit, denn die Gleichung $f(z) = c$, $c \in \mathbb{R}^k$, können wir auf die Form (*) bringen, indem wir f durch $f - c$ ersetzen.)

Wir schauen uns zuerst zwei Beispiele im instruktiven Spezialfall einer nichtlinearen Gleichung mit zwei Unbekannten ($k = 1$, $n = 2$, $f : \mathbb{R}^2 \to \mathbb{R}$) an.

Beispiel 1) $x^2 + y^2 - c = 0$, $c \in \mathbb{R}$. Offenbar ist die Lösungsmenge leer wenn $c < 0$, der Nullpunkt wenn $c = 0$, und eine Kreislinie wenn $c > 0$, d.h.

$$\text{Lösungsmenge} = \begin{cases} \varnothing, & c < 0 \\ \{0\}, & c = 0 \\ S_{\sqrt{c}}(0), & c > 0 \end{cases}$$

wobei $S_r(0)$ wie üblich die Kreislinie $\{z = (x, y) \in \mathbb{R}^2 : |z| = r\}$ vom Radius r bezeichnet.

Beispiel 2) $x^2 - y^2 - c = 0$, $c \in \mathbb{R}$. Falls $c < 0$, folgt $y^2 = x^2 + |c|$ und wir können nach y auflösen, $y = \pm\sqrt{x^2 + |c|}$ (geometrisch sind das zwei nach oben bzw. unten offene Hyperbeläste); falls $c > 0$, folgt $y^2 + |c| = x^2$ und wir können nach x auflösen, $x = \pm\sqrt{y^2 + |c|}$ (das sind zwei nach rechts bzw. links offene Hyperbeläste); für $c = 0$ erhalten wir zwei sich im Ursprung schneidende Geraden. Insgesamt folgt

$$\text{Lösungsmenge} = \begin{cases} \{y = \pm x\}, & c = 0 \\ \{y = \pm\sqrt{x^2 + |c|}\}, & c < 0 \\ \{x = \pm\sqrt{y^2 + |c|}\}, & c > 0. \end{cases}$$

Was lernen wir aus diesen Beispielen?
A) Global sind die Lösungsmengen *weder* Graphen über der x-Achse *noch* Graphen über der y-Achse. *D.h. wir können die Gleichung (*) global weder eindeutig nach x noch eindeutig nach y auflösen.*
B) Nahe $(x_0, y_0) \neq (0, 0)$ ist $f^{-1}(0)$ jeweils *entweder* ein Graph über der x-Achse *oder* ein Graph über der y-Achse. *D.h. wir können die Gleichung (*) lokal entweder*

eindeutig nach x oder eindeutig nach y auflösen.
C) Nahe $(x_0, y_0) = (0, 0)$ verhält sich die Lösungsmenge sonderbar. In Beispiel 1 hat sie die "falsche Dimension" (wir haben eine Gleichung mit zwei Unbekannten gelöst, erhalten aber nur einen einzigen Punkt, also ein "0-dimensionales" Objekt) und in Beispiel 2 erhalten wir zwei Graphen statt einem, und können also lokal weder eindeutig nach x noch eindeutig nach y auflösen.

Der Satz über implizite Funktionen (siehe unten) zeigt, dass das Verhalten in B) typisch ist, und liefert eine Erklärung für das sonderbare Verhalten in C).

Zur Vorbereitung teilen wir die Liste der Unbekannten $(z_1, ..., z_n)$ in Gleichung (*) auf in k Unbekannte $y_1, ..., y_k$, nach denen wir auflösen wollen, und $n-k$ verbleibende Unbekannte $x_1, ..., x_{n-k}$, also

$$z = (x, y) \in \mathbb{R}^{n-k} \times \mathbb{R}^k,$$

und schreiben (*) in der Form

$$\begin{array}{rcl} f_1(x_1, ..., x_{n-k}, y_1, ..., y_k) & = & 0 \\ \vdots & & \\ f_k(x_1, ..., x_{n-k}, y_1, ..., y_k) & = & 0. \end{array} \qquad (**)$$

Als nächstes führen wir die Teilmatrix der partiellen Ableitungen von f nach den Komponenten von y ein:

$$\frac{\partial f}{\partial y}(x_0, y_0) = \begin{pmatrix} \frac{\partial f_1}{\partial y_1}(x_0, y_0) & \cdots & \frac{\partial f_1}{\partial y_k}(x_0, y_0) \\ \vdots & & \vdots \\ \frac{\partial f_k}{\partial y_1}(x_0, y_0) & \cdots & \frac{\partial f_k}{\partial y_k}(x_0, y_0) \end{pmatrix}.$$

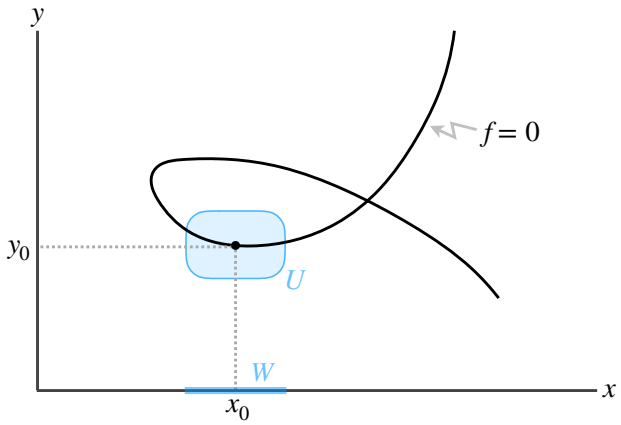
Während die gesamte Jacobimatrix $J_f(x_0, y_0)$ eine flache breite Matrix ist (...weniger Zeilen als Spalten...), ist die Teilmatrix $\frac{\partial f}{\partial y}(x_0, y_0)$ eine quadratische Matrix (...genausoviele Zeilen wie Spalten...).

Die Hauptvoraussetzung des nachfolgenden Satzes ist die Invertierbarkeit dieser Matrix. Diese besagt anschaulich, dass unsere k Gleichungen *erstens voneinander unabhängig sind* (also nicht z.B. $f_2 = 2f_1$ gelten darf, dann wäre nämlich die zweite Gleichung $f_2(x, y) = 0$ automatisch erfüllt, wenn die erste, $f_1(x, y) = 0$, erfüllt ist) und *zweitens wirklich von y abhängen* (also nicht z.B. $f_1(x, y) = \tilde{f}_1(x)$).

Satz 4.2 (Impliziter Funktionensatz) *Sei $\Omega \subseteq \mathbb{R}^{n-k} \times \mathbb{R}^k$ offen, $f : \Omega \to \mathbb{R}^k$ stetig differenzierbar, $1 \le k < n$, $(x_0, y_0) \in \Omega$ mit $f(x_0, y_0) = 0$, und $\frac{\partial f}{\partial y}(x_0, y_0)$ invertierbar. Dann gibt es ein offenes $W \subseteq \mathbb{R}^{n-k}$ mit $x_0 \in W$, ein offenes $U \subseteq \mathbb{R}^n$ mit $(x_0, y_0) \in U$, und eine stetig diff'bare Funktion $g : W \to \mathbb{R}^k$ sodass*

$$\underbrace{f^{-1}(0) \cap U}_{=\{(x,y) \in U : f(x,y)=0\}} = \underbrace{\operatorname{graph} g}_{=\{(x, g(x)) : x \in W\}}.$$

Der Implizite Funktionensatz liefert also die Existenz von 3 verschiedenen Objekten: W (offene Menge im $\mathbb{R}^{n-k}$), U (offene Menge im $\mathbb{R}^n$), g (Funktion von W nach $\mathbb{R}^k$). Eselsbrücke zur Memorisierung: What U Get.



Es ist nützlich, die Mengengleichheit $f^{-1}(0) \cap U = \operatorname{graph} g$ in zwei Inklusionen "$\supseteq$" und "$\subseteq$" aufzuteilen und sich klarzumachen, was diese bedeuten. "$\supseteq$" ist äquivalent zu

$$f(x, g(x)) = 0 \quad \text{für alle } x \in W$$

und besagt folglich die *Existenz* einer Lösung y des Gleichungssystems $f(x, y) = 0$ für gegebenes x (diese ist im Satz mit $g(x)$ bezeichnet). "$\subseteq$" besagt die *Eindeutigkeit*: es gibt (innerhalb der Menge $U \ni (x, y)$) ausser $g(x)$ keine weitere Lösung.

Informelle Kurzfassung des Impliziten Funktionensatzes: Die Lösungsmenge des Gleichungssystems $f(x, y) = 0$ bestehend aus k Gleichungen für n Unbekannte kann lokal mithilfe der Funktion g durch $n - k$ freie Parameter $x_1, ..., x_{n-k}$ beschrieben werden.

Die Funktion g heisst *implizit definierte* oder kurz *implizite* Funktion, da sie nicht – wie bisher bei Funktionen üblich – durch eine explizite Vorschrift, sondern "implizit" (d.h. indirekt), als Lösung des Gleichungssystems $f(x, g(x)) = 0$, definiert ist.

Der lineare Fall: Wir besprechen als nächstes, wie man den IFS für lineare Abbildungen f mit Hilfe von Linearer Algebra beweisen kann, und leiten in diesem Fall explizite Formeln für W, U und g her. Das liefert zwar keine Ideen für den Beweis im nichtlinearen Fall, illustriert aber, wieso die Bedingung $\frac{\partial f}{\partial y}(x_0, y_0)$ invertierbar eine (n-k)-dimensionale Lösungsmenge garantiert und warum diese ein Graph über $x_1, ..., x_{n-k}$ ist. Sei also $f : \mathbb{R}^n \to \mathbb{R}^k$ linear, d.h.

$$f(x, y) = \underbrace{\begin{pmatrix} A & B \end{pmatrix}}_{\substack{k\times(n-k) \\ k\times k}} \begin{matrix} x \\ y \end{matrix} \begin{matrix} \}n-k \\ \}k \end{matrix},$$

$(x_0, y_0) = (0, 0)$. Die Nullstellenmenge $f^{-1}(0)$ ist die Lösung des LGS

$$\begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0,$$

d.h. der Kern der Matrix $\begin{pmatrix} A & B \end{pmatrix}$. In unserem Fall sind die Jacobimatrix und die relevante Teilmatrix gegeben durch

$$J_f(0,0) = \begin{pmatrix} A & B \end{pmatrix} \in \mathbb{R}^{k\times n}, \quad \frac{\partial f}{\partial y}(0,0) = B \in \mathbb{R}^{k\times k}.$$

Die Voraussetzung des IFS sagt: B invertierbar. Dies impliziert, dass $\begin{pmatrix} A & B \end{pmatrix}$ Rang k hat, denn die letzten k Spalten sind linear unabhängig. Laut Rangsatz muss $\operatorname{Ker}\begin{pmatrix} A & B \end{pmatrix}$ ein $(n-k)$-dimensionaler Unterraum sein. Diesen können wir zudem mithilfe von x parametrisieren, indem wir das LGS nach y auflösen:

$$\begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0 \iff Ax + By = 0 \iff By = -Ax \iff y = -B^{-1}Ax. \qquad (*)$$

D.h. wählen wir

$$\begin{aligned} W &= \mathbb{R}^{n-k} \\ U &= \mathbb{R}^n \\ g(x) &= -B^{-1}Ax, \ \ g : W \to \mathbb{R}^k, \end{aligned}$$

so folgt

$$\operatorname{graph} g \underset{\text{Def. von } g}{=} \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^n : y = -B^{-1}Ax \right\} \underset{(*)}{=} \operatorname{Ker}\begin{pmatrix} A & B \end{pmatrix} \underset{\text{trivialerweise}}{=} \operatorname{Ker}\begin{pmatrix} A & B \end{pmatrix} \cap U,$$

wie im IFS behauptet.

Beweis des IFS Wir benutzen einen Trick. Wir nehmen $n-k$ Gleichungen dazu, die sowieso gelten, und betrachten statt (**) das Gleichungssystem

$$\Phi(x,y) = \begin{pmatrix} x \\ 0 \end{pmatrix}$$

wobei

$$\Phi(x,y) = \begin{pmatrix} x \\ f(x,y) \end{pmatrix}, \quad \Phi : \Omega \to \mathbb{R}^n.$$

Dieses System (von n Gleichungen für n Unbekannte) können wir ohne Mühe mithilfe des Satzes über inverse Funktionen (Satz 4.1) behandeln.

Details: Die Jacobimatrix der Abbildung Φ ist

$$J_\Phi(x_0,y_0) = \begin{pmatrix} I_{n-k} & 0 \\ \frac{\partial f}{\partial x}(x_0,y_0) & \frac{\partial f}{\partial y}(x_0,y_0) \end{pmatrix},$$

wobei I_{n-k} die $(n-k)\times(n-k)$ Einheitsmatrix bezeichnet. Wir behaupten: $J_\Phi(x_0,y_0)$ ist invertierbar. Dies folgt aus der vorausgesetzten Invertierbarkeit von $\frac{\partial f}{\partial y}(x_0,y_0)$ sowie dem folgenden

Lemma 4.2 Eine $n \times n$ Matrix der Form

$$\begin{pmatrix} I & 0 \\ A & B \end{pmatrix}, \quad A \in M^{k\times(n-k)},\ B \in M^{k\times k},\ B \text{ invertierbar}$$

ist invertierbar.

Ein-Zeilen-Beweis:

$$\begin{pmatrix} I & 0 \\ A & B \end{pmatrix}\begin{pmatrix} I & 0 \\ -B^{-1}A & B^{-1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ A - BB^{-1}A & BB^{-1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix};$$

also ist die erste Matrix auf der linken Seite invertierbar, und die zweite Matrix auf der linken Seite ihre Inverse.

Wir können also den Satz über inverse Funktionen im Punkt (x_0, y_0) anwenden, folglich gibt es offene Mengen $U_0 \ni (x_0, y_0)$, $V_0 \ni \Phi(x_0, y_0) = (x_0, 0)$ sodass $\Phi|_{U_0} : U_0 \to V_0$ bijektiv und die Umkehrabbildung $\Phi^{-1} : V_0 \to U_0$ stetig diff'bar. Wir bezeichnen die ersten $n - k$ Komponenten der Umkehrabbildung mit u und die restlichen k Komponenten mit v, d.h.

$$\Phi^{-1}(x', y') =: \begin{pmatrix} u(x', y') \\ v(x', y') \end{pmatrix} \in \mathbb{R}^{n-k} \times \mathbb{R}^k.$$

Da Φ^{-1} Umkehrabbildung von Φ, ist

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \Phi(\Phi^{-1}(x', y')) = \Phi(u(x', y'), v(x', y')) = \begin{pmatrix} u(x', y') \\ f(u(x', y'), v(x', y')) \end{pmatrix}.$$

(Die erste Gleichung, $u(x', y') = x'$, war zu erwarten, denn die ersten $n - k$ Komponenten von Φ waren nach Konstruktion die Identitätsabbildung, also sollten auch die ersten $n - k$ Komponenten der Umkehrabbildung die Identitätsabbildung sein.) Indem wir $y' = 0$ setzen, erhalten wir

$$\begin{aligned} &(1) \quad x' = u(x', 0) \\ &(2) \quad 0 = f(u(x', 0), v(x', 0)) \underset{(1)}{=} f(x', v(x', 0)). \end{aligned}$$

Setze $g(x') := v(x', 0)$, wähle offene Mengen $W \ni x_0$, $Z \ni 0$ sodass $W \times Z \subseteq V_0$, und nimm W als Definitionsbereich von g, d.h. $g : W \to \mathbb{R}^k$. Dann gilt $f(x, g(x)) = 0$ für alle $x \in W$ (wegen (2)), oder – in Mengenschreibweise –

$$f^{-1}(0) \cap U_0 \supseteq \operatorname{graph} g.$$

Darüber hinaus ist g stetig diff'bar (da Φ^{-1} stetig diff'bar), und $g(x_0) = y_0$ (da $\Phi(x_0, y_0) = (x_0, 0)$, folglich $\Phi^{-1}(x_0, 0) = (x_0, y_0)$, und somit $v(x_0, 0) = y_0$). Der einzige Schönheitsfehler ist, dass wir in obiger Aussage "$\supseteq$" statt "$=$" erhalten haben. Dies beheben wir, indem wir die Menge U_0 durch eine kleinere Menge ersetzen. Sei $U := \Phi^{-1}(W \times Z)$, so gilt $f^{-1}(0) \cap U = \Phi^{-1}(W \times \{0\}) = \operatorname{graph} g$. Damit ist der Beweis beendet.

Zusatz: Wir merken noch an, dass die Identität $f(x, g(x)) = 0\ \forall x \in W$ eine explizite Formel für die Jacobimatrix $J_g(x)$ liefert. Indem wir $\varphi(x) := f(x, g(x))$, $\psi(x) = (x, g(x))$ setzen, folgt $\varphi = f \circ \psi$ und die Identität nimmt die Form $\varphi \equiv 0$ an. Durch Ableiten nach x folgt

$$0 \equiv J_\varphi(x) = J_f(\psi(x))J_\psi(x) = \begin{pmatrix} \frac{\partial f}{\partial x}(x, g(x)) & \frac{\partial f}{\partial y}(x, g(x)) \end{pmatrix}\begin{pmatrix} I \\ J_g(x) \end{pmatrix} = \frac{\partial f}{\partial x}(x, g(x)) + \frac{\partial f}{\partial y}(x, g(x))J_g(x)$$

und somit durch Auflösen nach $J_g(x)$ (für alle Punkte x, an denen $\frac{\partial f}{\partial y}(x, g(x))$ invertierbar ist, also insbesondere für alle x nahe x_0)

$$J_g(x) = -\frac{\partial f}{\partial y}(x, g(x))^{-1}\frac{\partial f}{\partial x}(x, g(x)).$$

Der implizite Funktionensatz erklärt das Verhalten der Lösungsmengen in Beispielen 1 und 2.

Beispiele 1) und 2), Fortsetzung: Betrachte die Gleichung $f(x,y) = x^2+y^2-c = 0$, $c > 0$. Die Jacobimatrix von f an einem beliebigen Punkt (x_0, y_0) der Lösungsmenge ist die 1×2 Matrix

$$J_f(x_0,y_0) = \Big(\frac{\partial f}{\partial x}(x_0,y_0),\ \frac{\partial f}{\partial y}(x_0,y_0)\Big) = (2x_0,\ 2y_0).$$

Die Teilmatrix $\frac{\partial f}{\partial y}(x_0,y_0)$ ist die 1×1 Matrix

$$\frac{\partial f}{\partial y}(x_0,y_0) = 2y_0.$$

Um den IFS anwenden zu können, muss diese Matrix invertierbar sein. Das ist der Fall, wenn $y_0 \neq 0$, d.h. an allen Punkten (x_0, y_0) der Lösungsmenge, die nicht auf der x-Achse liegen. An solchen Punkten sagt der IFS: wir können die Gleichung lokal (d.h. in einer Umgebung von (x_0, y_0)) nach y auflösen, d.h. die Lösungsmege geschnitten mit einer offenen Umgebung U von (x_0, y_0) ist ein Graph über der x-Achse. In der Tat, wenn $y_0 \neq 0$, muss $x_0^2 < c$ sein, d.h. $x_0 \in (-\sqrt{c}, \sqrt{c})$, und mit

$$\begin{aligned} W &= (-\sqrt{c}, \sqrt{c}) \\ U &= \{(x,y) \in \mathbb{R}^2 : sgn(y) = sgn(y_0)\} \\ g(x) &= sgn(y_0)\sqrt{c-x^2}, \quad g : W \to \mathbb{R} \end{aligned}$$

gilt

$$f^{-1}(0) \cap U = \operatorname{graph} g.$$

Umgekehrt: ist die Voraussetzung $\frac{\partial f}{\partial y}(x_0,y_0)$ invertierbar verletzt, d.h. $y_0 = 0$, existieren solche g und U **nicht**, denn dann ist $x_0 = \pm\sqrt{c}$ und somit enthält jede offene Umgebung U von (x_0, y_0) für alle hinreichend nah bei x_0 liegenden $x \in (-\sqrt{c}, \sqrt{c})$ die beiden Lösungen $(x, \pm\sqrt{c-x^2})$.

Analog ist die Lösungsmenge lokal als Graph über der y-Achse darstellbar, wenn $\frac{\partial f}{\partial x}(x_0,y_0) \neq 0$, d.h. wenn $x_0 \neq 0$. Dies erklärt Aussage B) oben.

Nun zum Fall $c = 0$. Dann ist der Punkt $(x_0, y_0) = (0, 0)$ die einzige Lösung unserer Gleichung; dort verschwindet aber die gesamte Jacobimatrix und der IFS kann weder bezüglich x noch y angewandt werden. In der Tat ist die Lösungsmenge lokal kein Graph einer auf einer offenen Menge definierten stetigen Funktion g.

Insbesondere folgt: in unserem Beispiel ist die Voraussetzung der Invertierbarkeit der Jacobi-Teilmatrix nicht nur hinreichend, sondern auch **notwendig** für die lokale Darstellbarkeit der Lösungsmenge als Graph; das sonderbare Verhalten der Lösungsmenge im Fall $c = 0$, d.h. Aussage C) oben, hängt damit zusammen, dass dort – und nur dort – die Invertierbarkeitsvoraussetzung des IFS sowohl bzgl. x als auch bzgl. y verletzt ist.

Analoges gilt für Beispiel 2), d.h. $f(x,y) = x^2 - y^2 - c$: dann ist $J_f(x_0, y_0) = 2(x_0, -y_0)$, und somit die Invertierbarkeitsvoraussetzung des IFS nur am Nullpunkt sowohl bzgl. x als auch bzgl. y verletzt; aber die den Nullpunkt enthaltende Lösungsmenge ($c = 0$) ist lokal kein Graph, sondern der Schnittpunkt der beiden Geraden $x = \pm y$.

4.3 Singuläre Punkte; Untermannigfaltigkeiten

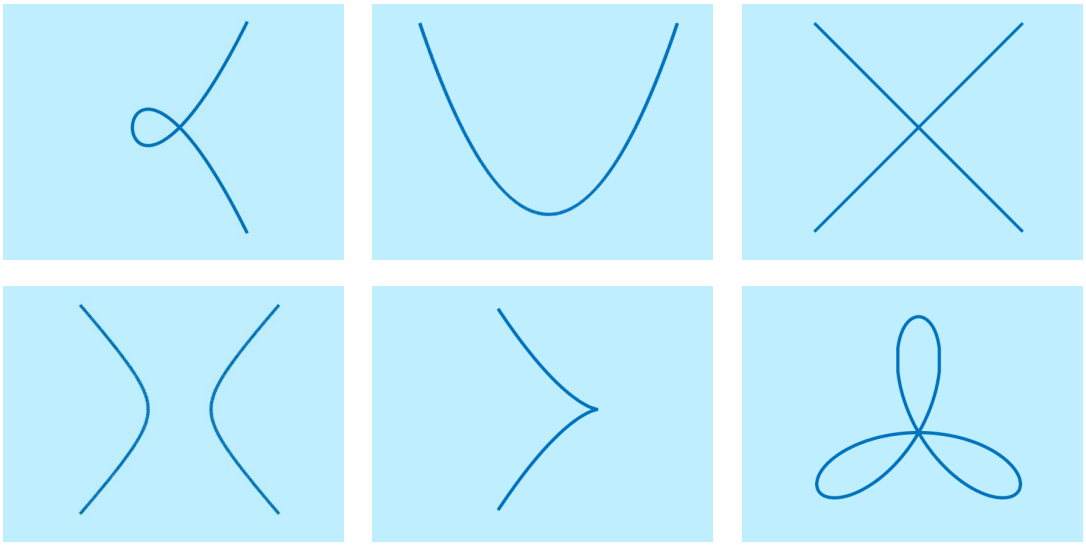
Der Punkt $(0,0)$ in obigen Beispielen, an dem beide Teilmatrizen der Jacobi-Matrix nicht invertierbar sind, ist ein sogennanter *singulärer Punkt*. Allgemeiner definiert man:

Def. 4.1 Sei $\Omega \subset \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}^k$ stetig differenzierbar, $1 \leq k < n$. Ein Punkt $z_0 \in \Omega$ mit $f(z_0) = 0$ heisst *singulärer Punkt* der Lösungsmenge $f^{-1}(0)$, wenn der Rang der Jacobimatrix $J_f(z_0)$ kleiner als k ist; anderenfalls heisst er *regulärer Punkt*.

Singuläre Punkte sind also genau diejenigen Punkte, an denen keine Wahl von k Koordinaten $y_1 = z_{i_1}, \ldots, y_k = z_{i_k}$ $(1 \leq i_1 < \ldots < i_k \leq n)$ möglich ist, sodass $\frac{\partial f}{\partial y}(z_0)$ invertierbar.

Geometrisch gesehen sind die regulären Punkte der Lösungsmenge "lokal langweilig", denn gemäss IFS ist die Lösungsmenge dort lokal als Graph darstellbar. An den singulären Punkten können hingegen interessante Dinge passieren.

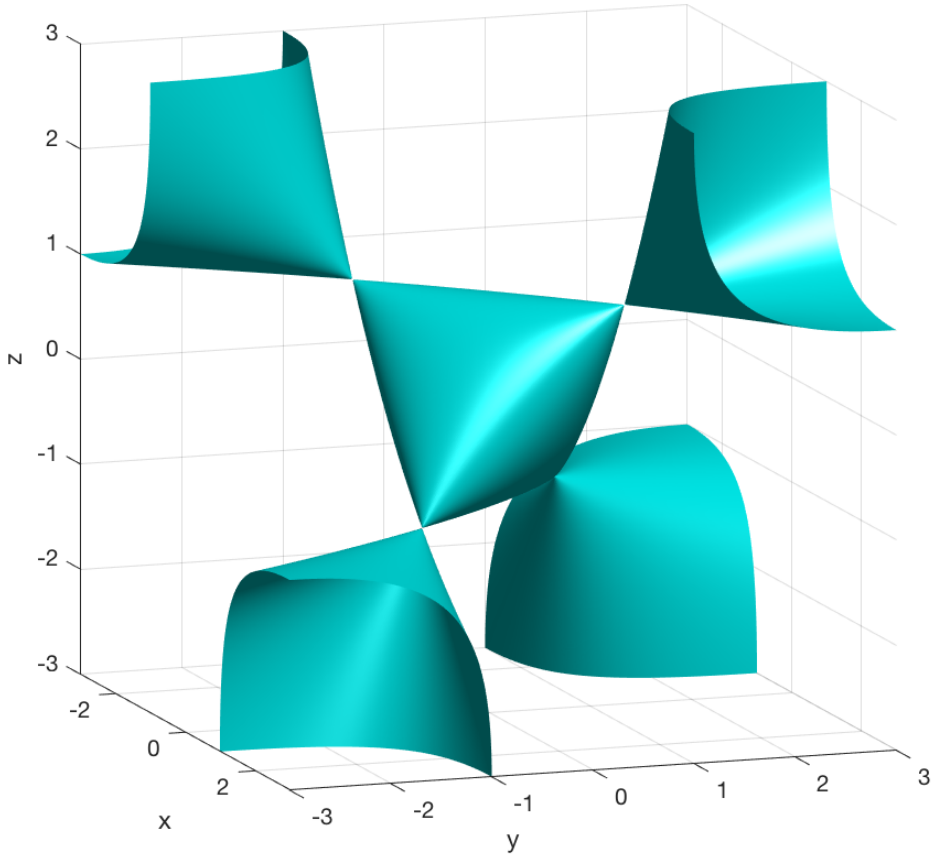
Beispielgalerie impliziter Kurven in $\mathbb{R}^2$ (d.h. Lösungsmengen von Gleichungen der Form $f(x,y) = 0$, $f : \mathbb{R}^2 \to \mathbb{R}$):



Gleichungen: $x^2 - y = 0$, $x^2 - y^2 = 0$, $x^2 + x^3 - y^2 = 0$, $x^2 - y^2 = 1$, $(x^2 + y^2)^2 + 3x^2y - y^3 = 0$, $x^3 + y^2 = 0$. Welche Gleichung gehört zu welchem Bild?

Mithilfe des IFS können wir die "Singularitäten" solcher Kurven (d.h. Punkte, an denen die Lösungsmenge nicht lokal als Graph einer stetig differenzierbaren Funktion darstellbar ist) vorhersagen. Die Strategie ist: *1. Rechne die Jacobimatrix aus. 2. Bestimme die Nullstellen der Jacobimatrix* (beachte: sie hat Rang < 1 g.d.w. sie Null ist). *3. Bestimme die singulären Punkte*, d.h. diejenigen Nullstellen, die auf der Kurve liegen. Nach IFS müssen "Singularitäten" notwendigerweise singuläre Punkte der impliziten Kurve sein; durch Inspektion der Bildergalerie stellen wir fest, dass in unseren Beispielen diese Bedingung jeweils auch hinreichend ist. Untersuchen wir z.B. die Gleichung für die "Schlaufe". Die Jacobimatrix der Funktion $f(x,y) = x^2 + x^3 - y^2$ ist $(x(2+3x), -2y)$. Ist sie gleich Null, muss $y = 0$ sein, und $x = 0$ oder $-\frac{2}{3}$. Der Punkt $(0,0)$ liegt in der Tat in der Lösungsmenge der Gleichung, der Punkt $(-\frac{2}{3}, 0)$ aber nicht. Also ist $(0,0)$ der einzige singuläre Punkt. Das Schaubild zeigt, dass dort die Schlaufe sich selbst kreuzt, d.h. die Lösungsmenge ist in der Tat lokal kein Graph.

Beispiel einer impliziten Fläche im $\mathbb{R}^3$ (d.h. einer Lösungsmenge einer Gleichung der Form $f(x,y,z)=0$, $f:\mathbb{R}^3\to\mathbb{R}$):



Nullstellenmenge der Funktion $f(x,y,z) = 1-(x^2+y^2+z^2)+2xyz$, $f:\mathbb{R}^3\to\mathbb{R}$. Auf dem Bild sehen wir 4 Punkte, an denen die Nullstellenmenge lokal kein Graph ist (sogenannte "Singularitäten" der Fläche). Mithilfe des impliziten Funktionensatzes können wir die Existenz von höchstens 4 solcher Punkte sowie deren Lage vorhersagen, indem wir die singulären Punkte bestimmen (siehe Text).

Gemäss IFS sind die einzigen Kandidaten für Singularitäten die singulären Punkte; wegen $k=1$ sind dies – genau wie im Fall ebener Kurven – die Punkte, an denen die gesamte Jacobimatrix $J_f(x,y,z)\in\mathbb{R}^{1\times 3}$ verschwindet. [Die Nullstellenmenge nahe (x,y,z) ist lokal als Graph über der (x,y)-Ebene darstellbar wenn $\frac{\partial f}{\partial z}(x,y,z)\neq 0$, und analog als Graph über der (x,z)- bzw. (y,z)-Ebene wenn $\frac{\partial f}{\partial y}(x,y,z)\neq 0$ bzw. $\frac{\partial f}{\partial x}(x,y,z)\neq 0$.] Die Jacobimatrix ist

$$J_f(x,y,z)=\begin{pmatrix}-2x+2yz & -2y+2xz & -2z+2xy\end{pmatrix},$$

und verschwindet somit genau dann, wenn $x = yz$, $y = xz$, und $z = xy$. Einsetzen der zweiten Gleichung in die erste liefert $x = xz^2$, mit den Lösungen $x = 0$, z beliebig oder $x \neq 0$, $z = \pm 1$. Im ersten Fall gilt wegen der zweiten und dritten Gleichung $x = y = z = 0$. Anderenfalls müssen – wegen der ersten Gleichung – alle drei Komponenten $\neq 0$ sein und – analog zu obigem Argument für z – in $\{\pm 1\}$ liegen. Nicht alle Vorzeichenwahlen liefern Lösungen, denn unser Gleichungssystem impliziert, dass das Vorzeichen jeder Variablen gleich dem Produkt der beiden anderen Vorzeichen sein muss. Schliesslich stellen wir noch fest, dass der erste kritische Punkt $(0,0,0)$ nicht in der Nullstellenmenge von f liegt, die anderen kritischen Punkte aber schon. Insgesamt folgt also:

$$\{(x,y,z) \in f^{-1}(0) : J_f(x,y,z) = 0\} = \left\{\begin{pmatrix}1\\1\\1\end{pmatrix}, \begin{pmatrix}1\\-1\\-1\end{pmatrix}, \begin{pmatrix}-1\\1\\-1\end{pmatrix}, \begin{pmatrix}-1\\-1\\1\end{pmatrix}\right\}.$$

Die Nullstellenmenge enthält insbesondere genau 4 kritische Punkte des Gradienten, und somit höchstens 4 Singularitäten. Die berechneten kritischen Punkte des Gradienten entsprechen – geometrisch gesehen – den Ecken eines Tetraeders, d.h. jeder zweiten Ecke des Würfels $[-1,1]^3$ im $\mathbb{R}^3$. Das Schaubild zeigt, dass die Nullstellenmenge dort tatsächlich singulär ist. Die innere Fläche entspricht optisch einer etwas aufgeblasenen Version dieses Tetraeders.

Der implizite Funktionensatz erlaubt keine Aussage darüber, wie die Nullstellenmenge einer Funktion in der Nähe kritischer Punkte des Gradienten aussieht. Diese Frage wird in der *algebraischen Geometrie* untersucht, sofern die Funktion f – wie in obigem Beispiel – ein Polynom ist. Für nichtpolynomiale f kann die Nullstellenmenge lokal extrem wild aussehen:

Beispiel Sei

$$f(x,y) = (x^2+y^2)^3 \sin \tfrac{1}{x^2+y^2}$$

$(= r^6 \sin \frac{1}{r^2}$ mit $r = |(x,y)|)$, $f : \mathbb{R}^2 \to \mathbb{R}$. Die Nullstellenmenge $f^{-1}(0)$ besteht aus dem Nullpunkt sowie allen Kreisen um 0 vom Radius $1/\sqrt{\pi n}$, $n \in \mathbb{N}$.

Wir wenden uns nun dem singularitätenfreien Fall zu. Wir beginnen mit einer einfachen aber nützlichen Umformulierung des IFS. Im IFS haben wir die Lösungsmenge lokal als Graph einer Funktion $g : W \subseteq \mathbb{R}^{n-k} \to \mathbb{R}^k$ beschrieben. Der Graph von g, $\{(x,y) \in \mathbb{R}^n : x \in W, y = g(x)\}$, ist aber exakt dasselbe wie das Bild $\psi(W)$ der Abbildung

$$\psi : W \to \mathbb{R}^n, \quad \psi(x) = \begin{pmatrix} x \\ g(x) \end{pmatrix}. \tag{1}$$

Die Abbildung ψ ist im Gegensatz zu g injektiv, denn man kann ja x aus $\psi(x)$ ablesen; sie "verbiegt" das flache Teilstück W des niedrigdimensionalen Unterraums

$\mathbb{R}^{n-k}$ und platziert es irgendwo im hochdimensionalen Raum $\mathbb{R}^n$. Dies liefert folgende alternative, koordinateninvariante Formulierung des IFS:

Korollar 4.1 [Impliziter Funktionensatz: Formulierung via Parametrisierung] Sei $\Omega \subset \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}^k$ stetig differenzierbar, $1 \leq k < n$, $z_0 \in \Omega$ mit $f(z_0) = 0$, und $J_f(z_0)$ habe maximalen Rang (d.h. Rang k). Dann gibt es ein offenes $W \subseteq \mathbb{R}^{n-k}$, ein offenes $U \subseteq \mathbb{R}^n$ mit $z_0 \in U$, und eine stetig diff'bare injektive Abbildung $\psi : W \to \mathbb{R}^n$ sodass

(i) $f^{-1}(0) \cap U = \psi(W)$
(ii) Rang $J_\psi(x)$ maximal (d.h. gleich $n-k$) für alle $x \in W$
(iii) $\psi^{-1} : \psi(W) \to W$ stetig.

Die Abbildung ψ heisst *lokale Parametrisierung* der Lösungsmenge.

Beweis Da $J_f(z_0)$ maximalen Rang hat, existieren Koordinaten $z_{i_1} =: y_1, ..., z_{i_k} =: y_k$ sodass die Spalten $\frac{\partial f}{\partial y_1}(z_0), ..., \frac{\partial f}{\partial y_k}(z_0)$ linear unabhängig sind. Indem wir den IFS anwenden und ψ gemäss (1) wählen, folgen alle Behauptungen bis auf (ii) und (iii). Letzteres ist aber trivial, denn $\psi^{-1}(x, y) = x$, und ersteres folgt aus

$$J_\psi(x) = \begin{pmatrix} I \\ J_g(x) \end{pmatrix}$$

(mit $I = (n-k) \times (n-k)$ Identitätsmatrix), denn dies impliziert, dass die ersten $n-k$ Zeilen von $J_\psi(x)$ linear unabhängig sind.

Mengen, die sich wie in Korollar 4.1 parametrisieren lassen, haben einen besonderen Namen (und spielen in Geometrie und Physik eine wichtige Rolle).

Def. 4.2 (Untermannigfaltigkeiten des $\mathbb{R}^n$) Sei $\ell \leq n$, $\alpha \in \mathbb{N}$. Nichtleere Teilmengen $M \subseteq \mathbb{R}^n$, für die zu *jedem* Punkt $z_0 \in M$ (wie in obigem Korollar für $M = f^{-1}(0)$ und $\ell = n - k$) ein offenes $W \subseteq \mathbb{R}^\ell$, ein offenes $U \subseteq \mathbb{R}^n$ mit $z_0 \in U$, und eine α-mal stetig diff'bare injektive Abbildung $\psi : W \to \mathbb{R}^n$ mit den Eigenschaften (i) $M \cap U = \psi(W)$, (ii) Rang $J_\psi(x)$ maximal(d.h. gleich ℓ) für alle $x \in W$, (iii) $\psi^{-1} : \psi(W) \to W$ stetig existieren, heissen **ℓ-dimensionale $\mathcal{C}^\alpha$-Untermannigfaltigkeit** des $\mathbb{R}^n$.

Informelle Zusammenfassung der Definition: ℓ-dimensionale Untermannigfaltigkeiten sehen lokal in der Nähe jedes Punktes wie ein verbogenes Stück des $\mathbb{R}^\ell$ aus.

Was ist der Grund für die Forderung (iii) in Def. 4.2? Lässt man sie fallen, kann man z.B. ein Stück des $\mathbb{R}$ so "verbiegen", dass das Ende fast die Mitte berührt, aber trotzdem (i) und (ii) gelten. Dieser technische Punkt ist in Beispiel 4) erklärt. Ob eine Menge eine Untermannigfaltigkeit des $\mathbb{R}^n$ ist, hat also nicht nur damit zu tun, wie "kleine Teile" (d.h. Bilder kleiner Teilmengen der Parametergebiete W unter den Abbildungen ψ) aussehen, sondern auch damit, wie sie global im $\mathbb{R}^n$ drinliegt.

Mit dieser Terminologie folgt sofort aus Korollar 4.1:

Korollar 4.2 Ist $\Omega \subset \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}^k$ stetig differenzierbar, $f^{-1}(0)$ nicht leer, und $\operatorname{Rang} J_f(z) = k$ für alle $z \in f^{-1}(0)$, so ist $f^{-1}(0)$ $(n-k)$-dimensionale $\mathcal{C}^1$-Untermannigfaltigkeit des $\mathbb{R}^n$.

Untermannigfaltigkeiten können also lokal auf zwei verschiedene Weisen beschrieben werden: durch eine Parametrisierung ψ mit Hilfe von $n-k$ freien Parametern gemäss Def. 4.2 ("primale Beschreibung"); oder als Lösungsmenge von k Gleichungen wie in Korollar 4.2 ("duale Beschreibung").

Beispiele 1) *Ebenen im dreidimensionalen Raum* sind 2-dimensionale $\mathcal{C}^1$-Untermannigfaltigkeiten. Aus der Schule kennen wir zwei alternative Beschreibungen einer Ebene:

(a) Punkt-Richtungs-Form: $E = \{x_0 + tv + sw : t, s \in \mathbb{R}\}$. Hierbei ist x_0 ein fester Punkt in der Ebene und v, w sind linear unabhängige Richtungsvektoren.

(b) Hesse'sche Normalform: $E = \{x \in \mathbb{R}^3 : n \cdot x = a\}$. Hierbei ist $n \in \mathbb{R}^3$ ($|n| = 1$) der Normalenvektor und $a \geq 0$ der Abstand vom Ursprung.

Hier haben wir die übliche Notation $n \cdot x = \sum_{i=1}^3 n_i x_i$ benutzt. (a) ist die "primale" Beschreibung der Ebene (durch eine Parametrisierung) gemäss Def. 4.2; (b) ist die "duale" Beschreibung (als Lösungsmenge einer Gleichung) gemäss Korollar 4.2. Setzt man nämlich

$$\psi(s,t) := x_0 + sv + tw, \quad \psi : \mathbb{R}^2 \to \mathbb{R}^3,$$

dann folgt $E = \psi(\mathbb{R}^2)$, ψ injektiv, und $\operatorname{Rang} J_\psi(s,t) = 2$ für alle s, t, denn

$$J_\psi(s,t) = \begin{pmatrix} | & | \\ v & w \\ | & | \end{pmatrix}$$

und v, w sind linear unabhängig; alternativ dazu führt man die Funktion

$$f(x) := n \cdot x - a, \quad f : \mathbb{R}^3 \to \mathbb{R}$$

ein, dann folgt $E = f^{-1}(0)$ und $\operatorname{Rang} J_f(x) = 1$ für alle x, denn $J_f(x) = (n_1, n_2, n_3) \neq 0$ wegen $|n| = 1$.

2) Die *Sphäre* $S^{n-1} := \{z \in \mathbb{R}^n : |z| = 1\}$ ist nach Korollar 4.1 eine $(n-1)$-dimensionale $\mathcal{C}^1$-Untermannigfaltigkeit des $\mathbb{R}^n$, denn für $\Omega = \mathbb{R}^n$ und $f(z) = |z|^2 - 1$ gilt $S^{n-1} = f^{-1}(0)$ und $J_f(z) = (2z_1, ..., 2z_n) \neq 0 \ \forall z \in S^{n-1}$, und somit $\operatorname{Rang} J_f(z) = 1 \ \forall z \in S^{n-1}$. Eine explizite Parametrisierung der Kreislinie $S^1 = \{z \in \mathbb{R}^2 : |z| = 1\}$ gemäss Definition 4.2 ist z.B. durch die beiden überlappenden "Dreiviertelkreise"

$$\psi^{(1)} : (\tfrac{\pi}{4}, \tfrac{7\pi}{4}) \to \mathbb{R}^2, \ \ \psi^{(1)}(x) = (\cos x, \sin x), \ \ U^{(1)} = \{z \in \mathbb{R}^2 : z_1 < \tfrac{1}{\sqrt{2}}\},$$
$$\psi^{(2)} : (\tfrac{5\pi}{4}, \tfrac{11\pi}{4}) \to \mathbb{R}^2, \ \ \psi^{(2)}(x) = (\cos x, \sin x), \ \ U^{(2)} = \{z \in \mathbb{R}^2 : z_1 > -\tfrac{1}{\sqrt{2}}\}$$

gegeben.

3) Wir behaupten: die *Gruppe der orthogonalen Matrizen* (oder kurz *orthogonale Gruppe*)

$$O(n) := \{X \in \mathbb{R}^{n\times n} : X^T X = I\}$$

ist $\frac{n(n-1)}{2}$-dimensionale $\mathcal{C}^1$-Untermannigfaltigkeit des $\mathbb{R}^{n\times n}$. Zum Beweis betrachten wir die Abbildung

$$f(X) = X^T X - I, \quad f : \underbrace{\mathbb{R}^{n\times n}}_{\cong \mathbb{R}^{n^2}} \to \underbrace{\mathbb{R}^{n\times n}_{sym}}_{\cong \mathbb{R}^{n(n+1)/2}},$$

wobei $\mathbb{R}^{n\times n}_{sym} = \{X \in \mathbb{R}^{n\times n} : X = X^T\}$ den Vektorraum der symmetrischen $n \times n$ Matrizen bezeichnet. Da symmetrische Matrizen durch ihren oberen Dreiecksteil eindeutig bestimmt sind, haben sie $n+(n-1)+(n-2)+\ldots+1$ unabhängige Komponenten und wir können $\mathbb{R}^{n\times n}_{sym}$ mit dem $\mathbb{R}^{n(n+1)/2}$ identifizieren. Wegen

$$\frac{n(n-1)}{2} = n^2 - \frac{n(n+1)}{2}$$

reicht es zu zeigen, dass Rang $J_f(X)$ maximal für alle $X \in O(n)$, dann folgt die Behauptung aus Korollar 4.2. Das versucht man besser nicht in Koordinaten, sondern rechnet abstrakt:

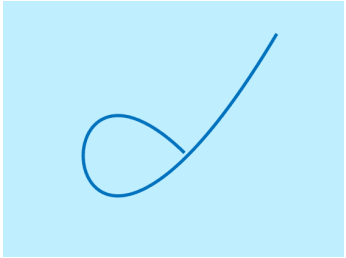
$$\begin{aligned} Df(X)(H) &= \lim_{t\to 0} \frac{1}{t} \underbrace{\Big(f(X+tH) - f(X)\Big)}_{=(X+tH)^T(X+tH)-X^TX = t(H^TX+X^TH)+t^2H^TH} \\ &= H^TX + X^TH. \end{aligned}$$

Da $Df(X)$ eine lineare Abbildung von einem höherdimensionalen in einen niedrigerdimensionalen Raum ist, ist ihr Rang maximal genau dann wenn sie surjektiv ist. $Df(X)$ ist aber surjektiv für jedes invertierbare X, denn zu gegebenem $B \in \mathbb{R}^{n\times n}_{sym}$ ist

$$H := \frac{1}{2}(X^{-1})^T B$$

Lösung der Gleichung $Df(X)(H) = B$. (Nachrechnen: $Df(X)(H) = \frac{1}{2}BX^{-1}X + X^T\frac{1}{2}(X^T)^{-1}B = B$.)

4) Die *"sich fast selbst berührende Schlaufe"*



ist keine Untermannigfaltigkeit des $\mathbb{R}^2$. Setze $M = \gamma((-1,\infty)) \subseteq \mathbb{R}^2$, $\gamma(t) = (t^2 - 1, t^3 - t)$. Dann erfüllt M zwar die Bedingungen (i), (ii) aus Def. 4.2 [solche Mengen gehören übrigens zur Klasse der sogenannten "Mannigfaltigkeiten"; diesen Begriff erklären wir hier nicht], denn $\gamma'(t) \neq 0$ für alle t, also können wir (für beliebiges $z_0 \in M$) $W = (-1,\infty)$, $\psi = \gamma$, $U = \mathbb{R}^2$ wählen. Aber es existieren keine W, ψ, U mit der Zusatzeigenschaft (iii), insbesondere ist $\gamma^{-1} : M \to (-1,\infty)$ unstetig, denn für $t_j \to -1$, $t_* = 1$ gilt $\gamma(t_j) \to \gamma(t_*)$ aber $t_j \not\to t_*$.

4.4 Lagrange'sche Multiplikatorregel

Wir wenden uns nun einer Frage zu, die auf den ersten Blick wenig mit dem impliziten Funktionensatz zu tun hat, zu deren Beantwortung wir den Satz aber heranziehen werden:

Wie findet man Maximums- und Minimumsstellen unter Nebenbedingungen?

Als Vorüberlegung betrachten wir folgende Situation: f, $g : \mathbb{R}^2 \to \mathbb{R}$ seien gegebene Funktionen, gesucht sind die Maximums- und Minimumsstellen von f unter der Nebenbedingung $g(x,y) = c$. (Deren Existenz wird unter sehr allgemeinen Voraussetzungen durch Satz 1.5 und das zugehörige Beispiel 2) sichergestellt.)

Anschaulich-geometrische (nicht mathematisch rigorose aber instruktive) Lösung: Die Menge $\{(x,y) \in \mathbb{R}^2 : g(x,y) = c\}$ ist Niveaulinie von g. Wir skizzieren diese Menge sowie – im selben Bild – die Niveaulinien von f. Sei (x,y) ein Punkt auf der Niveaulinie von g. Der Gradient von f zeigt (nach Satz 2.3) in Richtung des stärksten Anstiegs von f. Falls die Projektion des Gradienten auf die Niveaulinie von g nicht Null ist, können wir durch Entlanglaufen auf der Linie $g(x,y) = c$ den Wert von f erhöhen/erniedrigen. Also muss an einer Maximums- oder Minimumsstelle der Gradient von f senkrecht auf der Niveaulinie von g stehen. Dies ist aber (wegen der grundlegenden Tatsache, dass die Niveaulinie von g senkrecht zum Gradienten von g steht, siehe Satz 2.4) äquivalent dazu, dass ∇f in dieselbe Richtung wie ∇g zeigt. D.h. an einer Maximums- oder Minimumsstelle gilt $\nabla f(x,y) = \lambda \nabla g(x,y)$ für ein $\lambda \in \mathbb{R}$.

Derartige anschaulich-geometrische Betrachtungen sind immer lehrreich, aber nicht immer korrekt. Beispiel zweier einfacher Funktionen f und g, für die die obige notwendige Bedingung für Extremstellen trotz aller Plausibilität leider falsch ist: siehe Beispiel 2) unten. Was ist hier schiefgegangen? Wir haben uns bei obiger Argumentation stillschweigend und unberechtigterweise die Niveaumenge $g^{-1}(c)$ als schöne glatte Kurve vorgestellt. Dies dürfen wir – laut implizitem Funktionensatz – immerhin dann, wenn der Gradient von g an der Extremstelle ungleich Null ist. In der Tat lässt sich, für beliebige Raumdimension n, beweisen:

Satz 4.3 (Lagrange'sche Multiplikatorregel) *Sei $\Omega \subseteq \mathbb{R}^n$ offen, f, $g : \Omega \to \mathbb{R}$ stetig diff'bar. Sei $c \in \mathbb{R}$, sei $z_0 \in \Omega$ lokale Maximums- oder Minimumsstelle von f*

auf der Niveaumenge $\{z \in \Omega : g(z) = c\}$*, und sei* $\nabla g(z_0) \neq 0$*. Dann existiert ein* $\lambda \in \mathbb{R}$ *sodass*

$$\nabla f(z_0) = \lambda \nabla g(z_0).$$

Die Zahl λ heisst *Lagrange'scher Multiplikator*; daher auch der Name der Regel.

Der Vorteil der Regel besteht darin, dass man die $(n-1)$-dimensionale Lösungsmenge der Gleichung $g(x) = c$ nicht zu bestimmen braucht; stattdessen benötigt man nur die – in der Praxis typischerweise viel einfacher zu bestimmende – Ableitung von g.

Beispiel 1) Maximiere $f(x, y) = x$ bezüglich der Nebenbedingung $g(x, y) = x^2 + xy + y^2 = 1$.

Bemerkung (nicht zum Bestimmen der Maximumsstelle benötigt): Die Lösungsmenge der Nebenbedingung, d.h. die Niveaumenge $g^{-1}(1)$, ist eine Ellipse mit grosser Halbachse der Länge $\sqrt{2}$ in Richtung v_2 und kleiner Halbachse der Länge $\sqrt{2/3}$ in Richtung v_1, wobei

$$v_1 = \tfrac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad v_2 = \tfrac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

(Skizze.) Dies lässt sich mittels linearer Algebra herleiten. Die Extremwertaufgabe bedeutet geometrisch: wir suchen den Punkt auf dieser Ellipse, der am weitesten rechts liegt.

Wir berechnen

$$\nabla f(x, y) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \nabla g(x, y) = \begin{pmatrix} 2x + y \\ x + 2y \end{pmatrix}.$$

Falls (x, y) Maximumsstelle, gilt wegen Satz 4.3: $\nabla f(x, y) = \lambda \nabla g(x, y)$ für ein $\lambda \in \mathbb{R}$. Die erste Komponente dieser Gleichung liefert $1 = \lambda(2x + y)$ und insbesondere $\lambda \neq 0$; die zweite Komponente liefert $0 = \lambda(x + 2y)$ und somit (da $\lambda \neq 0$) $x + 2y = 0$, d.h. $y = -x/2$. Einsetzen in die Nebenbedingung ergibt $1 = g(x, -x/2) = x^2 - x^2/2 + x^2/4 = \frac{3}{4}x^2$, d.h. $x = \pm\frac{2}{\sqrt{3}}$. Insgesamt folgt also

$$\left\{(x, y) \in g^{-1}(1) : \nabla f(x, y) = \lambda \nabla g(x, y) \text{ für ein } \lambda \in \mathbb{R}\right\} = \left\{(\tfrac{2}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}), (-\tfrac{2}{\sqrt{3}}, \tfrac{1}{\sqrt{3}})\right\}.$$

Durch Einsetzen in f und Wertevergleich entscheiden wir, welcher Punkt die Maximums- und welcher die Minimumsstelle ist. In unserem Beispiel können wir die Werte von f sofort ablesen. Der erste Punkt ist die Maximumsstelle; der maximale Wert von f ist gleich $\frac{2}{\sqrt{3}}$; und der genau gegenüberliegende zweite Punkt ist die Minimumsstelle.

Das folgende Gegenbeispiel zeigt, dass die Bedingung $\nabla g(z_0) \neq 0$ in der Lagrange'schen Multiplikatorregel nicht weggelassen werden kann.

Beispiel 2) Minimiere $f(x, y) = y$ bezüglich der Nebenbedingung $g(x, y) = y^3 - x^2 =$

0, f, $g : \mathbb{R}^2 \to \mathbb{R}$. (Gesucht ist also der Punkt auf der Lösungsmenge der Nebenbedingung, der am weitesten unten liegt.) Da $x^2 \geq 0$, ist die Nebenbedingung $y^3 = x^2$ eindeutig nach y auflösbar, sie ist äquivalent zu $y = \sqrt[3]{x^2} = x^{2/3}$. (Graph.) Insbesondere hat die Lösungsmenge der Nebenbedingung eine scharfe nach unten zeigende "Spitze" bei $(x, y) = (0, 0)$. Nun zu unserem Minimierungsproblem. Offensichtlich ist $y > 0$ es sei denn $x = 0$, also $(0, 0)$ einzige Minimumsstelle. Es gilt aber (für beliebiges $\lambda \in \mathbb{R}$)

$$\nabla f(0,0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \neq \lambda \nabla g(0,0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Was geht hier schief? An der Minimumsstelle steht die Niveaulinie von f *senkrecht* auf derjenigen von g, statt parallel zu sein! Dies hat mit der Nichtglattheit der Niveaulinie von g zu tun. Wäre sie der Graph einer glatten (d.h. differenzierbaren) Funktion, müsste der Graph natürlich (wie schon aus Analysis 1 bekannt) an der Minimumsstelle eine waagerechte Tangente haben und somit parallel zur Niveaulinie von f liegen. Salopp gesprochen "versteckt" die Funktion g die waagerechte Tangente in ihrer Singularität.

Beweis (Lagrange'sche Multiplikatorregel).

Idee: Nach Voraussetzung ist $\nabla g(z_0) \neq 0$. Finde eine Komponente des Gradienten, die ungleich Null ist. Löse die Gleichung $g(z) = c$ in der Nähe von z_0 nach der entsprechenden Komponente von z auf (dies ist möglich wegen IFS). Damit können wir unsere Extremwertaufgabe mit Nebenbedingung in eine Extremwertaufgabe ohne Nebenbedingungen für die verbleibenden $n-1$ Komponenen von z umschreiben. Diese bearbeiten wir mithilfe der Theorie für solche Extremwertaufgaben aus Abschnitt 2.7.

Details: Nach Voraussetzung ist $\nabla g(z_0) \neq 0$, o.B.d.A. sei $\partial g/\partial z_n(z_0) \neq 0$. Teile die Variable z auf in $z = (x, y)$, $x \in \mathbb{R}^{n-1}$, $y \in \mathbb{R}$; entsprechend schreiben wir $z_0 = (x_0, y_0)$. Nach IFS können wir die Nebenbedingung nach y auflösen, genauer: es existieren offene Mengen $U \ni z_0$ und $W \ni x_0$, $W \subseteq \mathbb{R}^{n-1}$, und eine stetig diff'bare Funktion $h : W \to \mathbb{R}$ sodass $g^{-1}(c) \cap U = \operatorname{graph} h$. Folglich ist $x_0 \in W$ lokale Extremstelle der Verkettung

$$\alpha(x) = f(x, h(x)), \quad \alpha : W \subseteq \mathbb{R}^{n-1} \to \mathbb{R}.$$

D.h. wir haben unsere Extremwertaufgabe mit Nebenbedingung in eine entsprechende Aufgabe ohne Nebenbedingung umgewandelt. Die Optimalitätsbedingung erster Ordnung lautet (siehe Satz 2.7) $0 = J_\alpha(x_0)$. Um die Jacobimatrix von α auszurechnen, schreiben wir α als Verkettung,

$$\alpha(x) = f(\Phi(x)), \quad \Phi(x) = \begin{pmatrix} x \\ h(x) \end{pmatrix}, \quad \Phi : W \subseteq \mathbb{R}^{n-1} \to \mathbb{R}^n.$$

Nach Kettenregel gilt

$$\begin{aligned} J_\alpha(x_0) = J_f(\Phi(x_0))\, J_\Phi(x_0) &= \underbrace{\begin{pmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_{n-1}} & \frac{\partial f}{\partial y} \end{pmatrix}}_{1\times n \text{ Matrix}}(x_0, y_0) \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & 1 \\ \frac{\partial h}{\partial x_1}(x_0) & \frac{\partial h}{\partial x_2}(x_0) & \dots & \frac{\partial h}{\partial x_{n-1}}(x_0) \end{pmatrix}}_{n\times(n-1) \text{ Matrix}} \\ &= \underbrace{\begin{pmatrix} \frac{\partial f}{\partial x_1}(x_0, y_0) + \frac{\partial f}{\partial y}(x_0, y_0)\frac{\partial h}{\partial x_1}(x_0) & \cdots & \frac{\partial f}{\partial x_{n-1}}(x_0, y_0) + \frac{\partial f}{\partial y}(x_0, y_0)\frac{\partial h}{\partial x_{n-1}}(x_0) \end{pmatrix}}_{1\times(n-1) \text{ Matrix}} \end{aligned}$$

und folglich (mit der Notation $\nabla_x f$ für den Vektor der partiellen Ableitungen von f nach den Komponenten von x)

$$\nabla_x f(x_0, y_0) = -\frac{\partial f}{\partial y}(x_0, y_0)\nabla_x h(x_0). \tag{1}$$

Die Ableitung von h haben wir als Zusatz zum Beweis des IFS ausgerechnet, sie lautet im Allgemeinen

$$J_h(x) = -\big(\frac{\partial g}{\partial y}(x, h(x))\big)^{-1}\frac{\partial g}{\partial x}(x, h(x))$$

und folglich im Fall unserer skalaren Funktion h an der Stelle x_0

$$\nabla_x h(x_0) = -\big(\frac{\partial g}{\partial y}(x_0, y_0)\big)^{-1}\nabla_x g(x_0, y_0). \tag{2}$$

Einsetzen von (2) in (1) liefert

$$\nabla_x f(x_0, y_0) = \lambda\,\nabla_x g(x_0, y_0) \quad \text{mit } \lambda = \frac{\partial f}{\partial y}(x_0, y_0)\big(\frac{\partial g}{\partial y}(x_0, y_0)\big)^{-1}.$$

Dies ist fast die gewünschte Gleichung, aber noch nicht ganz, denn hier steht nur der Teil der Gradienten, der die Ableitungen nach den ersten $n-1$ Komponenten von z enthält. Für die letzte Komponente der Gradienten gilt aber trivialerweise wegen der obigen Formel für λ

$$\frac{\partial f}{\partial y}(x_0, y_0) = \lambda\frac{\partial g}{\partial y}(x_0, y_0). \tag{3}$$

Insgesamt folgt die Behauptung $\nabla f(z_0) = \lambda\nabla g(z_0)$.

Beispiel 3) Sei A reelle symmetrische $n \times n$ Matrix. Wir untersuchen folgende Extemwertaufgabe mit Nebenbedingung: Maximiere/Minimiere

$$f(x) = \langle x, Ax\rangle\Big(= \sum_{i=1}^{n} x_i(Ax)_i = \sum_{i,j=1}^{n} x_i A_{ij} x_j\Big), \quad f : \mathbb{R}^n \to \mathbb{R},$$

bezüglich der Nebenbedingung

$$g(x) = \langle x, x\rangle\Big(= |x|^2 = \sum_{i=1}^{n} x_i^2\Big) = 1.$$

Die Menge $g^{-1}(1)$ ist die Einheitssphäre im $\mathbb{R}^n$ und insbesondere kompakt, und f ist quadratisch und insbesondere stetig, also existiert eine Maximumsstelle/Minimumsstelle $x \in g^{-1}(1)$. Die Gradienten von f und g berechnen sich zu $\nabla f(x) = 2Ax$, $\nabla g(x) = 2x$. Nach Multiplikatorregel (Satz 4.3) existiert ein Multiplikator $\lambda \in \mathbb{R}$ sodass

$$\nabla f(x) = \lambda\nabla g(x) \Longleftrightarrow Ax = \lambda x.$$

Da $x \neq 0$, ist x also *Eigenvektor* von A, und λ zugehöriger Eigenwert! Des weiteren gilt: indem wir die Gleichung $Ax = \lambda x$ (für einen beliebigen Eigenwert und zugehörigen normierten Eigenvektor) mit x skalarmultiplizieren, ergibt sich $\langle x, Ax\rangle = \lambda$. Insgesamt folgt:

Korollar 4.3 (Rayleigh-Ritz'sches Variationsprinzip) Sei A eine reelle symmetrische $n \times n$ Matrix. Der maximale bzw. minimale Wert der quadratischen Funktion $x \mapsto \langle x, Ax \rangle$ auf der Einheitssphäre ist der grösste/kleinste Eigenwert von A. Maximums- bzw. Minimumsstellen sind genau die zugehörigen normierten Eigenvektoren.

Insbesondere folgt das Definitheitskriterium aus Lemma 2.5: eine symmetrische Matrix ist genau dann positiv/negativ definit, wenn alle Eigenwerte positiv/negativ sind. Als weitere Folgerung aus dem Rayleigh-Ritz'schen Variationsprinzip erhalten wir ein wichtiges Resultat der Linearen Algebra:

Korollar 4.4 (Spektralsatz für symmetrische Matrizen) Sei A eine reelle symmetrische $n \times n$ Matrix. Dann existiert eine Orthonormalbasis des $\mathbb{R}^n$ bestehend aus Eigenvektoren von A.

Um dies einzusehen, benötigen wir

Lemma 4.3 Sei $V \subset \mathbb{R}^n$ invarianter Unterraum von A, d.h. V Unterraum und $v \in V \Longrightarrow Av \in V$. Dann ist das orthogonale Komplement von V,

$$V^\perp := \{x \in \mathbb{R}^n : \langle v, x \rangle = 0 \; \forall v \in V\},$$

ebenfalls invarianter Unterraum von A.

Beweis: Sei $x \in V^\perp$. Zu zeigen: $Ax \in V^\perp$. Berechne für beliebiges $v \in V$

$$\langle v, Ax \rangle \underset{A \text{ symmetrisch}}{=} \langle Av, x \rangle \underset{Av \in V}{=} 0.$$

Beweis von Korollar 4.4: Wir müssen orthonormale Vektoren $v_1, ..., v_n$ finden sodass $Av_i = \lambda_i v_i$ (für geeignete $\lambda_i \in \mathbb{R}$), $i = 1, ..., n$. Dies tun wir mittels Induktion und Lemma 4.3: haben wir $v_1, ..., v_i$, $i < n$, bereits gefunden, setzen wir $V_i := \text{Span}\{v_1, ..., v_i\}$, dann ist V_i invarianter Unterraum von A, also wegen Lemma 4.3 auch $V_i^\perp$. Indem wir in $V_i^\perp$ irgendeine Orthonormalbasis wählen und $A|_{V_i^\perp}$ mit der entsprechenden Matrixdarstellung A_i identifizieren, folgt A_i symmetrisch, und Korollar 4.3 liefert einen neuen normierten Eigenvektor v_{i+1}, der zu $v_1, ..., v_i$ orthogonal ist.

Schreiben wir die n Eigenwertgleichungen $Av_i = \lambda_i v_i$ in der Form

$$A \underbrace{\begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix}}_{=:S} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix}$$

und benutzen die Tatsache, dass wegen Orthonormalität der v_i die Matrix S die Beziehung $S^T S = I$ erfüllt (d.h. eine orthogonale Matrix ist), erhalten wir durch

Multiplikation von links mit S^T die folgende alternative Formulierung von Korollar 4.4: für jede reelle symmetrische $n \times n$ Matrix A existiert eine orthogonale Matrix S und eine reelle Diagonalmatrix Λ sodass

$$S^T A\, S = \Lambda.$$

5 Systeme gewöhnlicher Differentialgleichungen

Eine Vielzahl wichtiger Prozesse in Physik, Biologie, Chemie, Ingenieurwissenschaften können mithilfe von *Systemen gewöhnlicher Differentialgleichungen* modelliert und analysiert werden. Umgekehrt zeigen die Lösungen solcher Systeme einen grossen Reichtum an Verhaltensweisen oder "Effekten", die das Verhalten des zugrundeliegenden physikalischen / biologischen / chemischen / technischen Systems wiederspiegeln.

In Analysis 1 Abschnitt 12 hatten wir bereits einige gewöhnliche Differentialgleichungen für eine einzelne Funktion $y : I \subseteq \mathbb{R} \to \mathbb{R}$ kennengelernt und mithilfe der Separationsmethode explizit gelöst, z.B. die Gleichungen

$$y' = ay \qquad \text{bzw.} \qquad y' = ay - by^2 \quad (a, b > 0)$$

für exponentielles bzw. logistisches Wachstum einer Population. Die Lösung ist die Exponentialfunktion bzw. eine sich der Konstanten a/b annähernde Funktion,

$$y(t) = y(0)e^{at} \qquad \text{bzw.} \qquad y(t) = \frac{a}{b + (\frac{a}{y(0)} - b)e^{-at}}.$$

In diesem Kapitel betrachten wir nun allgemeiner Systeme mehrerer miteinander gekoppelter gewöhnlicher Differentialgleichungen. Für Systeme ist explizite Lösbarkeit untypisch. Auf den ersten Blick mag das enttäuschen. Auf den zweiten Blick sollte es begeistern: Differentialgleichungen beschreiben "neue" Funktionen und Effekte. Wir

- beginnen mit einer ausführlichen Einleitung, in der wir typische Fragestellungen und die moderne Herangehensweise an solche Systeme anhand von drei sorgfältig ausgewählten Beispielen besprechen
- gehen als nächstes der – in Analysis 1 selbst im Fall einer einzigen Gleichung offengebliebenen – grundlegenden Frage nach, wie man unter möglichst allgemeinen Voraussetzungen Existenz und Eindeutigkeit von Lösungen nachweisen kann
- zeigen, wie man lineare Systeme explizit mithilfe einer Verallgemeinerung der Exponentialfunktion auf Matrizen lösen kann
- untersuchen das qualitative Verhalten von Lösungen linearer und nichtlinearer Systeme in der Nähe von Gleichgewichtspunkten.

5.1 Einleitung

Ein System gewöhnlicher Differentialgleichungen ist eine Gleichung der Form

$$y'(t) = f(y(t), t) \quad \text{für alle } t \in I,\ I \subseteq \mathbb{R} \text{ Intervall.} \tag{D}$$

Gegeben: Funktion

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} : \mathbb{R}^n \times I \to \mathbb{R}^n.$$

Gesucht: Funktion

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} : I \to \mathbb{R}^n.$$

Geometrisch gesehen ist y eine *Kurve im* $\mathbb{R}^n$, und f für jeden festen Zeitpunkt $t \in I$ ein *Vektorfeld im* $\mathbb{R}^n$ (man nennt Abbildungen f der obigen Form deshalb *zeitabhängige Vektorfelder*). Die Differentialgleichung (D) bedeutet geometrisch: der Tangentialvektor an die Kurve im Punkt $y(t)$ ist gleich dem zeitabhängigen Vektorfeld am Punkt $y(t) \in \mathbb{R}^n$ zur Zeit t, für alle t. Hängt f nicht von t ab, heisst das Differentialgleichungssystem *autonom*.

Eine typische Aufgabenstellung lautet: Löse die Differentialgleichung (D) zusammen mit einer Anfangsbedingung

$$y(t_0) = y_0 \tag{A}$$

mit gegebenem Anfangszeitpunkt $t_0 \in I$ und Anfangswert $y_0 \in \mathbb{R}^n$. Man interessiert sich zunächst dafür, ob (bzw. für welche Anfangsdaten und Zeitintervalle) überhaupt eine eindeutige Lösung existiert. Dann möchte man die qualitativen Eigenschaften des Systems verstehen. Z.B.: Was ist das Langzeitverhalten? Nähern sich Lösungen einer Gleichgewichtslösung, oszillieren sie periodisch, verhalten sie sich "chaotisch"? Wie ändert sich das Verhalten in Abhängigkeit von Anfangsbedingung oder Systemparametern (d.h. Parametern des Vektorfeldes f)?

Def. 5.1 Eine *Lösung von (D), (A) auf dem Intervall I* ist eine Funktion $y : I \to \mathbb{R}^n$, die differenzierbau auf I ist und die (D), (A) erfüllt.

Beispiel 1 (Räuber-Beute-Modell aus der Biologie)

$$\begin{aligned} H' &= aH - bH^2 - c\,H\,F, \\ F' &= -dF + e\,H\,F \end{aligned}$$

wobei a, b, c, d, e positive Konstanten, H, $F : [0, T) \to \mathbb{R}$. Hierbei ist $H(t)$ die Populationsgrösse der Beute zur Zeit t, und $F(t)$ diejenige der Räuber.

Biologische Erläuterung der Modellierung: Man kann sich z.B. Hasen und Füchse vorstellen. Der Term $H \cdot F$ modelliert die Wahrscheinlichkeit eines Räuber-Beute-Treffens, aber die Proportionalitätskonstanten, mit denen die Wahrscheinlichkeit solcher Treffen die langfristige zeitliche Änderung der Popoulationsgrössen beeinflusst, sind unterschiedlich, denn der Hase rennt um sein Leben, aber der Fuchs nur um sein Abendessen. In Abwesenheit von Räubern ($F = 0$) reduziert sich das Modell

auf die logistische Gleichung $H' = aH - bH^2$, die wir in Analysis 1 §12 hergeleitet und gelöst hatten; dann saturiert die Grösse der Hasenpopulation für $t \to \infty$ bei der Umweltkapazität a/b. In Abwesenheit von Beute ($H = 0$) reduziert sich das Modell auf $F' = -dF$; die Lösung ist dann (gemäss Analysis 1 §12) die abfallende Exponentialfunktion $F(t) = F(0)e^{-dt}$, d.h. die Füchse sterben aus.

Beispiel 2 (Bewegungsgleichungen für die Umlaufbahn von Planeten aus der Astronomie)

$$mx'' = -\alpha \frac{x}{|x|^3} \qquad (*)$$

wobei m, α positive Konstanten, $x : [0,T) \to \mathbb{R}^3$. Indem wir $p(t) := mx'(t) \in \mathbb{R}^3$ einführen, können wir das System in die Form (D) bringen:

$$(*) \Longleftrightarrow \begin{pmatrix} x' \\ p' \end{pmatrix} = \begin{pmatrix} \frac{1}{m}p \\ -\alpha \frac{x}{|x|^3} \end{pmatrix} \Longleftrightarrow \begin{pmatrix} x_1' \\ x_2' \\ x_3' \\ p_1' \\ p_2' \\ p_3' \end{pmatrix} = \begin{pmatrix} \frac{1}{m}p_1 \\ \frac{1}{m}p_2 \\ \frac{1}{m}p_3 \\ -\alpha x_1 (x_1^2 + x_2^2 + x_3^2)^{-3/2} \\ -\alpha x_2 (x_1^2 + x_2^2 + x_3^2)^{-3/2} \\ -\alpha x_3 (x_1^2 + x_2^2 + x_3^2)^{-3/2} \end{pmatrix}.$$

Hierbei ist $x(t) \in \mathbb{R}^3$ der Ort eines Planeten zur Zeit t und $x'(t)$ seine Geschwindigkeit.

Physikalische Erläuterung der Modellierung: Die Differentialgleichung (*) ist ein Spezialfall des 2. Newton'schen Gesetzes *Masse mal Beschleunigung = Kraft*. Der Parameter m ist die Masse des Planeten, $x''(t)$ ist die Beschleunigung zur Zeit t, und die rechte Seite von (*) ist die zur Zeit t wirkende Anziehungskraft der Sonne, deren Position hier zeitunabhängig am Nullpunkt angenommen ist. Die Hilfsgrösse p ist der Impuls (Impuls = Masse mal Geschwindigkeit). Gleichung (*) wurde 1667 von Isaac Newton – als erstes Differentialgleichungsmodell in der Geschichte der Naturwissenschaften – aufgestellt.

Beispiel 3 (Lorenz-Gleichung)

$$\begin{aligned} x' &= \sigma(y - x) \\ y' &= x(\rho - z) - y \\ z' &= -\beta z + xy \end{aligned}$$

wobei σ, ρ, β positive Konstanten und $x, y, z : [0,T) \to \mathbb{R}$.

Erläuterung der Modellierung: Dieses System wurde vom Mathematiker und Meteorologen Ed Lorenz als drastische Vereinfachung eines Systems nichtlinearer partieller Differentialgleichungen aus der Hydrodynamik aufgestellt, indem das volle System auf einen endlichdimensionalen Unterraum projiziert wurde. Die Lorenz-Gleichung ist ein Modellbeispiel dafür, dass unschuldig aussehende Systeme (mit

nur drei gekoppelten Differentialgleichungen und nur linearen und quadratischen Termen) kompliziertes Verhalten der Lösungen produzieren können. Siehe unten.

Elementare Beobachtungen (für alle drei Beispiele):

- Die Gleichungen sind jeweils gekoppelt. In Beispiel 2: x_1' benötigt p_1; p_1' benötigt x_1, x_2, x_3; x_2', x_3' benötigen p_2, p_3.
- Die Gleichungen sind nichtlinear.

Bevor wir uns die Beispiele genauer anschauen, führen wir noch etwas Terminologie ein.

Def. 5.2 Eine Lösung y von (D), die nicht von t abhängt, also zeitlich konstant ist, heisst *Gleichgewichtslösung* oder *stationäre Lösung*. Der entsprechende Punkt $\bar{y} \in \mathbb{R}^n$ sodass $y(t) \equiv \bar{y}$ für alle t heisst *Gleichgewichtspunkt* oder *stationärer Punkt*.

Im autononomen Fall sind die Gleichgewichtspunkte offensichtlich genau die Nullstellen von f.

Das Bild $\{y(t) : t \in I\} \subset \mathbb{R}^n$ einer Lösung von (D), (A) zu gegebenem Anfangswert y_0 heisst *Orbit* von y_0. Eine Zerlegung von $\mathbb{R}^n$ in Orbits, bzw. ein simultaner Plot einer repräsentativen Teilmenge von Orbits, heisst *Phasenporträt* von (D).

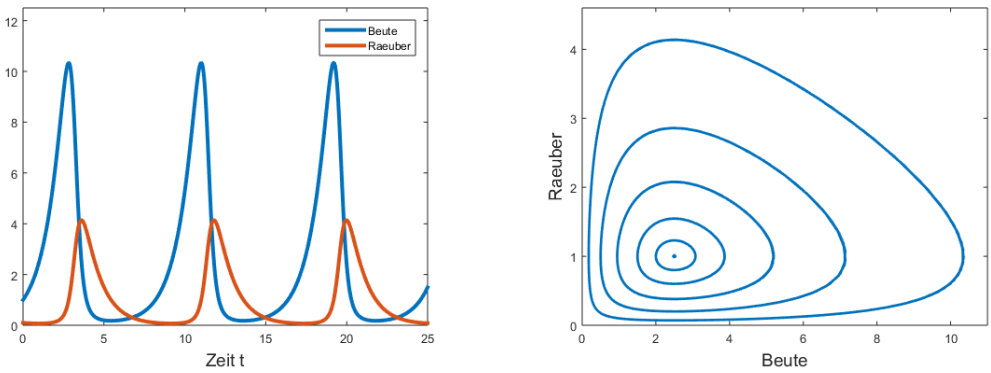
Beispiel 1 (Fortsetzung): Wir setzen der Einfachheit halber $b = 0$. Dann sind die stationären Punkte die Lösungen $(\bar{H}, \bar{F}) \in \mathbb{R}^2$ der beiden nichtlinearen Gleichungen

$$\begin{aligned} 0 &= a\bar{H} - c\bar{H}\,\bar{F}, \\ 0 &= -d\bar{F} + e\bar{H}\,\bar{F}, \end{aligned}$$

d.h.

$$(\bar{H}, \bar{F}) = (0,0) \text{ oder } (\tfrac{d}{e}, \tfrac{a}{c}).$$

Das erste Gleichgewicht ist biologisch trivial (wenn es anfangs weder Hasen noch Füchse gibt, bleibt das so), das zweite nicht (es gibt ein bestimmtes, aus den Modellparametern errechenbares, Gleichgewichtsverhältnis zwischen Räubern und Beutetieren). Wenn man vermutet, die Populationen würden sich für typische Anfangsdaten im Laufe der Zeit diesem Verhältnis annähern, liegt man falsch.



Links: Numerische Lösung des Räuber-Beute-Modells (Beispiel 1) für $a = c = d = 1$, $b = 0$, $e = 0.4$ mit Anfangsbedingung $H_0 = 1$, $F_0 = 0.1$.
Rechts: Phasenporträt des Modells.

Biologische Interpretation: Anfangs wächst die Hasenpopulation stark an. Dies ermöglicht eine zeitverzögerte starke Vermehrung der Füchse. Die Füchse dezimieren nun die Hasenpopulation so stark, dass sie am Ende selbst nicht mehr genug Beute finden und ihr Bestand stark zurückgeht. Danach wiederholt sich dieser Zyklus. Die von unserem einfachen Modell vorhergesagten zeitversetzen Oszillationen kann man in der Natur (näherungsweise) beobachten.

Beispiel 2 (Fortsetzung): Für dieses System gibt es keine Gleichgewichtspunkte, denn die rechte Seite der Gleichung für p' ist stets $\neq 0$. Aufgrund der speziellen Struktur des Systems gibt es aber viele Erhaltungsgrössen:

$$\begin{aligned} E &= \tfrac{1}{2m}|p|^2 - \tfrac{\alpha}{|x|} && \text{(Energie)} \\ \ell &= x \times p && \text{(Drehimpuls)} \\ \varepsilon &= \tfrac{x}{|x|} - \tfrac{1}{m\alpha} p \times \ell && \text{(Runge-Lenz-Vektor).} \end{aligned}$$

Nachrechnen, dass diese Grössen entlang Lösungen zeitlich konstant sind, ist eine exzellente Übung. Mithilfe dieser Erhaltungsgleichungen lässt sich zeigen: *Typische Orbits der Ortsvariablen x (d.h. $\{x(t) : t \in \mathbb{R}\}$) sind Ellipsen, deren einer Brennpunkt der Ursprung $x = 0$ ist.* Dieses Ergebnis geht auf Newton höchstpersönlich zurück, und erklärte die bereits vorher von Kepler anhand empirischer Daten gefundene Tatsache, das sich die Erde auf einer Ellipsenbahn bewegt, in deren einem Brennpunkt die Sonne steht ("erstes Kepler'sches Gesetz"). Da sich die Details in vielen Mechanik-Büchern finden (siehe z.B. V.I. Arnol'd, Mathematical methods in classical mechanics, Springer-Verlag) und wenig zum Verständnis allgemeiner Systeme beitragen (explizite Lösbarkeit ist untypisch), führen wir sie hier nicht aus. In der Tat ist das entsprechende System von Bewegungsgleichungen für drei oder mehr

Himmelskörper, deren Wechselwirkung durch die Gravitationskraft gegeben ist,

$$m_i x^{(i)\prime\prime} = -\sum_{j\in\{1,\dots,N\}\setminus\{i\}} G m_i m_j \frac{x^{(i)} - x^{(j)}}{|x^{(i)} - x^{(j)}|^3} \qquad (i = 1, \dots, N)$$

(mit m_i=Massen der Himmelskörper und G=Gravitationskonstante), nicht mehr explizit lösbar.

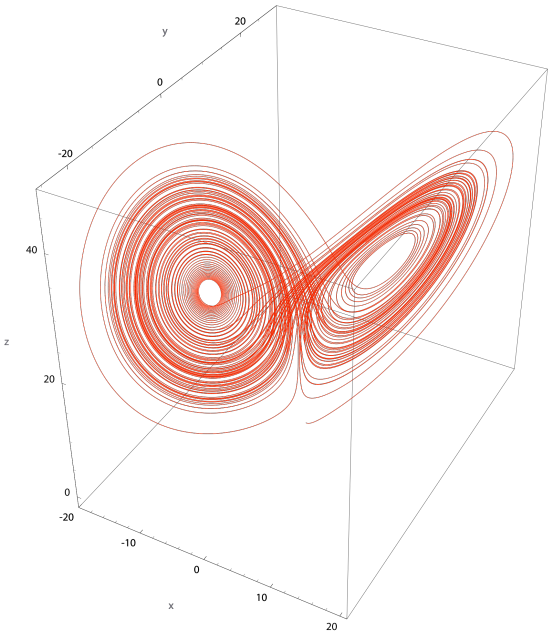
Beispiel 3 (Fortsetzung): Wir bestimmen zunächst die Gleichgewichtspunkte, d.h. die Lösungen des Gleichungssystems

$$\begin{aligned} y - x &= 0 \\ x\rho - xz - y &= 0 \\ -\beta z + xy &= 0. \end{aligned}$$

Elimininieren von y mithilfe der ersten Gleichung liefert $x(\rho - 1) = xz$, $x^2 = \beta z$. Zunächst gibt es die triviale Lösung $x = y = z = 0$. Falls eine (und deshalb beide) dieser Variablen $\neq 0$, ist z wegen der ersten Gleichung gleich $\rho - 1$ und zudem wegen der zweiten Gleichung > 0, also muss $\rho > 1$ gelten und in diesem Fall ist $x = \pm\sqrt{\beta(\rho-1)}$. Insgesamt folgt also: Für $\rho \leq 1$ ist $x = y = z = 0$ das einzige Gleichgewicht; für $\rho > 1$ gibt es drei Gleichgewichte,

$$\big(0,0,0\big), \quad \big(\pm\sqrt{\beta(\rho-1)}, \pm\sqrt{\beta(\rho-1)}, \rho - 1\big).$$

Wenn man vermutet, typische Lösungen würden gegen eines dieser Gleichgewichte konvergieren, liegt man falsch.



Ein numerisch berechneter Orbit der Lorenzgleichung (Beispiel 3) für $\sigma = 10$, $\rho = 28$, $\beta = 8/3$. Quelle: Wikipedia (detaillierte Quellen- und Lizenzangabe siehe Ende des Kapitels). Der Orbit startet nahe dem ersten Gleichgewichtspunkt und umkurvt anschliessend wieder und wieder die anderen beiden Gleichgewichtspunkte, ohne sich für einen der beiden (oder einen festen Umkurvungsabstand oder eine feste Umkurvungsreihenfolge) zu entscheiden.

Mathematische Interpretation: Die Bilder zeigen: die Lösungen unseres Systems verhalten sich extrem kompliziert. Explizite Lösungsformeln sind nicht zu erwarten, ebensowenig implizite Lösungsformeln im Sinne von Abschnitt 4.2 (d.h. Darstellungen der Orbits als Lösungsmengen expliziter Gleichungssysteme der Form $g(y) = 0$). Kleinste Änderungen der Anfangsdaten führen dazu, dass man sich zu irgendeinem bestimmten späteren Zeitpunkt "links" statt "rechts" aufhält oder umgekehrt. Und das nicht nur wenn man an einem speziellen stationären Punkt startet (auch bei der 1D Gleichung $y' = y$ wandern Anfangswerte $y_0 > 0$ nach rechts und $y_0 < 0$ nach links, aber alle $y_0 > 0$ und alle $y_0 < 0$ verhalten sich gleich), sondern egal wo man startet! Der Meteorologe und Mathematiker Ed Lorenz beobachtete dieses Phänomen zunächst in einem viel komplizierteren Wettermodell, dass er in den 1960er Jahren simulierte: kleinste Variationen von Anfangsdaten und Modellparametern riefen stark abweichende Wetterprognosen hervor ("Schmetterlingseffekt"). Anschliessend entdeckte er, dass dieser Effekt bereits in obigem, heutzutage nach ihm benannten, einfachen System von 3 Differentialgleichungen auftritt.

Solche Beispiele motivieren die Herangehensweise der modernen Mathematik an Differentialgleichungen. Als "Ersatz" für die historische Suche nach expliziten Lösungsformeln geht man wie folgt vor:
– Existenz- und Eindeutigkeitsbeweis
– Analyse des qualitativen Verhaltens (wenn möglich)
– Design konvergenter numerischer Verfahren.
Der allgemeine Existenz- und Eindeutigkeitsbeweis, den wir im nächsten Abschnitt führen, zeigt, dass die Lösungen der Lorenz-Gleichung wohldefinierte mathematische Objekte sind. Warnung: Aufgrund des "chaotischen" Verhaltens und der sensitiven Abhängigkeit von Anfangsdaten und Systemparametern ist aber beim Lorenz-System – wie beim echten Wetter – eine zuverlässige numerische Vorhersage nur für kurze, aber nicht lange Zeiten möglich.

Systematische Methoden zur Analyse des qualitativen Verhaltens werden im Gebiet der *Dynamischen Systeme* entwickelt (siehe die Vorlesung 'Introduction to Nonlinear Dynamics'); als prototypisches Beispielproblem untersuchen wir in Abschnitt 5.4 das qualitative Langzeitverhalten von Lösungen in der Nähe von Gleichgewichtspunkten.

Design und Analyse numerischer Verfahren mit (je nach Anwendung) geeigneter Genauigkeit und Effizienz sind Gegenstand der *numerischen Mathematik* (siehe die Vorlesung 'Numerics of Differential Equations').

5.2 Existenz- und Eindeutigkeitssatz

In diesem Abschnitt gehen wir der folgenden grundlegenden Frage nach: *Besitzt das System (D), (A) unter möglichst allgemeinen Voraussetzungen an f eine Lösung?*

Die Grundidee wurde bereits in der Einleitung von Abschnitt 3 angedeutet: wir suchen zunächst nur Funktionen $y^{(k)}$, die das System (D), (A) näherungsweise lösen, aber mit fortschreitendem k besser und besser.

Als Vorüberlegung stellen wir fest, dass das System (D), (A) äquivalent zur Integralgleichung

$$y(t) \;=\; y_0 \;+\; \int_{t_0}^{t} f(y(s), s)\, ds \quad \text{für alle } t \in I \tag{I}$$

ist. Der Clou bei dieser Umformulierung ist, dass die Integralgleichung bereits für Funktionen y *niedrigerer Regularität* Sinn macht - nämlich bloss stetige Funktionen. Genauer gilt:

Lemma 5.0 Sei $f : \mathbb{R}^n \times I \to \mathbb{R}^n$ stetig, $y : I \to \mathbb{R}^n$, $t_0 \in I$. Dann sind äquivalent:
(i) y ist differenzierbar auf I und löst (D), (A).
(ii) y ist stetig auf I und löst (I).

Beweis: Falls y (ii) erfüllt, ist wegen Stetigkeit von $t \mapsto f(y(t), t)$ und Hauptsatz die rechte Seite von (I) (und folglich auch die linke Seite) differenzierbar, und somit hat y automatisch die höhere in (i) vorausgesetzte Regularität. Für solche y folgt die Äquivalenz von (D), (A) und (I) sofort aus dem Hauptsatz.

Ein cleveres, erstaunlich einfaches Verfahren, dass die Integralgleichung (I) näherungsweise besser und besser löst, ist das

Picard'sches Iterationsverfahren: Definiere rekursiv eine Folge von Näherungslösungen $y^{(k)} : I \to \mathbb{R}^n$ durch

$$\begin{aligned} y^{(0)}(t) &:= y_0 \text{ für alle } t \\ y^{(k)}(t) &:= y_0 + \int_{t_0}^{t} f(y^{(k-1)}(s), s)\, ds \quad (k \geq 1). \end{aligned}$$

Hierbei benutzen wir

Def. 5.3 Das Integral einer stetigen vektorwertigen Funktion $z : [a, b] \to \mathbb{R}^n$ ist komponentenweise definiert:

$$\int_a^b \begin{pmatrix} z_1(s) \\ \vdots \\ z_n(s) \end{pmatrix} ds := \begin{pmatrix} \int_a^b z_1(s)\, ds \\ \vdots \\ \int_a^b z_n(s)\, ds \end{pmatrix}.$$

Beispiel zum Picard-Verfahren: $n = 1$, $f(y, t) = y$, $t_0 = 0$, d.h. wir lösen das 1D

Anfangswertproblem (D) $y' = y$, (A) $y(0) = y_0$. Die Rekursionsvorschrift liefert

$$\begin{aligned}
y^{(0)}(t) &= y_0 \text{ für alle } t,\\
y^{(1)}(t) &= y_0 + \int_0^t \underbrace{y^{(0)}(s)}_{=y_0}\, ds = y_0(1+t),\\
y^{(2)}(t) &= y_0 + \int_0^t \underbrace{y^{(1)}(s)}_{=y_0(1+s)}\, ds = y_0(1+t+\tfrac{t^2}{2}),\\
y^{(3)}(t) &= y_0 + \int_0^t y_0(1+s+\tfrac{s^2}{2})\, ds = y_0(1+t+\tfrac{t^2}{2}+\tfrac{t^3}{3!})
\end{aligned}$$

und folglich durch Iteration

$$y^{(n)}(t) = y_0\Big(1+t+\tfrac{t^2}{2}+\ldots+\tfrac{t^n}{n!}\Big).$$

Der Term in Klammern ist aber gerade die Partialsumme n-ter Ordnung der Taylorreihe der Exponentialfunktion (siehe Analysis 1)! Folglich $y^{(n)}(t) \to y_0 e^t$ für $n \to \infty$. Dieser Grenzwert ist genau die eindeutige Lösung von (D), (A) (siehe Analysis 1 Abschnitt 12). Dies ist kein Zufall, siehe der folgende Satz; der allgemeine Nachweis der Konvergenz des Verfahrens benötigt aber substantielle analytische Werkzeuge, nämlich den Banach'schen Fixpunktsatz und die Vollständigkeits des Raumes der stetigen beschränkten Funktionen mit Supremumsnorm (siehe Abschnitt 3). Wir erinnern an die Definition der gleichmässigen Konvergenz (Def. 3.5 in Abschnitt 3).

Satz 5.1 (Existenz- und Eindeutigkeitssatz (globale Version)) *Sei I ein beliebiges kompaktes Intervall, und sei t_0 ein beliebiger Punkt in I. Sei $f : \mathbb{R}^n \times I \to \mathbb{R}^n$ stetig, sowie Lipschitzstetig in der ersten Variablen (d.h. es existiert eine Konstante L sodass $|f(y,t) - f(z,t)| \le L|y-z|$ für alle $y, z \in \mathbb{R}^n$ und alle t). Dann gilt:*

a) Die Differentialgleichung (D), (A) besitzt eine eindeutige Lösung auf I.

b) Die Folge des Picard-Verfahrens konvergiert gleichmässig gegen diese Lösung.

Beweisidee Durch Betrachten von $\tilde{y}(t) := y(t_0 + t)$ und $\tilde{f}(y,t) = f(y, t_0 + t)$ können wir annehmen, dass $t_0 = 0$. Wir zeigen Konvergenz der Folge des Picard-Verfahrens, indem wir den Banach'schen Fixpunktsatz in einem geeigneten Banachraum anwenden. Die Kunst ist die Wahl des Banachraums. Wir nehmen:

$$X := \{z : I \to \mathbb{R}^n : z \text{ stetig und beschränkt}\},$$

versehen mit der Norm

$$\|z\| := \sup_{t \in I} e^{-b|t|}|z(t)|$$

mit $b > L$. Warum diese gewichtete Supremumsnorm eine gute Wahl ist, sehen Sie in Schritt 2 des Beweises.

Lemma 5.1 Für beliebiges $b \in \mathbb{R}$ ist der Raum X versehen mit der Norm $\|\cdot\|$ ein Banachraum.

Beweis Dies folgt aus der Äquivalenz der obigen Norm zur Supremumsnorm, z.B. ist für $b \geq 0$ und $T := \max\{|t| : t \in I\}$

$$\|z\| \leq \underbrace{\|z\|_\infty}_{=\sup_{t\in I}|z(t)|} \leq e^{bT}\|z\|,$$

und der Banachraumeigenschaft der stetigen beschränkten Funktionen mit Supremumsnorm (Satz 3.2).

Beweis von Satz 5.1: Betrachte die dem Picard-Verfahren zugrundeliegende Abbildung F, die eine Funktion $y : I \to \mathbb{R}^n$ derart auf eine andere Funktion $\tilde{y} = F(y) : I \to \mathbb{R}^n$ abbildet, sodass $y^{(k)} = F(y^{(k-1)})$:

$$F(y)(t) := y_0 + \int_0^t f(y(s), s)\, ds \quad (t \in I).$$

1. Zunächst behaupten wir: dieses F ist Selbstabbildung von X nach X. Die Stetigkeit der Funktion $F(y)$ ist offensichtlich; die Beschränktheit folgt aus der Tatsache, dass die Funktion $|F(y)|$ auf der kompakten Menge I ihr Maximum annimmt.

2. Man zeigt nun, dass F Kontraktion, genauer: λ-Lipschitz mit Konstante $\lambda = \frac{L}{b} < 1$. Der Beweis ist magisch, geniessen Sie ihn. Für $y, z \in X$ ist

$$\begin{aligned}
|F(y)(t) - F(z)(t)| &= \Big|(y_0 - y_0) + \int_0^t \Big(f(y(s), s) - f(z(s), s)\Big)\, ds\Big| \\
&\leq \Big|\int_0^t |f(y(s), s) - f(z(s), s)|\, ds\Big| \\
&\underset{f \text{ Lipschitz}}{\leq} \Big|\int_0^t \underbrace{L}_{=Le^{b|s|}e^{-b|s|}} |y(s) - z(s)|\, ds\Big| \\
&\leq L\Big|\int_0^t e^{b|s|}\|y - z\|\, ds\Big| \\
&= \frac{L}{b}(e^{b|t|} - 1)\|y - z\| \\
&\leq \frac{L}{b}e^{b|t|}\|y - z\| \qquad \Big|\cdot e^{-b|t|} \\
\Longrightarrow \quad e^{-b|t|}\,|F(y)(t) - F(z)(t)| &\leq \frac{L}{b}\|y - z\| \qquad \Big|\sup_t \\
\Longrightarrow \quad \|F(y) - F(z)\| &\leq \frac{L}{b}\|y - z\|.
\end{aligned}$$

(Die äusseren Betragsstriche in der 2. bis 4. Zeile sind überflüssig für $t \geq 0$ aber notwendig für $t < 0$, denn dann ist $\int_0^t |f| = -\int_t^0 |f| \leq 0$.)

3. Nach Banach'schem Fixpunktsatz konvergiert $y^{(k)}$ in X gegen eine Funktion $y \in X$, und letztere ist Fixpunkt, d.h. $F(y) = y$. Die Fixpunktgleichung ist aber gerade die Integralgleichung (I). Somit ist y wegen Lemma 5.0 differenzierbar auf I und löst (D), (A). Die Gleichmässigkeit der Konvergenz folgt aus der Konvergenz in X und $\|\cdot\|_\infty \leq e^{bT}\|\cdot\|$.

4. Es ist noch die Eindeutigkeit zu zeigen. Seien y, $\tilde{y}$ Lösungen von (I). Dann liegen y, $\tilde{y}$ in X und sind Fixpunkte der Abbildung F auf dem Raum X. Da F Kontraktion, gibt es laut Banach'schem Fixpunktsatz höchstens einen Fixpunkt, d.h. $y = \tilde{y}$.

Lokale Existenz und Eindeutigkeit. Die Forderung der globalen Lipschitzstetigkeit von f in Satz 5.1 ist für viele Anwendungen zu restriktiv, und schliesst z.B. quadratische rechte Seiten, etwa die 1D Differentialgleichung

$$y' = y^2$$

aus. Aus Analysis 1 Abschnitt 12 wissen wir, dass in diesem Beispiel das Anfangswertproblem (D), (A) nur "lokal", d.h. in einem vom Anfangswert abhängigen Zeitintervall, lösbar ist. Zur Erinnerung: die Lösung für $t_0 = 0$ und $y_0 > 0$ ist

$$y(t) = \frac{1}{\frac{1}{y_0} - t},$$

sie existiert nur auf dem Intervall $[0, 1/y_0)$ und geht für $t \to 1/y_0$ gegen ∞ ("Blow-Up"). Wir zeigen nun: eine solche lokale Existenzaussage gilt unter extrem allgemeinen, alle wichtigen Anwendungen abdeckenden, Voraussetzungen.

Satz 5.2 (Existenz- und Eindeutigkeitssatz (lokale Version)) *Sei I Intervall, und sei t_0 innerer Punkt von I. Sei $f : \mathbb{R}^n \times I \to \mathbb{R}^n$ stetig, sowie lokal Lipschitzstetig in der ersten Variablen (d.h. für jede kompakte Teilmenge $K \subset \mathbb{R}^n \times I$ existiert ein L sodass $|f(y,t) - f(z,t)| \leq L|y - z|$ für alle (y,t), $(z,t) \in K$). Dann existiert zu jedem Anfangswert $y_0 \in \mathbb{R}^n$ ein $\delta > 0$ sodass (D), (A) auf $[t_0 - \delta, t_0 + \delta]$ eine eindeutige Lösung besitzt.*

Beweisidee Wir können wieder annehmen, dass $t_0 = 0$. Wir gehen ähnlich vor wie beim Beweis von Satz 5.1, benutzen aber einen etwas anderen Funktionenraum. Wir zeigen Konvergenz der Folge des Picard-Verfahrens im Raum der stetigen Funktionen auf $[-\delta, \delta]$ mit (ungewichteter) Supremumsnorm, wobei Selbstabbildungs- und Kontraktionseigenschaft durch hinreichend kleine Wahl von δ erreicht werden.

Details: Wähle $R > 0$ und wähle $T_1 > 0$ hinreichend klein sodass $[-T_1, T_1] \subset I$. Definiere die Menge $\overline{B_R}(y_0) \times [-T_1, T_1]$ (wobei $\overline{B_R}(y_0)$ wie üblich die abgeschlossene Kugel $\{x \in \mathbb{R}^n : |x - y_0| \leq R\}$ bezeichnet). Wegen der lokalen Lipschitzstetigkeit von f existiert $L > 0$ sodass

$$|f(y,t) - f(z,t)| \leq L|y - z| \quad \text{für alle } y, z \in \overline{B_R}(y_0) \text{ und alle } t \in [-T_1, T_1]. \tag{*}$$

Wegen der Stetigkeit von f und des Satzes vom Maximum und Minimum existiert $C > 0$ sodass

$$|f(y,t)| \le C \quad \text{für alle } (y,t) \in \overline{B_R}(y_0) \times [-T_1, T_1]. \tag{**}$$

Für $\delta \le T_1$ definieren wir den Banachraum

$$X := \{z : [-\delta, \delta] \to \mathbb{R}^n : z \text{ stetig}\}$$

versehen mit der Supremumsnorm $\|z\| = \sup_{t\in[-\delta,\delta]} |z(t)|$, sowie die abgeschlossene Teilmenge

$$A := \{z \in X : |z(t) - y_0| \le R \text{ für alle } t \in [-\delta, \delta]\}.$$

Wir betrachten die aus Satz 5.1 bekannte Abbildung des Picard'schen Iterationsverfahrens, $F : A \to X$,

$$F(y)(t) := y_0 + \int_0^t f(y(s), s)\, ds \ (t \in [-\delta, \delta]).$$

Wir behaupten: Für

$$\delta = \min\{T_1, \frac{R}{C}, \frac{1}{2L}\}$$

ist F Selbstabbildung und Kontraktion auf A.

Selbstabbildung: Sei $y \in A$. Dann ist $F(y)$ offensichtlich stetig, und

$$|F(y)(t) - y_0| \le \Big| \int_0^t |f(y(s),s)|\, ds \Big| \underset{(**)}{\le} \delta C \underset{\text{Wahl von } \delta}{\le} R \quad \text{für alle } t \in [-\delta, \delta].$$

Kontraktion: Berechne für y, $z \in A$ und $t \in [-\delta, \delta]$

$$|F(y)(t) - F(z)(t)| \le \Big| \int_0^t \big|f(y(s),s) - f(z(s),s)\big| ds \Big| \underset{(*)}{\le} \Big| \int_0^t L \big|y(s) - z(s)\big| ds \Big| \le \delta L \|y - z\| \underset{\text{Wahl von } \delta}{\le} \frac{1}{2}\|y - z\|$$

und folglich, indem man das Supremum über $t \in [-\delta, \delta]$ nimmt,

$$\|F(y) - F(z)\| \le \frac{1}{2}\|y - z\|.$$

Nach Banach'schem Fixpunktsatz existiert ein eindeutiger Fixpunkt $y \in A$ von F. Nach Lemma 5.0 impliziert die Fixpunktgleichung $F(y) = y$, dass y Lösung von (D), (A) auf $[\delta, \delta]$.

Wir müssen noch die Eindeutigkeit zeigen. Sei $\tilde{y}$ eine beliebige Lösung von (D), (A) – bzw. der äquivalenten Integralgleichung (I) – auf $[-\delta, \delta]$. Wegen der Eindeutigkeit von Fixpunkten von F auf A reicht es zu zeigen: $\tilde{y} \in A$. Wir argumentieren indirekt. Angenommen $\tilde{y} \notin A$. Dann existiert wegen der Stetigkeit von $\tilde{y}$ und Zwischenwertsatz ein kleinstes $|t_1| < \delta$ sodass $|\tilde{y}(t_1) - y_0| = R$. Folglich

$$R = |\tilde{y}(t_1) - y_0| = \Big| \int_0^{t_1} F(\tilde{y}(s), s)\, ds \Big| \underset{|\tilde{y}(s)-y_0|\le R}{\le} \Big| \int_0^{t_1} C\, ds \Big| = C|t_1| < C\delta \le R,$$

Widerspruch. Damit ist die Eindeutigkeit bewiesen.

Wir beenden diesen Abschnitt mit einigen Bemerkungen.

1) Ohne viel Zusatzarbeit kann man folgendes zeigen: Ist $f : \mathbb{R}^n \times I \to \mathbb{R}^n$ stetig sowie lokal Lipschitzstetig in der ersten Variablen, so existiert zu jedem $y_0 \in \mathbb{R}^n$ ein *maximales* Intervall (a, b) mit $a = a(y_0) \in (-\infty, t_0) \cup \{-\infty\}$ und $b = b(y_0) \in (t_0, \infty) \cup \{+\infty\}$ sodass (D), (A) in (a, b) eindeutig lösbar ist, aber in keinem grösseren Intervall (a', b') eine Lösung besitzt. Siehe Übungen.

2) Eine einfache hinreichende Bedingung für lokale Lipschitzstetigkeit von f in der

ersten Variablen ist: f stetig partiell differenzierbar in der ersten Variablen. Beweis: Sei $K \subset \mathbb{R}^n \times I$ kompakt. Wähle eine kompakte Teilmenge $\overline{B_R}(0) \times I' =: K'$ von $\mathbb{R}^n \times I$, die K enthält. Setze $L = \max_{(z,t)\in K'} |\nabla_z f(z,t)|$. Dann gilt für alle (y,t), $(z,t) \in K'$:

$$|f(y,t) - f(z,t)| = \Big| \int_0^1 \underbrace{\tfrac{d}{ds} f(z + s(y-z), t)}_{= \left\langle \nabla_z f(z+s(y-z),t), y-z \right\rangle} \, ds \Big| \leq L|y-z|.$$

3) Lässt man die Voraussetzung der lokalen Lipschitzstetigkeit von f in Satz 5.2 weg und verlangt nur Stetigkeit von f, bleibt die lokale Existenzaussage erhalten ("Satz von Peano"); der Beweis liegt aber jenseits der Methoden dieses Skripts. Allerdings reicht Stetigkeit von f nicht, um lokale Eindeutigkeit zu garantieren. Dieses Phänomen ist uns bereits aus Analysis 1 bekannt; z.B. besitzt die Differentialgleichung $y' = 2\sqrt{|y|}$ mit Anfangsbedingung $y(0) = 0$ auf dem Zeitintervall $[0, \infty)$ die Lösungen

$$y(t) = 0, \quad y(t) = t^2, \quad y(t) = \begin{cases} 0 & t \leq a \\ (t-a)^2 & t > a \end{cases} \quad (a > 0 \text{ beliebig}).$$

5.3 Lineare Systeme

Wir betrachten nun lineare Systeme von Differentialgleichungen mit konstanten Koeffizienten. Die Lösung kann dann explizit durch Verallgemeinerung der Exponentialfunktion aus Analysis 1 auf Matrizen bestimmt werden.

Satz 5.3 (Lineare Differentialgleichungen und Matrixexponentialfunktion) *Sei $A \in \mathbb{R}^{n\times n}$, $y_0 \in \mathbb{R}^n$. Die eindeutige Lösung $y : \mathbb{R} \to \mathbb{R}^n$ der Differentialgleichung*

$$y' = Ay$$

mit Anfangsbedingung

$$y(0) = y_0$$

ist gegeben durch

$$y(t) = e^{tA} y_0,$$

mit der Matrixexponentialfunktion

$$e^B := \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!} B^k \quad (B \in \mathbb{R}^{n\times n}). \tag{*}$$

Hierbei ist B^k das k-fache Matrixprodukt $\underbrace{B\,B \cdots B}_{k\,mal}$, mit der Konvention $B^0 = I$.

Vorsicht: Die Funktionalgleichung $e^{A+B} = e^A e^B$ gilt nur für kommutierende Matrizen. Siehe Übungen.

Wir müssen zwei Dinge beweisen: die Konvergenz der Reihe in (*), d.h. die Tatsache, dass die Reihendefinition eine wohldefinierte $n \times n$ Matrix liefert, und die Tatsache, dass die Funktion y die Differentialgleichung löst.

Beweis der Konvergenz der Reihe (*): Wir zeigen dies ohne Mehraufwand für beliebige komplexe Matrizen $B \in \mathbb{C}^{n\times n}$. Als Norm auf dem $\mathbb{C}^{n\times n}$ benutzen wir wahlweise die *Hilbert-Schmidt-Norm*

$$||A|| = \sqrt{\sum_{i,j=1}^{n} |A_{ij}|^2} = \sqrt{\mathrm{tr}(\overline{A}^T A)}$$

oder die *von der euklidischen Norm* $|\cdot|$ *auf* $\mathbb{C}^n$ *induzierte Operatornorm*

$$||A|| = \sup_{x\in\mathbb{R}^n,\, x\neq 0} \frac{|Ax|}{|x|}$$

und folgendes Lemma:

Lemma 5.2 Die obigen Normen sind *submultiplikativ*, d.h. $||AB|| \leq ||A||\,||B||$, wobei AB das Matrizenprodukt von A, $B \in \mathbb{C}^{n\times n}$ ist.

Für die Operatornorm ist das trivial. Beweis für die Hilbert-Schmidt-Norm: $(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$, also folgt nach Cauchy-Schwarz

$$|(AB)_{ij}|^2 \leq \sum_{k=1}^{n} |A_{ik}|^2 \sum_{\ell=1}^{n} |B_{\ell j}|^2,$$

und Summieren über i, j liefert $||AB||^2 \leq ||A||^2||B||^2$.

Also folgt für beliebiges $B \in \mathbb{C}^{n\times n}$

$$\sum_{k=0}^{\infty} \left|\left(\tfrac{1}{k!}B^k\right)_{ij}\right| \leq \sum_{k=0}^{\infty} \|\tfrac{1}{k!}B^k\| \underset{\text{Lemma 5.2}}{\leq} \sum_{k=0}^{\infty} \frac{1}{k!}\|B\|^k < \infty$$

(die letzte Reihe ist nichts als die übliche Exponentialreihe an der Stelle $\|B\| \in \mathbb{R}$, deren Konvergenz wir in Analysis 1 mit Hilfe des Quotientenkriteriums hergeleitet hatten). Also ist die Reihe (*) komponentenweise absolut konvergent, und die Matrix $e^B \in \mathbb{C}^{n\times n}$ somit wohldefiniert.

Beweis von Satz 5.3: Die gliedweise Ableitung der Reihe

$$e^{tA} = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k$$

nach $t \in \mathbb{R}$ ist die Reihe

$$\sum_{k=1}^{\infty} \frac{k t^{k-1}}{k!} A^k \underset{\ell=k-1}{=} \sum_{\ell=0}^{\infty} \frac{t^\ell}{\ell!} A^{\ell+1} (= Ae^{tA} = e^{tA}A).$$

Diese Reihe ist – analog zur Reihe für e^{tA} – ebenfalls komponentenweise absolut konvergent, für alle $t \in \mathbb{R}$. Somit können wir Analysis 1 Satz 9.8 über die gliedweise Differentiation von Potenzreihen anwenden. Demnach ist die ursprüngliche Reihe diff'bar, und es gilt

$$\frac{d}{dt} e^{tA} = Ae^{tA} = e^{tA}A.$$

Anwenden der ersten beiden Ausdrücke auf den Vektor $y_0 \in \mathbb{R}^n$ liefert $t \mapsto e^{tA}y_0$ diff'bar,

$$\frac{d}{dt}\Big(e^{tA}y_0\Big) = A\Big(e^{tA}y_0\Big) \ \forall t \in \mathbb{R}.$$

Die Stetigkeit folgt aus der Diff'barkeit und die Anfangsbedingung aus der – gemäss Reihendefinition offensichtlichen – Tatsache

$$e^{tA}\Big|_{t=0} = I.$$

Beispiel 4 (Harmonischer Oszillator) Wir betrachten das System $y_1' = y_2$, $y_2' = -\omega^2 y_1$ mit Parameter $\omega > 0$, in Matrixschreibweise:

$$y' = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} y.$$

Interpretation als Schwingungsgleichung siehe unten. Da

$$\begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}^2 = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} = -\omega^2 I, \ \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}^3 = -\omega^2 \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}, \ \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}^4 = \omega^4 I,$$

ist die Matrixexponentialfunktion

$$\begin{aligned} &e^{t\begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}} \\ &= I + t\begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} - \frac{(\omega t)^2}{2!}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{\omega^2 t^3}{3!}\begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} + \frac{(\omega t)^4}{4!} I + \frac{\omega^4 t^5}{5!}\begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} - - + + \ldots \\ &= \begin{pmatrix} 1 - \frac{(\omega t)^2}{2!} + \frac{(\omega t)^4}{4!} - + \ldots & t - \frac{\omega^2 t^3}{3!} + \frac{\omega^4 t^5}{5!} - + \ldots \\ -\omega^2 t + \frac{\omega^4 t^3}{3!} - \frac{\omega^6 t^5}{5!} + - \ldots & 1 - \frac{(\omega t)^2}{2!} + \frac{(\omega t)^4}{4!} - + \ldots \end{pmatrix} = \begin{pmatrix} \cos(\omega t) & \frac{1}{\omega}\sin(\omega t) \\ -\omega \sin(\omega t) & \cos(\omega t) \end{pmatrix} \end{aligned}$$

und somit die Lösung von (D), (A)

$$y(t) = \begin{pmatrix} \cos(\omega t) & \frac{1}{\omega}\sin(\omega t) \\ -\omega \sin(\omega t) & \cos(\omega t) \end{pmatrix} y_0. \qquad \text{(L)}$$

Interpretation: unser System ist äquivalent zur (aus Analysis 1 bekannten) Schwingungsgleichung

$$z'' + \omega^2 z = 0, \quad z(0) = z_0 \text{ (Anfangsauslenkung)}, \; z'(0) = v_0 \text{ (Anfangsgeschwindigkeit)}$$

(setze $y_1 = z$, $y_2 = z'$, $(y_0)_1 = z_0$, $(y_0)_2 = v_0$). Physikalisch beschreibt $z(t)$ die Auslenkung eines ungedämpften Pendels aus der Ruhelage zur Zeit t; für eine Herleitung der Gleichung siehe Beispiel 5. Die Lösung der Schwingungsgleichung ist also gegeben durch die 1. Komponente der Lösung (L), d.h.

$$z(t) = z_0 \cos(\omega t) + \tfrac{v_0}{\omega} \sin(\omega t).$$

Dieses Ergebnis hatten wir bereits in Analysis 1 – mithilfe anderer Methoden – gefunden.

Im Allgemeinen kann man die Matrixexponentialfunktion nicht so einfach aus der Reihendefinition ablesen, dann hilft

Lemma 5.3 (Die Matrixexponentialfunktion vertauscht mit Basistransformationen) Seien A, $\Lambda \in \mathbb{R}^{n\times n}$, und sei

$$A = S\Lambda S^{-1}$$

für eine invertierbare Matrix $S \in \mathbb{R}^{n\times n}$. Dann gilt

$$e^{tA} = Se^{t\Lambda}S^{-1}.$$

Beweis: Das folgt sofort aus der Reihendefinition, denn ist $A = S\Lambda S^{-1}$, so folgt

$$A^2 = S\Lambda \underbrace{S^{-1}S}_{=I} \Lambda S^{-1} = S\Lambda^2 S^{-1}$$

und analog $A^k = S\Lambda^k S^{-1}$ für beliebiges $k \in \mathbb{N}$.

Aus der linearen Algebra wissen wir: typischerweise (z.B. wenn A symmetrisch oder orthogonal ist, oder n verschiedene Eigenwerte besitzt) ist A diagonalisierbar, d.h. es existieren n linear unabhängige Eigenvektoren. Indem man diese in die Spalten einer Matrix schreibt und die Matrix S nennt, folgt

$$A = S \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} S^{-1}; \tag{1a}$$

hierbei sind die λ_i die Eigenwerte der Matrix. Die Exponentialfunktion einer Diagonalmatrix lässt sich aber sofort aus der Reihendefinition ablesen und wegen Lemma

5.3 folgt

$$e^{tA} = S \begin{pmatrix} e^{\lambda_1 t} & 0 & \dots & 0 \\ 0 & e^{\lambda_2 t} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & e^{\lambda_n t} \end{pmatrix} S^{-1}. \tag{1b}$$

Die Matrixexponentialfunktion, und somit die Lösung des linearen Differentialgleichungssystems $y' = Ay$, $y(0) = y_0$, lässt sich also mithilfe der Eigenwerte und Eigenvektoren von A bestimmen. Siehe Beispiel 5) unten sowie Übungen.

Allgemeine Matrizen können wir auf Jordan'sche Normalform bringen. Genauer: Nach einem grundlegenden Resultat der Linearen Algebra gibt es für beliebiges $A \in \mathbb{R}^{n\times n}$ ein invertierbares S sodass

$$A = S \underbrace{\begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_m & & & \\ & & & J_1 & & \\ & & & & \ddots & \\ & & & & & J_\ell \end{pmatrix}}_{=:A_0} S^{-1} \tag{2a}$$

wobei die $\lambda_i \in \mathbb{C}$ und die J_i Jordanblöcke[11], d.h. von der Form

$$J = \begin{pmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix} = \lambda I + N \in \mathbb{C}^{k\times k}, \quad k \geq 2, \quad N = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}.$$

Auch für einen Jordanblock können wir e^{tJ} ausrechnen. Da die Matrizen λI und N vertauschen, ist $e^{tJ} = e^{t\lambda I} e^{tN} = e^{\lambda t} e^{tN}$, d.h wir müssen nur die Exponentialfunktion des nilpotenten Anteils ausrechnen. Im 2×2 Fall ist

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad N^2 = 0, \quad e^{tN} = I + tN = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

Analog ist die Exponentialreihe auch im $k \times k$ Fall wegen $N^k = 0$ endlich, und berechnet sich zu

$$e^{tN} = I + tN + \frac{t^2}{2!}N^2 + \dots + \frac{t^{k-1}}{(k-1)!}N^{k-1} = \begin{pmatrix} 1 & t & \dots & \frac{t^{k-1}}{(k-1)!} \\ & 1 & \ddots & \vdots \\ & & \ddots & t \\ & & & 1 \end{pmatrix}.$$

Insgesamt haben wir also

$$e^{tJ} = e^{\lambda t} \begin{pmatrix} 1 & t & \dots & \frac{t^{k-1}}{(k-1)!} \\ & 1 & \ddots & \vdots \\ & & \ddots & t \\ & & & 1 \end{pmatrix}, \quad e^{tA} = S \begin{pmatrix} e^{t\lambda_1} & & & & & \\ & \ddots & & & & \\ & & e^{t\lambda_m} & & & \\ & & & e^{tJ_1} & & \\ & & & & \ddots & \\ & & & & & e^{tJ_\ell} \end{pmatrix} S^{-1} \tag{2b}$$

Wir besprechen hierzu ein interessantes 2×2 Beispiel.

[11] Manche Autoren nennen die 1×1 Matrizen λ_i ebenfalls Jordanblöcke und die J_i nichttriviale Jordanblöcke

Beispiel 5 (Schwingungsgleichung mit Dämpfung) Wir betrachten die Gleichung

$$z'' + pz' + qz = 0, \quad p, q > 0 \tag{S}$$

(wobei $z : \mathbb{R} \to \mathbb{R}$), mit Anfangsbedingungen

$$z(0) = z_0,\ z'(0) = v_0.$$

Mit $y_1 = z$, $y_2 = z'$, $(y_0)_1 = z_0$, $(y_0)_2 = v_0$ ist dies äquivalent zu

$$y' = \underbrace{\begin{pmatrix} 0 & 1 \\ -q & -p \end{pmatrix}}_{=:A} y, \quad y(0) = \begin{pmatrix} z_0 \\ v_0 \end{pmatrix}$$

(wobei $y : \mathbb{R} \to \mathbb{R}^2$). Physikalische Interpretation: $z(t)$ =Auslenkung eines Pendels aus der Ruhelage zur Zeit t, z_0 =Anfangsauslenkung, v_0 =Anfangsgeschwindigkeit. Die Parameter p bzw. q beschreiben die Stärke der Dämpfungs- bzw. der rücktreibenden Kraft.

Herleitung der Schwingungsgleichung (S). Diese ist ein Spezialfall des zweiten Newton'schen Gesetzes $F = ma$. Genauer: sei $z(t)$ die Auslenkung eines an einem Faden- oder Federpendel befestigten Teilchens aus der Ruhelage zur Zeit t. Dann bedeutet $z'(t)$ die Geschwindigkeit und $z''(t)$ die Beschleunigung zur Zeit t. Nach dem zweiten Newton'schen Gesetz sowie der Annahme einer rückwirkenden Kraft proportional zur Auslenkung sowie einer Dämpfungs- oder Reibungskraft proportional zur Geschwindigkeit folgt

$$m\,z'' = F = -k\,z - \gamma\,z',$$

wobei m=Masse des Teilchens, F=Kraft, und k, γ Proportionalitätskonstanten. D.h. wir erhalten (nach Division durch m) eine Gleichung der Form (S). Der Fall $\gamma = 0$ entspricht einem ungedämpften Pendel.

Die Eigenwerte von A sind die Lösungen der Gleichung

$$0 = \det(A - \lambda I) = \det\begin{pmatrix} -\lambda & 1 \\ -q & -p - \lambda \end{pmatrix} = \lambda^2 + p\lambda + q.^{12}$$

Die Eigenwerte sind

(i)	$\lambda_{1/2} = -\frac{p}{2} \pm \alpha \in \mathbb{R},\ \alpha = \sqrt{(\frac{p}{2})^2 - q}$	wenn $(\frac{p}{2})^2 > q$ (starke Dämpfung)
(ii)	$\lambda = -\frac{p}{2} \in \mathbb{R}$	wenn $(\frac{p}{2})^2 = q$ (kritische Dämpfung)
(iii)	$\lambda_{1/2} = -\frac{p}{2} \pm i\omega \in \mathbb{C}\backslash\mathbb{R},\ \omega = \sqrt{q - (\frac{p}{2})^2}$	wenn $(\frac{p}{2})^2 < q$ (schwache Dämpfung).

Wir beschränken uns zunächst auf die "typischen" Fälle (i) und (iii). Dann ist A diagonalisierbar und wir können (1a), (1b) anwenden. Explizit:

$$A = S\begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} S^{-1}$$

[12] Die Eigenwerte von A sind also genau die Nullstellen der charakteristischen Gleichung $\lambda^2 + p\lambda + q = 0$ aus Analysis 1, die wir dort formal aus der Schwingungsgleichung durch die Substitution $z'' \rightsquigarrow \lambda^2$, $z' \rightsquigarrow \lambda$, $z \rightsquigarrow \lambda^0 = 1$ erhalten hatten. Das braucht uns aber hier nicht zu interessieren.

mit

$$S = \begin{pmatrix} | & | \\ v_1 & v_2 \\ | & | \end{pmatrix}, \quad v_1, v_2 \text{ Eigenvektoren zu } \lambda_1, \lambda_2.$$

Ist λ Eigenwert, so ist

$$\begin{pmatrix} 0 & 1 \\ -q & -p \end{pmatrix} \begin{pmatrix} 1 \\ \lambda \end{pmatrix} = \begin{pmatrix} \lambda \\ -q - p\lambda \end{pmatrix} \underset{\lambda^2+p\lambda+q=0}{=} \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ \lambda \end{pmatrix}$$

und somit sind die Eigenvektoren

$$v_1 = \begin{pmatrix} 1 \\ \lambda_1 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 1 \\ \lambda_2 \end{pmatrix}.$$

Die Inverse von S ist $\frac{1}{\lambda_2 - \lambda_1} \begin{pmatrix} \lambda_2 & -1 \\ -\lambda_1 & 1 \end{pmatrix}$ und Gleichung (1b) liefert

$$y(t) = e^{At} \begin{pmatrix} z_0 \\ v_0 \end{pmatrix} \quad \text{mit} \quad e^{At} = \begin{pmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix} \begin{pmatrix} e^{\lambda_1 t} & \\ & e^{\lambda_2 t} \end{pmatrix} \frac{1}{\lambda_2 - \lambda_1} \begin{pmatrix} \lambda_2 & -1 \\ -\lambda_1 & 1 \end{pmatrix}.$$

Insbesondere hat die Lösung die Form

$$y(t) = \alpha e^{\lambda_1 t} \; + \; \beta e^{\lambda_2 t}$$

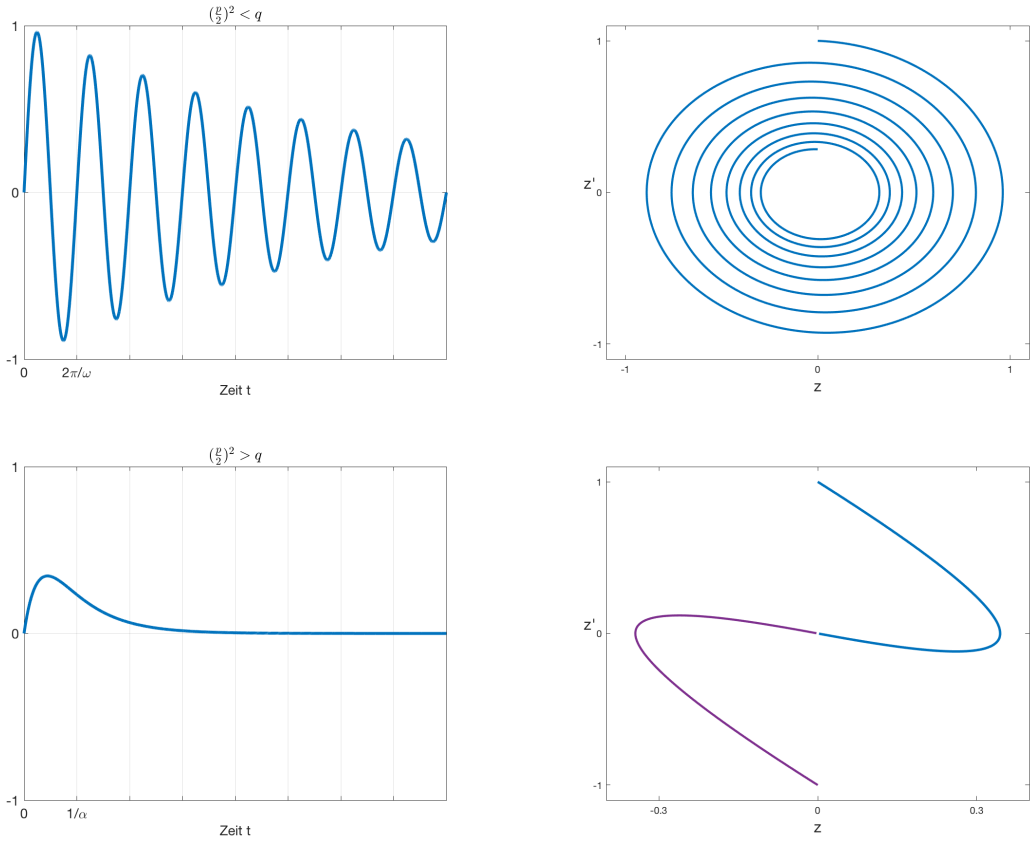
für geeignete Konstanten $\alpha, \beta \in \mathbb{C}$. Für starke Dämpfung (Fall (i)), und der Einfachheit halber $z_0 = 0$, ist also

$$z(t) = y_1(t) = e^{-\frac{p}{2}t} \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} e^{\alpha t} & \\ & e^{-\alpha t} \end{pmatrix} \frac{1}{-2\alpha} \begin{pmatrix} -v_0 \\ v_0 \end{pmatrix} = \frac{v_0}{\alpha} e^{-\frac{p}{2}t} \frac{e^{\alpha t} - e^{-\alpha t}}{2} = \frac{v_0}{\alpha} e^{-\frac{p}{2}t} \sinh(\alpha t)$$

und für schwache Dämpfung (Fall (iii)), und der Einfachheit halber $z_0 = 0$,

$$z(t) = y_1(t) = e^{-\frac{p}{2}t} \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} e^{i\omega t} & \\ & e^{-i\omega t} \end{pmatrix} \frac{1}{-2i\omega} \begin{pmatrix} -v_0 \\ v_0 \end{pmatrix} = \frac{v_0}{\omega} e^{-\frac{p}{2}t} \frac{e^{i\omega t} - e^{-i\omega t}}{2i} = \frac{v_0}{\omega} e^{-\frac{p}{2}t} \sin(\omega t).$$

Das Schaubild auf der nächsten Seite zeigt, wie die errechneten Lösungen aussehen.



Lösung der gedämpften Schwingungsgleichung mit Anfangsauslenkung 0 und Anfangsgeschwindigkeit 1.

Oben: Für schwache Dämpfung ist die Lösung eine abklingende Exponentialfunktion mal einem Sinus. Links: Lösung als Funktion der Zeit. Rechts: Orbit im Phasenraum.

Unten: Für starke Dämpfung ist die Lösung eine abklingende Exponentialfunktion mal einem Sinus hyperbolicus. Links: Lösung als Funktion der Zeit. Rechts: Orbit im Phasenraum (blau) sowie analoger Orbit mit Anfangsgeschwindigkeit -1 (lila).

Im Sonderfall (ii) (kritische Dämpfung) hat A *nur einen Eigenwert.* Da die Matrix A andererseits keine Diagonalmatrix ist, besitzt sie die Jordan'sche Normalform

$$A_0 = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

und folglich wegen (2b)

$$e^{tA_0} = e^{t\lambda} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

Um eine Basis $\{v, w\}$ des $\mathbb{R}^2$ zu bestimmen, in der A die Normalform A_0 besitzt, d.h. sodass

$$A = \begin{pmatrix} | & | \\ v & w \\ | & | \end{pmatrix} A_0 \begin{pmatrix} | & | \\ v & w \\ | & | \end{pmatrix}^{-1},$$

müssen wir die folgenden linearen Gleichungen lösen:

$$Av = \lambda v, \quad Aw = \lambda w + v.$$

Hierzu ist es nützlich, A mithilfe von λ auszudrücken (beachte $q = (p/2)^2 = \lambda^2$):

$$A = \begin{pmatrix} 0 & 1 \\ -\lambda^2 & 2\lambda \end{pmatrix}.$$

Indem wir o.B.d.A. $v_1 = 1$ annehmen, erhalten wir aus der ersten Gleichung $v_2 = \lambda$, und indem wir o.B.d.A. $w_1 = 0$ annehmen, folgt aus der zweiten Gleichung $w_2 = 1$; damit sind die Vektoren

$$v = \begin{pmatrix} 1 \\ \lambda \end{pmatrix}, \quad w = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

die gesuchte Basis. Somit folgt

$$\begin{aligned} A &= \begin{pmatrix} 1 & 0 \\ \lambda & 1 \end{pmatrix} A_0 \begin{pmatrix} 1 & 0 \\ -\lambda & 1 \end{pmatrix}, \\ e^{tA} &= \begin{pmatrix} 1 & 0 \\ \lambda & 1 \end{pmatrix} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\lambda & 1 \end{pmatrix} e^{\lambda t} = \begin{pmatrix} 1-\lambda t & t \\ -\lambda^2 t & \lambda t + 1 \end{pmatrix} e^{\lambda t} = \begin{pmatrix} 1 + \frac{p}{2} t & t \\ -(\frac{p}{2})^2 t & 1 - \frac{p}{2} t \end{pmatrix} e^{-\frac{p}{2} t} \end{aligned}$$

und, indem wir die erste Komponente des Vektors $e^{tA} y_0$ nehmen,

$$z(t) = z_0 \Big(1 + \frac{p}{2} t\Big) e^{-\frac{p}{2} t} + v_0 \, t \, e^{-\frac{p}{2} t}.$$

Übung: machen Sie die Probe, d.h. rechnen Sie nach, dass z die Schwingungsgleichung und die Anfangsbedingungen erfüllt. Diese Lösung wird in typischen Physik- und "Höhere Mathematik für Ingenieure"-Skripten (und manchen Analysis-2-Lehrbüchern) mithilfe des Ansatzes $z(t) = A\, e^{\lambda t} + B\, t\, e^{\lambda t}$ hergeleitet, der auf wundersame Weise vom Himmel fällt.

Wir beschäftigen uns nun mit der inhomogenen Gleichung $y' = Ay + f$, wobei f eine gegebene zeitabhängige Funktion ist. Wie im homogenen Fall ($f = 0$) kann man die Lösung explizit angeben.

Satz 5.4 (Variation der Konstanten) *Sei $A \in \mathbb{R}^{n\times n}$, $y_0 \in \mathbb{R}^n$, $f : \mathbb{R} \to \mathbb{R}^n$ stetig. Die eindeutige Lösung $y : [0, \infty) \to \mathbb{R}^n$ der Differentialgleichung*

$$y' = Ay + f$$

mit Anfangsbedingung

$$y(0) = y_0$$

ist gegeben durch

$$y(t) = e^{tA} \Big(y_0 + \int_0^t e^{-sA} f(s)\, ds \Big).$$

Der Name des Satzes ist in der Herleitung der Lösung erklärt.

Beweis Nach Produktregel und Hauptsatz gilt für obiges y

$$\frac{d}{dt} y(t) = A \underbrace{e^{tA} \Big(y_0 + \int_0^t e^{-sA} f(s)\, ds \Big)}_{= y(t)} + \underbrace{e^{tA} \Big(e^{-tA} f(t) \Big)}_{= f(t)}.$$

Die Eindeutigkeit folgt aus Satz 5.1.

Wie kommt man auf die Lösungsformel? Im homogenen Fall hat die Lösung gemäss Satz 5.3 die Form

$$y(t) = e^{At}c, \;\; c \text{ konstanter Vektor.}$$

Im inhomogenen Fall machen wir den Ansatz

$$y(t) = e^{At}c(t), \;\; c \text{ zeitabhängiger Vektor.} \tag{A}$$

Daher der Name "Variation der Konstanten". Angenommen es existiert ein solches diff'bares c, dann folgt wegen Produktregel $y'(t) = Ay(t) + e^{At}c'(t)$, d.h. y löst die gewünschte Gleichung falls

$$e^{At}c'(t) = f(t) \;\text{ für alle } t.$$

Aus dieser Bedingung können wir c bestimmen: indem wir von links mit der Matrix e^{-At} multiplizieren, folgt $c'(t) = e^{-At}f(t)$ und somit durch Integration

$$c(t) = c_0 + \int_0^t e^{-As}f(s)\,ds, \;\; c_0 \text{ konstanter Vektor.}$$

Aus der Anfangsbedingung ergibt sich schliesslich $c_0 = y_0$ und somit ist (A) genau die Lösung aus Satz 5.4.

Beispiel 6 (Erzwungene Schwingungen; Resonanz) Ein Anwendungsbeispiel, das man mit Hilfe von Variation der Konstanten ansatzfrei[13] lösen kann, ist die Gleichung für erzwungene Schwingungen

$$z'' + pz' + qz = F_0\cos(\omega_0 t). \tag{S_{inh}}$$

Physikalisch beschreibt die Gleichung – wie in Beispiel 5 – die Auslenkung $z(t)$ eines Pendels aus der Ruhelage zur Zeit t; die rechte Seite entspricht einer zusätzlichen von aussen einwirkenden zeitperiodischen Kraft, die hier der Einfachheit halber durch den Cosinus modelliert wird. Der Parameter $\omega_0 > 0$ heisst Anregungsfrequenz. Es ist mathematisch einfacher, zunächst die komplexe Glechung

$$z'' + pz' + qz = F_0e^{i\omega_0 t} \tag{S'$_{inh}$}$$

[13]In typischen Physik- und "Höhere Mathematik"-Skripten und manchen Analysis-Lehrbüchern wird diese Gleichung behandelt, indem zunächst eine spezielle Lösung vom Himmel fällt und man dann nur noch eine geeignete Lösung der homogenen Gleichung aus Beispiel 5 addieren muss, um gegebene Anfangsbedingungen zu erfüllen

zu lösen; der Realteil löst dann (S_{inh}). Wie in Beispiel 4 und 5 schreiben wir die Gleichung als System, indem wir $y_1 = z$, $y_2 = z'$ setzen:

$$y'(t) = \underbrace{\begin{pmatrix} 0 & 1 \\ -q & -p \end{pmatrix}}_{=:A} y(t) + \begin{pmatrix} 0 \\ F_0 e^{i\omega_0 t} \end{pmatrix}.$$

Variation der Konstanten liefert folgende eindeutige Lösung für gegebene Anfangsbedingungen $y(0)$

$$y(t) = e^{At}y(0) + \underbrace{\int_0^t e^{A(t-s)} \begin{pmatrix} 0 \\ F_0 e^{i\omega_0 s} \end{pmatrix} ds}_{=:\tilde{y}(t)}.$$

Wir beschränken uns auf den Fall schwacher Dämpfung, $(\frac{p}{2})^2 - q < 0$, dann besitzt A zwei verschiedene Eigenwerte λ_1 und λ_2 und durch Einsetzen der Formel für $e^{A\tau}$ aus Beispiel 5 erhalten wir

$$\begin{aligned} \tilde{y}(t) &= \int_0^t \begin{pmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix} \begin{pmatrix} e^{\lambda_1(t-s)} & \\ & e^{\lambda_2(t-s)} \end{pmatrix} \tfrac{1}{\lambda_2-\lambda_1} \begin{pmatrix} \lambda_2 & -1 \\ -\lambda_1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ F_0 e^{i\omega_0 s} \end{pmatrix} ds \\ &= \frac{F_0}{\lambda_2 - \lambda_1} \int_0^t \begin{pmatrix} e^{\lambda_2(t-s)} - e^{\lambda_1(t-s)} \\ \lambda_2 e^{\lambda_2(t-s)} - \lambda_1 e^{\lambda_1(t-s)} \end{pmatrix} e^{i\omega_0 s} ds. \end{aligned}$$

Folglich gilt für die erste Komponente $\tilde{z}(t)$ von $\tilde{y}(t)$

$$\tilde{z}(t) = \frac{F_0}{\lambda_2 - \lambda_1} \Big[e^{\lambda_2 t} \int_0^t e^{(-\lambda_2 + i\omega_0)s} ds - e^{\lambda_1 t} \int_0^t e^{(-\lambda_1 + i\omega_0)s} ds \Big].$$

Die Integrale können wir exakt auswerten: für beliebiges komplexes $a+ib$ ist $\int_0^t e^{(a+ib)s} ds = \frac{1}{a+ib}(e^{(a+ib)t} - 1)$ und somit

$$\tilde{z}(t) = \frac{F_0}{\lambda_2 - \lambda_1} \Big[\frac{e^{i\omega_0 t} - e^{\lambda_2 t}}{-\lambda_2 + i\omega_0} - \frac{e^{i\omega_0 t} - e^{\lambda_1 t}}{-\lambda_1 + i\omega_0} \Big].$$

Zusammen mit unserem Ergebnis für die erste Kompnente von $e^{At}y(0)$ aus Beispiel 5 ergibt sich die Lösung von (S'_{inh})

$$z(t) = \alpha e^{\lambda_1 t} + \beta e^{\lambda_2 t} + C e^{i\omega_0 t}, \qquad C = \frac{F_0}{\lambda_2 - \lambda_1} \Big[\frac{1}{-\lambda_2 + i\omega_0} - \frac{1}{-\lambda_1 + i\omega_0} \Big]$$

mit geeigneten, aus den Anfangsbedingungen zu bestimmenden Konstanten $\alpha, \beta \in \mathbb{C}$. Die explizite Konstante C hat sich zunächst in etwas unübersichtlicher Form

ergeben. Indem wir die beiden Brüche in der Klammer auf den Hauptnenner bringen, die Formel aus Beispiel 5 für die Eigenwerte von A benutzen, d.h.

$$\lambda_{1/2} = -\frac{p}{2} \pm i\omega, \quad \omega = \sqrt{q - (\tfrac{p}{2})^2}$$

(weshalb $\lambda_1 + \lambda_2 = -p$ und $\lambda_1\lambda_2 = q$), und die Grösse $\omega_u := \sqrt{q}$ einführen (Schwingungsfrequenz des ungedämpften homogenen Systems), folgt

$$C = \frac{F_0}{\lambda_1\lambda_2 - i\omega_0(\lambda_1 + \lambda_2) - \omega_0^2} = \frac{F_0}{\omega_u^2 - \omega_0^2 + ip\omega_0}.$$

Zum Bestimmen des Realteils der Lösung ist es vorteilhaft, die komplexe Zahl im Nenner in Polarkoordinaten anzugeben, folglich

$$C = \sigma e^{-i\delta}$$

mit

$$\sigma = \frac{F_0}{\sqrt{(\omega_u^2 - \omega_0^2)^2 + (p\omega_0)^2}}, \tag{1}$$

$$\delta = arg\Big((\omega_u^2 - \omega_0^2) + ip\omega_0\Big) \ (= \arctan \frac{p\omega_0}{\omega_u^2 - \omega_0^2} \text{ im 1. Quadranten}). \tag{2}$$

Einsetzen der expliziten Formeln für λ_1, λ_2 und C liefert die Lösung von (S'$_{\text{inh}}$) in der übersichtlichen Form

$$z(t) = e^{-\frac{p}{2}t}\Big(\alpha e^{i\omega t} \ + \ \beta e^{-i\omega t}\Big) \ + \ \sigma e^{i(\omega_0 t - \delta)}.$$

Indem wir den Realteil nehmen, erhalten wir schliesslich die Lösung von (S$_{\text{inh}}$),

$$z(t) = e^{-\frac{p}{2}t}\Big(\alpha' \cos\omega t \ + \ \beta' \sin\omega t\Big) \ + \ \underbrace{\sigma \cos(\omega_0 t - \delta)}_{=:z_*(t)},$$

mit geeigneten, aus den Anfangsbedingungen zu bestimenden Konstanten α', $\beta' \in \mathbb{R}$.

Noch interessanter als die explizite Lösungsformel ist das – durch die Formel offengelegte – qualitative Verhalten:

1. Wegen des exponentiell abklingenden Faktors vor dem ersten Term nähert sich die Lösung – unabhänging vom Wert der Konstanten α', β' bzw. den Anfangsbedingungen – für große Zeiten der speziellen Lösung z_*,

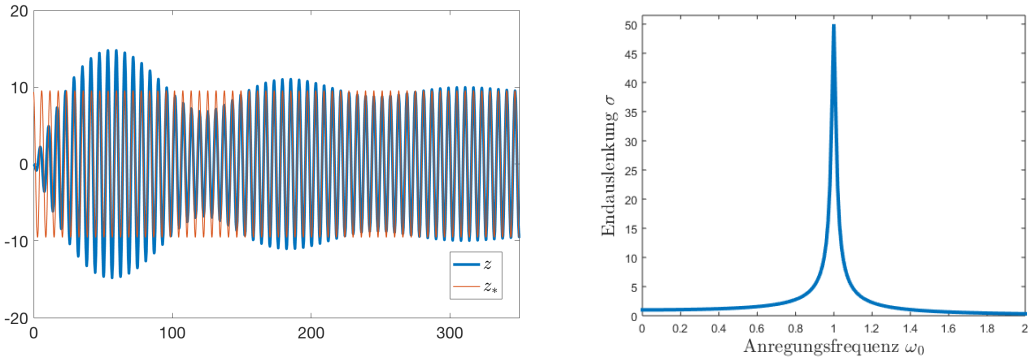
$$\lim_{t\to\infty}\Big(z(t) - z_*(t)\Big) = 0.$$

Siehe Schaubild. D.h. asymptotisch schwingt das System mit der Anregungsfrequenz ω_0 und der in Gleichung (1) berechneten Auslenkung σ. Diese ist also die *universelle* Endauslenkung.

2. Gleichung (1) zeigt: liegt die Anregungsfrequenz nahe an der intrinsischen (oder "Eigen"-)Frequenz des Systems (d.h. $\omega_0 \approx \omega_u$), wird die Endauslenkung gross. Dieser Effekt heisst *Resonanz*. (Alltagsbeispiel: nur wenn man ein Kind auf der Schaukel im richtigen Takt anschubst, erreicht man mit wenig Kraft eine grössere und grössere Auslenkung.) Im theoretischen Limes $p \to 0$ geht die Endauslenkung bei exakter Übereinstimmung beider Frequenzen sogar gegen Unendlich. Die Abhängigkeit der Endauslenkung von der Anregungsfrequenz ist von grossem Interesse in Physik und Ingenieurwissenschaften; der Graph der Funktion

$$\text{Anregungsfreuenz } \omega_0 \;\mapsto\; \text{Endauslenkung } \sigma$$

heisst Resonanzkurve (siehe Schaubild). Beim Design mechanischer Bauteile möchte man Resonanzen typischerweise verhindern; mögliche Anregungsfrequenzen und interne Frequenzen sollten nicht zu nah beieinanderliegen. Wer sich für ein Beispiel eines misslungenen Designs interessiert, kann "Tacoma Bridge" googeln.



Verhalten der inhomogenen Schwingungsgleichung (S_{inh}) für schwache Dämpfung (Parameterwerte $p = 0.02$, $q = 1$, $F_0 = 1$). *Links:* Lösung bei fast-resonanter Anregung und anfänglicher Ruhelage ($\omega_u = 1$, $\omega_0 = 1.05$, $z(0) = z'(0) = 0$). Nach anfänglichem Aufschaukeln und zwischenzeitlichen "Schwebungen" nähert sich die Lösung der speziellen periodischen Lösung z_*. *Rechts:* Resonanzkurve.

5.4 Stabilität

Eine wichtige Fragestellung lautet: was lässt sich über das Langzeitverhalten von Lösungen eines Systems gewöhnlicher Differentialgleichungen aussagen, also das Verhalten im Limes $t \to \infty$? Wir behandeln diese Frage in der Nähe von Gleichgewichtspunkten und untersuchen, ob Lösungen, die nah an einem Gleichgewicht starten, auch nah am Gleichtgewicht bleiben oder sogar gegen dieses konvergieren.

Def. 5.4 (Stabilitätsbegriffe) Sei $\bar{y}$ Gleichgewichtspunkt von $y' = f(y)$, $f : \mathbb{R}^n \to \mathbb{R}^n$ stetig diff'bar. $\bar{y}$ heisst

(a) **stabil**, falls es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt sodass für alle Lösungen $y : [0, \infty) \to \mathbb{R}^n$ mit $|y(0) - \bar{y}| < \delta$ gilt:

$$|y(t) - \bar{y}| < \varepsilon \ \forall t \geq 0.$$

Anderenfalls heisst $\bar{y}$ *instabil*.

(b) **asymptotisch stabil**, falls $\bar{y}$ stabil und $\delta > 0$ so gewählt werden kann, dass zusätzlich

$$\lim_{t \to \infty} |y(t) - \bar{y}| = 0.$$

Informelle Zusammenfassung:

stabil = nah starten, nah dranbleiben.

Für lineare Systeme lässt sich die Frage der Stabilität von Gleichgewichten vollständig beantworten.

Satz 5.5 (Stabilität für lineare Systeme) *Sei $A \in \mathbb{R}^{n \times n}$ beliebig. Für den Gleichgewichtspunkt $\bar{y} = 0$ von $y' = Ay$ gilt:*

a) $\bar{y}$ asyptotisch stabil	$\Longleftrightarrow$	*alle Eigenwerte von A haben Realteil < 0*
b) $\bar{y}$ stabil	$\Longleftrightarrow$	*alle Eigenwerte von A haben Realteil ≤ 0, und Eigenwerte mit Realteil $= 0$ treten nicht in nichttrivialen Jordanblöcken auf.*

Des weiteren gilt im asymptotisch stabilen Fall: Für alle $\alpha > 0$ mit $\max\{Re\,\lambda : \lambda$ ist Eigenwert von $A\} < -\alpha$ existiert eine Konstante C_α sodass

$$\|e^{tA}\| \leq C_\alpha e^{-\alpha t} \quad \forall t \in [0, \infty). \tag{*}$$

Die Bedingung an die Abklingrate α ist scharf; die Argumente im Beweis unten zeigen, dass ein solches C_α nicht existiert, wenn $\max\{\mathrm{Re}\,\lambda : \lambda \text{ ist Eigenwert von } A\} > -\alpha$.

Beweis Zunächst zeigen wir: die Spektralbedingung in a) impliziert (*), dann folgt auch die Implikation "$\Longleftarrow$" in a). Es reicht, Aussage (*) zu beweisen. Wir benutzen die Jordan'sche Normalform (2a), (2b) von A. Zunächst stellen wir fest:

$$\|e^{tA}\| \leq \|S\| \, \|e^{tA_0}\| \, \|S^{-1}\|.$$

Es reicht also, (*) für A_0 zu beweisen. Aus der expliziten Form von e^{tA_0} (siehe (2b)) folgt: alle Komponentenfunktionen von e^{tA_0} sind 0 oder von der Form

$$a(t) = e^{(\mathrm{Re}\,\lambda)t} e^{i(\mathrm{Im}\,\lambda)t} \frac{t^k}{k!} \text{ für einen Eigenwert } \lambda \text{ von } A \text{ und ein } k \in \mathbb{N} \cup \{0\}.$$

Das Produkt $e^{\alpha t}a(t) = e^{(\mathrm{Re}\,\lambda+\alpha)t}e^{i(\mathrm{Im}\,\lambda)t}\frac{t^k}{k!}$ ist aber auf dem Intervall $[0,\infty)$ beschränkt, denn $|e^{i(\mathrm{Im}\,\lambda)t}| = 1$ (siehe Analysis 1) und der erste Faktor ist (wegen $e^{-bt}t^k \to 0$ für $b > 0$ und $t \to \infty$) ebenfalls auf $[0,\infty)$ beschränkt. Insgesamt folgt $e^{\alpha t}\|e^{tA_0}\|$ beschränkt auf $[0,\infty)$, was zu zeigen war.

Als nächstes zeigen wir die Implikation "$\Longrightarrow$" in a). Wiederum reicht es, A_0 zu betrachten. Ist die Spektralbedingung verletzt, existiert entweder (i) ein Eigenwert λ mit Realteil > 0, der nicht in einem Jordanblock auftritt, oder (ii) ein Eigenwert mit Realteil ≥ 0, der in einem Jordanblock auftritt, oder (iii) ein Eigenwert mit Realteil 0, der nicht in einem Jordanblock auftritt. Ersterenfalls ist zu jedem $\delta > 0$ die Funktion $y(t) = \frac{\delta}{2}e^{\lambda t}v$, v zugehöriger normierter Eigenvektor, Lösung von $y' = A_0y$, und es gilt $|y(t)| \to \infty$ $(t \to \infty)$ obwohl $|y(0) - 0| < \delta$. Zweiterenfalls gibt es einen zugehörigen $k \times k$ Jordan-Block J von A_0 mit $k \geq 2$, und die Gleichung $y' = Jy$ besitzt die Lösung

$$\frac{\delta}{2}e^{\lambda t}\begin{pmatrix} \frac{t^{k-1}}{(k-1)!} \\ \vdots \\ t \\ 1 \end{pmatrix}.$$

Die entsprechende Lösung von $y' = A_0y$, deren andere Komponenten gleich Null sind, erfüllt $|y(t)| \to \infty$ $(t \to \infty)$ obwohl $|y(0) - 0| < \delta$. Im dritten Fall schliesslich gibt es einen normierten Eigenvektor v sodass $y(t) = \frac{\delta}{2}e^{i(\mathrm{Im}\,\lambda)t}v$ Lösung, also $|y(t)| \equiv \frac{\delta}{2} \not\to 0$.

Als letztes zeigen wir b). Die Implikation "$\Longleftarrow$" folgt, da aufgrund der Spektralbedingung alle Matrixelemente von e^{tA_0} beschränkt sind. Die Umkehrung folgt, da bei Verletzung der Spektralbedingung einer der beiden Fälle (i) und (ii) oben vorliegt.

Indem wir Satz 5.5 mit dem Konzept der Ableitung (§2) kombinieren, können wir auch im nichtlinearen Fall interessante Aussagen über die Stabilität von Gleichgewichten treffen. Hierbei kommt sowohl die konzeptionelle Seite des Ableitens (Begriff der *totalen Ableitung*) als auch die rechnerische Seite (Begriff der Jacobimatrix) zum Einsatz.

Satz 5.6 (Stabilität von Gleichgewichten nichtlinearer Systeme) *Sei $f : \mathbb{R}^n \to \mathbb{R}^n$ stetig diff'bar, und sei $\bar{y}$ Gleichgewicht von $y' = f(y)$. Dann gilt:*

$$\begin{array}{l} \textit{Alle Eigenwerte der Jacobimatrix} \\ J_f(\bar{y}) \textit{ haben Realteil} < 0 \end{array} \Longrightarrow \bar{y} \textit{ asymptotisch stabil.}$$

Des weiteren gilt im asymptotisch stabilen Fall: es gibt $\delta > 0$, $b > 0$, $C > 0$ sodass für alle Lösungen $y : [0,\infty) \to \mathbb{R}^n$ von $y' = f(y)$ mit $|y(0) - \bar{y}| < \delta$ gilt:

$$|y(t) - \bar{y}| \leq Ce^{-bt} \quad \forall t \geq 0.$$

Umgekehrt ist $\bar{y}$ instabil, wenn es einen Eigenwert mit positivem Realteil gibt. Das beweisen wir hier nicht. Im Fall $\max\{\mathrm{Re}\,\lambda : \lambda \text{ Eigenwert von } J_f(\bar{y})\} = 0$ können alle Möglichkeiten (asymptotische Stabilität, Stabilität aber keine asymptotische Stabilität, Instabilität) auftreten. Das sieht man z.B. an den eindimensionalen Gleichungen $y' = -y^3$, $y' = 0$, $y' = y^3$.

Informelle Beweisidee: Für y nahe $\bar{y}$ gilt wegen der totalen Diff'barkeit von f

$$f(y) = f(\bar{y}) + Df(\bar{y})(y-\bar{y}) + \eta(y-\bar{y}) \;\text{ mit }\; \frac{|\eta(k)|}{|k|} \to 0 \;(k \to 0).$$

Da $f(\bar{y}) = 0$ (denn $\bar{y}$ Gleichgewicht), folgt für $h(t) := y(t) - \bar{y}$

$$h' \approx D_f(\bar{y})(h)\big(= J_f(\bar{y})h\big).$$

Auf diese lineare Differentialgleichung können wir Satz 5.5 anwenden.

Um diese Idee rigoros zu machen, müssen wir den Fehlerterm $\eta(h)$ entlang Lösungen kontrollieren. Dies tun wir durch Kombination von 2 Ideen. Erstens, wir vergessen, dass der Fehlerterm von h abhängt, und behandeln ihn als zeitabhängigen Term, auf den wir die Methode der Variation der Konstanten (Satz 5.4) anwenden können. Zweitens, wir beweisen und benutzen folgendes Lemma.

Lemma 5.4 (Lemma von Gronwall) Sei $z : [0,T] \to \mathbb{R}$ stetig, $b \geq 0$, $a \in \mathbb{R}$. Sei

$$z(t) \leq a + b\int_0^t z(s)\,ds \;\; \forall t \in [0,T].$$

Dann gilt $z(t) \leq ae^{bt} \;\; \forall t \in [0,T]$.

Beweis $Z(t) := a + b\int_0^t z(s)\,ds$ ist diff'bar und löst

$$Z'(t) \leq bZ(t) \;\; \forall t.$$

Multiplikation mit e^{-bt} liefert wegen Produktregel

$$\frac{d}{dt}(e^{-bt}Z(t)) \leq 0.$$

Durch Integration von 0 bis t folgt

$$e^{-bt}Z(t) \leq Z(0) = a.$$

Folglich $Z(t) \leq ae^{bt}$. Die Behauptung folgt wegen $z(t) \leq Z(t)$.

Beweis von Satz 5.6 Ich folge hier der Argumentation im exzellenten Skript *Gewöhnliche Differentialgleichungen* meines Kollegen Martin Brokate. Schreibe $A = J_f(\bar{y})$. Wähle $\alpha > 0$ sodass

$$\max\{\mathrm{Re}\,\lambda : \lambda \text{ Eigenwert von } J_f(\bar{y})\} < -\alpha < 0.$$

Nach Satz 5.5 existiert $C > 0$ sodass $\|e^{tA}\| \le Ce^{-\alpha t}\ \forall t \ge 0$. Seien h und η wie in der informellen Diskussion des Beweises, dann nimmt die Differentialgleichung für y die Form

$$h' = J_f(\bar{y})h + \eta(h)$$

an und es gilt $h(0) = y(0) - \bar{y}$. Wegen der totalen Diff'barkeit von f gilt $|\eta(k)|/|k| \to 0$ ($|k| \to 0$) und somit existiert $\delta_* > 0$ sodass

$$|k| < \delta_* \implies |\eta(k)| \le \tfrac{\alpha}{2C}|k|.$$

Wir setzen nun

$$\delta := \min\{\tfrac{\delta_*}{2C},\ \tfrac{\delta_*}{2}\}$$

und nehmen an: $|h(0)| = |y(0) - \bar{y}| < \delta$. Es reicht zu zeigen (und wir behaupten):

$$|h(t)| \le \tfrac{\delta_*}{2} e^{-\frac{\alpha}{2}t} \tag{3}$$

für alle $t \in [0, \infty)$. Via Variation der Konstanten (Satz 5.4) schreiben wir die Differentialgleichung für h um als

$$h(t) = e^{At}h(0) + \int_0^t e^{A(t-s)}\eta(h(s))\,ds$$

und führen das folgende Zeitintervall ein:

$$I := \{t \in (0,\infty) : |h(s)| < \delta_*\ \forall s \in [0,t]\}.$$

Somit gilt für alle $t \in I$ wegen unserer Abschätzung an η und Satz 5.5 (*)

$$|h(t)| \le Ce^{-\alpha t}|h(0)| + \int_0^t Ce^{-\alpha(t-s)}\frac{\alpha}{2C}|h(s)|\,ds.$$

Multiplikation mit $e^{\alpha t}$ liefert

$$\underbrace{e^{\alpha t}|h(t)|}_{=:z(t)} \le C|h(0)| + \frac{\alpha}{2}\int_0^t \underbrace{e^{\alpha s}|h(s)|}_{=z(s)}\,ds.$$

Nach Lemma von Gronwall folgt $e^{\alpha t}h(t) \le C|h(0)|e^{(\alpha/2)t}$ und somit, indem wir mit $e^{-\alpha t}$ multiplizieren,

$$|h(t)| \le Ce^{-(\alpha/2)t}|h(0)|.$$

Nach Wahl von δ ist aber $|h(0)| < \frac{\delta_*}{2C}$, und somit gilt (3) für alle $t \in I$. Wir müssen noch zeigen: $I = [0,\infty)$, d.h. $|h(t)| < \delta_*$ für alle $t \ge 0$. Sei $t_* := \sup I$. Angenommen $t_* < \infty$. Dann ist $|h(t)| < \delta_*$ für alle $t \in [0, t_*)$, $|h(t_*)| = \delta_*$. Wegen (3) für alle $t \in [0, t_*)$ folgt aber $|h(t_*)| \le \delta_*/2$, Widerspruch. Somit ist $t_* = \infty$ und (3) gilt auf ganz $[0,\infty)$.

Als Anwendungsbeispiel zur Stabilitätsanalyse besprechen wir

Beispiel 6 (SIR-Model der Epidemiologie) Vielleicht interessieren sich ja neuerdings einige von Ihnen für Epidemiologie. Wie für viele andere Gebiete gilt: *Ohne Differentialgleichungen keine Epidemiologie.* Wir besprechen das Standardmodell, es heisst SIR wegen "susceptibles", "infectives", "recovered". Die Grundversion lautet

$$\begin{aligned} S' &= -\beta I \tfrac{S}{N} \\ I' &= \beta I \tfrac{S}{N} - \gamma I \\ R' &= \gamma I. \end{aligned}$$

$S(t)$ ist die Anzahl der Suszeptiblen für eine Krankheit zur Zeit t, $I(t)$ die der Infektiven (also derjenigen, die infiziert aber nicht quarantiniert sind, also andere anstecken können), und $R(t)$ die der Genesenen (englisch recovered). $N = S + I + R$ ist die Gesamtzahl der Population und ist Erhaltungsgrösse, denn

$$\tfrac{d}{dt}(S + I + R) \equiv 0.$$

Es ist sinnvoll, statt der absoluten Zahlen die relativen Anteile $s := S/N$, $i := I/N$, $r := R/N$ zu betrachten. Division der Gleichungen durch N liefert

$$\begin{aligned} s' &= -\beta i s \\ i' &= \beta i s - \gamma i \\ r' &= \gamma i, \end{aligned}$$

und es gilt $s+i+r \equiv 1$. Da zudem s, i, r Anteile $\in [0,1]$ sind, sollten Anfangswerte im Dreiecksgebiet $\{(s_0, i_0, r_0) : s_0 \geq 0,\, i_0 \geq 0,\, r_0 \geq 0,\, s_0 + i_0 + r_0 = 1\}$ vorgegeben werden (Lösungen können dieses Gebiet nicht verlassen).

Erläuterung der Modellierung. Der nichtlineare Term $i \cdot s$ modelliert die Anzahl infektionsrelevanter Kontakte pro Zeiteinheit [sagen wir, pro Tag]. Bei hinreichend "zufälligen" Kontakten in einer Population (also nicht: jeder trifft immer nur dieselben Leute) kann er als proportional zum Produkt der Anzahl Suszeptibler und der Anzahl Infektiver angenommen werden. Die Proportionalitätskonstante β ist die *Kontaktrate*, d.h. die Anzahl ansteckungsrelevanter Kontakte pro Person und Zeiteinheit [Tag]. Die Zahl $1/\gamma$ ist die durchschnittliche infektive Periode, also die durchschnittliche Anzahl Tage, in der ein Infizierter frei herumläuft und andere anstecken kann.

Weitere interessante Parameter sind: die *Kontaktzahl* $\sigma = \beta/\gamma$, also die typische Anzahl ansteckungsrelevanter Kontakte eines Infizierten während der gesamten infektiven Periode, und die *Reproduktionszahl* $\mathcal{R}(t) = \sigma s(t) = \beta \frac{1}{\gamma} s(t)$, also die durchschnittliche Anzahl Personen, die ein Infektiver insgesamt ansteckt.

γ ist eine intrinsische Konstante einer Krankheit, β ist verhaltensabhängig, und $s(t)$ gibt den Immunitätsstatus der Population wieder. Die Reproduktionszahl kann aus 2 völlig verschiedenen Gründen klein sein

1) s klein, oder äquivalent dazu r gross ("Herdenimmunität")
2) β klein ("social distancing", "politische Massnahmen").

Ziel des Modells ist die ungefähre Vorhersage des zeitlichen Epidemieverlaufs anhand der Parameter β und γ und des anfänglichen Immunitätsstatus der Population, also der Anfangswerte $s(0) = s_0$, $i(0) = i_0$, $r(0) = r_0$.[14]

Gleichgewichtszustände: Wir betrachten nur das System für s und i, da r die anderen Grössen nicht beeinflusst und hinterher elementar durch den Erhaltungssatz $s + i + r = 1$ aus diesen berechnet werden kann, also

$$\begin{pmatrix} s \\ i \end{pmatrix}' = f(s,i) \text{ mit } f(s,i) = \begin{pmatrix} -\beta is \\ \beta is - \gamma i \end{pmatrix}.$$

Die Gleichgewichte (s_0, i_0), also die Nullstellen von f, lösen

$$-\beta i_0 s_0 = 0, \quad i_0(\beta s_0 - \gamma) = 0.$$

Da s_0 nicht gleichzeitig 0 und γ/β sein kann, folgt $i_0 = 0$, also sind die Gleichgewichte die Punkte

$$(s_0, 0), \ s_0 \in [0,1] \text{ beliebig.}$$

Es gibt somit ein Kontinuum von Gleichgewichten: wenn niemand infektiv ist, steckt sich auch niemand an, egal wie viele Personen suszeptibel sind.

Linearisiertes System: Das am Startwert s_0 für den Anteil Suszeptibler sowie 0 Infektive, also am Gleichgewicht $(s_0, 0)$, linearisierte System lautet

$$\begin{pmatrix} s - s_0 \\ i - 0 \end{pmatrix}' = \underbrace{f(s_0,0)}_{=0} + \underbrace{J_f(s_0,0)}_{=\begin{pmatrix} 0 & -\beta s_0 \\ 0 & \beta s_0 - \gamma \end{pmatrix}} \begin{pmatrix} s - s_0 \\ i - 0 \end{pmatrix}.$$

Eigenwerte der Jacobimatrix: Wir berechnen

$$det(J_f(s_0,0) - \lambda I) = \lambda\Big(\lambda - (\beta s_0 - \gamma)\Big)$$

und somit sind die Eigenwerte

$$\lambda_1 = 0, \ \ \lambda_2 = \beta s_0 - \gamma =: \lambda_*.$$

Nach Satz 5.5 und Satz 5.6 wird die Stabilität des Gleichgewichts $(s_0, 0)$ unter der linearisierten Dynamik (bzw. der vollen Dynamik) durch das Vorzeichen des nichttrivialen Eigenwerts λ_* entschieden (bzw. nahegelegt):

$$\lambda_* = \beta s_0 - \gamma \begin{cases} < 0 & \Longrightarrow (s_0, 0) \text{ stabil} \\ > 0 & \Longrightarrow (s_0, 0) \text{ instabil.} \end{cases}$$

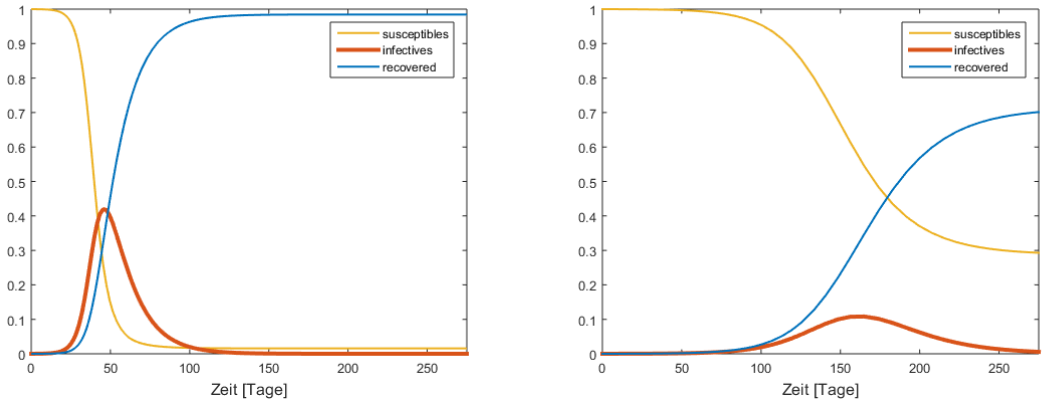
[14] Die Vorhersagen epidemiologischer Modelle sind natürlich weit weniger genau als die von Physikmodellen; insbesondere kann das wahre Kontaktverhalten einer Population nicht mit einem einzigen Parameter (hier β) zusammengefasst werden und selbst dieser Parameter ist nicht leicht zu schätzen.

Beachte: $\lambda_* > 0$ genau dann wenn die anfängliche Reproduktionszahl $\mathcal{R}(0) = \sigma s_0 > 1$. Die Lösung des linearisierten Systems mit Anfangswert $(0, i_0)$ (beachte: wir haben am Startwert s_0 linearisiert, die Null im ersten Eintrag bedeutet $s - s_0 = 0$) und $i_0 > 0$ ist

$$\begin{pmatrix} s(t) - s_0 \\ i(t) \end{pmatrix} = e^{tJ_f(s_0,0)} \begin{pmatrix} 0 \\ i_0 \end{pmatrix} = \begin{pmatrix} (1 - e^{t\lambda_*}) \frac{\beta s_0}{\lambda_*} i_0 \\ e^{t\lambda_*} i_0 \end{pmatrix}.$$

Insbesondere ist $i(t) = i_0 e^{\lambda_* t}$ (das kann man direkt aus der zweiten der ursprünglichen Gleichungen sehen, ohne die Matrixexponentialfunktion auszuwerten). Ist $\lambda_* < 0$, sinkt der Anteil Infektiver exponentiell (keine Epidemie) und für $\lambda_* > 0$ wächst sie exponentiell (Epidemie). Solange die Anzahl Infektiver im Vergleich zur Gesamtbevölkerung klein ist, also in der Anfangsphase (wenn $i(t) \ll 1$), und die Kontaktrate β zeitlich konstant ist, ist das linearisierte Modell eine gute Näherung des SIR-Modells und wir erwarten analoges Verhalten des vollen Modells.

Gesamter Zeitverlauf: Das Modell zeigt kein chaotisches Verhalten à la Lorenzgleichung und kann numerisch akkurat gelöst werden.



Epidemieverlauf gemäss SIR-Modell bei hoher Kontaktzahl, **links**, und reduzierter Kontaktzahl, **rechts**. In beiden Fällen sind die Parameter so gewählt, dass die anfängliche Reproduktionszahl $\mathcal{R}(0) > 1$ ist oder, *mathematisch gesagt*, der infektionsfreie Gleichgewichtszustand instabil ist, d.h. die dortige Jacobimatrix einen Eigenwert $\lambda_* > 0$ besitzt. (Links: infektive Periode $1/\gamma = 14$ Tage, Kontaktzahl $\beta = 0.3$; rechts: selbes $1/\gamma$, $\beta = 0.125$.) Anfangswerte: $i_0 = 10^{-4}$, $s_0 = 1 - 10^{-4}$ (neue Krankheit, d.h. niemand ist immun). In beiden Fällen beginnt die Epidemie zunächst – wie von der Stabilitätstheorie aus Satz 5.5 und Satz 5.6 vorhergesagt – mit einer exponentiellen Wachstumsphase. Wie aus dem Nichts wird aus dem anfänglich winzigen, in den obigen Graphen unsichtbaren Anteil Infizierter ein signifikanter Anteil der Gesamtpopulation. (Salopp gesagt: Epidemien sind Mist.) Der Buckel von Infektivfällen wird aber bei Reduktion sozialer Kontakte flacher und verschiebt sich zeitlich nach hinten, und Herdenimmunität wird bei einer niedrigeren Anzahl Genesener (recovered) erreicht.

Literatur: Ein klassisches Lehrbuch für das dynamische Verhalten gewöhnlicher Differentialgleichungen ist *John Guckenheimer and Philip Holmes, Nonlinear oscillations, Dynamical Systems, and Bifurcations of Vector Fields, Springer, 2002.* Dort werden allerdings einige fortgeschrittene mathematische Eigenschaften gewöhnlicher Differentiangleichungen (jenseits der hier behandelten Existenz-, Eindeutigkeits- und Stabilitätssätze) vorausgesetzt. Ein modernes, zugänglich geschriebenes Lehrbuch, das kein derartiges Vorwissen verlangt und demnächst erscheint, ist *Christian Kuehn, Essentials of Dynamical Systems, Band I: Introduction to ODEs and Nonlinear Dynamics.* Das Schaubild zur Lorenzgleichung ist folgender Quelle entnommen: *Wikipedia*, https://en.wikipedia.org/wiki/File:Lorenz_attractor_boxed.svg, Autor: D.328, Lizenz: CC BY-SA 3.0, https://creativecommons.org/licenses/by-sa/3.0/. Eine exzellente Übersicht über Modellierung und Analyse gewöhnlicher Differentialgleichungen aus der Mechanik findet sich im Standardlehrbuch *V. I. Arnold, Mathematical Methods of Classical Mechanics, Springer, 1989.* Eine Diskussion des SIR-Modells und einiger Varianten findet sich in dem Übersichtsartikel *Herbert W. Hethcote, The Mathematics of Infectious Diseases, SIAM Review Vol. 42 No. 4, 599-653, 2000.* Für weitergehende Informationen zu mathematischen Modellen in der Epidemiologie siehe Kapitel 4 des Lehrbuchs *Johannes Müller and Christina Kuttler, Methods and Models in Mathematical Biology, Springer, 2015.*

6 Einführung in die Topologie

Topologie ist das systematische Studium stetiger Abbildungen sowie von Eigenschaften, die unter stetigen Deformationen erhalten bleiben.

Wir erinnern an die elementare Definition von stetig: $f : M \subseteq \mathbb{R}^n \to \mathbb{R}^m$ heisst stetig, wenn gilt: $[x^{(\nu)} \in M, x \in M, x^{(\nu)} \to x \Longrightarrow f(x^{(\nu)} \to f(x)]$. Aber viele elementare Fragen sind auf der Basis dieser Definition nicht so einfach zu beantworten.

Beispiel: *Betrachte die Mengen*

d.h. $A = \{(x,y) \in \mathbb{R}^2 : x = 0\}$, $B = \{(x,y) \in \mathbb{R}^2 : x \cdot y = 0\}$, $C = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$, $D = (0,1) \times \{0,1\}$. *Zwei Teilmengen* M, M' *des* $\mathbb{R}^n$ *heissen homöomorph, wenn es eine bijektive, in beiden Richtungen stetige Abbildung zwischen ihnen gibt (d.h.* $f : M \to M'$, *bijektiv, stetig,* f^{-1} *stetig).*

Welche der obigen vier Mengen sind zueinander homömorph?

In der Topologie beginnen wir damit, grundlegende Begriffe wie abgeschlossen, konvergent, kompakt, stetig nicht nur in $\mathbb{R}$ (siehe Analysis 1) oder $\mathbb{R}^n$ (siehe Analysis 2), sondern viel allgemeiner und abstrakter zu definieren – natürlich so, dass im $\mathbb{R}$ und $\mathbb{R}^n$ wieder dasselbe herauskommt wie bisher. Diese Herangehensweise ist für viele Bereiche der modernen Mathematik grundlegend – und wird uns auch erlauben, obige (elementar formulierbare, aber gar nicht so triviale) Frage zu beantworten.

Die Grundidee ist, nicht mit einer Norm, also einem "Abstandsbegriff", anzufangen, sondern das Konzept der *offenen Menge* an den Anfang zu stellen und alle weiteren Konzepte hieraus abzuleiten. Gegenüber den anschaulichen, auf Normen beruhenden "Fussgängerdefinitionen" aus Analysis 1 und Analysis 2 werden die Konzepte dadurch mathematisch einfacher, allerdings auf Kosten von Anschaulichkeit.

Wie soll man aber definieren, was eine offene Menge ist? Die geniale Lösung dieses Problems besteht darin, offene Mengen durch ihre "Wechselwirkungseigenschaften untereinander" zu definieren! Dies ist ein Paradebeispiel des axiomatischen Standpunkts in der Mathematik. Wir umgehen die Frage, was die Objekte an der Spitze einer daraus abgeleiteten Begriffskette "sind", und definieren sie dadurch, was man mit ihnen machen kann.

Zur Vorbereitung der Definitionen beginnen wir mit zwei Wechselwirkungseigenschaften der offenen Mengen im $\mathbb{R}^n$. Wiederholung der Definition von offen mithilfe

der euklidischen Norm im $\mathbb{R}^n$: $A \subseteq \mathbb{R}^n$ heisst offen, wenn gilt: zu jedem Punkt in A existiert eine Kugel um diesen Punkt, die ebenfalls in A liegt; in Formelsprache: $\forall x \in A\ \exists \varepsilon > 0 : B_\varepsilon(x) := \{y \in \mathbb{R}^n : |x-y| < \varepsilon\} \subseteq A$.

Lemma 6.0 Für offene Mengen im $\mathbb{R}^n$ gilt:
a) Beliebige Vereinigungen offener Mengen sind offen. D.h. A_i offen $\forall i \in I$, I Indexmenge $\Longrightarrow \bigcup_{i\in I} A_i$ offen.
b) Endliche Durchschnitte offener Mengen sind offen. D.h. A_i offen $\forall i \in I$, I endlich $\Longrightarrow \bigcap_{i\in I} A_i$ offen.

Beweis a):

$$x \in \bigcup_i A_i \Longrightarrow \exists i_0 : x \in A_{i_0} \underset{A_{i_0} \text{ offen}}{\Longrightarrow} \exists B_\varepsilon(x) \subseteq A_{i_0} \Longrightarrow B_\varepsilon(x) \subseteq \bigcup_i A_i.$$

b):

$$x \in \bigcap_i A_i \underset{A_i \text{ offen}}{\Longrightarrow} \forall i \exists B_{\varepsilon_i}(x) \subseteq A_i \Longrightarrow B_{\min_{i\in I}\varepsilon_i}(x) \subseteq A_i \forall i, \text{ d.h. } \subseteq \bigcap_i A_i.$$

6.1 Der Begriff des topologischen Raumes

Die Hauptzutat in der Definition eines topologischen Raumes sind die beiden soeben hergeleiteten Wechselwirkungseigenschaften a) und b) offener Mengen im $\mathbb{R}^n$.

Für eine beliebige Menge X bezeichne $\mathcal{P}(X) := \{A : A \subseteq X\}$ die Potenzmenge. Wir erinnern daran, dass $\varnothing, X \in \mathcal{P}(X)$.

Def. 6.1 Ein *topologischer Raum* ist ein Paar $(X, \mathcal{T})$, wobei X eine Menge und $\mathcal{T} \subseteq \mathcal{P}(X)$ eine Familie von Teilmengen von X ist, sodass folgende Axiome gelten:
(1) $X, \varnothing \in \mathcal{T}$
(2) $X_i \in \mathcal{T}$ für alle $i \in I \Longrightarrow \bigcup_{i\in I} X_i \in \mathcal{T}$
(3) $X_i \in \mathcal{T}$ für alle $i \in I$, I endlich $\Longrightarrow \bigcap_{i\in I} X_i \in \mathcal{T}$.
Eine solche Familie $\mathcal{T}$ heisst *Topologie* auf X. Die Elemente von $\mathcal{T}$ heissen *offene Mengen* (bezüglich der Topologie $\mathcal{T}$).

In Worten: Ganzraum und leere Menge müssen offen sein; beliebige Vereinigungen und endliche Durchschnitte offener Mengen müssen offen sein.

Beispiele 1) $X = \mathbb{R}$, $|\cdot|$=Absolutbetrag (d.h. $|x| = x$ wenn $x \geq 0$ und $-x$ wenn $x < 0$), $T \in \mathcal{T} :\Longleftrightarrow$ zu jedem $x_0 \in T$ existiert ein $\varepsilon > 0$ sodass $B_\varepsilon(x_0) := \{x \in \mathbb{R} : |x - x_0| < \varepsilon\} \subseteq T$.

2) $X = \mathbb{R}^n$, $|\cdot|$=euklidische Norm (d.h. $|x| = (x_1^2 + ... + x_n^2)^{1/2}$), $T \in \mathcal{T} :\Longleftrightarrow$ zu jedem $x_0 \in T$ existiert ein $\varepsilon > 0$ sodass $B_\varepsilon(x_0) := \{x \in \mathbb{R} : |x - x_0| < \varepsilon\} \subseteq T$.

3) $X = C([a,b])$, $\|\cdot\|_\infty$=L^∞ Norm (d.h. $\|f\|_\infty = \sup_{x\in[a,b]} |f(x)|$), $T \in \mathcal{T} :\Longleftrightarrow$ zu jedem $f_0 \in T$ existiert ein $\varepsilon > 0$ sodass $B_\varepsilon(f_0) := \{f \in C([a,b]) : \|f - f_0\|_\infty < \varepsilon\} \subseteq T$.

4) $X = C([a,b])$, $\|\cdot\|_2$=L^2 Norm (d.h. $\|f\|_2 = (\int_a^b |f(x)|^2 dx)^{1/2}$), $\mathcal{T}$ analog

5) X normierter $\mathbb{K}$-Vektorraum mit Norm $\|\cdot\|$, $\mathbb{K}$ beliebiger Körper, $\mathcal{T}$ analog

Alle obigen Beispiele sind von der Form 5). Der Nachweis, dass 5) die Axiome (1), (2), (3) erfüllt und damit $(X,\mathcal{T})$ topologischer Raum ist, geht genauso wie in Lemma 6.0 im Fall des $\mathbb{R}^n$ vorgerechnet. Die Familie $\mathcal{T}$ heisst *Normtopologie* auf X. Beachte: der dem Vektorraum zugrundeliegende Körper braucht keineswegs $\mathbb{R}$ oder $\mathbb{C}$ zu sein; $\mathbb{Q}$, endliche Körper, etc. sind erlaubt.

6) Die "kleinstmögliche" Topologie auf einer Menge X ist $\mathcal{T} = \{X, \varnothing\}$. Sie heisst *triviale Topologie*, und erfüllt offensichtlich die Axiome (1), (2), (3).

7) Die "grösstmögliche" Topologie auf einer Menge X ist $\mathcal{T} = \mathcal{P}(X)$. Sie heisst *diskrete Topologie*, und erfüllt offensichlich die Axiome (1), (2), (3).

8) Sei $(X,\mathcal{T})$ topologischer Raum, $A \subseteq X$. Die Familie $\mathcal{T}_A := \{T \cap A : T \in \mathcal{T}\}$ heisst *Relativtopologie* auf A, und macht A auf natürliche Weise ebenfalls zu einem topologischen Raum. Mengen in $\mathcal{T}_A$ heissen *offen bezüglich* $\mathcal{T}_A$ oder *relativ offen bezüglich* A oder kurz A-*offen*.

9) Das Paar $(\mathbb{R}^n,\mathcal{T})$ mit $\mathcal{T}$ =abgeschlossene Mengen des $\mathbb{R}^n$ ist keine Topologie. Es gelten zwar die Axiome (1) und (3), aber (2) ist verletzt, denn z.B. ist die Vereinigung $\bigcup_{n\in\mathbb{N}}[\frac{1}{n},1] = (0,1]$ nicht abgeschlossen.

10) X metrischer Raum mit Metrik d, $T \in \mathcal{T} :\Longleftrightarrow$ zu jedem $x_0 \in T$ existiert ein $\varepsilon > 0$ sodass $B_\varepsilon(x_0) := \{x \in X : d(x,x_0) < \varepsilon\} \subseteq T$.

11) Der Raum $(X,\mathcal{T})$ mit $X = \{0,1\}$ und $\mathcal{T} = \{\varnothing, \{1\}, X\}$ heisst Sierpinski-Raum. Er ist das einfachste nichttriviale Beispiel eines topologischen Raumes, dessen Topologie nicht von einer Metrik induziert ist.

Beispiele zur Relativtopologie 1) $X = \mathbb{R}$ (versehen mit der üblichen Normtopologie aus Beispiel 1)), $A = [0,1]$, $B = [0,b)$ $(b \in (0,1))$. Die Menge B ist nicht offen, aber relativ offen bezüglich $[0,1]$, denn es gibt eine offene Menge T sodass $B = T \cap [0,1]$, z.B. $B = (-1,b) \cap [0,1]$.

2) $X = \mathbb{R}$ wie in Beispiel 1), $A = \mathbb{Z}$. Die Relativtopologie auf A ist die diskrete Topologie, d.h. $\mathcal{T}_A = \mathcal{P}(\mathbb{Z})$. Für jedes $m \in \mathbb{Z}$ gilt nämlich $\{m\}$ offen, denn $\{m\} = (m-\frac{1}{2}, m+\frac{1}{2}) \cap \mathbb{Z}$, und wegen Axiom (2) folgt hieraus die Offenheit bzgl. $\mathcal{T}_A$ *jeder* Teilmenge $M \subseteq \mathbb{Z}$, denn

$$M = \bigcup_{m\in M} \{m\}.$$

Topologie des $\mathbb{R}^n$. Als nächstes kommen wir zu einer Besonderheit des $\mathbb{R}^n$, nämlich der (uns aus Satz 3.1 bekannten) Äquivalenz aller Normen, und besprechen diese aus topologischer Sicht. Um diese Besonderheit angemessen würdigen zu können, sei betont, dass selbst in "scheinbar einfacheren" Vektorräumen wie $\mathbb{Q}$ *nicht* alle Normen äquivalent sind. (Die Nichtäquivalenz von Normen in unendlichdimen-

sionalen Vektorräumen wie $C([0,T])$ hatten wir schon in Abschnitt 3.3 kennengelernt.)

Beispiel (p-adische Normen auf $\mathbb{Q}$) Sei $X = \mathbb{Q}$ (aufgefasst als $\mathbb{Q}$-Vektorraum), p Primzahl. Die p-adische Norm ist definiert durch

$$|x|_p := \begin{cases} \frac{1}{p^n} & \text{wenn } x = p^n \frac{a}{b},\ a \in \mathbb{Z}\backslash\{0\},\ b \in \mathbb{N},\ n \in \mathbb{Z},\ a,b,p \text{ teilerfremd} \\ 0 & \text{wenn } x = 0. \end{cases}$$

Dies ist wohldefiniert, da sich jede von Null verschiedene rationale Zahl eindeutig in der obigen Form darstellen lässt, und erfüllt die Normaxiome (siehe Übungen). Diese Norm ist klein, wenn der Zähler den Primfaktor p oft enthält, und kann sinnvoll zum Studium von Teilbarkeitseigenschaften rationaler Zahlen benutzt werden. Die p-adische Norm ist hochgradig nichtäquivalent zur üblichen Norm (dem Absolutbetrag), und führt zu bemerkenswerten Konvergenzeigenschaften von Folgen, z.B.

$$|3^k|_3 \to 0\ (k \to \infty), \quad |\tfrac{1}{3^k}|_3 \to \infty\ (k \to \infty), \quad |2^k|_3 = |\tfrac{1}{2^k}|_3 = 1 \text{ für alle } k.$$

Lemma 6.1 Die Normtopologie auf $\mathbb{R}^n$ hängt nicht von der Wahl der Norm ab; d.h. $\|\cdot\|_1$, $\|\cdot\|_2$ Normen $\Longrightarrow \mathcal{T}_{\|\cdot\|_1} = \mathcal{T}_{\|\cdot\|_2}$.

Die von verschiedenen Normen induzierten Topologien sind also nicht nur in irgendeinem schwammigen Sinne "ähnlich", oder in irgendeinem präzisen aber komplizierten Sinne "äquivalent", sondern IDENTISCH!

Beweis Dies folgt aus der Äquivalenz aller Normen auf dem $\mathbb{R}^n$ (Satz 3.1).

Zusammenfassend halten wir fest: Die Äquivalenz aller Normen liefert auf dem $\mathbb{R}^n$ eine natürliche Topologie – *die* Normtopologie.

6.2 Abgeschlossene Mengen, Abschluss, Inneres, Rand

In allgemeinen topologischen Räumen können diese Begriffe allein mithilfe der elementaren Mengenoperationen "Vereinigung", "Schnitt", "Komplement" erklärt werden. Wir erinnern an die De Morgan'schen Regeln für diese Mengenoperationen, die man sich sofort anhand von Skizzen klarmachen kann: falls X beliebige Menge und A, B Teilmengen von X, gilt

$$X\backslash(A \cup B) = (X\backslash A) \cap (X\backslash B),$$
$$X\backslash(A \cap B) = (X\backslash A) \cup (X\backslash B).$$

In Worten: Das Komplement der Vereinigungsmenge ist die Schnittmenge der Komplemente; das Komplement der Schnittmenge ist die Vereinigungsmenge der Komplemente.

Def. 6.2 Sei $(X, \mathcal{T})$ topologischer Raum, $A \subseteq X$.

$$A \textbf{ abgeschlossen} \ :\Longleftrightarrow \ X \backslash A \text{ offen.}$$

Darüber hinaus definieren wir:

$$\begin{aligned} \overline{A} &:= \bigcap_{B \supseteq A,\, B \text{ abgeschlossen}} B \quad \text{(Abschluss von } A\text{)} \\ int\, A &:= \bigcup_{B \subseteq A,\, B \text{ offen}} B \quad \text{(Inneres von } A\text{)} \\ \partial A &:= \overline{A} \setminus int\, A \quad \text{(Rand von } A\text{).} \end{aligned}$$

Der Abschluss ist also die kleinste abgeschlossene Menge, die A enthält; das Innere ist die grösste offene Menge, die in A enthalten ist.

Eine offensichtliche Folgerung aus der Definition von abgeschlossen, den Axiomen (1), (2), (3), und den De Morgan'schen Regeln ist:

(1)' X, $\varnothing$ abgeschlossen
(2)' Beliebige Durchschnitte $\bigcap_{i \in I} A_i$ abgeschlossener Mengen A_i sind abgeschlossen
(3)' Endliche Vereinigungen $\bigcup_{i \in I} A_i$ abgeschlossener Mengen A_i sind abgeschlossen.

Im $\mathbb{R}^n$ (mit der Normtopologie) ist abgeschlossen gemäss obiger Definition äquivalent zur anschaulichen "Fussgängerdefinition" aus Abschnitt 1.5 ($A \subseteq \mathbb{R}^n$ heisst abgeschlossen, wenn für jede Folge (x_n) in A mit $x_n \to x \in \mathbb{R}^n$ gilt: $x \in A$; in Worten: Grenzwerte müssen ebenfalls zur Menge gehören); siehe Satz 1.2. Beispiele zu Abschluss, Innerem und Rand im $\mathbb{R}^n$ hatten wir schon in Abschnitt 1.5 kennengelernt.

Zurück zum Fall beliebiger Teilmengen eines beliebigen topologischen Raumes. Wir behaupten:

Lemma 6.2
a) $\overline{A}$ ist abgeschlossen.
b) $int\, A$ ist offen.
c) ∂A ist abgeschlossen.

Beweis Das Innere $int\, A$ ist laut Definition eine Vereinigung offener Mengen, und daher nach Axiom (2) offen. Um den Abschluss $\overline{A}$ zu untersuchen, benutzen wir zunächst die de Morgan'schen Regeln, um $\overline{A}$ mithilfe einer Vereinigungsmenge statt einer Schnittmenge auszudrücken:

$$\overline{A} = X \setminus \Big(X \backslash \overline{A} \Big) = X \setminus \Big(X \backslash \bigcap_{B \supseteq A,\, \text{B abg.}} B \Big) = X \setminus \Big(\bigcup_{B \supseteq A,\, \text{B abg.}} (X \backslash B) \Big).$$

Den Ausdruck auf der rechten Seite können wir nun analysieren: $X \backslash B$ ist offen (nach Def. von abgeschlossen), also $\bigcup (X \backslash B)$ offen (nach Axiom (2)), und folglich

$X\backslash(\bigcup(X\backslash B))$ abgeschlossen (nach Def. von abgeschlossen). Schliesslich zum Rand ∂A. Es gilt

$$\partial A = \overline{A} \setminus int\, A = \underbrace{\overline{A}}_{\text{abgeschlossen wg. b)}} \cap \underbrace{(X \setminus int\, A)}_{\text{abgeschlossen wg. a)}}$$

und somit ist ∂A, als Durchschnitt zweier abgeschlossener Mengen, abgeschlossen.

Korollar 6.1
a) A abgeschlossen $\Longleftrightarrow A = \overline{A}$; insbesondere $\overline{\overline{A}} = \overline{A}$
b) A offen $\Longleftrightarrow A = int\, A$; insbesondere $int\, int\, A = int\, A$.

Beweis a): "$\Longrightarrow$" ist offensichtlich, denn dann gehört A selbst zu der Familie der Mengen B, als deren Durchschnitt $\overline{A}$ definiert ist. "$\Longleftarrow$": Sei $A = \overline{A}$. Nach Lemma 10.2 ist $\overline{A}$ abgeschlossen, und folglich (wegen $A = \overline{A}$) auch A selbst. b): analog.

6.3 Konvergenz

In einem beliebigen topologischen Raum können wir diesen Begriff wie folgt einführen.

Def. 6.3 Sei $(X, \mathcal{T})$ topologischer Raum, $x \in X$. Eine Teilmenge $U \subseteq X$ heisst **Umgebung** von x, wenn gilt:

$$\text{es gibt eine offene Menge } \Omega \text{ sodass } x \in \Omega \subseteq U.$$

Eine Folge (x_n) in X heisst **konvergent gegen** x, Schreibweise: $x_n \to x$ oder $\lim_{n\to\infty} x_n = x$, wenn gilt:

$$\text{jede Umgebung } U \text{ von } x \text{ enthält alle bis auf endlich viele } x_n.$$

Beispiele 1) Sei X normierter Vektorraum, $\mathcal{T}$ die Normtopologie. Dann ist (x_n) genau dann konvergent gegen x gemäss Def. 6.3, wenn zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ existiert sodass $\|x_n - x\| < \varepsilon$ für alle $n \geq N$ ("Fussgängerdefinition" von Konvergenz aus Analysis 1 für $X = \mathbb{R}$ bzw. aus Analysis 2 für $X = \mathbb{R}^n$).

2) Sei X Menge, $\mathcal{T}$ die diskrete Topologie. Dann ist (x_n) genau dann konvergent, wenn (x_n) stationär ist, d.h. wenn es ein $x \in X$ gibt sodass $x_n = x$ für alle bis auf endlich viele x_n.

3) Sei X Menge, $\mathcal{T}$ die triviale Topologie. Dann konvergiert *jede* Folge gegen *jedes* Element $x \in X$, denn die einzige Umgebung von x ist der ganze Raum X, und dieser enthält alle Folgenglieder. Grenzwerte brauchen also nicht eindeutig zu sein!

4) Sei $(X, \mathcal{T})$ der Sierpinski-Raum (siehe Beispiel 11) oben). Jede Folge konvergiert gegen 0, da X die einzige Umgebung von 0 ist, aber nur diejenigen Folgen mit nur endlich vielen von 1 verschiedenen Gliedern konvergieren gegen 1, da $\{1\}$ Umgebung

von 1 ist.

Beispiele 3) und 4) motivieren die Einführung eines zusätzlichen Axioms, das die Eindeutigkeit von Grenzwerten garantiert:

Satz 6.1 *Sei $(X, \mathcal{T})$ topologischer Raum. Falls $\mathcal{T}$ das Hausdorff'sche Trennungsaxiom erfüllt, d.h.:*

$$\begin{gathered} \textit{Zu je zwei Punkten } x,\, y \in X \textit{ mit } x \neq y \textit{ existieren} \\ \textit{offene Mengen } A \ni x,\, B \ni y \textit{ mit } A \cap B = \varnothing, \end{gathered} \tag{H}$$

sind Grenzwerte eindeutig.

Beweis Sei $x_n \to x$, $x_n \to y$, $x \neq y$. Wegen (H) existieren offene Mengen $A \ni x$, $B \ni y$, mit $A \cap B = \varnothing$. Wegen $x_n \to x$ enthält A alle bis auf endlich viele x_n. Wegen $x_n \to y$ enthält andererseits B alle bis auf endlich viele x_n. Widerspruch.

Lemma 6.3 Sei $\mathbb{K}$ Körper, X normierter $\mathbb{K}$-Vektorraum. Die Normtopologie erfüllt das Hausdorff'sche Trennungsaxiom (H).

Beweis Sei $x \neq y$. Wegen Positivität der Norm ist $\|x - y\| =: \varepsilon > 0$. Dann haben die Mengen $A = B_{\varepsilon/2}(x)$, $B = B_{\varepsilon/2}(y)$ die erforderlichen Eigenschaften, denn sie sind offen und wegen Dreiecksungleichung gilt für beliebige $x' \in B_{\varepsilon/2}(x)$, $y' \in B_{\varepsilon/2}(y)$

$$\|x' - y'\| = \|(x' - x) + (x - y) + (y - y')\| \geq \underbrace{\|x - y\|}_{=\varepsilon} - \Big(\underbrace{\|x - x'\| + \|y - y'\|}_{<\varepsilon/2 \,+\, \varepsilon/2}\Big) > 0,$$

und folglich wegen Positivität der Norm $x' \neq y'$; somit sind A und B disjunkt.

6.4 Kompaktheit

In einem topologischen Raum ist dieser Begriff wie folgt erklärt.

Def. 6.4 Sei X topologischer Raum, $K \subseteq X$. Eine Familie $\{U_i\}_{i \in I}$ von Mengen $U_i \subseteq X$ heisst *Überdeckung* von K, wenn gilt:

$$K \subseteq \bigcup_{i \in I} U_i.$$

Eine Teilüberdeckung von $\{U_i\}_{i \in I}$ ist eine Familie $\{U_i\}_{i \in J}$ mit $J \subseteq I$, die immer noch eine Überdeckung von K ist. K heisst **kompakt**, wenn gilt:

jede offene Überdeckung von K besitzt eine endliche Teilüberdeckung
(*Heine-Borel Eigenschaft*).

Satz 6.2 *Sei $X = \mathbb{R}^n$ mit der Normtopologie, $K \subseteq X$. Dann sind äquivalent:*
(1) Jede offene Überdeckung von K besitzt eine endliche Teilüberdeckung (Heine-Borel Eigenschaft)
(2) Jede Folge in K besitzt einen Häufungspunkt in K (Bolzano-Weierstrass Eigenschaft)
(3) K ist abgeschlossen und beschränkt.

Die Äquivalenz (2) $\Longleftrightarrow$ (3) ist der Satz von Bolzano-Weierstrass (Satz 1.1). Die Methoden der Topologie liefern einen neuen, effizienten Beweis.

Beweis Es reicht zu zeigen: (1) $\Longrightarrow$ (2) $\Longrightarrow$ (3) $\Longrightarrow$ (1).

(1) $\Longrightarrow$ (2): Sei (x_k) Folge in K, $A = \{x_k : k \in \mathbb{N}\}$. Falls A endlich, besitzt (x_k) sogar eine konstante Teilfolge. Sei also A unendlich. Falls (x_k) keinen Häufungspunkt hat, besitzt jedes $x \in K$ eine offene Umgebung $U(x)$ sodass $U(x)$ nur endlich viele Punkte von A enthält. Nun ist aber $\{U(x) : x \in K\}$ offene Überdeckung von K. Wegen (1) existiert eine endliche Teilüberdeckung $\{U(x_i) : i = 1, ..., N\}$ von K. Nach Konstruktion enthält jedes $U(x_i)$ nur endlich viele Elemente von A; damit ist A endlich. Widerspruch.

(2) $\Longrightarrow$ (3): K besitze die Bolzano-Weierstrass Eigenschaft. Wäre K unbeschränkt, gäbe es eine Folge (x_k) in K sodass $|x_{k+1}| \geq |x_k| + 1$ für alle k. Diese Folge besitzt keinen Häufungspunkt. Widerspruch. Nun zur Abgeschlossenheit. Sei (x_k) eine Folge in K mit $x_k \to x \in \mathbb{R}^n$. Wegen Eindeutigkeit von Grenzwerten ist x der einzige Häufungspunkt der Folge, und die Bolzano-Weierstrass Eigenschaft liefert $x \in K$.

(3) $\Longrightarrow$ (1): K sei abgeschlossen und beschränkt. Um (1) zu zeigen, argumentieren wir indirekt. Wir nehmen an, $\{U_i : i \in I\}$ sei eine offene Überdeckung von K, die *keine* endliche Teilüberdeckung besitzt. Da K beschränkt, gibt es einen Würfel W_0 der Kantenlänge s mit $K \subseteq W_0$. Wir zerlegen W_0 in 2^n Teilwürfel der halben Kantenlänge $s/2$ und finden einen Teilwürfel W_1, sodass $K \cap W_1$ nicht von endlich vielen U_i überdeckt wird. Durch Iteration finden wir eine Folge von Würfeln $W_0 \supset W_1 \supset W_2 \supset ...$ sodass W_k Kantenlänge $s/2^k$ hat und folgendes gilt:

$$\text{kein } K \cap W_k \text{ wird von endlich vielen } U_i \text{ überdeckt.} \qquad (*)$$

Wir wählen nun Elemente $x_k \in K \cap W_k$. Nach Konstruktion ist (x_k) Cauchyfolge. Wir benutzen nun entscheidend die Vollständigkeit von $\mathbb{R}^n$. Diese liefert die Existenz eines $x \in \mathbb{R}^n$ sodass $x_k \to x$. Wegen K abgeschlossen folgt $x \in K$. Da $\{U_i : i \in I\}$ Überdeckung von K, existiert ein U_{i_0} mit $U_{i_0} \ni x$. Da U_{i_0} offen, liegen alle bis auf endlich viele W_k in U_{i_0}. Dies widerspricht (*). Damit ist Satz 6.2 bewiesen.

Wie unser Beweis zeigt, gilt die Implikation (1) $\Longrightarrow$ (2) in *beliebigen* topologischen Räumen. Man kann zeigen, dass die Umkehrung (2) $\Longrightarrow$ (1) immerhin in beliebigen Banachräumen (d.h. vollständigen aber nicht notwendig endlichdimensionalen $\mathbb{R}$-Vektorräumen, siehe Def. 3.3) gilt; der Beweis ist aber wesentlich komplizierter

als im $\mathbb{R}^n$. Die Implikation (2) $\Longleftrightarrow$ (3) beruht hingegen entscheidend auf der Endlichdimensionalität des $\mathbb{R}^n$, und ist in jedem unendlichdimensionalen normierten $\mathbb{R}$-Vektorraum falsch. Ein explizites Beispiel einer abgeschlossenen und beschränkten, aber weder folgen- noch überdeckungskompakten Menge sind die *trigonometrischen Monome* $x \mapsto (e^{ix})^k$ im Vektorraum $X = C([-\pi, \pi])$ der stetigen Funktionen auf dem Intervall $[-\pi, \pi]$ versehen mit der L^2 Norm,

$$K = \{e_k : k \in \mathbb{Z}\}, \;\; e_k(x) = \frac{e^{ikx}}{\sqrt{2\pi}},$$

denn wegen der Orthogonalität der e_k gilt nach Pythagoras $||e_k - e_\ell||_2 = \sqrt{2}$ für alle $k \neq \ell$. Somit enthält z.B. die offene Überdeckung von K bestehend aus den Einheitskugeln um jeden Punkt, $\{B_1(e_k) : k \in \mathbb{Z}\}$, keine endliche Teilüberdeckung, und die Folge $\{e_k\}_{k\in\mathbb{N}}$ besitzt keine konvergente Teilfolge.

6.5 Stetigkeit

Von einem abstrakten mathematischen Standpunkt aus sind "stetige Abbildungen" zwischen topologischen Räumen das Analogon von "Gruppenhomomorphismen" zwischen Gruppen, oder "linearen Abbildungen" zwischen Vektorräumen. Wie kommt man auf diese Klassen von Abbildungen? Wenn die zugrundeliegenden Mengen zusätzliche Struktur besitzen (z.B. Gruppe, Vektorraum, topologischer Raum), ist es natürlich, Abbildungen zu betrachten, die mit dieser Struktur in geeigneter Weise kompatibel sind. Im Falle topologischer Räume ist die zusätzliche Struktur die Topologie $\mathcal{T}$, Abbildungen sollten also kompatibel mit den Topologien auf Definitions- und Wertebereich sein.

Def. 6.5 (Stetig) Seien $(X, \mathcal{T})$, $(X', \mathcal{T}')$ topologsiche Räume.
a) $f : X \to X'$ heisst stetig, wenn gilt:

die Urbilder offener Mengen sind offen.

b) Sei $A \subset X$. $f : A \to X'$ heisst stetig, wenn gilt:

die Urbilder offener Mengen sind offen bezüglich der Relativtopologie $\mathcal{T}_A$.

Im $\mathbb{R}^n$ bedeutet diese abstrakte Eigenschaft nichts anderes als die anschauliche Eigenschaft der Kompatibilität von Funktionsanwendung mit Grenzwertbildung. Genauer:

Lemma 6.4 Für $X = \mathbb{R}^n$, $X' = \mathbb{R}^m$, $A \subseteq \mathbb{R}^n$ gilt: $f : A \to \mathbb{R}^m$ stetig genau dann wenn gilt:

für jede Folge (x_n) in A mit $x_n \to x \in A$ gilt $f(x_n) \to f(x)$. (F)

In allgemeinen topologischen Räumen ist für die Eigenschaft (F) die Bezeichnung "Folgenstetigkeit" üblich, um sie von der in Def. 6.5 geforderten Eigenschaft zu unterscheiden.

Beweis Für $A = \mathbb{R}^n$ ist dies genau das abstrakte ε-δ-Kriterium aus Satz 1.4. Der Beweis für $A \subset \mathbb{R}^n$ verläuft analog.

Zurück zu allgemeinen topologischen Räumen. Aus den De Morgan'schen Regel folgt sofort, unter den Voraussetzungen von Def. 6.5:
a)' $f : X \to X'$ stetig $\Longleftrightarrow$ die Urbilder abgeschlossener Mengen sind abgeschlossen
b)' $f : A \to X'$ stetig $\Longleftrightarrow$ die Urbilder abgeschlossener Mengen sind abgeschlossen bezüglich der Relativtopologie $\mathcal{T}_A$.

Auch in allgemeinen topologischen Räumen bleibt Stetigkeit unter Verkettung erhalten.

Satz 6.3 *Die Verkettung stetiger Funktionen ist stetig, d.h. wenn X, Y, Z topogische Räume und $f : Y \to Z$ und $g : X \to Y$ stetig, so ist auch die durch $h(x) := f(g(x))$ definierte Abbildung $h : X \to Z$ stetig.*

Beweis Dies folgt sofort aus $h^{-1}(A) = g^{-1}\big(f^{-1}(A)\big)$.

Zur mathematischen Allgemeinbildung gehört folgende einfache Aussage über das Zusammenspiel zwischen Kompaktheit und Stetigkeit.

Satz 6.4 *Seien X, X' topologische Räume, $K \subseteq X$ kompakt, $f : X \to X'$ stetig. Dann ist das Bild $f(K)$ kompakt.*

In Worten: Bilder kompakter Mengen unter stetigen Funktionen sind kompakt.

Beweis Sei $\{U_i\}_{i\in I}$ offene Überdeckung von $f(K)$. Dann ist $\{f^{-1}(U_i)\}_{i\in I}$ offene Überdeckung von K. Da K kompakt, gibt es eine endliche Teilüberdeckung, d.h. $K \subseteq \big(f^{-1}(U_{i_1}) \cup \ldots \cup f^{-1}(U_{i_N})\big)$. Folglich $f(K) \subseteq \big(U_{i_1} \cup \ldots \cup U_{i_N}\big)$.

Als Korollar verallgemeinern wir den Satz vom Maximum und Minimum (Satz 1.5) auf kompakte topologische Räume.

Korollar 6.2 Sei X topologischer Raum, $K \subseteq X$ kompakt, $f : X \to \mathbb{R}$ stetig. Dann besitzt f auf K eine Maximums- und eine Minimumsstelle.

Beweis Das Bild $f(K)$ ist kompakt, also nach Satz 6.2 abgeschlossen und beschränkt. Wegen der Beschränktheit existieren Supremum $M := \sup f(K) \in \mathbb{R}$ und Infimum $m := \inf f(K) \in \mathbb{R}$. Wegen der Abgeschlossenheit liegen M und m in $f(K)$. Deren Urbilder sind Maximums- bzw. Minimumsstellen.

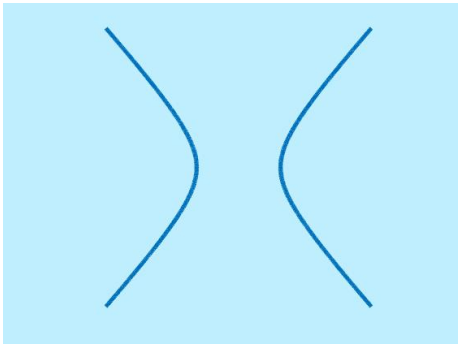
6.6 Zusammenhängend

Anschaulich weiss jeder, dass man ein Blatt Papier, ein Gummiband, ein Seil durch stetiges Verformen nicht in zwei Teile zerreissen kann. *Aber was heisst das eigent-*

lich – ein Objekt besteht nur aus einem Teil? Bemerkenswerterweise kann man diese Eigenschaft mathematisch präzise formulieren; die Tatsache, dass sie unter stetigen Abbildungen erhalten bleibt, wird dann zu einem mathematischen Satz. Wir formulieren die Eigenschaft und den Satz nicht nur im Spezialfall anschaulicher Mengen im dreidimensionalen Raum (Blatt Papier, Gummiband, Seil), sondern gleich ohne Mehraufwand in allgemeinen topologischen Räumen.

Def. 6.6 (Zusammenhängend) Sei $(X, \mathcal{T})$ topologischer Raum.
a) X heisst **zusammenhängend**, wenn es keine Zerlegung $X = U \cup V$ gibt mit U, V offen, disjunkt, nichtleer.
b) Eine Teilmenge $A \subseteq X$ heisst zusammenhängend, wenn A zusammenhängend bezüglich der Relativtopologie $\mathcal{T}_A$.

Beispiel 1) $A = \{(x, y) \in \mathbb{R}^2 : x^2 - y^2 = 1\}$. A ist eine Hyperbel, siehe Skizze.



Diese Menge ist nicht zusammenhängend, denn

$$\mathbb{R}^2 \cap A = \underbrace{\{(x, y) \in \mathbb{R}^2 : x < 0\} \cap A}_{=:U \text{ (offen bzgl. } \mathcal{T}_A)} \cup \underbrace{\{(x, y) \in \mathbb{R}^2 : x > 0\} \cap A}_{=:V \text{ (offen bzgl. } \mathcal{T}_A)}.$$

Um zu zeigen, dass die beiden Teilstücke U und V nicht weiter zerlegt werden können, also zusammenhängend sind, braucht man etwas Theorie.

Satz 6.5 *Sei $A \subseteq \mathbb{R}$ mit mindestens zwei Elementen. Dann gilt:*

$$A \text{ zusammenhängend} \iff A \text{ Intervall.}$$

Diese Aussage klärt den in Analysis 1 ein wenig ad hoc eingeführten Begriff des Intervalls (Wiederholung: ein Intervall ist eine Teilmenge I von $\mathbb{R}$, die mindestens zwei Punke enthält und für die gilt: wenn $x, y \in I$, dann folgt auch $z \in I$ für alle $z \in [x, y]$. Dies sind die Mengen $[a, b]$ oder $[a, b)$ oder $(a, b]$ oder (a, b) oder $(-\infty, b)$ oder (a, ∞) oder $[a, \infty)$ oder $(-\infty, b]$ oder $\mathbb{R}$).

Den Beweis besprechen wir weiter unten.

Satz 6.6 *X, X' topologische Räume, X zusammenhängend, $f : X \to X'$ stetig $\Longrightarrow$ $f(X)$ zusammenhängend.*

Beweis Anderenfalls existieren nichtleere $f(X)$-offene disjunkte U', V' mit

$$f(X) = U' \cup V'.$$

Hieraus folgt eine analoge Zerlegung von X, nämlich

$$X = f^{-1}(U') \cup f^{-1}(V').$$

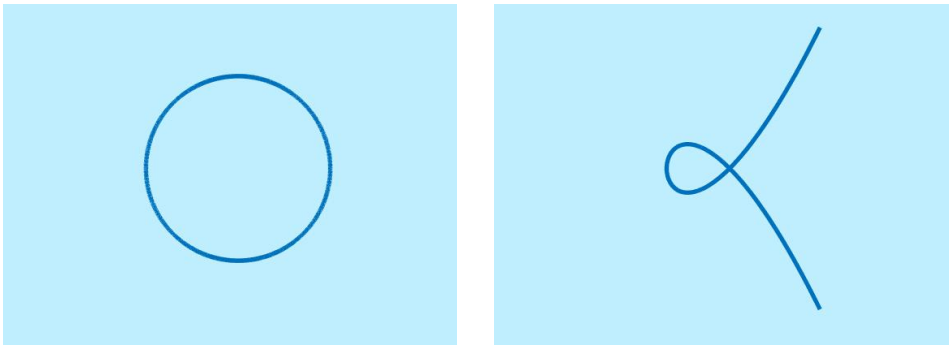
Widerspruch zu X zusammenhängend.

Beispiel 1), Fortsetzung Wir können jetzt zeigen, dass die beiden Hyperbeläste U und V zusammenhängend sind. Wir betrachten z.B. V. Wir können V als Graph über der y-Achse darstellen, indem wir die definierende Gleichung nach x auflösen: $V = \{(x, y) \in \mathbb{R}^2 : x = \sqrt{y^2 + 1}\}$. Folglich gilt

$$V = \varphi(\mathbb{R}) \quad \text{mit } \varphi : \mathbb{R} \to \mathbb{R}^2,\ \varphi(y) = \begin{pmatrix} \sqrt{y^2+1} \\ y \end{pmatrix}.$$

V ist also Bild einer gemäss Satz 6.4 zusammenhängenden Menge unter der stetigen Abbildung φ, und damit nach Satz 6.6 zusammenhängend.

Beispiel 2) $B = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$, d.h. B Kreislinie (links); $C = \{(x, y) \in \mathbb{R}^2 : x^2 + x^3 - y^2 = 0\}$ Seil mit Schlaufe (rechts).



Wir behaupten: B und C sind zusammenhängend. Zunächst zur ersten Menge. B ist gleich $\varphi(\mathbb{R})$ für die stetige Abbildung

$$\varphi : \mathbb{R} \to \mathbb{R}^2,\ \ \varphi(t) = (\cos t, \sin t),$$

also das Bild einer gemäss Satz 6.5 zusammenhängenden Menge unter einer stetigen Abbildung, und damit nach Satz 6.6 zusammenhängend. Eine weniger elegante, aber elementarere Darstellung von B als Bild eines Intervalls unter einer stetigen

Abbildung erhält man, indem man die Gleichung zunächst nach einer Variablen auflöst, z.B. nach y. Wegen $x^2 + y^2 = 1$ gilt $x^2 \leq 1$, also $x \in [-1, 1]$. Auflösen nach y liefert $B = \{(x, y) \in \mathbb{R}^2 : x \in [-1, 1],\ y = \pm\sqrt{1 - x^2}\}$. Indem man beide Zweige der Lösung nacheinander durchläuft, also z.B. $x = |t| - 1$ mit $t \in [-2, 2]$ setzt, erhält man $B = \varphi([-2, 2])$ mit

$$\varphi(t) = \begin{cases} \left(|t| - 1,\ \sqrt{1 - (|t| - 1)^2}\right) & t \geq 0 \\ \left(|t| - 1,\ -\sqrt{1 - (|t| - 1)^2}\right) & t < 0. \end{cases}$$

Nun zur zweiten Menge. Da $y^2 \geq 0$, gilt für alle Punkte in C: $x^2 + x^3 \geq 0$, folglich (wegen $x^2 + x^3 = x^2(1 + x)$) $x \geq -1$. Auflösen nach y ergibt $C = \{(x, y) \in \mathbb{R}^2 : x \geq -1,\ y = \pm\sqrt{x^2 + x^3}\}$. Indem wir wie oben beide Zweige der Lösung nacheinander durchlaufen, also z.B. $x = |t| - 1$ mit $t \in \mathbb{R}$ setzen, erhalten wir wiederum eine Darstellung der Lösungsmenge als Bild $\varphi(\mathbb{R})$ von $\mathbb{R}$ unter einer stetigen Abbildung:

$$\varphi(t) = \begin{cases} \left(|t| - 1,\ \sqrt{(|t| - 1)^2 + (|t| - 1)^3}\right), & t \geq 0 \\ \left(|t| - 1,\ -\sqrt{(|t| - 1)^2 + (|t| - 1)^3}\right), & t < 0. \end{cases}$$

Wir haben hier absichtlich eine elementare Parametrisierung hergeleitet, die für unsere Zwecke ausreicht. Eleganter wäre z.B. $\varphi : \mathbb{R} \to \mathbb{R}^2$, $\varphi(t) = (t^2 - 1, t^3 - t)$, aber systematisch auf solche Formeln zu kommen liegt jenseits der Methoden von Analysis 2 (dazu bräuchte man etwas algebraische Geometrie).

Beweis von Satz 6.5. Dies beruht auf speziellen Eigenschaften von $\mathbb{R}$, nämlich den Anordnungsaxiomen sowie dem Vollständigkeitsaxiom.

Das folgende Korollar ist unter dem Namen "verallgemeinerter Zwischenwertsatz" bekannt und zeigt, dass die Eigenschaft eines topologischen Raumes, zusammenhängend zu sein, die Lösbarkeit gewisser Gleichungen der Form $f(x) = c$, $f : X \to \mathbb{R}$, nach sich zieht.

Korollar 6.3 (Verallgemeinerter Zwischenwertsatz) Sei X topologischer Raum, X zusammenhängend, $f : X \to \mathbb{R}$ stetig, $a, b \in X \Longrightarrow f$ nimmt jeden Wert zwischen $f(a)$ und $f(b)$ an.

Beweis Im Fall $f(a) \neq f(b)$ ist $f(X)$ wegen Satz 6.6 zusammenhängend, und somit wegen Satz 6.5 ein Intervall.

Wichtig: vom praktischen Standpunkt aus ist Korollar 6.3 allerdings bei weitem keine gleichwertige, sondern nur eine sehr schwache Verallgemeinerung des ZWS. Sei nämlich X eine offene zusammenhängende Teilmenge des $\mathbb{R}^n$, $n \geq 2$. Dann besagt Korollar 6.3 nur, dass *eine* Gleichung mit *mehr als einer* Unbekannten mindestens eine Lösung besitzt. Dies ist weit davon entfernt, die gesamte Lösungsmenge zu verstehen oder zu charakterisieren. Eine zwar nur lokale, aber viel stärkere Aussage ist

der *Satz über inverse Funktionen* (Satz 4.1), der unter geeigneten Voraussetzungen die Lösbarkeit von n Gleichungen mit n Unbekannten garantiert.

Eine weitere interessante Folgerung aus "zusammenhängend" ist die folgende Charakterisierung reellwertiger Funktionen mehrerer Variablen, deren Gradient gleich Null ist.

Satz 6.7 *Sei $\Omega \subseteq \mathbb{R}^n$ offen. Dann gilt:*

$$\Omega \text{ zusammenhängend} \iff \begin{array}{l}\text{jede diff'bare Funktion } f : \Omega \to \mathbb{R} \text{ mit} \\ \nabla f(x) = 0 \text{ für alle } x \in \Omega \text{ ist konstant.}\end{array}$$

Beweis "$\Longleftarrow$": Falls Ω nicht zusammenhängend, ist $\Omega = U \cup V$ für offene disjunkte nichtleere Mengen U und V. Die nichtkonstante Funktion

$$f(x) = \begin{cases} 1 & x \in U \\ 0 & x \in V \end{cases}$$

ist offensichtlich differenzierbar, mit $\nabla f(x) = 0$ für alle $x \in \Omega$.

"$\Longrightarrow$": Diese zwar anschaulich plausible aber nichttriviale Aussage können wir mit unseren bisher entwickelten topologischen Methoden elegant und effizient beweisen. Die Beweismethode ist genauso interessant wie die Aussage selbst, und kann vielfältig verwendet werden. Sie ist eine Art *kontinuierliches Analogon zu vollständiger Induktion:* man beweist eine Aussage über Punkte im $\mathbb{R}^n$ durch sukzessive Vergrösserung des Gültigkeitsbereiches, genau wie man eine Aussage über natürliche Zahlen durch sukzessive Vergrösserung $n \to n+1$ ihres Gültigkeitsbereiches beweist. Sei $x_0 \in \Omega$, $c := f(x_0)$. Wir betrachten die Menge $M = \{x \in \Omega : f(x) = c\}$. M ist nichtleer, da $x_0 \in M$. Wir behaupten: M ist offen. Sei nämlich $x \in M$. Da Ω offen, gibt es eine Kugel $B_\varepsilon(x) \subseteq \Omega$, $\varepsilon > 0$. Für $y \in B_\varepsilon(x)$ betrachten wir die Funktion $c : [0,1] \to B_\varepsilon(x)$, $c(t) = (1-t)x + ty$. Nach Hauptsatz und Kettenregel (siehe Lemma 2.1 für den hier ausreichenden Spezialfall) gilt

$$f(\underbrace{c(1)}_{=y}) - f(\underbrace{c(0)}_{=x}) = \int_0^1 \tfrac{d}{dt} f(c(t))\,dt = \int_0^1 \Big\langle \underbrace{\nabla f(c(t))}_{=0}, c'(t) \Big\rangle\,dt = 0,$$

d.h. $y \in M$. Dies beweist die Offenheit von M. Andererseits ist M Ω-abgeschlossen, denn M ist das Urbild der abgeschlossenen Menge $\{c\}$ unter der stetigen Funktion f (beachte: der Gradient von f ist Null, insbesondere stetig, und damit die Implikation "stetig diff'bar $\Longrightarrow$ stetig" aus Abschnitt 2.2 anwendbar). Also ist auch das Komplement $\Omega \backslash M$ Ω-offen. Da Ω zusammenhängend ist, muss $\Omega \backslash M$ leer sein, denn sonst hätten wir Ω in zwei offene disjunkte nichtleere Mengen zerlegt.

6.7 Was ist eine topologische Eigenschaft (oder topologische Invariante)?

Informell versteht man darunter eine Eigenschaft, die unter "stetigen Deformationen" erhalten bleibt. Z.B. ist die genaue Gestalt einer Kurve im $\mathbb{R}^3$ (stellen Sie sich ein Drahtstück vor) durch stetige Deformation (Verbiegen) veränderbar, nicht aber die Tatsache, dass die Kurve nur aus einem einzigen Stück besteht. Diese Idee formalisieren wir wie folgt.

Definition 6.7 Seien X, X' topologische Räume. Ein *Homöomorphismus* ist eine bijektive, in beide Richtungen stetige Abbildung, d.h. eine Abbildung $f : X \to X'$ sodass f bijektiv, f stetig, f^{-1} stetig. Existiert ein solcher Homöomorphismus, heissen X und X' homöomorph.

Beispiel Die Abbildung $f : x \mapsto e^x$ ist Homöomorphismus zwischen $\mathbb{R}$ und $(0, \infty)$; $g : y \mapsto \frac{y}{1+y}$ ist Homöomorphismus zwischen $(0, \infty)$ und $(0, 1)$; die Verkettung $g \circ f : x \mapsto e^x/(1 + e^x)$ ist Homöomorphismus zwischen $\mathbb{R}$ und $(0, 1)$. Insbesondere ist die unbeschränkte Menge $\mathbb{R}$ homöomorph zur beschränkten Menge $(0, 1)$.

Definition 6.8 Eine *topologische Eigenschaft* (oder *topologische Invariante*) ist eine Eigenschaft eines topologischen Raumes, die unter beliebigen Homöomorphismen erhalten bleibt.

Zwei solche Eigenschaften haben wir schon kennengelernt: *kompakt* und *zusammenhängend*. Mithilfe topologischer Invarianten kann man z.B. untersuchen, ob zwei gegebene topologische Räume homöomorph sind. Unterscheiden sie sich bzgl. einer Invariante, können sie nicht homöomorph sein.

Beispiel 1 Die Einheitskreislinie $S^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ und das Winkelintervall $[0, 2\pi)$ sind nicht homöomorph, denn S^1 ist kompakt, aber $[0, 2\pi)$ nicht. (Beispiel zum Beispiel: die Polarkoordinatenabbildung $\varphi \mapsto (\cos\varphi, \sin\varphi)$ von $[0, 2\pi)$ nach S^1 ist bijektiv und stetig, die Umkehrabbildung muss also wegen der Nichthomöomorphie beider Mengen unstetig sein. In der Tat ist sie unstetig im Punkt $(1, 0) \in S^1$.)

Beispiel 2 Keine der Mengen $|$, $\bigcirc$ und $=$ aus dem einführenden Beispiel sind zueinander homöomorph. Die mittlere Menge (also die Einheitskreislinie) ist kompakt, die beiden anderen aber nicht. Die linke Menge ist zusammenhängend, denn sie ist das Bild von $\mathbb{R}$ unter der stetigen Abbildung $x \mapsto (0, x)$, aber die rechte Menge $D = (0, 1) \times \{0, 1\}$ nicht, denn

$$\mathbb{R}^2 \cap D = \underbrace{\{(x, y) \in \mathbb{R}^2 : y > \tfrac{1}{2}\} \cap D}_{=:U \text{ (offen bzgl. } \mathcal{T}_D)} \cup \underbrace{\{(x, y) \in \mathbb{R}^2 : y < \tfrac{1}{2}\} \cap D}_{=:V \text{ (offen bzgl. } \mathcal{T}_D)}.$$

Eine bessere Invariante als zusammenhängend, an der wir insbesondere ablesen

können, ob ein topologischer Raum zusammenhängend ist, ist die Anzahl Zusammenhangskomponenten oder 0-*te Betti-Zahl*

$$b_0(X) := \sup\Big\{ \sharp\, I \,:\, \{U_i\}_{i\in I} \text{ ist eine Zerlegung von } X \text{ in offene, disjunkte, nichtleere Mengen}\Big\}.$$

Hierbei bedeutet $\sharp\, I \in \{0\} \cup \mathbb{N} \cup \{+\infty\}$ die Anzahl Elemente einer Menge I. Zusammehängende Mengen $X \cap U_i$, U_i offen, heissen Zusammenhangskomponenten von X. Offenbar gilt:

$$X \text{ zusammenhängend} \iff b_0(X) = 1.$$

Unsere Ergebnisse in Beispiel 1) und 2) für die Hyperbel A und den Kreis B lassen sich mithilfe der 0-ten Betti-Zahl elegant formulieren als

$$b_0(A) = 2; \quad b_0(B) = 1.$$

Mithilfe von b_0 können wir auch den noch offenen Teil der eingangs gestellten Frage (ist $+$ homöomorph zu irgendeiner der anderen drei Mengen?) aufklären: wäre nämlich f ein entsprechender Homöomorphismus, so wäre die Einschränkung $\tilde{f}$ von f auf $+$ ohne den Mittelpunkt ein Homöomorphismus auf eine der anderen drei Mengen ohne einen Punkt; dies widerspräche aber der Tatsache, dass $\tilde{f}$ die Invariante b_0 erhalten muss. Details siehe Übungen.

Unser Beweis von Satz 6.7 für offene Teilmengen Ω des $\mathbb{R}^n$ liefert auch ein interessantes Ergebnis wenn Ω nicht zusammenhängend, nämlich

$$b_0(\Omega) = \dim \operatorname{Ker} \nabla$$

(wobei $\nabla : \{f : \Omega \to \mathbb{R} : f \text{ diff'bar}\} \to \{v : \Omega \to \mathbb{R}^n\}$). Dies ist eine tiefliegende Charakterisierung von $b_0(\Omega)$ als Dimension der Lösungsmenge des Kerns eines Differentialoperators, und ein Spezialfall einer sehr allgemeinen Aussage namens *Satz von de Rham*.

Die Invariante b_0 ist nur die erste in einer ganzen Serie topologischer Invarianten b_k, $k \in \mathbb{N} \cup \{0\}$ (k-te Betti-Zahl). Die erste Betti-Zahl b_1 zählt anschaulich gesehen die Anzahl Löcher der Kodimension 2 in einer zusammenhängenden Menge, z.B.

$$b_1(\mathbb{R}^2) = 0,$$
$$b_1(\mathbb{R}^2 \backslash \{0\}) = 1,$$
$$b_1(\mathbb{R}^2 \backslash \{a_1, ..., a_N\}) = N \text{ für paarweise verschiedene } a_1, ..., a_N \in \mathbb{R}^2.$$

Wie man b_1 mathematisch präzise definiert, erklären wir hier nicht. Die Konstruktion solcher Invarianten und das systematische Studium invarianter Eigenschaften topologischer Räume ist Gegenstand der *Algebraischen Topologie*.

7 Kurvenintegral

Die vielleicht einfachste der vielen Verallgemeinerungen des Integrals $\int_a^b f$ einer 1D Funktion $f : [a,b] \to \mathbb{R}$ ist das Integral eines Vektorfeldes über eine Kurve. Hier bleibt der Definitionsbereich (über den wir integrieren) eindimensional, und Definition und wesentliche Eigenschaften ergeben sich ohne grosse Mühe durch Verbinden des Riemann-Integrals aus Analysis 1 mit unseren Kenntnissen über Vektorfelder und Kurven aus Kapiteln 1 und 2.[15] Das Kurvenintegral erlaubt die Bestimmung der Länge einer Kurve, und ermöglicht ein tieferes Verständnis der beiden Ableitungsoperationen "Gradient" und "Rotation".

Def. 7.1 Eine **reguläre Kurve** im $\mathbb{R}^n$ ist eine stetig differenzierbare Abbildung $\gamma : [a,b] \to \mathbb{R}^n$ mit $\gamma'(t) \neq 0 \ \forall t \in [a,b]$.

Oft wird das Bild $\Gamma = \gamma([a,b]) \subset \mathbb{R}^n$ einer Kurve ebenfalls Kurve genannt. Reguläre Kurven müssen nicht injektiv sein; sind sie es, hängt der Tangentialvektor $\gamma'(t)$ an Γ im Punkt $\gamma(t)$ nur vom Fusspunkt $\gamma(t)$ ab.

Def. 7.2 Sei $\gamma : [a,b] \to \mathbb{R}^n$ reguläre Kurve, $v : \gamma([a,b]) \to \mathbb{R}^n$ stetiges Vektorfeld. Dann heisst

$$\int_\gamma v \cdot ds := \int_a^b \big\langle v(\gamma(t)), \gamma'(t) \big\rangle dt$$

Kurvenintegral von v über γ.

Hier ist $\langle \cdot, \cdot \rangle$ das euklidische Skalarprodukt im $\mathbb{R}^n$. Man integriert also das Skalarprodukt zwischen Vektorfeld und Tangentialvektor.

Physikalische Bedeutung des Kurvenintegrals: $-\int_\gamma v \cdot ds$ = Arbeit, die verrichtet werden muss, um ein Teilchen im Kraftfeld v entlang γ zu transportieren. Als Beispiel kann man sich Schwimmen in einem Strömungsfeld vorstellen. Entlang der Trajektorie γ gegen die Strömung schwimmen ($\gamma' \cdot v < 0$) kostet Arbeit (Kurvenintegral < 0); mit der Strömung schwimmen ($\gamma' \cdot v > 0$) spart Arbeit (Kurvenintegral > 0).

Der Clou bei dieser Definition ist die Invarianz unter Umparametrisierung.

Def. 7.3 Sei $\gamma : [a,b] \to \mathbb{R}^n$ reguläre Kurve, $\varphi : [a',b'] \to [a,b]$ bijektiv und stetig diff'bar mit $\varphi'(s) \neq 0 \ \forall s$, und

$$\tilde{\gamma} := \gamma \circ \varphi : [a',b'] \to \mathbb{R}^n.$$

[15] Integration über mehrdimensionale Gebiete (insbesondere offene und beschränkte Teilmengen des $\mathbb{R}^n$) lernen Sie erst in Analysis 3 kennen. Dies beruht darauf, dass der Riemann'sche Zugang (...Integrationsgebiet in Intervalle zerlegen...) keine offensichtliche Verallgemeinerung besitzt (...Quader? Simplizes? Welcher Grösse? Aber schon für einfache Gebiete wie Kreisscheiben oder Kugeln kommt man ja dann nie mit endlich vielen Quadern oder Simplizes aus...) Wenn man auch im Mehrdimensionalen einen schönen Integrationskalkül mit Analoga von Hauptsatz, Substitutionsregel etc. entwickeln möchte, ist der wesentlich abstraktere aber schlagkräftige Zugang des *Lebesgue-Integrals* sinnvoll.

Dann heisst $\tilde{\gamma}$ **Umparametrisierung** von γ. Ist $\varphi' > 0$ (bzw. < 0), heisst $\tilde{\gamma}$ orientierungserhaltend (bzw. orientierungsumkehrend).

Lemma 7.1 Sei $\gamma : [a,b] \to \mathbb{R}^n$ reguläre Kurve, $\tilde{\gamma} : [a',b'] \to \mathbb{R}^n$ Umparametrisierung von γ, $v : \gamma([a,b]) \to \mathbb{R}^n$ stetiges Vektorfeld. Dann gilt:

$$\int_{\tilde{\gamma}} v \cdot ds = \begin{cases} \int_\gamma v \cdot ds & \text{falls } \tilde{\gamma} \text{ orientierungserhaltend} \\ -\int_\gamma v \cdot ds & \text{falls } \tilde{\gamma} \text{ orientierungsumkehrend} \end{cases}$$

Für injektive reguläre Kurven γ hängt das Kurvenintegral also bis auf ein Vorzeichen nur von der Bildmenge $\Gamma := \gamma([a,b])$ ab, man schreibt deshalb oft

$$\int_\Gamma v \cdot ds$$

statt $\int_\gamma v \cdot ds$. Diese Schreibweise unterdrückt aber, dass zwecks Festlegung des Vorzeichens zusätzlich eine "Durchlaufrichtung", alias Orientierung, von Γ vorgegeben werden muss.

Beweis Wir verwenden die Substitution $t = \varphi(s)$, $dt = \varphi'(s)ds$ sowie die Tatsache, dass nach Kettenregel $\tilde{\gamma}'(s) = \gamma'(\varphi(s))\varphi'(s)$. Sei zunächst φ orientierungserhaltend, dann ist $\varphi(a') = a$, $\varphi(b') = b$. Berechne

$$\begin{aligned} \int_\gamma v \cdot ds &\underset{\text{Def. des Kurvenint.}}{=} \int_a^b \big\langle v(\gamma(t)), \gamma'(t) \big\rangle \, dt \\ &\underset{\text{Substitutionsregel}}{=} \int_{a'}^{b'} \big\langle v(\gamma(\varphi(s))), \gamma'(\varphi(s)) \big\rangle \varphi'(s) \, ds \\ &\underset{\text{Kettenregel}}{=} \int_{a'}^{b'} \big\langle v(\tilde{\gamma}(s)), \tilde{\gamma}'(s) \big\rangle \, ds \underset{\text{Def. des Kurvenint.}}{=} \int_{\tilde{\gamma}} v \cdot ds. \end{aligned}$$

Falls φ orientierungsumkehrend, ist $\varphi(a') = b$, $\varphi(b') = a$, und die Behauptung folgt aus der Eigenschaft des Riemann-Integrals, dass $\int_{b'}^{a'} f = -\int_{a'}^{b'} f$ (siehe Analysis 1 Abschnitt 11).

Beispiel 1 (Länge einer Kurve) Sei die reguläre Kurve γ injektiv. Wähle als Vektorfeld v das Einheitstangentenfeld

$$v(\gamma(t)) := \frac{\gamma'(t)}{|\gamma'(t)|}.$$

Dann ist

$$\int_\gamma v \cdot ds = \int_a^b \Big\langle \frac{\gamma'(t)}{|\gamma'(t)|}, \gamma'(t) \Big\rangle \, dt = \int_a^b |\gamma'(t)| dt.$$

Dieses Integral heisst (und kann geometrisch interpretiert werden als) **Länge** (oder Bogenlänge) der Kurve. Wegen Lemma 7.1 hängt es nicht von der Wahl der Parametrisierung ab.

Spezialfall: $n = 2$, $f : [a,b] \to \mathbb{R}$ stetig diff'bar,

$$\gamma(t) = \begin{pmatrix} t \\ f(t) \end{pmatrix} \quad (t \in [a,b]).$$

Dann ist das Bild $\gamma([a,b]) = \operatorname{graph} f$, und

$$\int_a^b |\gamma'(t)|\,dt = \int_a^b \left|\begin{pmatrix} 1 \\ f'(t) \end{pmatrix}\right| dt = \int_a^b \sqrt{1 + f'(t)^2}\,dt,$$

d.h. wir erhalten gerade die Formel für die Länge von $\operatorname{graph} f$ aus Analysis 1 Abschnitt 11.6.

Die geometrische Interpretation des obigen Integrals als Länge erklärt sich wie folgt. Wir teilen (wie bei der Diskussion der Länge eines Graphen in Analysis 1) das Intervall $[a,b]$ in N Teilintervalle $[t_{j-1}, t_j]$ $(j = 1, ..., N)$ mit $t_j = a + jh$, $h = \frac{b-a}{N}$, und approximieren γ durch den Streckenzug γ_N, der aus den geraden Verbindungsstrecken zwischen $\gamma(t_{j-1})$ und $\gamma(t_j)$ besteht, d.h.

$$\gamma_N\Big((1-s)t_{j-1} + st_j\Big) := (1-s)\gamma(t_{j-1}) + s\gamma(t_j).$$

Die elementargeometrische Länge von γ_N ist

$$L(\gamma_N) = \sum_{j=1}^N |\gamma(t_j) - \gamma(t_{j-1})| = \sum_{j=1}^N \Big|\frac{1}{h}\Big(\gamma(t_j) - \gamma(t_{j-1})\Big)\Big| h \approx \sum_{j=1}^N |\gamma'(t_j)|\,h \to \int_a^b |\gamma'(t)|\,dt,$$

wobei wir in der vorletzten Zeile den Differenzenquotienten durch die Ableitung von γ an der Stelle t_j approxmiert und die so entstandene Summe als *Riemann-Summe* für das Integral auf der rechten Seite erkannt haben. Die Konvergenz gegen das Integral folgt sofort aus der Tatsache, dass eine solche Summe gegen das Integral konvergiert (Analysis 1 Satz 11.1).

Die Näherung in der vorletzten Zeile lässt sich wie folgt streng rechtfertigen. Sei $\varepsilon > 0$. Wegen der gleichmässigen Stetigkeit von γ' existiert $\delta > 0$ sodass $|\gamma'(t) - \gamma'(t')| < \varepsilon$ $\forall t, t' \in [a,b]$ mit $|t - t'| < \delta$. Für $h < \delta$ folgt somit

$$\begin{aligned} \Big|\frac{\gamma(t_j) - \gamma(t_{j-1})}{h} - \gamma'(t_j)\Big| &\underset{\text{Hauptsatz}}{=} \Big|\frac{1}{h}\int_{t_{j-1}}^{t_j} \big(\gamma'(s) - \gamma'(t_j)\big)\,ds\Big| \\ &\leq \frac{1}{h}\int_{t_{j-1}}^{t_j} \underbrace{|\gamma'(s) - \gamma'(t_j)|}_{<\varepsilon}\,ds < \frac{1}{h}\cdot\varepsilon\cdot\underbrace{(t_j - t_{j-1})}_{=h} \end{aligned}$$

und deshalb

$$\Big|L(\gamma_N) - \sum_{j=1}^N |\gamma'(t_j)|\,h\Big| < N\cdot h\cdot\varepsilon = (b-a)\cdot\varepsilon \quad \forall h < \delta.$$

Insgesamt folgt also

$$\lim_{N\to\infty} L(\gamma_N) = \int_a^b |\gamma'(t)|\,dt.$$

Beispiel 2 (Gradientenfelder) Als zweites Beispiel betrachten wir Vektorfelder der Form

$$v = \nabla f$$

für eine skalare Funktion f. Dann hängt das Kurvenintegral nur von Anfangs- und Endpunkt der Kurve ab:

Satz 7.1 (Wegunabhängigkeit für Gradientenfelder) *Sei $\Omega \subseteq \mathbb{R}^n$ offen, $f : \Omega \to \mathbb{R}$ stetig diff'bar, $\gamma : [a,b] \to \Omega$ reguläre Kurve mit $\gamma(a) = A$, $\gamma(b) = B$. Dann gilt*

$$\int_\gamma \nabla f \cdot ds = f(B) - f(A).$$

Insbesondere gilt für jede andere reguläre Kurve $\tilde{\gamma} : [a',b'] \to \Omega$ mit $\tilde{\gamma}(a') = A$, $\tilde{\gamma}(b') = B$:

$$\int_\gamma \nabla f \cdot ds = \int_{\tilde{\gamma}} \nabla f \cdot ds \quad \textit{(Wegunabhängigkeit).}$$

Die erste Identität kann als mehrdimensionales Analogon des Hauptsatzes aufgefasst werden.

Beweis Nach Kettenregel und Hauptsatz ist

$$\int_\gamma \nabla f \cdot ds = \int_a^b \underbrace{\left\langle \nabla f(\gamma(t)), \gamma'(t) \right\rangle}_{= \frac{d}{dt} f(\gamma(t))} dt = f(\gamma(b)) - f(\gamma(a)).$$

Als Anwendung des Kurvenintegrals untersuchen wir, welche Vektorfelder von der Form

$$v = \nabla f \text{ für eine skalare Funktion } f \tag{P}$$

sind.

Def. 7.4 Falls eine solche skalare Funktion existiert, heisst sie **Potential** des Vektorfelds.

(In der Physik ist die Konvention $v = -\nabla f$ üblich.) Mathematisch gesehen ist (P) ein System partieller Differentialgleichungen (gegeben: Vektorfeld $v : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$, gesucht: Lösung $f : \Omega \to \mathbb{R}$). Eine notwendige Bedingung für die Lösbarkeit in der Dimension $n = 3$ ist $\operatorname{rot} v = 0$; siehe §2.5 Lemma 2.3. Wir werden zeigen, dass für solche v in "gutartigen" Mengen Ω eine Lösung f existiert, und die (bis auf eine additive Konstante eindeutige) Lösung explizit als Kurvenintegral konstruieren. Dies kann als weitreichende Verallgemeinerung der Tatsache aus Analysis 1 Abschnitt 12 angesehen werden, dass die gewöhnliche Differentialgleichung $f' = h$, $h : [a,b] \to \mathbb{R}$ gegeben, durch das Integral $f(x) := \int_a^x h$ gelöst wird.

Def. 7.5 Eine Menge $\Omega \subseteq \mathbb{R}^n$ heisst **sternförmig**, wenn ein Punkt $x_0 \in \Omega$ existiert, sodass für alle $x \in \Omega$ die gerade Verbindungsstrecke $[x_0, x] = \{(1-t)x_0 + tx : t \in [0,1]\}$ in Ω liegt.

Satz 7.2 *Sei $\Omega \subseteq \mathbb{R}^n$ offen und sternförmig, $v : \Omega \to \mathbb{R}^n$ stetig diff'bares Vektorfeld. Dann gilt:*

a)

$$v = \nabla f \text{ für ein } f : \Omega \to \mathbb{R} \iff Dv \text{ symmetrisch.}$$

Insbesondere gilt in der Dimension $n = 3$:

$$v = \nabla f \text{ für ein } f : \Omega \to \mathbb{R} \iff \operatorname{rot} v = 0.$$

b) Falls die Bedingung in a) erfüllt ist, ist

$$f(x) := \int_{\gamma_x} v \cdot ds \text{ mit } \gamma_x(t) := (1-t)x_0 + tx \ (t \in [0,1])$$

das bis auf eine additive Konstante eindeutige Potential von v. Hierbei ist der Punkt x_0 gemäss Def. 7.5 gewählt.

Wir erhalten also das Potential am Punkt x, indem wir das Kurvenintegral des Vektorfeldes von einem festen Punkt x_0 nach x ausrechnen. Nach Satz 7.1 ist es übrigens egal, ob wir den obigen geraden Weg verwenden.

Beweis Die Implikation "$\Longrightarrow$" in a) folgt aus dem Satz von Schwarz. Die umgekehrte Implikation folgt durch Ausrechnen des Gradienten der Funktion f aus b). Die Eindeutigkeit des Potentials bis auf eine additive Konstante folgt aus Satz 7.1.

Die Voraussetzung Ω sternförmig kann nicht weggelassen werden. Sei z.B. in Dimension $n = 2$

$$v(x) = \frac{1}{x_1^2 + x_2^2} \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix}, \ v : \mathbb{R}^2 \backslash \{0\} \to \mathbb{R}^2$$

bzw. in Dimension $n = 3$

$$v(x) = \frac{1}{x_1^2 + x_2^2} \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix}, \ v : \mathbb{R}^3 \backslash x_3\text{-Achse} \to \mathbb{R}^3.$$

Dann ist Dv symmetrisch bzw. $\operatorname{rot} v = 0$, aber für den (jeweils nach oben durchlaufenen) rechten bzw. linken Halbkreis um 0 in der (x_1, x_2)-Ebene ist

$$\int_{\gamma_{\text{rechts}}} v \cdot ds \neq \int_{\gamma_{\text{links}}} v \cdot ds,$$

denn γ_{rechts} läuft mit der Strömung und γ_{links} gegen die Strömung; somit kann v wegen Satz 7.1 kein Gradient sein.

Index

abgeschlossen
  im $\mathbb{R}^n$, 11
  in topologischen Räumen, 164
Ableitung, 25
  höhere partielle, 44
  höhere totale, 48
  partielle, 25
  totale, 29
Ableitungsmatrix, 27
Abschluss (einer Menge)
  im $\mathbb{R}^n$, 13
  in topologischen Räumen, 164
Aequivalenz von Normen, 90
Algebraische Topologie, 175
Anfangsbedingung, 75, 128
asymptotisch stabil, 152
autonom, 128

Backpropagation, 80
Banach'scher Fixpunktsatz, 97
Banachraum, 92
Bergsteigerin, 39
beschränkt, 10
Betti-Zahl
  0-te, 175
  1-te, 175
Bogenlänge, 177
Bolzano-Weierstrass Eigenschaft, 167

Cauchy-Schwarz-Ungleichung, 7

Diskretisierung, 5
Divergenz, 49
Drehimpuls, 131
Dynamische Systeme (Gebiet), 133

Energie, 131
epsilon-delta-Kriterium, 21
erzwungene Schwingungen, 148
euklidische Norm, 6
euklidisches Skalarprodukt, 6
Existenz- und Eindeutigkeitssatz (für Systeme gewöhnlicher Differentialgleichungen), 135, 137
Extremstelle, 59
  lokale, 59

Feedforward-Netz, 77
Funktionalanalysis, 87
funktionalanalytischer Standpunkt, 87

gewöhnliche Differentialgleichungen, 71
  Systeme, 127
Gleichgewichtslösung, 130
Gleichgewichtspunkt, 130
gleichmässige Konvergenz, 95
Gradient
  elementare Definition, 36
  geometrische Bedeutung, 40
  koordinateninvariante Definition, 36
Gradientenverfahren, 67
Grenzwert, 9

Häufungspunkt, 10
harmonische Funktionen, 71
harmonischer Oszillator, 141
Hausdorff'sches Trennungsaxiom, 166
Heine-Borel Eigenschaft, 167
Hesse-Matrix, 44
Hilbertraum, 94
homöomorph, 160

impliziter Funktionensatz, 109
  Formulierung via Parametrisierung, 118
innerer Produktraum, 94
Inneres (einer Menge)
  im $\mathbb{R}^n$, 13
  in topologischen Räumen, 164
inneres Produkt, 93

Jacobi-Matrix, 27
Jordan'sche Normalform, 143
Jordanblock, 143

Kepler'sche Gesetze, 131
Kettenregel, 33
kompakt
    im $\mathbb{R}^n$, 22
    in topologischen Räumen, 166
konvergent
    im $\mathbb{R}^n$, 9
    in topologischen Räumen, 165
konvex, 60
Kreuzentropie, 80
Kurve, 17
Kurvenintegral, 176
    Wegunabhängigkeit für Gradientenfelder, 179

Länge (einer Kurve), 177
Lagrange'sche Multiplikatorregel, 121
Lagrange'scher Multiplikator, 122
Laplace-Operator, 49
Laplacegleichung, 71
Lemma von Gronwall, 154
Lineare Regression, 64
Lipschitzstetig, 67
Lorenz-Gleichung, 129

maschinelles Lernen, 64, 77
Matrixexponentialfunktion, 139
Matrixinversionsabbildung, 106
Maximumsnorm, 88
Maximumsstelle, 23, 59
    lokale, 59
Metrik, 7
metrischer Raum, 94
Minimumsstelle, 23, 59
    lokale, 59
Multi-index, 56

neuronales Netz, 77
Niveauflächen, 15
Niveaulinien, 14
Norm, 6

offen
    im $\mathbb{R}^n$, 11
    in topologischen Räumen, 161
Optimalitätsbedingungen
    erster Ordnung, 59
    zweiter Ordnung, 60
Orbit, 130
orthogonale Gruppe, 120

p-adische Norm, 163
p-Norm, 88
partielle Differentialgleichung, 71
Phasenporträt, 130
Picard'sches Iterationsverfahren, 134
Poissongleichung, 71
Polarkoordinaten, 107
Potential (eines Vektorfeldes), 179
punktweise Konvergenz, 95

Quelle, 19

Räuber-Beute-Modell, 128
Rand (einer Menge)
    im $\mathbb{R}^n$, 13
    in topologischen Räumen, 164
Rayleigh-Ritz'sches Variationsprinzip, 125
regulärer Punkt (einer Lösungsmenge), 114
ReLU, 78
Resonanz, 151
Restglied, 52
Richtungsableitung, 38
Riesz'scher Darstellungssatz, 36
Rotation, 49
Runge-Lenz-Vektor, 131

Satz über inverse Funktionen, 101
Satz vom Maximum und Minimum, 23
Satz von Bolzano-Weierstrass, 11
Satz von de Rham, 175

Satz von Peano, 139
Satz von Schwarz, 44
Schmetterlingseffekt, 133
Schrödingergleichung, 72
Schwingungsgleichung, 144
Serpentinen, 42
Sierpinski-Raum, 162
singulärer Punkt (einer Lösungsmenge), 114
SIR-Modell, 156
skalare Funktion, 14
Softmax, 79
Spektralsatz (für symmetrische Matrizen), 125
stabil, 152
stationäre Lösung, 130
stationärer Punkt, 130
sternförmig, 180
stochastic gradient descent, 85
submultiplikativ, 140
Supremumsnorm, 89, 94

Tangentialvektor (an eine Kurve), 28
Taylorentwicklung, 51
Taylorpolynom, 52
Topologie, 161
  diskrete, 162
  Normtopologie, 162
  Relativtopologie, 162
  triviale, 162
Topologie (Gebiet), 160
topologische Eigenschaft, 174
topologische Invariante, 174
topologischer Raum, 161
Training (eines neuronalen Netzes), 80

Ueberdeckung, 166
Umgebung, 165
Umlaufbahn von Planeten, 129
Umparametrisierung (einer Kurve), 177
Untermannigfaltigkeit des $\mathbb{R}^n$, 118

Variation der Konstanten, 147
Vektorfeld, 19
vereinfachtes Newtonverfahren, 102
Verkettung, 20
vollständig, 92

Wärmeleitungsgleichung, 72
Wasserstein-Metrik, 8
Wellengleichung, 71
Wirbel, 19

zusammenhängend, 170

Adresse des Autors:

Gero Friesecke
Department of Mathematics
School of Information, Computation and Technology
Technische Universtät München
Boltzmannstr. 3
D-85748 Garching b. München

gf@ma.tum.de
https://www.math.cit.tum.de/math/personen/professuren/friesecke-gero