Michael Růžička

Analysis

aufbauend auf die Skripte von
Prof. em. Dr. Dr. h.c. Rolf Schneider

Inhaltsverzeichnis

Analysis I

Analysis II

8 Funktionenreihen

In diesem Kapitel wollen wir etwas ausführlicher die Möglichkeit studieren, Funktionen durch unendliche Reihen darzustellen. Wir hatten schon im Anschluß an die Taylorformel kurz die Möglichkeit erörtert, bei beliebig oft differenzierbaren Funktionen von den Taylorpolynomen zur Taylorreihe überzugehen. Schon wesentlich früher hatten wir unendliche Reihen benutzt, um gewisse spezielle Funktionen einzuführen, zum Beispiel die Exponentialfunktion durch

$$\exp x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Hierbei ergibt sich unter anderem die Frage, wie man von Eigenschaften der Reihenglieder auf Eigenschaften der Funktion schließen kann. Wir können zum Beispiel zur Berechnung der Ableitung versuchen, „gliedweise“ zu differenzieren. Formal vorgehend, erhält man

$$\exp' x = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{dx^k}{dx} = \sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!} = \sum_{n=0}^{\infty} \frac{x^n}{n!} = \exp x.$$

Das Ergebnis ist richtig; aber führt ein solches Vorgehen in jedem Fall zum richtigen Ergebnis? Diese und ähnliche Fragen werden im folgenden beantwortet. Wir betrachten zunächst allgemein konvergente Folgen von Funktionen und erst dann spezielle Reihen.

8.1 Konvergenz von Funktionenfolgen

Wir kennen bereits zwei verschiedene Konvergenzbegriffe für Funktionenfolgen, und an diese sei zunächst erinnert.

Sei $(f_n)_{n\in\mathbb{N}}$ eine Folge von reellen Funktionen, die sämtlich denselben Definitionsbereich D haben. Für jedes $x \in D$ ist dann $(f_n(x))_{n\in\mathbb{N}}$ eine Folge reeller Zahlen, und für diese ist ein Konvergenzbegriff wohldefiniert. Wenn für jedes $x \in D$ die Folge $(f_n(x))_{n\in\mathbb{N}}$ konvergiert, ist durch

$$f(x) := \lim_{n\to\infty} f_n(x) \qquad (x \in D)$$

eine neue Funktion f auf D erklärt, die wir als *Grenzfunktion* der Folge bezeichnen können.

Definition. Seien f, f_n $(n \in \mathbb{N})$ reelle Funktionen auf D. Die Folge $(f_n)_{n\in\mathbb{N}}$ *konvergiert* (*punktweise*) gegen f, geschrieben

$$\lim_{n\to\infty} f_n = f \qquad \text{oder} \qquad f_n \to f \ \ (n \to \infty),$$

wenn $\lim_{n\to\infty} f_n(x) = f(x)$ für alle $x \in D$ gilt.

Beispiele. (1) Sei $D = \mathbb{R}$ und

$$f_n(x) := \sum_{k=0}^{n} \frac{x^k}{k!}.$$

Dann konvergiert $(f_n)_{n\in\mathbb{N}}$ gegen exp.

(2) Sei $D = [0, 1]$ und

$$f_n(x) := x^n.$$

Dann konvergiert $(f_n)_{n\in\mathbb{N}}$ gegen die durch

$$f(x) := \begin{cases} 1 & \text{für } x = 1, \\ 0 & \text{für } 0 \le x < 1. \end{cases}$$

erklärte Funktion f.

Bei diesem Beispiel fällt auf, dass zwar jede Funktion f_n der Folge stetig ist, dass aber die Grenzfunktion unstetig ist. Stetigkeit überträgt sich also bei punktweiser Konvergenz i.a. nicht auf die Grenzfunktion. Um den oft wünschenswerten Schluß von der Stetigkeit der Folgenglieder auf die Stetigkeit der Grenzfunktion zu ermöglichen, braucht man einen Konvergenzbegriff, der schärfer ist als punktweise Konvergenz. Dies ist die bereits in Abschnitt 7.2 benutzte gleichmäßige Konvergenz, die wir jetzt etwas allgemeiner definieren wollen.

Definition. Seien $f, f_n (n \in \mathbb{N})$ reelle Funktionen auf D, sei $D' \subset D$. Die Folge $(f_n)_{n\in\mathbb{N}}$ *konvergiert gleichmäßig in D' gegen f*, wenn gilt

$$\forall \varepsilon \in \mathbb{R}^+ \ \ \exists n_0 \in \mathbb{N} \ \ \forall n \ge n_0 \ \ \forall x \in D' : |f_n(x) - f(x)| \le \varepsilon.$$

Statt „gleichmäßig in D“ sagt man kurz „gleichmäßig“.

Ist ein fester Definitionsbereich D gegeben, so können wir wie früher für Intervalle die *Supremumsnorm* einer beschränkten Funktion $f : D \to \mathbb{R}$ erklären durch

$$\|f\| := \sup_{x\in D} |f(x)|.$$

Dann gilt also (für Funktionen f, f_n auf D):

$$\begin{aligned}&(f_n)_{n\in\mathbb{N}} \text{ konvergiert gleichmäßig gegen } f\\ &\Leftrightarrow \forall \varepsilon \in \mathbb{R}^+ \ \exists n_0 \in \mathbb{N} \ \forall n \geq n_0 : \|f_n - f\| \leq \varepsilon\\ &\Leftrightarrow \lim_{n\to\infty} \|f_n - f\| = 0.\end{aligned}$$

Man beachte, dass $\|f_n - f\| < \varepsilon$ impliziert, dass $\|f_n - f\|$ definiert, also $f_n - f_m$ beschränkt ist. Es wird aber nicht vorausgesetzt, dass f, f_n beschränkt sind.

Für die Konvergenz von Folgen reeller Zahlen kennen wir das Kriterium von Cauchy. Ein ganz analoges Kriterium gilt auch für die gleichmäßige Konvergenz von Funktionenfolgen. Im folgenden liege stets ein fester Definitionsbereich D zugrunde.

Definition. Die Folge $(f_n)_{n\in\mathbb{N}}$ von Funktionen auf D heißt *Cauchy-Folge* genau dann, wenn

$$\forall \varepsilon \in \mathbb{R}^+ \ \exists n_0 \in \mathbb{N} \ \forall n, m \geq n_0 : \|f_n - f_m\| < \varepsilon.$$

1.1 Satz. *Die Folge $(f_n)_{n\in\mathbb{N}}$ konvergiert genau dann gleichmäßig, wenn sie eine Cauchy-Folge ist.*

Beweis. „⇒“: Sei $(f_n)_{n\in\mathbb{N}}$ gleichmäßig konvergent gegen f. Sei $\varepsilon \in \mathbb{R}^+$. Es gibt ein $n_0 \in \mathbb{N}$ mit $\|f_n - f\| < \varepsilon/2$ für $n \geq n_0$. Für alle $n, m \geq n_0$ gilt also

$$\|f_n - f_m\| \leq \|f_n - f\| + \|f - f_m\| < \varepsilon,$$

wobei die Dreiecksungleichung für die Supremumsnorm benutzt wurde.

„⇐“: Sei $(f_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge. Sei $\varepsilon \in \mathbb{R}^+$. Nach Voraussetzung existiert ein $n_0 \in \mathbb{N}$ mit $\|f_n - f_m\| < \varepsilon$ für alle $n, m \geq n_0$. Insbesondere gilt also für jedes $x \in D$

$$|f_n(x) - f_m(x)| < \varepsilon \qquad \text{für } n, m \geq n_0.$$

Die Folge $(f_n(x))_{n\in\mathbb{N}}$ ist also eine Cauchy-Folge reeller Zahlen und daher nach dem gewöhnlichen Cauchy-Kriterium konvergent gegen eine Zahl, die wir $f(x)$ nennen. Da $x \in D$ beliebig war, ist damit eine Funktion $f : D \to \mathbb{R}$ erklärt. Wir behaupten, dass $(f_n)_{n\in\mathbb{N}}$ gleichmäßig gegen f konvergiert.

Sei $\varepsilon \in \mathbb{R}^+$ und dazu n_0 wie oben. Für beliebiges $x \in D$ gilt dann

$$|f_n(x) - f_m(x)| < \varepsilon \qquad \text{für } n, m \geq n_0.$$

Der Grenzübergang $m \to \infty$ liefert

$$|f_n(x) - f(x)| \leq \varepsilon \qquad \text{für } n \geq n_0.$$

Da dies für alle $x \in D$ gilt, folgt $\|f_n - f\| \leq \varepsilon$ für alle $n \geq n_0$. ∎

Der folgende Satz rückt die Bedeutung der gleichmäßigen Konvergenz ins rechte Licht.

1.2 Satz. *Seien f_n, f Funktionen auf D $(n \in \mathbb{N})$, sei $a \in D$. Sind alle f_n stetig in a und konvergiert $(f_n)_{n\in\mathbb{N}}$ gleichmäßig gegen f, so ist f stetig in a.*

Beweis. Sei $\varepsilon \in \mathbb{R}^+$. Es gibt ein $m \in \mathbb{N}$ mit $\|f_m - f\| < \varepsilon/3$, also

$$|f_m(x) - f(x)| < \frac{\varepsilon}{3} \qquad \text{für alle } x \in D.$$

Da f_m in a stetig ist, existiert ein $\delta \in \mathbb{R}^+$ mit

$$|f_m(x) - f_m(a)| < \frac{\varepsilon}{3} \quad \text{für alle } x \in D \text{ mit } |x - a| < \delta.$$

Sei jetzt $x \in D$ und $|x - a| < \delta$. Dann gilt

$$\begin{aligned} |f(x) - f(a)| &\le |f(x) - f_m(x)| + |f_m(x) - f_m(a)| + |f_m(a) - f(a)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

∎

Bemerkung. Natürlich kann man nicht umgekehrt schließen, d.h. wenn $(f_n)_{n\in\mathbb{N}}$ punktweise gegen f konvergiert und alle f_n sowie f stetig sind, braucht keineswegs die Konvergenz gleichmäßig zu sein.

Kann man in Satz 1.2 „stetig“ durch „differenzierbar“ ersetzen? Das ist nicht der Fall.

Beispiel. Sei $D = \mathbb{R}$, $f_n(x) := \sqrt{x^2 + \frac{1}{n}}$ $(n \in \mathbb{N})$ und $f(x) := |x|$ für $x \in \mathbb{R}$. Dann konvergiert $(f_n)_{n\in\mathbb{N}}$ gleichmäßig gegen f, wie aus

$$|f_n(x) - f(x)| = \left|\sqrt{x^2 + \frac{1}{n}} - \sqrt{x^2}\right| = \frac{x^2 + \frac{1}{n} - x^2}{\sqrt{x^2 + \frac{1}{n}} + \sqrt{x^2}} \le \frac{1}{\sqrt{n}}$$

folgt. Jede Funktion f_n ist differenzierbar, aber f ist in 0 nicht differenzierbar.

Es kann auch sein, dass zwar die Grenzfunktion f der Folge $(f_n)_{n\in\mathbb{N}}$ differenzierbar ist, aber die Folge $(f_n')_{n\in\mathbb{N}}$ nicht gegen f' konvergiert.

Beispiel. Sei $f_n(x) = \frac{1}{n}\sin nx$ und $f(x) = 0$ für $x \in \mathbb{R}$. Dann konvergiert $(f_n)_{n\in\mathbb{N}}$ gleichmäßig gegen f, alle f_n sowie f sind differenzierbar, aber $(f_n')_{n\in\mathbb{N}}$ konvergiert nicht punktweise gegen f'. Es ist nämlich $f_n'(x) = \cos nx$ und z.B. $\lim_{n\to\infty} f_n'(0) = 1$, $f'(0) = 0$.

Um wirklich Grenzübergang und Differentiation vertauschen zu können, braucht man stärkere Voraussetzungen, z.B. die folgenden:

1.3 Satz. *Sei $(f_n)_{n\in\mathbb{N}}$ eine Folge differenzierbarer Funktionen auf $[a, b]$. Die Folge $(f_n')_{n\in\mathbb{N}}$ konvergiere gleichmäßig, und $(f_n)_{n\in\mathbb{N}}$ konvergiere an wenigstens einer Stelle $x_0 \in [a, b]$. Dann konvergiert $(f_n)_{n\in\mathbb{N}}$ gleichmäßig gegen eine differenzierbare Funktion f und $(f_n')_{n\in\mathbb{N}}$ gleichmäßig gegen f'.*

Beweis. Wir zeigen zuerst, dass $(f_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge ist. Sei $\varepsilon \in \mathbb{R}^+$ gegeben. Nach Voraussetzung und Satz 1.1 existiert ein $n_1 \in \mathbb{N}$ mit

$$\|f'_m - f'_n\| < \frac{\varepsilon}{2(b-a)} \qquad \text{für alle } m, n \geq n_1.$$

Da $(f_n(x_0))_{n\in\mathbb{N}}$ eine Cauchy-Folge ist, existiert ein $n_2 \in \mathbb{N}$ mit

$$|f_m(x_0) - f_n(x_0)| < \frac{\varepsilon}{2} \qquad \text{für } m, n \geq n_2.$$

Sei $x \in [a, b]$. Es ist

$$|f_m(x) - f_n(x)| \leq |(f_m - f_n)(x) - (f_m - f_n)(x_0)| + |f_m(x_0) - f_n(x_0)|.$$

Nach dem Mittelwertsatz 2.2 aus Kapitel 6 existiert ein $z \in [a, b]$ mit

$$(f_m - f_n)(x) - (f_m - f_n)(x_0) = (f'_m(z) - f'_n(z))(x - x_0).$$

Für alle $m, n \geq n_0 := \max\{n_1, n_2\}$ folgt

$$\begin{aligned} |f_m(x) - f_n(x)| &\leq |f'_m(z) - f'_n(z)||x - x_0| + \frac{\varepsilon}{2} \\ &\leq \|f'_m - f'_n\|(b-a) + \frac{\varepsilon}{2} < \varepsilon. \end{aligned}$$

Da $x \in [a, b]$ beliebig war, folgt $\|f_m - f_n\| \leq \varepsilon$ für $m, n \geq n_0$. Also ist $(f_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge und daher nach Satz 1.1 gleichmäßig konvergent gegen eine Funktion f.

Wir zeigen jetzt die Differenzierbarkeit von f. Sei $c \in [a, b]$. Setze

$$g_n(x) := \begin{cases} \dfrac{f_n(x) - f_n(c)}{x - c} - f'_n(c) & \text{für } x \in [a, b] \setminus \{c\}, \\ 0 & \text{für } x = c. \end{cases}$$

Wegen der Differenzierbarkeit von f_n in c ist g_n in c stetig. Wir zeigen zunächst, dass $(g_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge ist. Sei $\varepsilon \in \mathbb{R}^+$ vorgegeben. Nach Voraussetzung und Satz 1.1 existiert ein $n_0 \in \mathbb{N}$ mit

$$\|f'_m - f'_n\| < \frac{\varepsilon}{2} \qquad \text{für alle } m, n \geq n_0.$$

Sei $x \in [a, b]$. Im Fall $x \neq c$ ist

$$g_m(x) - g_n(x) = \frac{(f_m - f_n)(x) - (f_m - f_n)(c)}{x - c} - (f_m - f_n)'(c).$$

Nach dem Mittelwertsatz existiert ein $z \in [a, b]$ mit

$$\frac{(f_m - f_n)(x) - (f_m - f_n)(c)}{x - c} = (f_m - f_n)'(z).$$

Für $m, n \geq n_0$ folgt

$$\begin{aligned} |g_m(x) - g_n(x)| &\leq |f'_m(z) - f'_n(z)| + |f'_m(c) - f'_n(c)| \\ &\leq 2\|f'_m - f'_n\| \leq \varepsilon. \end{aligned}$$

Dies gilt trivialerweise auch für $x = c$. Da $x \in [a, b]$ beliebig war, ist also $\|g_m - g_n\| \leq \varepsilon$ für $m, n \geq n_0$. Also ist $(g_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge.

Nach Satz 1.1 ist $(g_n)_{n\in\mathbb{N}}$ gleichmäßig konvergent gegen eine Funktion g, die $g(c) = 0$ erfüllt und nach Satz 1.2 in c stetig ist. Nun gilt für alle $x \in [a, b] \setminus \{c\}$

$$\frac{f_n(x) - f_n(c)}{x - c} - f'_n(c) = g_n(x),$$

woraus durch Grenzübergang

$$\frac{f(x) - f(c)}{x - c} - \lim_{n\to\infty} f'_n(c) = g(x)$$

folgt. Wegen $\lim_{x\to c} g(x) = g(c) = 0$ folgt

$$\lim_{x\to c} \frac{f(x) - f(c)}{x - c} = \lim_{n\to\infty} f'_n(c).$$

Also ist f in c differenzierbar und $f'(c) = \lim_{n\to\infty} f'_n(c)$. Damit ist die Differenzierbarkeit von f gezeigt, ferner die punktweise Konvergenz von $(f'_n)_{n\in\mathbb{N}}$ gegen f'. Diese Konvergenz ist nach Voraussetzung gleichmäßig. ∎

8.2 Potenzreihen

Nachdem wir früher Reihen reeller Zahlen und jetzt Funktionenfolgen betrachtet haben, ist klar, was allgemein unter einer Funktionenreihe zu verstehen ist. Sei $(f_k)_{k\in\mathbb{N}_0}$ eine Funktionenfolge auf D. Dann verstehen wir unter

$$\sum_{k=0}^{\infty} f_k$$

die Funktionenfolge der Partialsummen $(\sum_{k=0}^{n} f_k)_{n\in\mathbb{N}}$, und wir bezeichnen $\sum_{k=0}^{\infty} f_k$ als Funktionenreihe. Ist die Folge punktweise konvergent, so bezeichnen wir mit $\sum_{k=0}^{\infty} f_k$ auch die Grenzfunktion. Die Schreibweise

$$\sum_{k=0}^{\infty} f_k = f \qquad \text{in } D$$

bedeutet also definitionsgemäß:

$$\lim_{n\to\infty} \sum_{k=0}^{n} f_k(x) = f(x) \qquad \text{für alle } x \in D.$$

Konvergiert die Folge $(\sum_{k=0}^{n} f_k)_{n\in\mathbb{N}}$ in $D' \subset D$ gleichmäßig (gegen f), so sagen wir, dass die Reihe $\sum_{k=0}^{\infty} f_k$ in D' gleichmäßig (gegen f) konvergiert.

Die Begriffe der punktweisen oder gleichmäßigen Konvergenz von Reihen sind also nichts Neues gegenüber den Funktionenfolgen. Auch die Sätze aus Abschnitt 8.1 gelten natürlich sinngemäß für Funktionenreihen. Neu gegenüber der Folgenkonvergenz ist lediglich - wie schon bei Zahlenreihen - der Begriff der absoluten Konvergenz. Definitionsgemäß konvergiert die Reihe $\sum f_k$ absolut, wenn die Reihe $\sum |f_k|$ konvergiert.

In Verallgemeinerung von Satz 2.8 aus Kapitel 3 haben wir das folgende wichtige Kriterium für gleichmäßige und absolute Konvergenz.

2.1 Satz (Majorantenkriterium)**.** *Sei $f_k : D \to \mathbb{R}$, $c_k \in \mathbb{R}^+$ $(k \in \mathbb{N}_0)$ gegeben. Gilt*

$$\|f_k\| \le c_k \qquad \textit{für alle } k \in \mathbb{N}_0$$

und ist die Reihe $\sum_{k=0}^{\infty} c_k$ konvergent, so konvergiert die Reihe $\sum_{k=0}^{\infty} f_k$ absolut und gleichmäßig.

Beweis. Sei $\varepsilon \in \mathbb{R}^+$. Da $\sum c_k$ konvergiert, existiert ein $n_0 \in \mathbb{N}$ mit

$$\sum_{k=m}^{m+p} c_k < \varepsilon$$

für alle $m \ge n_0, p \in \mathbb{N}_0$. Für diese m, p gilt also

$$\left\| \sum_{k=m}^{m+p} f_k \right\| \le \left\| \sum_{k=m}^{m+p} |f_k| \right\| \le \sum_{k=m}^{m+p} \|f_k\| \le \sum_{k=m}^{m+p} c_k < \varepsilon.$$

Die Folge der Partialsummen von $\sum f_k$ und die Folge der Partialsummen von $\sum |f_k|$ sind also Cauchyfolgen und daher nach Satz 1.1 gleichmäßig konvergent. ∎

Das Vorstehende und die Ergebnisse aus Abschnitt 8.1 wollen wir jetzt anwenden auf den besonders wichtigen Spezialfall der Potenzreihen.

Unter einer *Potenzreihe zur Stelle a* versteht man eine Funktionenreihe

$$\sum_{k=0}^{\infty} f_k \qquad \text{mit } f_k(x) = a_k(x-a)^k \qquad \text{für } x \in \mathbb{R},$$

wo $a \in \mathbb{R}$ und $(a_k)_{k\in\mathbb{N}_0}$ eine Folge reeller Zahlen ist. Hier ist die folgende ungenaue, aber bequeme Sprechweise üblich. Man sagt

$$\text{„die Potenzreihe } \sum_{k=0}^{\infty} a_k(x-a)^k \text{ “}$$

statt

$$\text{„die Potenzreihe } \sum_{k=0}^{\infty} f_k \text{ mit } f_k(x) = a_k(x-a)^k \text{ für } x \in \mathbb{R} \text{ “.}$$

Wir fragen jetzt nach der Menge aller $x \in \mathbb{R}$, für die eine gegebene Potenzreihe

$$\sum_{k=0}^{\infty} a_k(x-a)^k \tag{2.2}$$

konvergiert. Auf jeden Fall konvergiert sie trivialerweise für $x = a$ (es gibt Potenzreihen, die für kein anderes x konvergieren). Allgemein gibt das Wurzelkriterium 2.10 aus Kapitel 3 Auskunft. Nach ihm ist (für gegebenes $x \in \mathbb{R}$) die Reihe (2.2) absolut konvergent, wenn

$$\limsup_{n\to\infty} \sqrt[n]{|a_n(x-a)^n|} = |x-a| \limsup_{n\to\infty} \sqrt[n]{|a_n|} < 1$$

ist, und sie ist divergent, wenn $|x-a| \limsup_{n\to\infty} \sqrt[n]{|a_n|} > 1$ ist.

Zur Vermeidung lästiger Fallunterscheidungen schreiben wir $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ und definieren

$$-\infty < y < \infty \qquad \text{für } y \in \mathbb{R}.$$

Außerdem definieren wir $\frac{1}{\infty} = 0$ und $\frac{1}{0} = \infty$, ferner $y + \infty = \infty$, $y - \infty = -\infty$ für $y \in \mathbb{R}$. Es sei daran erinnert, dass wir in Abschnitt 3.2 $\limsup \sqrt[n]{|a_n|} = \infty$ definiert hatten im Fall, dass die Folge $(\sqrt[n]{|a_n|})_{n\in\mathbb{N}}$ nicht beschränkt ist.

Für unsere gegebene Potenzreihe setzen wir jetzt

$$r := \frac{1}{\limsup\limits_{n\to\infty} \sqrt[n]{|a_n|}}.$$

Dann gilt also: Für alle $x \in \mathbb{R}$ mit $|x-a| < r$ ist (2.2) absolut konvergent, für alle $x \in \mathbb{R}$ mit $|x-a| > r$ ist (2.2) divergent. Das Intervall $(a-r, a+r)$ heißt daher *Konvergenzintervall* der Potenzreihe.

Wie steht es mit gleichmäßiger Konvergenz? Sei (2.2) absolut konvergent für ein x_0. Für alle $x \in \mathbb{R}$ mit $|x-a| \le |x_0-a|$ gilt dann

$$|a_k(x-a)^k| \le |a_k||x_0-a|^k \qquad \text{für } k \in \mathbb{N}_0,$$

und die Reihe $\sum^{\infty} |a_k||x_0-a|^k$ ist konvergent. Nach dem Majorantenkriterium 2.1 folgt, dass die Reihe (2.2) in $[a-|x_0-a|, a+|x_0-a|]$ absolut und gleichmäßig konvergiert. Wir fassen zusammen:

2.3 Satz. *Zu der Potenzreihe*

$$\sum_{k=0}^{\infty} a_k (x-a)^k \tag{2.4}$$

gibt es ein $r \in \overline{\mathbb{R}}$, $r \geq 0$, *so dass die Reihe (2.4) für* $|x-a| < r$ *absolut konvergiert und für* $|x-a| > r$ *divergiert. Die Zahl* r *heißt* Konvergenzradius *der Reihe (2.4) und ist gegeben durch*

$$r = \frac{1}{\limsup\limits_{n\to\infty} \sqrt[n]{|a_n|}}.$$

Das offene Intervall $(a-r, a+r)$ *heißt* Konvergenzintervall *der Reihe (2.4). In jedem kompakten Teilintervall des Konvergenzintervalls konvergiert die Reihe (2.4) gleichmäßig.*

Bemerkung. Achtung! Über die Konvergenz in den Endpunkten des Konvergenzintervalls wird hier nichts ausgesagt (und läßt sich auch allgemein nichts sagen). Es gibt Potenzreihen, die in keinem, einem oder beiden Endpunkten des Konvergenzintervalls konvergieren.

Ferner beachte man, dass i.a. nicht im ganzen Konvergenzintervall gleichmäßige Konvergenz vorliegt, sondern nur in kompakten Teilintervallen.

Beispiel.

$$\sum_{k=1}^{\infty} \frac{1}{k^m} x^k \qquad \text{mit einem } m \in \mathbb{N}_0. \tag{2.5}$$

Wegen $\lim_{n\to\infty} \sqrt[n]{n} = 1$ (vgl. Behauptung 1.11 aus Kapitel 5) ist der Konvergenzradius $= 1$. Ferner gilt:

für $m = 0$ ist (2.5) divergent in 1 und -1
für $m = 1$ ist (2.5) konvergent in -1, divergent in 1
für $m = 2$ ist (2.5) konvergent in -1 und 1.

Nehmen wir an, die Potenzreihe

$$\sum_{k=0}^{\infty} a_k (x-a)^k \tag{2.6}$$

habe den Konvergenzradius $0 < r < \infty$ und konvergiere etwa auch noch im Endpunkt $a + r$ des Konvergenzintervalls. Wir wollen zeigen, dass sie dann im Intervall $[a, a+r]$ gleichmäßig konvergiert. Der Unterschied zur früheren Argumentation ist, dass jetzt nicht notwendig absolute Konvergenz vorliegt, daher ist das Majorantenkriterium nicht anwendbar. O.B.d.A. können wir uns auf den Fall $a = 0$, $r = 1$ beschränken, der durch eine einfache Transformation erreichbar ist.

2.7 Satz. *Ist $\sum_{k=0}^{\infty} a_k$ konvergent, so ist $\sum_{k=0}^{\infty} a_k x^k$ gleichmäßig konvergent in [0,1].*

Beweis. Betrachte die Restglieder $b_k := \sum_{j=k+1}^{\infty} a_j$ für $k \in \mathbb{N}_0$. Dann ist $(b_k)_{k\in\mathbb{N}}$ eine Nullfolge, und es ist $b_{k-1} - b_k = a_k$. Also ergibt sich für $n > m$

$$\sum_{k=m+1}^{n} a_k x^k = b_m x^{m+1} - b_n x^n + \sum_{k=m+1}^{n-1} b_k (x^{k+1} - x^k).$$

Sei $\varepsilon \in \mathbb{R}^+$. Es gibt ein $n_0 \in \mathbb{N}$ mit $|b_k| < \varepsilon/3$ für $k \geq n_0$. Sei jetzt $m \geq n_0$, $p \in \mathbb{N}$. Dann gilt für beliebiges $x \in [0,1]$

$$\left|\sum_{k=m+1}^{m+p} a_k x^k\right| \leq |b_m| + |b_{m+p}| + \sum_{k=m+1}^{m+p-1} |b_k||x^{k+1} - x^k|$$

$$\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \underbrace{\sum_{k=m+1}^{m+p-1} (x^k - x^{k+1})}_{= x^{m+1} - x^{m+p} \leq 1} \leq \varepsilon.$$

Aus dem Cauchy-Kriterium 1.1 folgt jetzt die behauptete gleichmäßige Konvergenz. ∎

Aus Satz 2.7 und Satz 1.2 folgt insbesondere:

2.8 Folgerung. *Ist*

$$f(x) = \sum_{k=0}^{\infty} a_k x^k \qquad \text{für } x \in [0,1],$$

so ist f in $[0,1]$ stetig.

Eine durch eine konvergente Potenzreihe dargestellte Funktion ist also stetig. Allgemein sagen wir, falls

$$f(x) = \sum_{k=0}^{\infty} a_k (x-a)^k \qquad \text{für } x \in (a-r, a+r) \tag{2.9}$$

mit $r > 0$ gilt, die Funktion f sei durch die obige Potenzreihe *dargestellt*, oder sie sei in eine Potenzreihe um a *entwickelt*.

Wir untersuchen jetzt die Differenzierbarkeit einer derart dargestellten Funktion. Dazu betrachten wir die durch gliedweise Differentiation von (2.9) entstehende Potenzreihe

$$\sum_{k=1}^{\infty} k a_k (x-a)^{k-1} \tag{2.10}$$

Wegen $\limsup_{n\to\infty} \sqrt[n]{(n+1)|a_{n+1}|} = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$ hat sie denselben Konvergenzradius wie (2.9). In jedem kompakten Teilintervall des Konvergenzintervalls $(a-r, a+r)$ ist also (2.10) nach Satz 2.3 gleichmäßig konvergent; aus Satz 1.3 folgt daher die Differenzierbarkeit von f und die Konvergenz von (2.10) gegen f'. Auf f' kann man natürlich wieder denselben Schluß anwenden, usw. Auf diese Weise erhalten wir den folgenden Satz:

2.11 Satz. *Sei*

$$f(x) = \sum_{k=0}^{\infty} a_k (x-a)^k \qquad \text{für } x \in (a-r, a+r)$$

mit positivem Konvergenzradius r. Dann ist f beliebig oft differenzierbar, und für $n \in \mathbb{N}$ gilt

$$f^{(n)}(x) = \sum_{k=n}^{\infty} k(k-1)\cdots(k-n+1) a_k (x-a)^{k-n},$$

wobei die rechts stehende Reihe denselben Konvergenzradius r hat. Speziell ist

$$a_n = \frac{f^{(n)}(a)}{n!}.$$

Die letzte Aussage zeigt insbesondere, dass zwei verschiedene Potenzreihen (zur selben Stelle a) nicht dieselbe Funktion darstellen können. Mit anderen Worten: Aus

$$\sum_{k=0}^{\infty} a_k (x-a)^k = \sum_{k=0}^{\infty} b_k (x-a)^k$$

mit Konvergenz in einem Intervall um a folgt $a_k = b_k$ für $k \in \mathbb{N}_0$ (Eindeutigkeitssatz für Potenzreihen).

Taylorreihen

Wir haben eben gezeigt: Wenn die Funktion f durch eine (in einer Umgebung von a konvergente) Potenzreihe um a dargestellt wird, so ist diese gegeben durch

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x-a)^k.$$

Hier sei kurz an Abschnitt 6.3 erinnert: Ist f in einer Umgebung von a beliebig oft differenzierbar, so hatten wir die Reihe

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x-a)^k$$

als die *Taylorreihe* von f zur Stelle a bezeichnet. Wenn also f überhaupt durch eine Potenzreihe um a dargestellt werden kann, dann nur durch die Taylorreihe. Im konkreten Fall kommt es also darauf an, den Konvergenzradius der Taylorreihe zu ermitteln. Ist er positiv, so folgt aber allein daraus noch nicht, dass die Taylorreihe gegen die Funktion konvergiert, wie ein Beispiel in Abschnitt 6.3 zeigte. Um zu zeigen, dass eine gegebene Funktion im Konvergenzintervall wirklich durch ihre Taylorreihe dargestellt wird, muß man also entweder zeigen, dass das Restglied gegen Null konvergiert, oder, falls dies nicht gelingt, auf andere Weise schließen. Hierfür im folgenden einige Beispiele. Für die Abschätzung des Restgliedes haben wir die durch die Taylorformel gegebenen Darstellungen zur Verfügung: Wird

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + R_{n+1}(x)$$

gesetzt (und ist f auf einem Intervall definiert), so gilt nach Satz 3.2 aus Kapitel 6

$$R_{n+1}(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}$$

mit einem (von n und x abhängenden) c zwischen a und x, und nach Satz 3.3 aus Kapitel 7 gilt

$$R_{n+1}(x) = \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n\,dt.$$

In manchen, aber nicht in allen Fällen kann man hiermit die gewünschte Konvergenz der Taylorreihe gegen f zeigen.

Wir wollen als Beispiele die Taylorreihen für einige der im Kapitel über spezielle Funktionen betrachteten Funktionen untersuchen. Für die Funktionen exp, sin, cos wurde (für $a = 0$) bereits früher gezeigt, dass das Restglied gegen Null geht.

Wir betrachten die Logarithmus-Funktion. Als Entwicklungsstelle kommt nur ein Punkt des Definitionsbereiches $\mathbb{R}^+$ in Frage. Wir wählen $a = 1$. Für $f = \ln$ beweist man leicht durch vollständige Induktion

$$f^{(k)}(x) = (-1)^{k-1}(k-1)!\,x^{-k}, \qquad (k \in \mathbb{N})$$

speziell $f^{(k)}(1) = (-1)^{k-1}(k-1)!$. Die Taylorreihe zur Stelle 1 lautet also

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k}(x-1)^k.$$

Der Konvergenzradius dieser Reihe ist offenbar 1; sie stellt also in (0,2) eine Funktion g dar. Eine Abschätzung des Restgliedes stößt auf Schwierigkeiten: Da man nichts über c_n weiß, kann man nicht ausschließen, dass $c_n = 1/2$ für alle n ist. Wegen

$$R_{n+1}(x) = \frac{(-1)^n}{n+1}\frac{(x-1)^{n+1}}{c_n^{n+1}}$$

würde dann aber für $0 < x < 1/2$ gelten:

$$\left|\frac{x-1}{c_n}\right| > 1,$$

also $R_{n+1}(x) \not\to 0$. Man kann aber auf andere Weise leicht zeigen, dass g in (0,2) mit der Funktion ln übereinstimmt. Dazu differenzieren wir die Funktion $\ln - g$: Es ist

$$\ln' x - g'(x) = \frac{1}{x} - \sum_{k=1}^{\infty}(-1)^{k-1}(x-1)^{k-1} = \frac{1}{x} - \frac{1}{1+(x-1)} = 0.$$

Die Funktion $\ln - g$ ist also in (0,2) konstant; da sie an der Stelle $x = 1$ gleich Null ist, ist sie überall Null. Damit ist

$$\ln x = \sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k}(x-1)^k \qquad \text{für } x \in (0,2)$$

gezeigt. Die rechts stehende Reihe konvergiert nach dem Leibnizkriterium auch für $x = 2$. Da die dargestellte Funktion nach Folgerung 2.8 (passend transformiert) in 2 noch stetig ist, stimmt sie dort mit $\ln 2$ überein. Wir können das Ergebnis auch in der Form

$$\ln(1+x) = \sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k}x^k \qquad \text{für } -1 < x \le 1$$

schreiben.

Wir wollen diese Gleichung für $|x| < 1$ noch auf andere Weise herleiten. Es ist

$$\ln(1+x) = \int_0^x \frac{1}{1+t}dt = \int_0^x \sum_{k=0}^{\infty}(-t)^k dt.$$

Die geometrische Reihe $\sum_{k=0}^{\infty}(-t)^k$ ist nach dem Majorantenkriterium für $t \in [-|x|, |x|]$ gleichmäßig konvergent; nach Satz 2.4 aus Kapitel 7 darf man also gliedweise integrieren und erhält

$$\ln(1+x) = \sum_{k=0}^{\infty}\int_0^x (-t)^k dt = \sum_{k=0}^{\infty}\left[\frac{(-1)^k}{k+1}t^{k+1}\right]_0^x = \sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k}x^k.$$

Speziell haben wir die hübsche Formel

$$\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$$

Zur numerischen Berechnung ist diese Reihe aber ungeeignet; besser konvergiert die Taylorreihe der Funktion

$$\ln \frac{1+x}{1-x}$$

zur Stelle 0. Man bekommt sie aus

$$\begin{aligned}\ln \frac{1+x}{1-x} &= \ln(1+x) - \ln(1-x) \\ &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots - \left(-x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} \dots\right) \\ &= 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \dots\dots\right) \qquad \text{für } |x| < 1.\end{aligned}$$

Als nächstes Beispiel betrachten wir die Funktion $f = \arctan$, die wir in eine Potenzreihe um 0 entwickeln wollen. Zur Berechnung der n-ten Ableitung an der Stelle 0 kann man den Ansatz

$$f^{(n)}(x) = \frac{P_n(x)}{(1+x^2)^n}$$

machen und findet für P_n eine Rekursionsformel, aus der sich herleiten läßt, dass

$$f^{(2n)}(0) = 0, \quad f^{(2n+1)}(0) = (-1)^n (2n)!$$

ist. Die Taylorreihe der Funktion arctan lautet also

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1}.$$

Der Konvergenzradius ist offenbar 1. Sei f die in $(-1,1)$ dargestellte Funktion. Es gilt

$$f'(x) = \sum_{k=0}^{\infty} (-1)^k x^{2k} = \frac{1}{1+x^2} = \arctan' x,$$

ferner $f(0) = 0 = \arctan 0$. Also ist $f(x) = \arctan x$. Wir haben also

$$\begin{aligned}\arctan x &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots \\ &= \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1} \qquad \text{in } (-1,1).\end{aligned}$$

Nach dem Leibniz-Kriterium konvergiert die Reihe auch für $x = 1$ und $x = -1$. Nach Folgerung 2.8 ist die dargestellte Funktion dort stetig, stimmt also mit der Funktion arctan überein. Speziell haben wir $\tan \frac{\pi}{4} = 1$, also $\arctan 1 = \frac{\pi}{4}$, somit

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$$

Auch diese Reihe konvergiert aber zu langsam, um für numerische Rechnungen brauchbar zu sein.

Als letztes Beispiel wollen wir die allgemeine Potenz $x \mapsto x^\alpha$ betrachten. Als Entwicklungsstelle wählen wir ebenfalls 1, aber es schreibt sich bequemer, die Funktion $x \mapsto (1+x)^\alpha$ zu nehmen und um 0 zu entwickeln. Für $f(x) = (1+x)^\alpha \ \ (x > -1)$ berechnet man

$$f^{(k)}(x) = \alpha(\alpha-1)\cdots(\alpha-k+1)(1+x)^{\alpha-k}.$$

Zur übersichtlicheren Schreibweise wollen wir wie früher den Binomialkoeffizienten $\binom{\alpha}{k}$ erklären durch

$$\binom{\alpha}{k} := \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}.$$

Dann ist die Taylorreihe der Funktion f zur Stelle 0 gegeben durch

$$\sum_{k=0}^{\infty} \binom{\alpha}{k} x^k.$$

Im Fall $\alpha \in \mathbb{N}_0$ sind fast alle Koeffizienten 0, es liegt also ein Polynom vor. Wir setzen jetzt $\alpha \notin \mathbb{N}_0$ voraus.

Aus dem Quotientenkriterium 2.9 aus Kapitel 3 folgt sofort, dass der Konvergenzradius gleich 1 ist. Die Reihe stellt also in $(-1,1)$ eine Funktion g dar. Setzen wir

$$h(x) := \frac{g(x)}{(1+x)^\alpha} \qquad \text{für } -1 < x < 1,$$

so gilt

$$h'(x) = \frac{g'(x)(1+x)^\alpha - g(x)\alpha(1+x)^{\alpha-1}}{(1+x)^{2\alpha}} = \frac{g'(x)(1+x) - \alpha g(x)}{(1+x)^{\alpha+1}}.$$

Nun ist

$$\begin{aligned}(1+x)g'(x) &= (1+x)\sum_{k=1}^{\infty}\binom{\alpha}{k}kx^{k-1}\\ &= \sum_{k=0}^{\infty}\binom{\alpha}{k+1}(k+1)x^k + \sum_{k=1}^{\infty}\binom{\alpha}{k}kx^k\\ &= \sum_{k=0}^{\infty}\left\{\binom{\alpha}{k+1}(k+1) + \binom{\alpha}{k}k\right\}x^k\\ &= \alpha\sum_{k=0}^{\infty}\binom{\alpha}{k}x^k = \alpha g(x),\end{aligned}$$

also $h' = 0$ und daher $h = \text{const}$. Wegen $h(0) = 1$ folgt

$$(1+x)^\alpha = \sum_{k=0}^{\infty}\binom{\alpha}{k}x^k \qquad \text{für } -1 < x < 1.$$

Man nennt diese Reihe auch die „binomische Reihe“. Das Konvergenzverhalten an den Stellen ± 1 ist etwas schwieriger zu beurteilen:

Bemerkung. Sei $\alpha \in \mathbb{R} \setminus \mathbb{N}_0$. Die binomische Reihe $\sum_{n=0}^{\infty}\binom{\alpha}{n}x^n$ ist für

	in $x = 1$	in $x = -1$
$\alpha > 0$	konvergent	konvergent
$-1 < \alpha < 0$	konvergent	divergent
$\alpha \le -1$	divergent	divergent

8.3 Fourierreihen

Ein wichtiger Typ von Reihenentwicklungen wird gegeben durch die Fourierreihen. Fourier war Physiker, und wir wollen zunächst kurz die physikalische Fragestellung, bei deren Behandlung Fourier (1807) erstmals die heute nach ihm benannten Reihen verwendete, heuristisch erläutern.

Fourier befaßte sich mit Problemen der Wärmeleitung in homogenen Medien. Als besonders einfaches (idealisiertes) Wärmeleitungsproblem betrachten wir einen geraden Stab der Länge l, den wir als „unendlich dünn“ annehmen. Seine Punkte können also durch eine Koordinate x beschrieben werden, die das Intervall $[0, l]$ durchläuft. In diesem Stab herrsche an der Stelle x zur Zeit $t \ge 0$ die Temperatur $T(x, t)$. Wir nehmen an, die beiden Enden des Stabes würden von außen ständig auf der Temperatur Null gehalten, und im übrigen sei der Stab vollkommen isoliert. Zur Zeit $t = 0$ herrsche eine bekannte Temperaturverteilung, gegeben durch eine Funktion $f : [0, l] \to \mathbb{R}$. Von der Temperaturverteilung $T(x, t)$ ist also bekannt, dass

$$T(x, 0) = f(x) \qquad \text{für } x \in [0, l]$$

und

$$T(0,t) = T(l,t) = 0 \qquad \text{für } t \geq 0$$

ist. Im Lauf der Zeit werden sich nun auf Grund der Wärmeleitung die Temperaturunterschiede ausgleichen. Die Frage ist, welche Temperaturverteilung zu einer Zeit $t > 0$ herrscht, also wie man die Funktion $T(x,t)$ aus der Anfangsverteilung $T(x,0)$ berechnen kann.

Physikalische Grundannahmen und Überlegungen führen zu der Folgerung, dass die Funktion T (unter, wie üblich, geeigneten Differenzierbarkeitsannahmen) der partiellen Differentialgleichung

$$\frac{\partial T}{\partial t} = \mu \frac{\partial^2 T}{\partial x^2}$$

genügen muß, wobei $\mu > 0$ eine Materialkonstante ist. Wir stellen nun zunächst die Frage der Anfangsverteilung zurück und suchen allgemeiner eine Lösung $\varphi : [0,l] \times [0,\infty) \to \mathbb{R}$ des Problems

$$\frac{\partial \varphi}{\partial t} = \mu \frac{\partial^2 \varphi}{\partial x^2}, \qquad \varphi(0,t) = \varphi(l,t) = 0. \tag{3.1}$$

Um spezielle Lösungen zu finden, kann man untersuchen, ob es Lösungen der Form

$$\varphi(x,t) = h(t)g(x)$$

gibt („Separationsansatz“, „Trennung der Veränderlichen“). Hierzu muß wegen

$$\frac{\partial \varphi}{\partial t} = h'(t)g(x), \qquad \frac{\partial^2 \varphi}{\partial x^2} = h(t)g''(x)$$

also

$$\frac{h'(t)}{h(t)} = \mu \frac{g''(x)}{g(x)}$$

sein. Da die linke Seite nicht von x und die rechte nicht von t abhängt, handelt es sich um eine Konstante c. Somit werden wir auf die Gleichungen

$$\begin{aligned} h'(t) &= ch(t), \\ \mu g''(x) &= cg(x) \end{aligned}$$

geführt. Die erste Differentialgleichung läßt sich sofort lösen:

$$h(t) = be^{ct}$$

mit einer Konstanten b. Bei der zweiten ist die Randbedingung

$$g(0) = g(l) = 0$$

zu berücksichtigen. Es liegt daher der Ansatz $g(x) = \sin ax$ nahe. Die Differentialgleichung ist erfüllt, wenn $al = k\pi$ mit ganzzahligem k ist. Für die Konstanten ergibt sich also

$$a = \frac{k\pi}{l}, \qquad c = -\frac{\mu k^2 \pi^2}{l^2}.$$

Insgesamt erhalten wir, dass bei beliebigem $k \in \mathbb{N}$ durch

$$\varphi_k(x,t) = b_k e^{-\frac{\mu k^2 \pi^2}{l^2} t} \sin \frac{k\pi x}{l}$$

mit $b_k = \text{const.}$ eine Lösung des Problems (3.1) gegeben ist.

Natürlich kann hierdurch i.a. noch nicht die gesuchte Temperaturverteilung $T(x,t)$ gegeben sein, denn es sollte ja $T(x,0)$ eine vorgeschriebene Funktion $f(x)$ sein. Wir erhalten aber

$$\varphi_k(x,0) = b_k \sin \frac{k\pi x}{l},$$

also für jedes k eine sehr spezielle Funktion.

Nun beachte man aber, dass unser Problem „linear“ ist. Dies hat zur Folge, dass die Summe von Lösungen auch eine Lösung ist, also ist durch

$$\varphi(x,t) := \sum_{k=1}^{n} \varphi_k(x,t) = \sum_{k=1}^{n} b_k e^{-\frac{\mu k^2 \pi^2}{l^2} t} \sin \frac{k\pi x}{l}$$

($n \in \mathbb{N}$) ebenfalls eine Lösung gegeben. Für diese Lösung ist

$$\varphi(x,0) = \sum_{k=1}^{n} b_k \sin \frac{k\pi x}{l},$$

und im allgemeinen wird es immer noch nicht möglich sein, die Zahl n und die Koeffizienten b_k so zu wählen, dass $\varphi(x,0) = f(x)$ ist. Hier setzt nun Fouriers entscheidende (und damals kühne) Überlegung ein. Er geht davon aus, dass man die weitgehend willkürliche Funktion f durch die *unendliche* Reihe

$$\sum_{k=1}^{\infty} b_k \sin \frac{k\pi x}{l}$$

darstellen kann, wenn man die Koeffizienten b_k geeignet bestimmt. Alsdann setze man

$$T(x,t) := \sum_{k=1}^{\infty} b_k e^{-\frac{\mu k^2 \pi^2}{l^2} t} \sin \frac{k\pi x}{l}.$$

Wenn die obere Reihe für jedes $x \in [0,l]$ absolut konvergiert, dann nach dem Majorantenkriterium auch die untere. Die „Anfangsbedingung“ $T(x,0) = f(x)$ (für $x \in [0,l]$) und die „Randbedingung“ $T(0,t) = T(l,t) = 0$ (für $t \geq 0$) sind dann erfüllt. Falls es erlaubt ist, die Reihe gliedweise zu differenzieren (einmal nach t, zweimal nach x), erhält man, dass T auch der vorgeschriebenen partiellen Differentialgleichung genügt. Das Problem der Temperaturverteilung ist damit als gelöst anzusehen.

Der springende Punkt bei dieser Argumentation ist natürlich die Frage, ob es möglich ist, eine gegebene Funktion $f : [0,l] \to \mathbb{R}$ mit $f(0) = f(l) = 0$ darzustellen als Reihe der Form

$$f(x) = \sum_{k=1}^{\infty} b_k \sin \frac{k\pi x}{l}.$$

Mit einer Variante dieser Fragestellung befassen wir uns im folgenden. Es bedeutet dabei keine Einschränkung der Allgemeinheit, $l = \pi$ anzunehmen. Andererseits soll die Voraussetzung $f(0) = f(l) = 0$ fallengelassen werden, und wir fragen allgemeiner nach der Entwicklungsmöglichkeit in Reihen der Form

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx).$$

Es ist bequem, Funktionen zu betrachten, die auf ganz $\mathbb{R}$ definiert sind. Wenn eine solche Funktion durch die obige Reihe dargestellt wird, ist $f(x + 2\pi) = f(x)$ für $x \in \mathbb{R}$. Funktionen mit dieser Eigenschaft nennen wir 2π-periodisch.

Wir nehmen nun an, für eine gegebene Funktion $f : \mathbb{R} \to \mathbb{R}$ existiere in der Tat eine Darstellung der Form

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx), \quad x \in \mathbb{R},$$

wobei die Reihe sogar gleichmäßig konvergent sei. Dann lassen sich die Koeffizienten a_k, b_k folgendermaßen berechnen. Multiplikation mit $\cos mx$ und Integration über $[0, 2\pi]$ (beachte, dass f wegen der gleichmäßigen Konvergenz stetig ist) ergibt

$$\begin{aligned} &\int_0^{2\pi} f(x) \cos mx \, dx \\ &= \frac{a_0}{2} \int_0^{2\pi} \cos mx \, dx + \int_0^{2\pi} \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx) \cos mx \, dx. \end{aligned}$$

Die rechts stehende Reihe ist ebenfalls gleichmäßig konvergent (nämlich gegen $f(x)\cos mx$; beachte $\|fg\| \le \|f\|\|g\|$), darf also nach Satz 2.4 aus Kapitel 7 gliedweise integriert werden. Es ist also

$$\int_0^{2\pi} f(x)\cos mx\,dx$$
$$= \frac{a_0}{2}\int_0^{2\pi}\cos mx\,dx + \sum_{k=1}^{\infty}\left(a_k\int_0^{2\pi}\cos kx\cos mx\,dx + b_k\int_0^{2\pi}\sin kx\cos mx\,dx\right).$$

Nun findet man leicht durch partielle Integration:

$$\int_0^{2\pi}\cos kx\cos mx\,dx = \int_0^{2\pi}\sin kx\sin mx\,dx = 0 \qquad \text{für } k \ne m$$
$$\int_0^{2\pi}\sin kx\cos mx\,dx = 0,$$
$$\int_0^{2\pi}\cos^2 kx\,dx = \int_0^{2\pi}\sin^2 kx\,dx = \pi \qquad \text{für } k \ge 1.$$

Damit erhält man

$$a_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos kx\,dx \qquad \text{für } k \in \mathbb{N}_0.$$

Ganz analog beweist man

$$b_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin kx\,dx \qquad \text{für } k \in \mathbb{N}.$$

Man bezeichnet nun ganz allgemein für jede 2π-periodische Funktion f und für $k \in \mathbb{N}_0$ die Zahlen

$$a_k := \frac{1}{\pi}\int_0^{2\pi} f(x)\cos kx\,dx,$$
$$b_k := \frac{1}{\pi}\int_0^{2\pi} f(x)\sin kx\,dx,$$

($k \in \mathbb{N}_0$) falls diese Integrale existieren, als die *Fourierkoeffizienten* von f und nennt die Funktionenreihe

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

die *Fourierreihe* von f. Dabei ist (analog wie bei Taylorreihen) nichts darüber ausgesagt, ob diese Reihe überhaupt (und für welche x bzw. in welchem Sinne) konvergiert und ob sie im Konvergenzfall gegen $f(x)$ konvergiert. Die Beantwortung dieser Fragen erfordert eingehendere Untersuchungen.

Bevor wir hierauf eingehen, wollen wir aber die hier sehr zweckmäßige komplexe Schreibweise einführen. Es ist ja

$$e^{ix} = \cos x + i \sin x, \qquad \cos x = \frac{1}{2}(e^{ix} + e^{-ix}), \qquad \sin x = \frac{1}{2i}(e^{ix} - e^{-ix}).$$

Damit ergibt sich für $k \in \mathbb{N}_0$

$$\begin{aligned} a_k \cos kx + b_k \sin kx &= a_k \cdot \frac{1}{2}(e^{ikx} + e^{-ikx}) - b_k \cdot \frac{i}{2}(e^{ikx} - e^{-ikx}) \\ &= \frac{1}{2}(a_k - ib_k)e^{ikx} + \frac{1}{2}(a_k + ib_k)e^{-ikx} \\ &= c_k e^{ikx} + c_{-k} e^{-ikx} \end{aligned}$$

mit

$$c_k := \frac{1}{2}(a_k - ib_k), \qquad c_{-k} := \frac{1}{2}(a_k + ib_k),$$

also für $n \in \mathbb{N}$

$$\begin{aligned} &\frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx) \\ &= c_0 + \sum_{k=1}^{n} (c_k e^{ikx} + c_{-k} e^{-ikx}) \\ &= \sum_{k=-n}^{n} c_k e^{ikx} \end{aligned}$$

Die Folge $\left(\sum_{k=-n}^{n} c_k e^{ikx}\right)_{n\in\mathbb{N}}$ kürzt man ab mit

$$\sum_{k=-\infty}^{\infty} c_k e^{ikx}.$$

Dies ist also ebenfalls die Fourierreihe von f, nur in komplexer Schreibweise.

Es ist weiter zweckmäßig, auch komplexwertige Funktionen $f : \mathbb{R} \to \mathbb{C}$ zuzulassen. Jede solche Funktion läßt sich eindeutig darstellen in der Form

$f = u + iv$ mit reellen Funktionen u, v. Man definiert dann (falls u und v, eingeschränkt auf $[a, b]$, Regelfunktionen sind)

$$\int_a^b f(x)\,dx := \int_a^b u(x)\,dx + i\int_a^b v(x)\,dx.$$

Mit dieser Konvention ist für $k \geq 0$

$$\begin{aligned} c_k &= \frac{1}{2}(a_k - ib_k) \\ &= \frac{1}{2\pi}\int_0^{2\pi} f(x)(\cos kx - i\sin kx)\,dx \\ &= \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,dx \end{aligned}$$

und

$$c_{-k} = \frac{1}{2}(a_k + ib_k) = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{ikx}\,dx,$$

somit

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,dx \qquad \text{für } k \in \mathbb{Z}.$$

Wir wollen hier auch komplexwertige Funktionen f zulassen, erklären dann die Fourierkoeffizienten c_k von f durch diese Gleichung (falls die Integrale existieren) und nennen die Reihe

$$\sum_{k=-\infty}^{\infty} c_k e^{ikx}$$

die *Fourierreihe* von f.

Die Konvergenztheorie der Fourierreihen wird dann besonders einfach und übersichtlich, wenn man nicht nach der (i.a. selbst bei stetigen reellen Funktionen nicht vorliegenden) punktweisen Konvergenz fragt, sondern nach Konvergenz in einem schwächeren Sinne. Mit dieser auch für Anwendungen wichtigen Konvergenz wollen wir uns hier befassen.

Konvergenz im quadratischen Mittel

Wir betrachten die Menge V der Funktionen $f : \mathbb{R} \to \mathbb{C}$ mit $f(x+2\pi) = f(x)$ für $x \in \mathbb{R}$ und der Eigenschaft, dass Real- und Imaginärteil von f, eingeschränkt auf $[0, 2\pi]$, Regelfunktionen sind. Offenbar ist V (mit den üblichen Verknüpfungen) ein Vektorraum über $\mathbb{C}$.

Definition. Für $f, g \in V$ sei

$$\langle f, g\rangle := \frac{1}{2\pi}\int_0^{2\pi} f(x)\overline{g(x)}\,dx.$$

Die komplexe Zahl $\langle f, g\rangle$ heißt das *Skalarprodukt* von f und g.

3.2 Satz. *Die Abbildung $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{C}$ hat die folgenden Eigenschaften (für $f, g, h \in V$ und $\lambda \in \mathbb{C}$):*

(1) $\langle f + g, h\rangle = \langle f, h\rangle + \langle g, h\rangle$
(2) $\langle f, g + h\rangle = \langle f, g\rangle + \langle f, h\rangle$
(3) $\langle \lambda f, g\rangle = \lambda\langle f, g\rangle$
(4) $\langle f, \lambda g\rangle = \overline{\lambda}\langle f, g\rangle$
(5) $\langle g, f\rangle = \overline{\langle f, g\rangle}$
(6) $\langle f, f\rangle \geq 0$ *(d.h. $\langle f, f\rangle$ ist reellwertig und nichtnegativ)*
(7) $|\langle f, g\rangle|^2 \leq \langle f, f\rangle\langle g, g\rangle$
(8) $\sqrt{\langle f + g, f + g\rangle} \leq \sqrt{\langle f, f\rangle} + \sqrt{\langle g, g\rangle}$.

Beweis. (1) – (6) folgen unmittelbar aus der Definition; bei (6) beachte man $f\overline{f} = |f|^2 \geq 0$. Wegen (6) sind die Wurzeln in (8) im Reellen erklärt.

(7) Nach (6) gilt für $f, g \in V$ und $\lambda \in \mathbb{C}$

$$\langle f + \lambda g, f + \lambda g\rangle \geq 0,$$

also unter Verwendung von (1) – (5)

$$\langle f, f\rangle + \overline{\lambda}\langle f, g\rangle + \lambda\overline{\langle f, g\rangle} + \lambda\overline{\lambda}\langle g, g\rangle \geq 0.$$

Ist $\langle g, g\rangle \neq 0$, so kann man

$$\lambda = -\frac{\langle f, g\rangle}{\langle g, g\rangle}$$

einsetzen und erhält die behauptete Ungleichung. Ist $\langle g, g\rangle = 0$, so setze man $\lambda = -n\langle f, g\rangle$; es folgt $|\langle f, g\rangle|^2 \leq 1/2n\langle f, f\rangle$, also (da $n \in \mathbb{N}$ beliebig ist) $|\langle f, g\rangle| \leq 0 = \langle f, f\rangle\langle g, g\rangle$.

(8) Aus (7) folgt

$$\mathrm{Re}\langle f,g\rangle \le |\langle f,g\rangle| \le \sqrt{\langle f,f\rangle\langle g,g\rangle},$$

also

$$\begin{aligned}\langle f+g,f+g\rangle &= \langle f,f\rangle + \langle f,g\rangle + \overline{\langle f,g\rangle} + \langle g,g\rangle \\ &= \langle f,f\rangle + \langle g,g\rangle + 2\mathrm{Re}\langle f,g\rangle \\ &\le \langle f,f\rangle + \langle g,g\rangle + 2\sqrt{\langle f,f\rangle\langle g,g\rangle} \\ &= \left(\sqrt{\langle f,f\rangle} + \sqrt{\langle g,g\rangle}\right)^2,\end{aligned}$$

woraus (8) folgt. ∎

Definition. Für $f \in V$ ist

$$\|f\|_2 := \sqrt{\langle f,f\rangle}$$

die *L_2-Norm* oder *Hilbert-Norm* von f.

Die Ungleichungen (7) und (8) aus Satz 3.2 schreiben sich damit in der Form

$$|\langle f,g\rangle| \le \|f\|_2\|g\|_2$$

(*Cauchy-Schwarzsche Ungleichung*) und

$$\|f+g\|_2 \le \|f\|_2 + \|g\|_2;$$

letzteres ist die Dreiecksungleichung für die L_2-Norm. Für die L_2-Norm gilt auch $\|\lambda f\|_2 = |\lambda|\ \|f\|_2$ und $\|f\|_2 = 0$ für $f = 0$. Allerdings folgt aus $\|f\|_2 = 0$ nicht, dass f die Nullfunktion ist.

Wir erinnern uns daran, dass wir schon früher Konvergenz im Sinne einer Norm erklärt hatten, und definieren:

Definition. Seien $f, f_n \in V$ $(n \in \mathbb{N})$. Die Folge $(f_n)_{n\in\mathbb{N}}$ heißt *konvergent im quadratischen Mittel* (oder konvergent in der L_2-Norm) gegen f, wenn

$$\lim_{n\to\infty} \|f - f_n\|_2 = 0$$

ist.

Die Bezeichnung „im quadratischen Mittel“ ist naheliegend wegen

$$\|f - f_n\|_2^2 = \frac{1}{2\pi} \int_0^{2\pi} |f(x) - f_n(x)|^2 \, dx;$$

der rechts stehende Ausdruck läßt sich als Mittelwert der quadratischen Abweichung von f und f_n ansehen. Man beachte, dass eine im quadratischen Mittel konvergierende Funktionenfolge nicht punktweise konvergent zu sein braucht.

Wir betrachten jetzt speziell die durch

$$e_k(x) := e^{ikx}$$

definierten Elemente $e_k \in V$ $(k \in \mathbb{Z})$ und berechnen $\langle e_k, e_m \rangle$. Nach Definition ist

$$\begin{aligned}
\langle e_k, e_m \rangle &= \frac{1}{2\pi} \int_0^{2\pi} e^{ikx} \overline{e^{imx}} \, dx = \frac{1}{2\pi} \int_0^{2\pi} e^{i(k-m)x} \, dx \\
&= \frac{1}{2\pi} \left(\int_0^{2\pi} \cos(k-m)x \, dx + i \int_0^{2\pi} \sin(k-m)x \, dx \right) \\
&= \begin{cases} 1 & \text{für } k = m, \\ 0 & \text{für } k \neq m. \end{cases}
\end{aligned}$$

Allgemein nennt man zwei Elemente $f, g \in V$ mit $\langle f, g \rangle = 0$ *orthogonal*, und eine Folge $(b_k)_{k \in \mathbb{Z}}$ in V heißt *Orthonormalsystem*, wenn

$$\langle b_k, b_m \rangle = \begin{cases} 1 & \text{für } k = m \\ 0 & \text{für } k \neq m \end{cases}$$

ist.

Im folgenden sei jetzt zunächst ein beliebiges Orthonormalsystem $(b_k)_{k \in \mathbb{Z}}$ gegeben. Ist $f \in V$, so nennt man die durch

$$c_k := \langle f, b_k \rangle$$

definierten komplexen Zahlen die *Fourierkoeffizienten* von f *bezüglich des gegebenen Orthonormalsystems*. Die früher definierten speziellen Fourierkoeffizienten sind also genau diejenigen bezüglich des speziellen Orthonormalsystems $(e_k)_{k \in \mathbb{Z}}$. Die Reihe

$$\sum_{k \in \mathbb{Z}} c_k b_k$$

nennt man die *Fourierreihe* von f *bezüglich des Orthonormalsystems* $(b_k)_{k\in\mathbb{Z}}$.

Sei jetzt $f \in V$ gegeben, und seien $\alpha_{-n}, \ldots, \alpha_n$ beliebige komplexe Zahlen. Wir wollen

$$\left\| f - \sum_{k=-n}^{n} \alpha_k b_k \right\|_2^2$$

berechnen und aus dem Ergebnis wichtige Folgerungen ziehen. Es ist (Summation jeweils von $-n$ bis n)

$$\begin{aligned} \left\| f - \sum \alpha_k b_k \right\|_2^2 &= \left\langle f - \sum \alpha_k b_k, f - \sum \alpha_j b_j \right\rangle \\ &= \langle f, f\rangle - \sum \alpha_k \underbrace{\langle b_k, f\rangle}_{\overline{c_k}} - \sum \overline{\alpha_j} \underbrace{\langle f, b_j\rangle}_{c_j} + \left\langle \sum \alpha_k b_k, \sum \alpha_j b_j \right\rangle \end{aligned}$$

und

$$\left\langle \sum \alpha_k b_k, \sum \alpha_j b_j \right\rangle = \sum_{k,j} \alpha_k \overline{\alpha_j} \langle b_k, b_j\rangle = \sum_k \alpha_k \overline{\alpha_k}.$$

Ferner ist

$$\begin{aligned} \sum |c_k - \alpha_k|^2 &= \sum (c_k - \alpha_k)(\overline{c_k} - \overline{\alpha_k}) \\ &= \sum c_k \overline{c_k} - \sum \alpha_k \overline{c_k} - \sum \overline{\alpha_k} c_k + \sum \alpha_k \overline{\alpha_k}. \end{aligned}$$

Damit ergibt sich

$$\left\| f - \sum \alpha_k b_k \right\|_2^2 = \|f\|_2^2 - \sum |c_k|^2 + \sum |c_k - \alpha_k|^2.$$

Hieran lesen wir folgendes ab:

(1) $\|f - \sum_{k=-n}^{n} \alpha_k b_k\|_2^2$ wird genau dann minimal, wenn $\alpha_k = c_k$ für $k = -n, \ldots, n$) ist. Dies ist also eine Charakterisierung der Fourierkoeffizienten durch eine Extremaleigenschaft: f wird durch eine Linearkombination der Vektoren $b_{-n}, \ldots, b_n$ im Sinne der L_2-Norm genau dann am besten approximiert, wenn man als Koeffizienten die Fourierkoeffizienten wählt.

(2) Wählen wir jetzt $\alpha_k = c_k$, so lautet die obige Gleichung

$$\|f\|_2^2 - \sum_{k=-n}^{n} |c_k|^2 = \left\| f - \sum_{k=-n}^{n} c_k b_k \right\|_2^2,$$

und dies ist ≥ 0. Es folgt die Konvergenz der Reihe $\sum_{k=-\infty}^{\infty} |c_k|^2$ und die Ungleichung

$$\sum_{k=-\infty}^{\infty} |c_k|^2 \leq \|f\|_2^2.$$

Genau dann gilt hier das Gleichheitszeichen, wenn

$$\lim_{n\to\infty} \left\| f - \sum_{k=-n}^{n} c_k b_k \right\|_2 = 0$$

ist.

Wir fassen die erhaltenen Ergebnisse zusammen und ergänzen sie durch einige neue Bezeichnungen.

3.3 Satz. *Sei $(b_k)_{k\in\mathbb{Z}}$ ein Orthonormalsystem in V, sei $f \in V$, und seien $c_k = \langle f, b_k\rangle$, $k \in \mathbb{Z}$, die Fourierkoeffizienten von f bezüglich des Orthonormalsystems $(b_k)_{k\in\mathbb{Z}}$. Dann gilt*

$$\sum_{k=-\infty}^{\infty} |c_k|^2 \leq \|f\|_2^2 \qquad \textit{(Besselsche Ungleichung).}$$

Die Fourierreihe von f konvergiert genau dann im quadratischen Mittel gegen f, d.h. es gilt

$$\lim_{n\to\infty} \left\| f - \sum_{k=-n}^{n} c_k b_k \right\|_2 = 0,$$

wenn

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = \|f\|_2^2 \qquad \textit{(Parsevalsche Vollständigkeitsrelation)} \tag{3.4}$$

gilt. Das Orthonormalsystem $(b_k)_{k\in\mathbb{Z}}$ heißt vollständig, *wenn (3.4) für jedes $f \in V$ gilt.*

Nach diesen allgemeinen Betrachtungen kehren wir zum speziellen Orthonormalsystem $(e_k)_{k\in\mathbb{Z}}$ zurück und zeigen:

3.5 Satz. *Das Orthonormalsystem $(e_k)_{k\in\mathbb{Z}}$ ist vollständig.*

Für jede Funktion $f \in V$ konvergiert also die Fourierreihe

$$\sum_{k=-\infty}^{\infty} c_k e^{ikx}$$

mit

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx}\, dx$$

im quadratischen Mittel gegen $f(x)$, d.h. es ist

$$\lim_{n\to\infty} \int_0^{2\pi} \left| f(x) - \sum_{k=-n}^{n} c_k e^{ikx} \right|^2 dx = 0$$

oder, anders geschrieben,

$$\lim_{n\to\infty} \|f - S_n[f]\|_2 = 0,$$

wenn wir (wie auch im folgenden) mit

$$S_n[f] := \sum_{k=-n}^{n} c_k e^{ikx}$$

die n-te Partialsumme der Fourierreihe von f bezeichnen. Nach Satz 3.3 ist die Konvergenz der Fourierreihe gleichwertig mit (3.4), oder anders geschrieben mit

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = \frac{1}{2\pi} \int_0^{2\pi} |f(x)|^2\, dx. \tag{3.6}$$

Zum Beweis von (3.6) nehmen wir zunächst f von einer sehr speziellen Gestalt an. Sei $0 \le a \le b \le 2\pi$ und

$$f(x) = \chi_{[a,b]}(x) := \begin{cases} 1 & \text{für } a \le x \le b, \\ 0 & \text{für } 0 \le x < a \text{ und } b < x \le 2\pi; \end{cases}$$

im übrigen sei $f(x+2\pi) = f(x)$. Für die komplexen Fourierkoeffizienten von f erhalten wir

$$c_0 = \frac{1}{2\pi} \int_0^{2\pi} f(x)\, dx = \frac{b-a}{2\pi}$$

und für $k \neq 0$

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx}\, dx = \frac{1}{2\pi} \left[\frac{-1}{ik} e^{-ikx} \right]_a^b = \frac{i}{2\pi k} (e^{-ikb} - e^{-ika}),$$

also

$$|c_k|^2 = \frac{1-\cos k(b-a)}{2\pi^2 k^2}.$$

Somit ist

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = \frac{(b-a)^2}{4\pi^2} + \frac{1}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{k^2} - \frac{1}{\pi^2}\sum_{k=1}^{\infty}\frac{\cos k(b-a)}{k^2}.$$

Zur Berechnung der letzten Summe folgt eine längere Zwischenrechnung.

Behauptung.

$$\sum_{k=1}^{\infty}\frac{\sin kx}{k} = \frac{\pi - x}{2} \qquad \text{für } 0 < x < 2\pi.$$

Beweis.

$$\begin{aligned} 1 + 2\sum_{k=1}^{n}\cos kt &= \sum_{k=-n}^{n} e^{ikt} = e^{-int}\sum_{k=0}^{2n} e^{ikt} \\ &= e^{-int}\frac{e^{i(2n+1)t}-1}{e^{it}-1} = \frac{e^{i(n+\frac{1}{2})t} - e^{-i(n+\frac{1}{2})t}}{e^{\frac{i}{2}t} - e^{-\frac{i}{2}t}} \\ &= \frac{\sin(n+\frac{1}{2})t}{\sin\frac{1}{2}t}, \end{aligned}$$

also

$$\sum_{k=1}^{n}\cos kt = \frac{\sin(n+\frac{1}{2})t}{2\sin\frac{1}{2}t} - \frac{1}{2}.$$

Integration von π bis x ergibt

$$\sum_{k=1}^{n}\frac{\sin kx}{k} = \int_{\pi}^{x}\frac{\sin(n+\frac{1}{2})t}{2\sin\frac{1}{2}t}\,dt - \frac{x-\pi}{2}.$$

Für $n \to \infty$ konvergiert das Integral gegen 0, da die reellen Fourierkoeffizienten einer Regelfunktion wegen der Besselschen Ungleichung gegen 0 konvergieren. ∎

Behauptung. Für $\delta \in (0,\pi)$ ist die Reihe $\sum_{k=1}^{\infty}\sin kx/k$ gleichmäßig konvergent in $[\delta, 2\pi - \delta]$.

Beweis. Setze

$$s_n(x) := \sum_{k=1}^{n} \sin kx = \operatorname{Im}\left(\sum_{k=1}^{n} e^{ikx}\right).$$

Für $\delta \le x \le 2\pi - \delta$ gilt

$$\begin{aligned} |s_n(x)| &\le \left|\sum_{k=1}^{n} e^{ikx}\right| = |e^{ix}| \left|\frac{e^{inx}-1}{e^{ix}-1}\right| \\ &\le \frac{2}{\left|e^{\frac{ix}{2}} - e^{-\frac{ix}{2}}\right|} = \frac{1}{\sin\frac{x}{2}} \le \frac{1}{\sin\frac{\delta}{2}}. \end{aligned}$$

Für $m > n > 0$ folgt

$$\begin{aligned} \left|\sum_{k=n}^{m} \frac{\sin kx}{k}\right| &= \left|\sum_{k=n}^{m} \frac{s_k(x) - s_{k-1}(x)}{k}\right| \\ &= \left|\sum_{k=n}^{m} s_k(x)\left(\frac{1}{k} - \frac{1}{k+1}\right) + \frac{s_m(x)}{m+1} - \frac{s_{n-1}(x)}{n}\right| \\ &\le \frac{1}{\sin\frac{\delta}{2}}\left(\frac{1}{n} - \frac{1}{m+1} + \frac{1}{m+1} + \frac{1}{n}\right) = \frac{2}{n\sin\frac{\delta}{2}}. \end{aligned}$$

Nach dem Cauchy-Kriterium 1.1 folgt die gleichmäßige Konvergenz. ∎

Behauptung.

$$\sum_{k=1}^{\infty} \frac{\cos kx}{k^2} = \left(\frac{x-\pi}{2}\right)^2 - \frac{\pi^2}{12} \qquad \text{für } x \in [0, 2\pi].$$

Beweis. Setze $F(x) := \sum_{k=1}^{\infty} \cos kx/k^2$. Diese Reihe ist nach dem Majorantenkriterium gleichmäßig konvergent. Für beliebiges $\delta > 0$ ist die Reihe $-\sum_{k=1}^{\infty} \sin kx/k$ auf $[\delta, 2\pi - \delta]$ nach der 2. Behauptung gleichmäßig konvergent. Nach Satz 1.3 und der 1. Behauptung folgt

$$F'(x) = \frac{x-\pi}{2}, \qquad \text{also } F(x) = \left(\frac{x-\pi}{2}\right)^2 + c.$$

Nun ist

$$\int_0^{2\pi} F(x)\,dx = \frac{\pi^3}{6} + 2\pi c,$$

und dies ist andererseits nach Satz 2.4 aus Kapitel 7

$$= \int_0^{2\pi} \sum_{k=1}^{\infty} \frac{\cos kx}{k^2} dx = \sum_{k=1}^{\infty} \frac{1}{k^2} \int_0^{2\pi} \cos kx dx = 0.$$

Also ist $c = -\frac{\pi^2}{12}$, womit die 3. Behauptung bewiesen ist. ∎

Jetzt können wir unsere oben begonnene Rechnung fortsetzen und erhalten

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = \frac{(b-a)^2}{4\pi^2} + \frac{1}{\pi^2} \cdot \frac{\pi^2}{6} - \frac{1}{\pi^2} \left(\frac{b-a-\pi}{2} \right)^2 + \frac{1}{\pi^2} \cdot \frac{\pi^2}{12}$$
$$= \frac{b-a}{2\pi} = \|f\|_2^2.$$

Die Gleichung (3.6) ist für diese spezielle Funktion f also bewiesen.

Nun sei $f : [0, 2\pi] \to \mathbb{R}$ eine Treppenfunktion. Dann gibt es endlich viele Funktionen $f_1, \ldots, f_r$ der oben betrachteten Art und reelle Zahlen $\alpha_1, \ldots, \alpha_r$ mit $f = \sum_{j=1}^r \alpha_j f_j$. Allgemein sei mit $S_n[g]$ die n-te Partialsumme der Fourierreihe von g bezeichnet, also

$$S_n[g](x) := \sum_{k=-n}^{n} c_k[g] e^{ikx}, \qquad c_k[g] := \frac{1}{2\pi} \int_0^{2\pi} g(x) e^{-ikx}\, dx.$$

Dann gilt

$$\|f - S_n[f]\|_2 = \Big\| \sum_{j=1}^r \alpha_j f_j - \sum_{j=1}^r \alpha_j S_n[f_j] \Big\|_2$$
$$\leq \sum_{j=1}^r |\alpha_j| \|f_j - S_n[f_j]\|_2 \to 0 \qquad \text{für } n \to \infty$$

nach dem bereits Bewiesenen und Satz 3.3.

Schließlich sei $f : [0, 2\pi] \to \mathbb{R}$ eine Regelfunktion. Zu $\varepsilon \in \mathbb{R}^+$ existiert eine Treppenfunktion t auf $[0, 2\pi]$ mit $|f(x) - t(x)| \leq \varepsilon/2$ für $x \in [0, 2\pi]$. Für $g := f - t$ folgt

$$\|g - S_n[g]\|_2^2 = \|g\|_2^2 - \sum_{k=-n}^{n} |c_k[g]|^2 \leq \|g\|_2^2 \leq \left(\frac{\varepsilon}{2} \right)^2.$$

Nach dem bereits Bewiesenen existiert ein $n_0 \in \mathbb{N}$ mit $\|t - S_n[t]\|_2 < \varepsilon/2$ für $n \geq n_0$. Für $n \geq n_0$ gilt also

$$\|f - S_n[f]\|_2 \leq \|g - S_n[g]\|_2 + \|t - S_n[t]\|_2 < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Somit ist

$$\lim_{n\to\infty} \|f - S_n[f]\|_2 = 0.$$

Die Ausdehnung auf $f \in V$ ist jetzt trivial und wir haben (3.6) für alle $f \in V$ bewiesen.

Gegenüber der Konvergenz im quadratischen Mittel ist die Beurteilung der punktweisen Konvergenz einer Fourierreihe wesentlich schwieriger. Selbst für stetiges f braucht die Folge $(S_n[f])_{n\in\mathbb{N}}$ der Partialsummen der Fourierreihe von f nicht punktweise gegen f zu konvergieren. Erstaunlicherweise gilt aber, dass die arithmetischen Mittel dieser Partialsummen sogar gleichmäßig gegen f konvergieren, wie der folgende Satz zeigt.

3.7 Satz (Fejér)**.** *Ist $f : \mathbb{R} \to \mathbb{R}$ 2π-periodisch und stetig und wird*

$$\sigma_n[f] := \frac{1}{n}(S_0[f] + \cdots + S_{n-1}[f])$$

gesetzt, so konvergiert die Folge $(\sigma_n[f])_{n\in\mathbb{N}}$ gleichmäßig gegen f.

Beweis. Es ist

$$\begin{aligned} S_n[f](x) &= \sum_{k=-n}^{n} \left(\frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-ikt}\, dt \right) e^{ikx} \\ &= \frac{1}{2\pi} \int_0^{2\pi} f(t) \sum_{k=-n}^{n} e^{ik(x-t)}\, dt \\ &= \frac{1}{2\pi} \int_0^{2\pi} f(t) D_n(x-t)\, dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-t) D_n(t)\, dt \qquad \text{[Substitution und Periodizität]} \end{aligned}$$

mit

$$D_n(x) := \sum_{k=-n}^{n} e^{ikx}.$$

Wir bemerken

$$\frac{1}{2\pi} \int_0^{2\pi} D_n(x)\, dx = 1$$

und

$$(e^{ix}-1)D_n(x) = e^{i(n+1)x} - e^{-inx}.$$

Multiplikation mit $e^{-\frac{ix}{2}}$ ergibt

$$D_n(x) = \frac{\sin(n+\frac{1}{2})x}{\sin\frac{x}{2}}.$$

Setze

$$F_n := \frac{1}{n}(D_0 + \cdots + D_{n-1}),$$

dann ist

$$\frac{1}{2\pi}\int_0^{2\pi} F_n(x)\,dx = 1$$

und

$$F_n(x) = \frac{1}{n}\sum_{k=0}^{n-1}\frac{\sin(k+\frac{1}{2})x}{\sin\frac{x}{2}}.$$

Aus

$$\begin{aligned}\sum_{k=0}^{n-1} e^{i(k+\frac{1}{2})x} &= e^{\frac{ix}{2}}\sum_{k=0}^{n-1} e^{ikx} = e^{\frac{ix}{2}}\frac{e^{inx}-1}{e^{ix}-1}\\ &= \frac{e^{inx}-1}{e^{\frac{ix}{2}}-e^{-\frac{ix}{2}}} = \frac{\cos nx - 1 + i\sin nx}{2i\sin\frac{x}{2}}\end{aligned}$$

folgt

$$\sum_{k=0}^{n-1}\sin(k+\tfrac{1}{2})x = \frac{1-\cos nx}{2\sin\frac{x}{2}} = \frac{\sin^2\frac{nx}{2}}{\sin\frac{x}{2}},$$

also

$$F_n(x) = \frac{1}{n}\left(\frac{\sin\frac{nx}{2}}{\sin\frac{x}{2}}\right)^2,$$

was insbesondere $F_n(x) \geq 0$ impliziert. Nun ist

$$\begin{aligned}\sigma_n[f](x) &= \sum_{k=0}^{n-1}\frac{1}{n}S_k[f](x) = \frac{1}{n}\sum_{k=0}^{n-1}\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-t)D_k(t)\,dt\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-t)F_n(t)\,dt,\end{aligned}$$

also

$$\begin{aligned}
&|f(x) - \sigma_n[f](x)| \\
&= \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} F_n(t)\,dt \cdot f(x) - \frac{1}{2\pi} \int_{0}^{2\pi} f(x-t) F_n(t)\,dt \right| \\
&\le \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x-t) - f(x)| F_n(t)\,dt.
\end{aligned}$$

Zu gegebenem $\varepsilon > 0$ existiert ein $\delta > 0$ mit $\delta < \pi$ und

$$|f(x-t) - f(x)| \le \varepsilon \qquad \text{für } |t| < \delta,$$

da f nach Satz 3.5 aus Kapitel 4 gleichmäßig stetig ist; ferner ist

$$|f(x-t) - f(x)| \le M \qquad \text{für alle } x \text{ und } t$$

mit passendem M. Es folgt

$$\begin{aligned}
&\int_{-\pi}^{\pi} |f(x-t) - f(x)| F_n(t)\,dt \\
&= \int_{|t|\le\delta} |f(x-t) - f(x)| F_n(t)\,dt + \int_{|t|\ge\delta} |f(x-t) - f(x)| F_n(t)\,dt \\
&\le \varepsilon\, 2\pi + M \int_{|t|\ge\delta} \frac{1}{n} \left[\frac{\sin\frac{nt}{2}}{\sin\frac{t}{2}} \right]^2 dt \le \varepsilon + \frac{M}{n} \frac{2\pi}{(\sin\frac{\delta}{2})^2}.
\end{aligned}$$

Für $n \ge n_0$ mit passendem n_0 ist der letzte Summand kleiner als $\varepsilon\, 2\pi$, also ist $|f(x) - \sigma_n[f](x)| < 2\varepsilon$ für $n \ge n_0$. ∎

Wir wollen an den Satz von Fejér noch zwei Bemerkungen anfügen. Zunächst soll noch einmal kurz das Wesen der dort verwendeten Summierung herausgestellt werden. Für eine Reihe

$$\sum_{k=0}^{\infty} a_k$$

bildet man die Mittel der Partialsummen, also

$$\begin{aligned}\sigma_n &= \frac{s_0 + s_1 + \cdots + s_n}{n+1} \\ &= \frac{a_0 + (a_0 + a_1) + (a_0 + a_1 + a_2) + \cdots + (a_0 + a_1 + \cdots + a_n)}{n+1} \\ &= a_0 + \frac{n}{n+1} a_1 + \frac{n-1}{n+1} a_2 + \cdots + \frac{1}{n+1} a_n \\ &= \sum_{k=0}^{n} \left(1 - \frac{k}{n+1}\right) a_k.\end{aligned}$$

Gegenüber der Bildung der Partialsumme $\sum_{k=0}^{n} a_k$ werden hierbei also die Summanden a_k mit Gewichten versehen, die von n abhängen und mit k abnehmen.

Sodann wollen wir darauf hinweisen, dass man aus dem Satz von Fejér leicht einen wichtigen Satz über die gleichmäßige Approximierbarkeit stetiger reeller Funktionen durch Polynome herleiten kann.

3.8 Satz (Approximationssatz von Weierstraß)**.** *Sei $g : [a, b] \to \mathbb{R}$ stetig und $\varepsilon > 0$. Dann existiert ein Polynom P mit*

$$|g(x) - P(x)| \leq \varepsilon \quad \textit{für alle } x \in [a, b].$$

Beweis. O.B.d.A. sei $a = -1$ und $b = 1$. Setze

$$f(\varphi) := g(\cos \varphi) \qquad \text{für } -\pi \leq \varphi \leq \pi$$

und ergänze f zu einer 2π-periodischen Funktion. Wegen $f(-\pi) = f(\pi)$ ist f stetig. Nach dem Satz von Fejér existiert ein $n \in \mathbb{N}$ mit

$$|f(\varphi) - \sigma_n[f](\varphi)| \leq \varepsilon \qquad \text{für alle } \varphi \in \mathbb{R}.$$

Die reelle Fourierreihe von f ist von der Form

$$f(\varphi) \sim \frac{a_0}{2} + \sum_{k=0}^{\infty} a_k \cos k\varphi,$$

denn wegen $f(\varphi) = f(-\varphi)$ verschwinden alle b_k. Da $\cos k\varphi$ sich als Polynom in $\cos \varphi$ ausdrücken läßt, ist

$$\sigma_n[f](\varphi) = P(\cos \varphi)$$

mit einem Polynom P. Es ist also

$$|g(\cos \varphi) - P(\cos \varphi)| \leq \varepsilon \qquad \text{für alle } \varphi$$

und daher

$$|g(x) - P(x)| \leq \varepsilon \quad \text{für alle } x \in [-1, 1]. \qquad \blacksquare$$

Analysis II

9 Metrische Räume

In diesem Kapitel sollen einige Aussagen über Konvergenz, Stetigkeit und damit zusammenhängende Begriffsbildungen in einem allgemeineren Rahmen betrachtet werden. Diese Verallgemeinerung geschieht nicht um ihrer selbst willen, sondern ist zweckmäßig, damit man in verschiedenartigen konkreten Fällen nicht gleichartige Sachverhalte stets neu beweisen muß.

Zur Motivation soll zunächst an einige Definitionen aus Analysis I erinnert werden. In Analysis I, Abschnitt 3.1, wurde definiert:

Definition. Sei $(a_n)_{n\in\mathbb{N}}$ eine Folge in $\mathbb{R}$, sei $a \in \mathbb{R}$. Die Folge $(a_n)_{n\in\mathbb{N}}$ heißt *konvergent gegen* a, wenn gilt:

$$\forall \varepsilon \in \mathbb{R}^+ \ \exists n_0 \in \mathbb{N} \ \forall n \geq n_0 : |a_n - a| < \varepsilon.$$

In Abschnitt 5.2 haben wir in völlig gleichlautender Weise die Konvergenz einer Folge komplexer Zahlen erklärt. Dabei bezeichnete $|z|$ den Betrag einer komplexen Zahl z.

In Abschnitt 8.1 tauchte eine ähnliche Definition auf. Sei $D \subseteq \mathbb{R}$. Für jede beschränkte Funktion $f : D \to \mathbb{R}$ ist durch

$$\|f\| := \sup_{x\in D} |f(x)|$$

die *Supremumsnorm* von f erklärt. Seien jetzt f, f_n $(n \in \mathbb{N})$ beschränkte reelle Funktionen auf D.

Definition. Die Folge $(f_n)_{n\in\mathbb{N}}$ *konvergiert gleichmäßig* (oder *im Sinne der Supremumsnorm*) gegen f, wenn gilt:

$$\forall \varepsilon \in \mathbb{R}^+ \ \exists n_0 \in \mathbb{N} \ \forall n \geq n_0 : \|f_n - f\| < \varepsilon.$$

Die Definition sieht formal wieder genau gleich aus wie oben, obwohl die Folgenglieder keine Zahlen mehr sind, sondern Funktionen, und die Norm

$\|\cdot\|$ daher eine andere Bedeutung hat. Trotzdem erinnern wir uns, dass einige einfache Aussagen über diese Konvergenz sich völlig analog wie im Bereich der reellen Zahlen beweisen ließen. Wie eine genauere Analyse dieser Beweise zeigt, werden dabei nur wenige formale Eigenschaften des „Abstandes“ $\|f_n - f\|$ benutzt. Man kann daher eine abstrakte Konvergenztheorie entwickeln, wenn man nur auf einer Grundmenge einen Abstandsbegriff mit entsprechenden formalen Eigenschaften zur Verfügung hat. Einen solchen Abstand nennt man eine „Metrik“ und eine Menge mit einer Metrik dann einen „metrischen Raum“.

Nicht nur Konvergenz läßt sich in diesem allgemeinen Rahmen behandeln. In Analysis I, Abschnitt 5.2, haben wir erklärt:

Definition. Sei $f : D \to \mathbb{R}$ eine Funktion (mit $D \subseteq \mathbb{R}$) und $a \in D$. Die Funktion f heißt *stetig in* a, wenn gilt:

$$\forall \varepsilon \in \mathbb{R}^+ \ \exists \delta \in \mathbb{R}^+ \ \forall x \in D : (|x - a| < \delta \Rightarrow |f(x) - f(a)| < \varepsilon).$$

Auch hier kommen nur die „Abstände“ $|x - a|$ und $|f(x) - f(a)|$ vor. Die Beweise einiger einfacher Aussagen über stetige Funktionen benutzen nur die formalen Eigenschaften dieser Abstände. Dementsprechend läßt sich auch Stetigkeit allgemein in metrischen Räumen behandeln.

Im Zusammenhang mit der Untersuchung stetiger Funktionen hatten wir ferner in Analysis I sogenannte topologische Begriffe benutzt wie *Umgebung, offen, abgeschlossen, kompakt*. Auch diese Begriffe erfordern zu ihrer Definition nur den Abstandsbegriff (oder noch weniger) und lassen sich daher in wesentlich allgemeinerem Rahmen erfolgreich einsetzen.

9.1 Metrische und topologische Grundbegriffe

1.1 Definition. *Sei M eine nichtleere Menge. Eine* Metrik *auf M ist eine Abbildung $d : M \times M \to \mathbb{R}$ mit folgenden Eigenschaften. Für alle $x, y, z \in M$ gilt*

(a) $d(x, y) \geq 0$; $d(x, y) = 0 \Leftrightarrow x = y$,
(b) $d(x, y) = d(y, x)$,
(c) $d(x, z) \leq d(x, y) + d(y, z)$ *(„Dreiecksungleichung“).*

Ist d eine Metrik auf M, so heißt die Zahl $d(x, y)$ der Abstand *von x und y, und das Paar (M, d) heißt* metrischer Raum.

Bemerkung. Sei (M, d) ein metrischer Raum. Die Elemente von M werden häufig als „Punkte“ bezeichnet. Oft wird auch einfach M als metrischer Raum bezeichnet, wenn aus dem Zusammenhang klar ist, welche Metrik auf M vorgegeben ist.

Beispiele. (1) $M = \mathbb{R}$, $d(x, y) := |x - y|$,

(2) $M = \mathbb{C}$, $d(x, y) := |x - y|$,

(3) Sei M die Menge aller endlichen 0-1-Folgen der Länge n. Für $x = (\xi_1, \ldots, \xi_n)$, $y = (\eta_1, \ldots, \eta_n) \in M$, $\xi_i, \eta_i \in \{0, 1\}$, sei $d(x, y)$ die Anzahl der Indizes $i \in \{1, \ldots, n\}$ mit $\xi_i \neq \eta_i$. Dann ist d eine Metrik auf M, und $d(x, y)$ heißt der *Hamming-Abstand* von x und y,

(4) $M = \mathbb{R}^n$,

$$d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) := \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

(„euklidische Metrik“, „ℓ^2-Metrik“),

(5) $M = \mathbb{R}^2$,

$$d((x_1, x_2), (y_1, y_2)) := |x_1 - y_1| + |x_2 - y_2|$$

(„Taxi-Metrik“, „ℓ^1-Metrik“),

(6) $M \neq \emptyset$,

$$d(x, y) := \begin{cases} 1, & \text{wenn } x \neq y, \\ 0, & \text{wenn } x = y. \end{cases}$$

(„diskrete Metrik“),

(7) Sei V ein Vektorraum über $\mathbb{R}$ oder $\mathbb{C}$. Eine *Norm* auf V ist eine Abbildung $\| \cdot \| : V \to \mathbb{R}$ mit folgenden Eigenschaften. Für alle $x, y \in V$ gilt

(a) $\|x\| \geq 0; \|x\| = 0 \Leftrightarrow x = 0$ (Nullvektor),
(b) $\|\lambda x\| = |\lambda| \|x\|$ für $\lambda \in \mathbb{R}$ bzw. $\lambda \in \mathbb{C}$,
(c) $\|x + y\| \leq \|x\| + \|y\|$.

Ist $\| \cdot \|$ eine Norm auf V, so wird durch

$$d(x, y) := \|x - y\| \qquad \text{für } x, y \in V$$

eine Metrik auf V induziert.

(8) Sei M die Menge aller Folgen reeller Zahlen. Auf M wird eine Metrik erklärt durch

$$d(x,y) := \sum_{j=1}^{\infty} \frac{1}{2^j} \frac{|x_j - y_j|}{1 + |x_j - y_j|}$$

für $x = (x_j)_{j\in\mathbb{N}}$, $y = (y_j)_{j\in\mathbb{N}} \in M$.

(9) Seien (M,d), (M',d') metrische Räume. Dann werden auf dem Produkt $M \times M'$ Metriken erklärt durch

$$d_1((x,x'),(y,y')) := d(x,y) + d'(x',y')$$

und durch

$$d_2((x,x'),(y,y')) := \sqrt{d(x,y)^2 + d'(x',y')^2}.$$

Unmittelbar einzusehen ist auch die Richtigkeit der folgenden Definition und Behauptung.

1.2 Definition. *Sei (M,d) ein metrischer Raum und $\emptyset \neq M' \subseteq M$ eine Teilmenge. Dann wird durch die Einschränkung*

$$d' := d|_{M'\times M'}$$

auf M' eine Metrik d' gegeben. (M',d') *heißt* Unterraum *oder* Teilraum *von* (M,d).

Im folgenden sei (M,d) ein gegebener metrischer Raum.

1.3 Lemma. Für alle $x,\overline{x},y,\overline{y} \in M$ gilt die „Viereckungleichung“

$$|d(x,\overline{x}) - d(y,\overline{y})| \leq d(x,y) + d(\overline{x},\overline{y}).$$

Beweis. Aus der Dreiecksungleichung folgt

$$\begin{aligned} d(x,\overline{x}) &\leq d(x,y) + d(y,\overline{x}) \\ &\leq d(x,y) + d(y,\overline{y}) + d(\overline{y},\overline{x}) \\ &= d(x,y) + d(y,\overline{y}) + d(\overline{x},\overline{y}), \end{aligned}$$

also

$$d(x,\overline{x}) - d(y,\overline{y}) \leq d(x,y) + d(\overline{x},\overline{y}).$$

Analog ergibt sich

$$d(y,\overline{y}) - d(x,\overline{x}) \leq d(x,y) + d(\overline{x},\overline{y})$$

und damit die Behauptung. ∎

Wir übertragen nun einige aus Analysis I bekannte Begriffsbildungen metrischer und topologischer Art in den allgemeineren Rahmen metrischer Räume und beweisen darüber eine Reihe einfacher Aussagen, die im Spezialfall bereits vertraut sind.

1.4 Definition. *Eine Teilmenge $A \subseteq M$ heißt* beschränkt*, wenn es eine Zahl $c \in \mathbb{R}^+$ gibt mit $d(x,y) < c$ für alle $x, y \in A$. Ist A beschränkt, so heißt die reelle Zahl*

$$\delta(A) := \sup_{x,y \in A} d(x,y)$$

der Durchmesser *von A.*

1.5 Satz. *Die Vereinigung von endlich vielen beschränkten Mengen ist beschränkt.*

Der Beweis ist eine einfache Übungsaufgabe.

1.6 Definition. *Sei $x \in M$, $\varepsilon \in \mathbb{R}^+$. Die Menge*

$$U(x,\varepsilon) := \{y \in M \mid d(x,y) < \varepsilon\}$$

heißt offene Kugel *um x mit Radius ε.*

1.7 Definition. *Eine Teilmenge $A \subseteq M$ heißt* Umgebung *des Punktes $x \in M$, wenn ein $\varepsilon \in \mathbb{R}^+$ existiert mit $U(x,\varepsilon) \subseteq A$. Die Menge A heißt* offen*, wenn zu jedem $x \in A$ ein $\varepsilon \in \mathbb{R}^+$ existiert mit $U(x,\varepsilon) \subseteq A$.*

Eine Menge ist also genau dann offen, wenn sie Umgebung für jeden ihrer Punkte ist.

1.8 Satz. *Jede offene Kugel ist offen.*

Beweis. Die offene Kugel $U(x,\varepsilon)$ in M sei gegeben. Sei $y \in U(x,\varepsilon)$. Dann ist $d(x,y) < \varepsilon$, also $\varepsilon' := \varepsilon - d(x,y) > 0$. Ist $z \in U(y,\varepsilon')$, so ist $d(x,z) \leq d(x,y) + d(y,z) < d(x,y) + \varepsilon' = \varepsilon$, also $z \in U(x,\varepsilon)$. Da $z \in U(y,\varepsilon')$ beliebig war, ist $U(y,\varepsilon') \subseteq U(x,\varepsilon)$. Da $y \in U(x,\varepsilon)$ beliebig war, ist $U(x,\varepsilon)$ also offen. ∎

Das System aller offenen Teilmengen von M hat die folgenden Eigenschaften.

1.9 Satz. *Es gilt*
(a) Die Vereinigung von beliebig vielen offenen Mengen ist offen.
(b) Der Durchschnitt von endlich vielen offenen Mengen ist offen.
(c) $\emptyset$ und M sind offen.

Beweis. (a) Sei $(A_i)_{i\in I}$ (I eine Indexmenge) eine Familie offener Teilmengen von M. Setze $A := \bigcup_{i\in I} A_i$. Sei $x \in A$. Dann gibt es ein $i \in I$ mit $x \in A_i$. Da A_i offen ist, existiert ein $\varepsilon \in \mathbb{R}^+$ mit $U(x,\varepsilon) \subseteq A_i \subseteq A$. Da $x \in A$ beliebig war, ist A offen.

(b) Sei $(A_i)_{i\in E}$ (E eine endliche Indexmenge) eine endliche Familie offener Teilmengen von M. Setze $B := \bigcap_{i\in E} A_i$. Sei $x \in B$. Für jedes $i \in E$ gilt $x \in A_i$; da A_i offen ist, existiert ein $\varepsilon_i \in \mathbb{R}^+$ mit $U(x,\varepsilon_i) \subseteq A_i$. Setze $\varepsilon := \min\{\varepsilon_i \mid i \in E\}$. Da E endlich ist, existiert das Minimum und ist positiv. Es gilt $U(x,\varepsilon) \subseteq U(x,\varepsilon_i) \subseteq A_i$ für alle $i \in E$, also $U(x,\varepsilon) \subseteq B$. Da $x \in B$ beliebig war, ist B offen.

(c) ist trivial. ∎

1.10 Definition. *Sei $A \subseteq M$ und $x \in M$. Der Punkt x heißt* Berührpunkt *von A, wenn jede Umgebung von x einen Punkt von A enthält, und x heißt* Häufungspunkt *von A, wenn jede Umgebung von x einen Punkt aus $A \setminus \{x\}$ enthält. Die Menge A heißt* abgeschlossen*, wenn sie alle ihre Berührpunkte enthält.*

Es sei daran erinnert, dass man (unter Bezugnahme auf eine bestimmte vorliegende Grundmenge M) unter dem *Komplement* einer Teilmenge $A \subseteq M$ die Menge

$$A^c := M \setminus A$$

versteht und dass $(A^c)^c = A$ ist.

1.11 Satz. *Die Teilmenge $A \subseteq M$ ist genau dann abgeschlossen, wenn ihr Komplement A^c offen ist.*

Beweis. Sei A abgeschlossen. Sei $x \in A^c$. Dann ist x nicht Berührpunkt von A, also existiert ein $\varepsilon \in \mathbb{R}^+$ mit $U(x,\varepsilon) \cap A = \emptyset$, das heißt mit $U(x,\varepsilon) \subseteq A^c$. Da $x \in A^c$ beliebig war, ist A^c offen.

Sei A^c offen. Sei x Berührpunkt von A. Angenommen, es wäre $x \notin A$, also $x \in A^c$. Da A^c offen ist, existiert ein $\varepsilon \in \mathbb{R}^+$ mit $U(x,\varepsilon) \subseteq A^c$, also mit $U(x,\varepsilon) \cap A = \emptyset$, folglich ist x kein Berührpunkt von A. Aus dem Widerspruch folgt $x \in A$. Da x ein beliebiger Berührpunkt von A war, ist A abgeschlossen. ∎

Beispiel. In einem Raum mit diskreter Metrik d ist jede Teilmenge offen, denn mit $x \in A$ ist $U(x,1) = \{x\} \subseteq A$. Nach Satz 1.11 ist daher auch jede Teilmenge abgeschlossen.

Im allgemeinen ist eine Teilmenge eines metrischen Raumes weder offen noch abgeschlossen.

Unter Verwendung von Satz 1.11 und den de Morganschen Regeln $(\bigcup A_i)^c = \bigcap A_i^c$ und $(\bigcap A_i)^c = \bigcup A_i^c$ ergibt sich aus Satz 1.9 sofort die folgende duale Aussage.

1.12 Satz. *Es gilt*
(a) Der Durchschnitt von beliebig vielen abgeschlossenen Mengen ist abgeschlossen.
(b) Die Vereinigung von endlich vielen abgeschlossenen Mengen ist abgeschlossen.
(c) M und $\emptyset$ sind abgeschlossen.

1.13 Definition. *Sei $A \subseteq M$ und $x \in M$. Der Punkt x heißt* innerer Punkt *von A, wenn ein $\varepsilon \in \mathbb{R}^+$ existiert mit $U(x,\varepsilon) \subseteq A$ (also wenn A Umgebung von x ist). Der Punkt x heißt* Randpunkt *von A, wenn jede Umgebung von x Punkte aus A und aus A^c enthält. Die Menge A^o aller inneren Punkte von A heißt das* Innere *oder der* offene Kern *von A; die Menge $\overline{A}$ aller Berührpunkte von A heißt die* abgeschlossene Hülle *von A, und die Menge ∂A aller Randpunkte von A heißt der* Rand *von A.*

Bemerkung. Unmittelbar aus den Definitionen folgt $\overline{A} = A \cup \partial A$ und $\partial A = \overline{A} \setminus A^o$.

1.14 Satz. *Sei $A \subseteq M$. Das Innere A^o ist die Vereinigung aller offenen Teilmengen von A; A^o ist offen. $\overline{A}$ ist abgeschlossen und ist der Durchschnitt aller abgeschlossenen Teilmengen von M, die A enthalten.*

Der Beweis kann als Übungsaufgabe dienen.

1.15 Definition. *Eine Teilmenge $A \subseteq M$ heißt* dicht, *wenn $\overline{A} = M$ ist.*

Beispiel. $\mathbb{Q}$ (vgl. Satz 4.8, Kapitel 1) und $\mathbb{R} \setminus \mathbb{Q}$ sind dicht in $\mathbb{R}$.

9.2 Konvergenz und Vollständigkeit

Im folgenden liegt wieder ein fester metrischer Raum (M,d) zugrunde.

Die Konvergenz einer Punktfolge in einem metrischen Raum kann wörtlich so wie für Folgen reeller Zahlen erklärt werden, wenn der Abstand reeller Zahlen durch den allgemeinen Abstand ersetzt wird.

2.1 Definition. *Sei $(x_n)_{n\in\mathbb{N}}$ eine Folge in M, sei $x \in M$. Die Folge $(x_n)_{n\in\mathbb{N}}$* konvergiert gegen *x, und x heißt Grenzwert (*Grenzpunkt, Limes*) der Folge, geschrieben*

$$\lim_{n\to\infty} x_n = x \qquad \textit{oder} \qquad x_n \to x \ \ (n \to \infty),$$

wenn zu jeder Umgebung U von x ein $n_0 \in \mathbb{N}$ existiert mit

$$x_n \in U \qquad \textit{für alle } n \geq n_0.$$

Die Folge $(x_n)_{n\in\mathbb{N}}$ heißt konvergent*, wenn sie konvergent gegen ein $x \in M$ ist.*

Man beachte, dass bei Verwendung der Schreibweisen $\lim x_n = x$ etc. stets klar sein muß, auf welche Metrik sich diese Konvergenz bezieht.

Wir geben noch zwei offensichtlich äquivalente Umformulierungen der Konvergenzdefinition an:

$$\begin{aligned}\lim_{n\to\infty} x_n = x &\Leftrightarrow \forall \varepsilon \in \mathbb{R}^+ \ \exists n_0 \in \mathbb{N} \ \forall n \geq n_0 : d(x_n, x) < \varepsilon \\ &\Leftrightarrow \lim_{n\to\infty} d(x_n, x) = 0.\end{aligned}$$

Einige elementare Aussagen über Konvergenz lassen sich analog wie in Analysis I beweisen.

2.2 Satz. *Der Limes einer konvergenten Folge ist eindeutig bestimmt.*

Beweis. Sei $(x_n)_{n\in\mathbb{N}}$ eine Folge in M, die sowohl gegen x als auch gegen y konvergiert. Angenommen, es wäre $x \neq y$. Dann ist $d(x, y) > 0$. Zu $\varepsilon := \frac{1}{2}d(x, y)$ gibt es nach Voraussetzung ein $n_0 \in \mathbb{N}$ mit $d(x, x_n) < \varepsilon$ für $n \geq n_0$ und ein $n_1 \in \mathbb{N}$ mit $d(y, x_n) < \varepsilon$ für $n \geq n_1$. Für $n \geq \max\{n_0, n_1\}$ gilt also

$$d(x, y) \leq d(x, x_n) + d(x_n, y) < \varepsilon + \varepsilon = d(x, y).$$

Aus diesem Widerspruch folgt $x = y$. ∎

Die Folge $(x_n)_{n\in\mathbb{N}}$ heißt *beschränkt*, wenn die Menge $\{x_n : n \in \mathbb{N}\}$ beschränkt ist.

2.3 Satz. *Jede konvergente Folge ist beschränkt.*

Beweis. Sei $(x_n)_{n\in\mathbb{N}}$ konvergent gegen x. Dann gibt es ein $n_0 \in \mathbb{N}$ mit $d(x, x_n) < 1$ für $n \geq n_0$. Mit

$$c := \max\{1, d(x, x_1), \ldots, d(x, x_{n_0-1})\}$$

gilt also für beliebige $n, m \in \mathbb{N}$

$$d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) \leq 2c.$$ ∎

Bemerkung. Der wichtige Satz von Bolzano-Weierstraß (Satz 1.3, Kapitel 4) aus Analysis I besagt, dass in $\mathbb{R}$ jede beschränkte Folge reeller Zahlen eine konvergente Teilfolge besitzt. Diese Aussage läßt sich nicht auf allgemeine metrische Räume übertragen. Zum Beispiel besitzt eine injektive Folge in einem Raum mit diskreter Metrik keine konvergente Teilfolge, ist aber beschränkt.

2.4 Satz. *Jeder Berührpunkt von $A \subseteq M$ ist Limes einer Folge in A.*

Beweis. Sei x Berührpunkt von A. Für jedes $n \in \mathbb{N}$ ist dann $U(x, \frac{1}{n}) \cap A \neq \emptyset$; wir können also einen Punkt $x_n \in U(x, \frac{1}{n}) \cap A$ auswählen. Damit ist eine Folge $(x_n)_{n\in\mathbb{N}}$ definiert, die wegen $d(x, x_n) < \frac{1}{n}$ gegen x konvergiert. ∎

Hieraus folgt insbesondere: Die Menge $A \subseteq M$ ist genau dann abgeschlossen, wenn der Grenzwert jeder konvergenten Folge in A ebenfalls in A liegt.

Wenn zwei Folgen konvergieren, so konvergieren auch die Abstände entsprechender Punkte:

2.5 Lemma. *Aus $\lim x_n = x$ und $\lim y_n = y$ folgt $\lim d(x_n, y_n) = d(x, y)$.*

Beweis. Nach der Viereckungleichung (Lemma 1.3) ist

$$|d(x_n, y_n) - d(x, y)| \leq d(x_n, x) + d(y_n, y),$$

woran man die Behauptung abliest. ∎

2.6 Definition. *Die Folge $(x_n)_{n\in\mathbb{N}}$ in M heißt* Cauchy-Folge*, wenn zu jedem $\varepsilon \in \mathbb{R}^+$ ein $n_0 \in \mathbb{N}$ existiert mit*

$$d(x_m, x_n) < \varepsilon \qquad \text{für alle } m, n \geq n_0.$$

2.7 Satz. *Jede konvergente Folge ist eine Cauchy-Folge.*

Beweis. Sei $\lim x_n = x$ und $\varepsilon \in \mathbb{R}^+$. Dann existiert ein $n_0 \in \mathbb{N}$ mit $d(x, x_n) < \varepsilon/2$ für $n \geq n_0$. Für $n, m \geq n_0$ gilt also

$$d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) < \varepsilon.$$ ∎

dass umgekehrt nicht jede Cauchy-Folge konvergiert, zeigt schon das Beispiel des Raumes der rationalen Zahlen. Andererseits gilt in $\mathbb{R}$ das Cauchysche Konvergenzkriterium, und hierauf (bzw. auf äquivalenten Aussagen) beruhten letzten Endes die meisten wesentlichen Sätze aus Analysis I. Die metrischen Räume, in denen jede Cauchy-Folge konvergiert, sind daher besonders wichtig.

2.8 Definition. *Ein metrischer Raum heißt* vollständig*, wenn in ihm jede Cauchy-Folge konvergiert. Eine Teilmenge A eines metrischen Raumes heißt vollständig, wenn sie als Unterraum (mit der induzierten Metrik) vollständig ist, wenn also jede Cauchy-Folge in A gegen einen Punkt von A konvergiert.*

2.9 Satz. *Jede vollständige Teilmenge eines metrischen Raumes ist abgeschlossen. In einem vollständigen metrischen Raum ist jede abgeschlossene Teilmenge vollständig.*

Beweis. Sei M ein metrischer Raum und $A \subseteq M$ eine vollständige Teilmenge. Sei x ein Berührpunkt von A. Nach Satz 2.4 gibt es eine Folge in A, die gegen x konvergiert. Nach Satz 2.7 ist diese Folge eine Cauchy-Folge, wegen der Vollständigkeit von A ist sie also konvergent gegen einen Punkt $y \in A$. Nach Satz 2.2 ist $x = y$, also $x \in A$. Somit ist A abgeschlossen.

Sei M ein vollständiger metrischer Raum und $A \subseteq M$ eine abgeschlossene Teilmenge. Sei $(x_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge in A. Da M vollständig ist, konvergiert sie gegen einen Punkt $x \in M$. Dieser ist Berührpunkt von A und daher Element von A. Somit ist A vollständig. ∎

2.10 Definition. *Ein normierter Vektorraum, der (mit der durch die Norm induzierten Metrik) vollständig ist, heißt* Banachraum*.*

Das folgende Beispiel wird im nächsten Abschnitt von Bedeutung sein. Wir betrachten eine Menge $\emptyset \neq D \subseteq \mathbb{R}$ und die Menge $B(D)$ aller beschränkten reellen Funktionen $f\colon D \to \mathbb{R}$. Mit den wie üblich (d.h. punktweise) für Funktionen erklärten Operationen der Addition und der Multiplikation mit Skalaren ist $B(D)$ ein reeller Vektorraum, und die Supremumsnorm $\|\cdot\|$ ist darauf eine Norm. Eine Folge $(f_n)_{n\in\mathbb{N}}$ in $B(D)$ ist nach der obigen Festsetzung eine Cauchy-Folge, wenn gilt:

$$\forall \varepsilon \in \mathbb{R}^+ \ \exists n_0 \in \mathbb{N} \ \forall n, m \geq n_0 : \|f_n - f_m\| < \varepsilon.$$

Diese Definition einer Cauchy-Folge von Funktionen auf D haben wir auch schon in Analysis I, Abschnitt 8.1, verwendet. Dort haben wir gezeigt: Die Folge $(f_n)_{n\in\mathbb{N}}$ konvergiert genau dann gleichmäßig, wenn sie eine Cauchy-Folge ist. Gleichmäßige Konvergenz ist aber dasselbe wie Konvergenz im Sinne der Supremumsnorm. Wir haben also:

2.11 Satz. *$B(D)$ ist ein Banachraum.*

Mit $C(D)$ bezeichnen wir nun die Teilmenge der stetigen Funktionen in $B(D)$. Natürlich ist $C(D)$ ein Untervektorraum. Wir behaupten, dass er abgeschlossen ist. Sei also f ein Berührpunkt von $C(D)$. Nach Satz 2.4 ist f Limes einer Folge $(f_n)_{n\in\mathbb{N}}$ in $C(D)$. Die Konvergenz bezieht sich hier auf die Supremumsnorm, die Folge $(f_n)_{n\in\mathbb{N}}$ konvergiert also gleichmäßig gegen f.

Da alle f_n stetig sind, ist f nach Satz 1.2 aus Kapitel 8, Analysis I ebenfalls stetig, also Element von $C(D)$. Somit ist $C(D)$ abgeschlossen. Aus Satz 2.11 und Satz 2.9 folgt, dass $C(D)$ vollständig ist. Wir haben also gezeigt:

2.12 Satz. *$C(D)$ ist ein Banachraum.*

9.3 Der Banachsche Fixpunktsatz

Eine wesentliche Aufgabe der Analysis, vor allem in den Anwendungen, besteht im „Lösen von Gleichungen“. Hierbei kann es sich um die Bestimmung von Nullstellen komplizierter Funktionen, Auflösung linearer Gleichungssysteme, Lösen von gewöhnlichen oder partiellen Differentialgleichungen, Berechnen von implizit definierten Funktionen, Integralgleichungen und vieles andere handeln. In manchen Fällen kann man hier das Verfahren der „sukzessiven Approximationen“ verwenden, das sowohl die Existenz einer Lösung zu beweisen gestattet als auch eine schrittweise Berechnung von Näherungslösungen mit vorgeschriebener Genauigkeit ermöglicht. Das zugrundeliegende Prinzip läßt sich allgemein formulieren und beweisen als ein Fixpunktsatz für gewisse Abbildungen eines vollständigen metrischen Raumes in sich. In diesem Satz, den wir jetzt beweisen wollen, haben wir ein besonders eindrucksvolles Beispiel für die Bedeutung und Auswirkung der Vollständigkeit von metrischen Räumen.

3.1 Definition. *Sei (M, d) ein metrischer Raum. Eine Abbildung $f : M \to M$ heißt* kontrahierend*, wenn es eine reelle Zahl $c < 1$ gibt mit*

$$d(f(x), f(y)) \leq cd(x, y) \qquad \text{für alle } x, y \in M.$$

Jede Zahl c mit dieser Eigenschaft heißt eine Lipschitzkonstante *der Abbildung f. Ein Punkt $x \in M$ heißt* Fixpunkt *von f, wenn $f(x) = x$ ist.*

3.2 Satz (Banachscher Fixpunktsatz)**.** *Sei (M, d) ein vollständiger metrischer Raum und $f : M \to M$ eine kontrahierende Abbildung. Dann hat f genau einen Fixpunkt.*

Ist $x_0 \in M$ beliebig und wird rekursiv

$$x_{n+1} := f(x_n) \qquad \text{für } n \in \mathbb{N}_0$$

definiert, so konvergiert die Folge $(x_n)_{n\in\mathbb{N}}$ gegen den Fixpunkt x, und es gilt die Fehlerabschätzung

$$d(x_n, x) \leq \frac{c}{1-c}\, d(x_{n-1}, x_n) \leq \frac{c^n}{1-c}\, d(x_0, x_1),$$

wenn $c < 1$ eine Lipschitzkonstante der Abbildung f ist.

Beweis. Sei M vollständig und $f : M \to M$ eine Abbildung mit einer Lipschitzkonstanten $c < 1$. dass f höchstens einen Fixpunkt hat, ist klar: Gilt $f(x) = x$ und $f(y) = y$, so ist

$$d(x, y) = d(f(x), f(y)) \leq cd(x, y),$$

wegen $c < 1$ also $d(x, y) = 0$ und daher $x = y$.

Wir wählen nun $x_0 \in M$ beliebig und definieren rekursiv $x_{n+1} := f(x_n)$ für $n \in \mathbb{N}_0$. Wir wollen zeigen, dass $(x_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge ist. Für $n \geq 1$ gilt

$$d(x_n, x_{n+1}) = d(f(x_{n-1}), f(x_n)) \leq cd(x_{n-1}, x_n).$$

Durch vollständige Induktion bekommt man daraus

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

und hieraus für $n \geq 1$ und $p \geq 1$ unter mehrfacher Verwendung der Dreiecksungleichung

$$\begin{aligned} d(x_n, x_{n+p}) &\leq d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+2}) + \cdots + d(x_{n+p-1}, x_{n+p}) \\ &\leq (c^n + c^{n+1} + \cdots + c^{n+p-1}) d(x_0, x_1) \\ &= c^n \frac{1 - c^p}{1 - c}\, d(x_0, x_1) \\ &\leq \frac{c^n}{1 - c}\, d(x_0, x_1). \end{aligned}$$

Zu gegebenem $\varepsilon \in \mathbb{R}^+$ gibt es also ein $n_0 \in \mathbb{N}$ mit $d(x_n, x_m) < \varepsilon$ für alle $n, m \geq n_0$; somit ist $(x_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge. Da M als vollständig vorausgesetzt ist, existiert ein Punkt $x \in M$ mit $\lim x_n = x$. Wegen $d(f(x), f(x_n)) \leq cd(x, x_n)$ gilt

$$f(x) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x.$$

Also ist x ein Fixpunkt von f.

Die Fehlerabschätzung ergibt sich aus

$$\begin{aligned} d(x_n, x_{n+p}) &\leq d(x_n, x_{n+1}) + \cdots + d(x_{n+p-1}, x_{n+p}) \\ &\leq (c + c^2 + \cdots + c^p)\, d(x_{n-1}, x_n) \end{aligned}$$

durch den Grenzübergang $p \to \infty$, wobei Lemma 2.5 zu beachten ist. Es folgt

$$d(x_n, x) \leq \frac{c}{1 - c}\, d(x_{n-1}, x_n) \leq \frac{c^n}{1 - c}\, d(x_0, x_1).$$

Damit ist alles bewiesen. ∎

Bei konkreter Anwendung wird man, um einen Fixpunkt der kontrahierenden Abbildung f näherungsweise zu berechnen, so vorgehen: Man wählt x_0 beliebig (aber möglichst nahe beim Fixpunkt, wenn man eine Vermutung über dessen ungefähre Lage hat), berechnet $x_1 = f(x_0)$ (wir unterstellen, dass dies möglich ist – meist wird es auch nur näherungsweise möglich sein), berechnet $x_2 = f(x_1)$, und so weiter. Wegen der schrittweisen Annäherung an den Fixpunkt spricht man von einem „Iterationsverfahren“ oder von „sukzessiven Approximationen“. Nach dem n-ten Schritt liefert die Ungleichung

$$d(x_n, x) \leq \frac{c}{1-c}\, d(x_{n-1}, x_n)$$

eine Information darüber, wie weit man höchstens vom Fixpunkt entfernt ist. Wenn man zu Beginn des Verfahrens schon abschätzen will, wieviele Schritte man höchstens braucht, um den Faktor unter eine vorgegebene Schranke zu drücken, so erhält man eine solche Information (nach Berechnung von x_1) aus der Ungleichung

$$d(x_n, x) \leq \frac{c^n}{1-c}\, d(x_0, x_1).$$

Bemerkung. In Satz 3.2 ist die Voraussetzung wesentlich, dass die Abbildung f eine Lipschitzkonstante $c < 1$ besitzt. Die schwächere Voraussetzung

$$d(f(x), f(y)) < d(x, y) \qquad \text{für alle } x, y \in M \text{ mit } x \neq y$$

reicht nicht aus, um die Existenz eines Fixpunktes zu zeigen. Zum Beispiel sei $M = \mathbb{R}$ (mit der üblichen Metrik) und $f : \mathbb{R} \to \mathbb{R}$ definiert durch

$$f(x) := x - \frac{\pi}{2} - \arctan x \qquad \text{für } x \in \mathbb{R}.$$

Es ist

$$f'(x) = 1 - \frac{1}{1+x^2},$$

also $0 \leq f'(x) < 1$ für alle $x \in \mathbb{R}$. Nach dem Mittelwertsatz der Differentialrechnung folgt daraus

$$|f(x) - f(y)| < |x - y| \qquad \text{für alle } x, y \in \mathbb{R} \text{ mit } x \neq y.$$

Da aber $f(x) < x$ für alle $x \in \mathbb{R}$ gilt, hat f keinen Fixpunkt.

Wir behandeln zwei Anwendungsbeispiele für den Banachschen Fixpunktsatz.

Beispiele. Zunächst betrachten wir eine reelle Funktion $f : [a, b] \to [a, b]$. Es sei eine Lösung der Gleichung $f(x) = x$ zu finden. Ist f stetig, so existiert

nach dem Zwischenwertsatz jedenfalls eine Lösung, denn es ist $f(x) - x \geq 0$ für $x = a$ und ≤ 0 für $x = b$. Für die Eindeutigkeit der Lösung und die näherungsweise Berechenbarkeit durch sukzessive Approximationen ist zum Beispiel hinreichend, dass f differenzierbar ist und eine Konstante $c < 1$ existiert mit

$$|f'(x)| \leq c \qquad \text{für alle } x \in [a, b].$$

Hieraus folgt nämlich nach dem Mittelwertsatz der Differentialrechnung

$$|f(x) - f(y)| \leq c|x - y| \qquad \text{für alle } x, y \in [a, b],$$

so dass Satz 3.2 anwendbar ist mit $M = [a, b]$.

Das nächste Beispiel ist etwas schwieriger, aber interessanter und besonders wichtig. Es handelt sich um das sogenannte *Anfangswertproblem für eine gewöhnliche Differentialgleichung erster Ordnung.* In der üblichen Schreibweise lautet es

$$y' = f(x, y), \quad y(x_0) = y_0.$$

Wir präzisieren und erläutern das folgendermaßen. Es seien $I, J \subseteq \mathbb{R}$ zwei offene Intervalle und $f : I \times J \to \mathbb{R}$ eine gegebene reelle Funktion. Ferner sei ein Punkt $(x_0, y_0) \in I \times J$ gegeben. Gesucht ist eine in einer Umgebung $U \subseteq I$ von x_0 definierte differenzierbare reelle Funktion y (mit $y(U) \subseteq J$), für die

$$\begin{aligned} y'(x) &= f(x, y(x)) \qquad \text{für alle } x \in U, \\ y(x_0) &= y_0 \end{aligned}$$

gilt (man veranschauliche sich die Aufgabenstellung durch ein „Richtungsfeld“).

Wir wollen zeigen, dass man die Existenz einer solchen Lösung aus dem Banachschen Fixpunktsatz erhält, wenn die Funktion f die folgende Voraussetzung erfüllt.

Voraussetzung. f sei stetig in $I \times J$ (die Erklärung erfolgt erst in Abschnitt 9.4) und beschränkt und genüge einer Lipschitzbedingung in der zweiten Veränderlichen, das heißt es gebe eine Zahl $L \in \mathbb{R}$ mit

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

für alle $x \in I$ und alle $y_1, y_2 \in J$.

Wenn wir eine (in einem Intervall um x_0 erklärte) stetige Funktion y haben mit

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\, dt, \tag{3.3}$$

so ist y nach dem Hauptsatz der Differential- und Integralrechnung (Analysis I, Kapitel 7, Satz 3.2) differenzierbar, und es gilt

$$\begin{aligned} y'(x) &= f(x, y(x)), \\ y(x_0) &= y_0, \end{aligned} \tag{3.4}$$

das heißt y löst unser Anfangswertproblem. Diese Tatsache werden wir im folgenden ausnutzen.

Wir müssen nun eine geeignete Situation herstellen, in der wir den Banachschen Fixpunktsatz anwenden können. Nach Voraussetzung existiert eine Konstante $a \in \mathbb{R}$ mit

$$|f(x, y)| < a \qquad \text{für } (x, y) \in I \times J.$$

Wir können $\delta > 0$ wählen, so dass $\delta L < 1$ und

$$[x_0 - \delta, x_0 + \delta] \times [y_0 - \delta a, y_0 + \delta a] \subseteq I \times J$$

ist. Sodann sei $C([x_0 - \delta, x_0 + \delta])$ der reelle Vektorraum aller stetigen reellen Funktionen auf dem Intervall $[x_0 - \delta, x_0 + \delta]$ mit der Supremumsnorm $\| \cdot \|$ und

$$\begin{aligned} M := \{ g \in C([x_0 - \delta, x_0 + \delta]) \mid \\ g(x_0) = y_0 \text{ und } |g(x) - y_0| \leq \delta a \text{ für } |x - x_0| \leq \delta \}. \end{aligned}$$

Der Raum $C([x_0 - \delta, x_0 + \delta])$ ist nach Satz 2.12 vollständig. Die Teilmenge M ist offenbar abgeschlossen und daher nach Satz 2.9 ebenfalls vollständig. Wir definieren nun eine Abbildung $F : M \to M$ in der folgenden Weise. Für $g \in M$ sei die Funktion $F(g)$ erklärt durch

$$F(g)(x) := y_0 + \int_{x_0}^{x} f(t, g(t))\, dt \qquad \text{für } x \in [x_0 - \delta, x_0 + \delta].$$

Die Funktion $F(g)$ ist stetig, sie erfüllt $F(g)(x_0) = y_0$, und für jedes $x \in [x_0 - \delta, x_0 + \delta]$ gilt

$$|F(g)(x) - y_0| = \left| \int_{x_0}^{x} f(t, g(t))\, dt \right| \leq \int_{x_0}^{x} |f(t, g(t))|\, dt \leq \delta a.$$

Also ist $F(g) \in M$, so dass in der Tat eine Abbildung $F : M \to M$ erklärt worden ist. Wir zeigen, dass F kontrahierend ist. Sei $g, h \in M$. Für $x \in [x_0 - \delta, x_0 + \delta]$ gilt

$$\begin{aligned}|F(g)(x) - F(h)(x)| &= \Big| \int_{x_0}^{x} [f(t, g(t)) - f(t, h(t))]\, dt \Big| \\ &\le \int_{x_0}^{x} |f(t, g(t)) - f(t, h(t))|\, dt \\ &\le L \int_{x_0}^{x} |g(t) - h(t)| dt \\ &\le L\delta \, \|g - h\|.\end{aligned}$$

Da dies für alle $x \in [x_0 - \delta, x_0 + \delta]$ gilt, ist

$$\|F(g) - F(h)\| \le L\delta \|g - h\|.$$

Nach Wahl von δ ist $L\delta < 1$, also F in der Tat kontrahierend. Jetzt folgt aus Satz 3.2, dass F einen Fixpunkt besitzt. Es gibt also ein Element $y \in M$ mit $F(y) = y$. Die Funktion y ist auf $[x_0 - \delta, x_0 + \delta]$ erklärt, dort stetig, und sie erfüllt

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\, dt \qquad \text{für } x \in [x_0 - \delta, x_0 + \delta].$$

Wie oben bemerkt, folgt hieraus

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

woraus insbesondere folgt, dass y stetig differenzierbar ist. Aus der eindeutigen Bestimmtheit des Fixpunktes folgt auch, dass y als Lösung des Anfangswertproblems auf $[x_0 - \delta, x_0 + \delta]$ eindeutig bestimmt ist. Wir fassen das gerade Bewiesene zusammen.

3.5 Satz (Picard, Lindelöf)**.** *Seien $I, J \subset \mathbb{R}$ zwei offene Intervalle und $f : I \times J \to \mathbb{R}$ eine gegebene stetige reelle Funktion. Diese sei beschränkt und genüge einer Lipschitzbedingung in der zweiten Veränderlichen, das heißt es gibt Zahlen $a, L \in \mathbb{R}^+$ mit*

$$\begin{aligned} |f(x, y_1)| &\le a\,, \\ |f(x, y_1) - f(x, y_2)\ &\le L\,|y_1 - y_2| \end{aligned}$$

für alle $x \in I$ und alle $y_1, y_2 \in J$. Dann existiert für alle $(x_0, y_0) \in I \times J$ ein $\delta > 0$ und genau eine stetig differenzierbare Funktion $y\colon [x_0 - \delta, x_0 + \delta] \to \mathbb{R}$, die das Anfangswertproblem (3.4) im Intervall $[x_0 - \delta, x_0 + \delta]$ löst.

9.4 Stetigkeit und Zusammenhang

Im folgenden seien (M, d), (M', d') metrische Räume. Wir wollen Abbildungen $f : M \to M'$ betrachten. Im Spezialfall $M' \subseteq \mathbb{R}$ sprechen wir auch von reellen Funktionen auf M. Die Definition der Stetigkeit einer Abbildung läßt sich wörtlich aus dem in Analysis I betrachteten Spezialfall übernehmen.

4.1 Definition. *Sei $f : M \to M'$ eine Abbildung, sei $x \in M$. Die Abbildung f heißt* stetig *in x, wenn zu jeder Umgebung V von $f(x)$ eine Umgebung U von x existiert mit $f(U) \subseteq V$. Die Abbildung f heißt* stetig*, wenn sie stetig in x ist für alle $x \in M$.*

Gemäß der Definition der Umgebungen können wir die Stetigkeit in äquivalenter Weise auch folgendermaßen erklären:

$$f \text{ stetig in } x \Leftrightarrow$$
$$\forall \varepsilon \in \mathbb{R}^+ \ \exists \delta \in \mathbb{R}^+ \ \forall y \in M : (d(x, y) < \delta \Rightarrow d'(f(x), f(y)) < \varepsilon).$$

4.2 Satz. *Die Abbildung $f : M \to M'$ ist genau dann stetig, wenn für jede offene Menge $A \subseteq M'$ das Urbild $f^{-1}(A)$ offen ist.*

Beweis. Sei f stetig. Sei $A \subseteq M'$ offen. Sei $x \in f^{-1}(A)$. Dann ist $f(x) \in A$. Da A offen ist, ist A Umgebung von $f(x)$. Da f in x stetig ist, existiert eine Umgebung $U \subseteq M$ von x mit $f(U) \subseteq A$, also mit $U \subseteq f^{-1}(A)$. Somit ist $f^{-1}(A)$ offen.

Für jede offene Menge $A \subseteq M'$ sei $f^{-1}(A)$ offen. Sei $x \in M$. Sei V eine Umgebung von $f(x)$; sie enthält eine offene Umgebung V_0 von $f(x)$. Dann ist $U := f^{-1}(V_0)$ offen, also Umgebung von x, und es gilt $f(U) \subseteq V_0 \subseteq V$. Also ist f stetig in x. ∎

Einige Aussagen über stetige Abbildungen zwischen metrischen Räumen lassen sich ganz analog wie im Spezialfall $M = M' = \mathbb{R}$ beweisen.

4.3 Satz. *Die Abbildung $f : M \to M'$ ist genau dann stetig in x, wenn für jede Folge $(x_n)_{n\in\mathbb{N}}$ in M aus $\lim_{n\to\infty} x_n = x$ stets $\lim_{n\to\infty} f(x_n) = f(x)$ folgt.*

Beweis. Sei $f : M \to M'$ stetig in x. Sei $(x_n)_{n\in\mathbb{N}}$ eine Folge in M mit $\lim_{n\to\infty} x_n = x$. Sei V eine Umgebung von $f(x)$. Wegen der Stetigkeit von f in x existiert eine Umgebung U von x mit $f(U) \subseteq V$. Wegen $\lim x_n = x$ existiert ein $n_0 \in \mathbb{N}$ mit $x_n \in U$ für $n \geq n_0$. Für alle $n \geq n_0$ gilt also $f(x_n) \in f(U) \subseteq V$. Da V eine beliebige Umgebung von $f(x)$ war, ist damit $\lim_{n\to\infty} f(x_n) = f(x)$ gezeigt.

Umgekehrt folge aus $\lim_{n\to\infty} x_n = x$ stets $\lim_{n\to\infty} f(x_n) = f(x)$. Angenommen, f wäre nicht stetig in x. Dann gibt es eine Umgebung V von $f(x)$ derart, dass für alle Umgebungen U von x gilt $f(U) \not\subseteq V$. Insbesondere können wir zu jedem $n \in \mathbb{N}$ einen Punkt $x_n \in U(x, 1/n)$ auswählen mit $f(x_n) \notin V$. Damit ist eine Folge $(x_n)_{n\in\mathbb{N}}$ in M definiert, die wegen $d(x, x_n) < 1/n$ gegen x konvergiert, für die aber die Folge $(f(x_n))_{n\in\mathbb{N}}$ wegen $f(x_n) \notin V$ nicht gegen $f(x)$ konvergiert. Das ist ein Widerspruch zur Voraussetzung; also ist f stetig in x. ∎

4.4 Satz. *Seien M, M', M'' metrische Räume. Sei $f : M \to M'$ stetig in x und $g : M' \to M''$ stetig in $f(x)$. Dann ist $g \circ f$ stetig in x.*

Beweis. Zu einer vorgegebenen Umgebung V von $(g \circ f)(x) = g(f(x))$ gibt es eine Umgebung W von $f(x)$ mit $g(W) \subseteq V$. Zu W gibt es eine Umgebung U von x mit $f(U) \subseteq W$. Es folgt $(g \circ f)(U) = g(f(U)) \subseteq g(W) \subseteq V$. ∎

Ebenfalls in Analogie zu Analysis I definieren wir noch:

4.5 Definition. *Die Abbildung $f : M \to M'$ heißt* gleichmäßig stetig, *wenn zu jedem $\varepsilon \in \mathbb{R}^+$ ein $\delta \in \mathbb{R}^+$ existiert mit*

$$d'(f(x), f(y)) < \varepsilon \quad \text{für alle } x, y \in M \text{ mit } d(x, y) < \delta.$$

Offensichtlich ist jede gleichmäßig stetige Abbildung auch stetig. Insbesondere (aber nicht nur in diesem Fall) ist eine Abbildung $f\colon M \to M'$ gleichmäßig stetig, wenn es eine Konstante c gibt mit

$$d'(f(x), f(y)) \le c d(x, y) \qquad \text{für alle } x, y \in M.$$

Abbildungen mit dieser Eigenschaft nennt man *Lipschitzabbildungen* (für $(M', d') = (M, d)$ und $c < 1$ kamen sie bereits im Banachschen Fixpunktsatz vor).

Wir betrachten jetzt speziell stetige Abbildungen $f : M \to \mathbb{R}$, also reellwertige stetige Funktionen, und wir wollen Satz 1.2 aus Analysis I, Kapitel 8 verallgemeinern. Wir können ganz allgemein (d.h. für beliebige Mengen M) für beschränkte Funktionen $f : M \to \mathbb{R}$ die Supremumsnorm erklären durch

$$\|f\| := \sup_{x\in M} |f(x)|.$$

Wir sagen dann, völlig analog zu dem früher betrachteten Spezialfall $M \subseteq \mathbb{R}$, dass die Folge $(f_n)_{n\in\mathbb{N}}$ von Funktionen auf M *gleichmäßig gegen f konvergiert*, wenn $\lim_{n\to\infty} \|f_n - f\| = 0$ ist. Offenbar impliziert gleichmäßige Konvergenz punktweise Konvergenz.

4.6 Satz. *Die Folge $(f_n)_{n\in\mathbb{N}}$ von stetigen reellen Funktionen auf M konvergiere gleichmäßig gegen die Funktion f. Dann ist f stetig.*

Beweis. Sei $x \in M$ und $\varepsilon \in \mathbb{R}^+$. Da $(f_n)_{n\in\mathbb{N}}$ gleichmäßig gegen f konvergiert, existiert ein $n_0 \in \mathbb{N}$ mit

$$|f_n(y) - f(y)| < \frac{\varepsilon}{3} \quad \text{für } n \geq n_0 \text{ und alle } y \in M.$$

Da f_{n_0} stetig ist, existiert eine Umgebung U von x mit

$$|f_{n_0}(x) - f_{n_0}(y)| < \frac{\varepsilon}{3} \qquad \text{für alle } y \in U.$$

Für $y \in U$ gilt also

$$|f(x) - f(y)| \leq |f(x) - f_{n_0}(x)| + |f_{n_0}(x) - f_{n_0}(y)| + |f_{n_0}(y) - f(y)| < \varepsilon.$$

Da $\varepsilon \in \mathbb{R}^+$ beliebig war, ist f stetig in x. Da $x \in M$ beliebig war, ist f stetig. ∎

Ein besonders wichtiger Satz aus Analysis I über stetige Funktionen ist der Zwischenwertsatz: Eine auf einem Intervall definierte stetige reelle Funktion nimmt jeden Wert zwischen zwei Funktionswerten an. Wenn man versuchen will, diesen Satz auf stetige Funktionen auf einem metrischen Raum zu verallgemeinern, steht man vor dem folgenden Problem. Wie triviale Gegenbeispiele zeigen, ist für die Gültigkeit des Zwischenwertsatzes die Voraussetzung unentbehrlich, dass der Definitionsbereich ein Intervall ist. Aber ein Intervall ist unter Verwendung der Anordnungsrelation für reelle Zahlen definiert worden; diese Definition läßt sich also nicht auf Teilmengen beliebiger metrischer Räume übertragen. Es zeigt sich aber, dass die für die Gültigkeit des Zwischenwertsatzes wesentliche Eigenschaft der Intervalle auch ohne Verwendung der Anordnungsrelation und damit in verallgemeinerungsfähiger Weise formuliert werden kann. Die hier gemeinte wesentliche Eigenschaft besteht darin, dass ein Intervall sozusagen „aus einem Stück besteht“. Man bezeichnet diese Eigenschaft als Zusammenhang und definiert sie allgemein folgendermaßen.

4.7 Definition. *Der metrische Raum M heißt* zusammenhängend*, wenn es keine offenen Teilmengen $A_1, A_2 \subseteq M$ gibt mit*

$$A_1 \neq \emptyset, \;\; A_2 \neq \emptyset, \;\; A_1 \cap A_2 = \emptyset, \;\; A_1 \cup A_2 = M.$$

Eine Teilmenge $A \subseteq M$ heißt zusammenhängend*, wenn sie als Teilraum (mit der induzierten Metrik) zusammenhängend ist.*

Wenn $M = A_1 \cup A_2$ und $A_1 \cap A_2 = \emptyset$ ist (man sagt dann, M sei in die Mengen A_1, A_2 *zerlegt*), so sind A_1 und A_2 nach Satz 1.11 genau dann beide offen, wenn sie abgeschlossen sind. Man kann also die Definition des Zusammenhangs auch äquivalent folgendermaßen fassen: Der metrische Raum M ist genau dann zusammenhängend, wenn $\emptyset$ und M die einzigen zugleich offenen und abgeschlossenen Teilmengen von M sind.

dass die Definition 4.7 das Gewünschte leistet, zeigt der folgende Satz.

4.8 Satz. *Eine stetige reelle Funktion auf einem zusammenhängenden metrischen Raum nimmt jeden Wert zwischen zwei Funktionswerten an.*

Beweis. Sei M zusammenhängend und $f : M \to \mathbb{R}$ stetig. Sei $c \in \mathbb{R}$ eine Zahl, die zwischen zwei von f angenommenen Funktionswerten liegt. Dann sind die Mengen $A_1 := f^{-1}((-\infty, c))$ und $A_2 := f^{-1}((c, \infty))$ nicht leer und nach Satz 4.2 offen, ferner ist $A_1 \cap A_2 = \emptyset$. Würde c nicht als Funktionswert angenommen, so wäre $A_1 \cup A_2 = M$, also wäre M nicht zusammenhängend, ein Widerspruch. ∎

Der Zwischenwertsatz aus Analysis I (Satz 3.4, Kapitel 4) ergibt sich allerdings nicht unmittelbar als Spezialfall von Satz 4.8, da wir noch nicht wissen, dass ein Intervall zusammenhängend ist. Dies kann man umgekehrt gerade mit dem Zwischenwertsatz beweisen.

4.9 Satz. *Jedes Intervall in $\mathbb{R}$ ist zusamenhängend.*

Beweis. Allgemeiner sei M ein metrischer Raum mit der Eigenschaft, dass jede stetige Funktion $f : M \to \mathbb{R}$ jeden Wert zwischen zwei Funktionswerten annimmt. Seien $A_1, A_2 \subseteq M$ nichtleere offene Teilmengen mit $A_1 \cap A_2 = \emptyset$. Angenommen, es wäre $A_1 \cup A_2 = M$. Setze

$$f(x) := \begin{cases} 1 & \text{für } x \in A_1, \\ 0 & \text{für } x \in A_2. \end{cases}$$

Die damit erklärte Funktion $f : M \to \mathbb{R}$ ist stetig, weil Urbilder offener Mengen offen sind. Sie nimmt, da A_1 und A_2 nicht leer sind, die Werte 0 und 1 an, aber keinen Wert, der echt zwischen 0 und 1 liegt. Aus dem Widerspruch folgt, dass M zusammenhängend ist. ∎

9.5 Kompaktheit

Unter den möglichen Eigenschaften metrischer Räume sind zwei von herausragender Bedeutung in der Analysis, da sie direkt oder indirekt in die Vor-

aussetzungen vieler wichtiger Sätze eingehen. Dies sind die Vollständigkeit und die jetzt zu behandelnde Kompaktheit.

In Analysis I (Abschnitt 4.1) hatten wir eine Teilmenge von $\mathbb{R}$ als *kompakt* bezeichnet, wenn sie beschränkt und abgeschlossen ist. Die Voraussetzung der Kompaktheit der Menge $K \subseteq \mathbb{R}$ spielte in den folgenden wichtigen Sätzen aus Analysis I eine wesentliche Rolle.

- Analysis I, Kapitel 4, Satz 1.6: Jede Folge in K besitzt eine Teilfolge, die gegen ein Element von K konvergiert.
- Analysis I, Kapitel 4, Satz 3.2: Sei $f : K \to \mathbb{R}$ stetig. Dann nimmt f ein Maximum an.
- Analysis I, Kapitel 4, Satz 3.5: Sei $f : K \to \mathbb{R}$ stetig. Dann ist f gleichmäßig stetig.

Wir möchten diese Sätze ausdehnen auf allgemeine metrische Räume. Wenn wir auch dort unter kompakten Mengen solche verstehen, die beschränkt und abgeschlossen sind, so sind aber die Sätze im allgemeinen falsch! Zum Beispiel ist der metrische Raum $(\mathbb{N}, d)$, wo d die diskrete Metrik ist, beschränkt und abgeschlossen. Die Folge $(n)_{n\in\mathbb{N}}$ besitzt aber keine konvergente Teilfolge, da keine Teilfolge eine Cauchy-Folge ist. Und die durch $f(n) = n$ erklärte Funktion $f : (\mathbb{N}, d) \to (\mathbb{R}, |\cdot|)$ ist stetig, da in $(\mathbb{N}, d)$ jede Teilmenge offen ist, sie nimmt aber kein Maximum an. Diese Beispiele zeigen, dass die Verallgemeinerung der obigen Sätze höchstens dann gelten kann, wenn im allgemeinen Fall die Kompaktheit einschneidender definiert wird. Wie dies zweckmäßig zu geschehen hat, wird durch den Überdeckungssatz von Heine-Borel nahegelegt. Nach diesem Satz (Analysis I, Kapitel 4, Satz 1.8) und seiner Umkehrung (Analysis I, Kapitel 4, Satz 1.9) ist eine Teilmenge $A \subseteq \mathbb{R}$ genau dann kompakt, wenn jede offene Überdeckung von A eine endliche Teilüberdeckung enthält. Diese Überdeckungseigenschaft verwenden wir nun im allgemeinen Fall als Definition der Kompaktheit.

5.1 Definition. *Sei (M, d) ein metrischer Raum, $A \subseteq M$ eine Teilmenge und $(U_i)_{i\in I}$ eine Familie von Teilmengen von M. Die Familie $(U_i)_{i\in I}$ heißt* Überdeckung *der Menge A, wenn $A \subseteq \bigcup_{i\in I} U_i$ gilt, und* offene Überdeckung *von A, wenn außerdem alle U_i offene Mengen sind.*

5.2 Definition. *Die Teilmenge $A \subseteq M$ heißt* kompakt (überdeckungskompakt)*, wenn jede offene Überdeckung von A eine endliche Teilüberdeckung von A enthält.*

5.3 Satz. *Jede kompakte Teilmenge eines metrischen Raumes ist abgeschlossen und beschränkt. In einem kompakten metrischen Raum ist jede abgeschlossene Teilmenge kompakt.*

Beweis. Sei (M,d) ein metrischer Raum. Sei $A \subseteq M$ kompakt. Sei x ein Berührpunkt von A. Angenommen, $x \notin A$. Dann ist $(U(y, \frac{1}{2}d(x,y)))_{y\in A}$ eine offene Überdeckung von A, enthält also eine endliche Teilüberdeckung $(U(y_i, \frac{1}{2}d(x,y_i)))_{i=1,\dots,n}$. Setze $\varepsilon := \min\{d(x,y_i) \mid i = 1,\dots,n\}$. Dann gilt

$$U\left(x, \tfrac{1}{2}\varepsilon\right) \cap U\left(y_i, \tfrac{1}{2}d(x,y_i)\right) = \emptyset \qquad \text{für } i = 1,\dots,n,$$

also $U(x, \frac{1}{2}\varepsilon) \cap A = \emptyset$. Somit ist x nicht Berührpunkt von A, ein Widerspruch. Damit ist die Abgeschlossenheit von A gezeigt.

Da die Überdeckung $(U(y,1))_{y\in A}$ von A eine endliche Überdeckung von A enthält, ist A nach Satz 1.5 beschränkt.

Sei jetzt M kompakt und $A \subseteq M$ abgeschlossen. Sei $(U_i)_{i\in I}$ eine offene Überdeckung von A. Dann ist $(U_i \cup A^c)_{i\in I}$ eine offene Überdeckung von M, enthält also eine endliche Teilüberdeckung $(U_i \cup A^c)_{i\in E}$ ($E \subseteq I$ endlich). $(U_i)_{i\in E}$ ist eine endliche Überdeckung von A. Damit ist die Kompaktheit von A gezeigt. ∎

Wir müssen nun zeigen, dass unser allgemeiner Kompaktheitsbegriff wirklich das Gewünschte leistet, also eine Ausdehnung der am Anfang dieses Abschnitts zitierten Sätze aus Analysis I ermöglicht.

5.4 Satz. *Ist $f : M \to M'$ eine stetige Abbildung eines kompakten metrischen Raumes M in einen metrischen Raum M', so ist das Bild $f(M)$ kompakt.*

Beweis. Sei $(U_i)_{i\in I}$ eine offene Überdeckung von $f(M)$. Nach Satz 4.2 ist $(f^{-1}(U_i))_{i\in I}$ eine offene Überdeckung von M, enthält also eine endliche Teilüberdeckung $(f^{-1}(U_i))_{i\in E}$ ($E \subseteq I$ endlich). Dann ist $(U_i)_{i\in E}$ eine endliche Überdeckung von $f(M)$. ∎

5.5 Folgerung. *Eine stetige Funktion $f : M \to \mathbb{R}$ auf einem kompakten metrischen Raum M nimmt ein Maximum an.*

Beweis. Nach Satz 5.4 ist $f(M)$ kompakt, also abgeschlossen und beschränkt. Jede abgeschlossene, beschränkte und nichtleere Menge reeller Zahlen enthält ein größtes Element. ∎

5.6 Satz. *Jede stetige Abbildung $f : M \to M'$ eines kompakten metrischen Raumes (M,d) in einen metrischen Raum (M',d') ist gleichmäßig stetig.*

Beweis. Sei M kompakt und $f : M \to M'$ stetig. Sei $\varepsilon \in \mathbb{R}^+$ gegeben. Zu jedem $z \in M$ existiert, da f in z stetig ist, ein $\delta(z) \in \mathbb{R}^+$ mit

$$d'(f(x), f(z)) < \frac{\varepsilon}{2} \quad \text{für alle } x \in M \text{ mit } d(x, z) < \delta(z).$$

Das System $(U(z, \frac{1}{2}\delta(z)))_{z\in M}$ ist eine offene Überdeckung von M, enthält also eine endliche Teilüberdeckung $(U(z_i, \frac{1}{2}\delta(z_i)))_{i=1,\dots,n}$. Setze $\delta := \min\{\frac{1}{2}\delta(z_i) \mid i = 1, \dots, n\}$. Seien jetzt $x, y \in M$ beliebige Punkte mit $d(x, y) < \delta$. Es gibt ein $i \in \{1, \dots, n\}$ mit $x \in U(z_i, \frac{1}{2}\delta(z_i))$. Dann ist $d(x, z_i) < \frac{1}{2}\delta(z_i) < \delta(z_i)$, ferner $d(y, z_i) \le d(y, x) + d(x, z_i) < \delta + \frac{1}{2}\delta(z_i) \le \delta(z_i)$. Es folgt

$$d'(f(x), f(y)) \le d'(f(x), f(z_i)) + d'(f(z_i), f(y)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Also ist f gleichmäßig stetig. ∎

5.7 Satz. *Sei $A \subseteq M$ kompakt. Dann besitzt jede Folge in A eine Teilfolge, die gegen einen Punkt von A konvergiert.*

Beweis. Sei A kompakt und $(x_i)_{i\in\mathbb{N}}$ eine Folge in A. Für $n \in \mathbb{N}$ setzen wir $A_n := \{x_i \mid i \ge n\}$. Wir zeigen zuerst, dass es eine Element $z \in A$ gibt mit $z \in \bigcap_{n\in\mathbb{N}} \overline{A}_n$ gibt. Anderenfalls gäbe es für alle $y \in A$ einen Index $m \in \mathbb{N}$ mit $y \notin \overline{A}_n$, d.h. $A \subseteq \bigcup_{n\in\mathbb{N}}(M \setminus \overline{A}_n)$. Also ist die Familie $(M \setminus \overline{A}_n)_{n\in\mathbb{N}}$ eine offene Überdeckung von A, enthält also eine endliche Teilüberdeckung von A. Es gibt daher ein $m \in \mathbb{N}$ mit $A \subseteq \bigcup_{n=1}^{m}(M \setminus \overline{A}_n) \subseteq \bigcup_{n=1}^{m}(M \setminus A_n)$. Wegen $x_m \in A$ und $x_m \in A_n$ für $n = 1, \dots, m$ ist das ein Widerspruch. Somit existiert ein Punkt $z \in A \cap \bigcap_{n\in\mathbb{N}} \overline{A}_n$. Wir definieren nun rekursiv eine Teilfolge von $(x_i)_{i\in\mathbb{N}}$. Da z Berührpunkt von $A_1 = \{x_i \mid i \ge 1\}$ ist, existiert ein $i_1 \in \mathbb{N}$ mit $x_{i_1} \in U(z, 1)$. Seien $i_1, \dots, i_{k-1}$ schon definiert. Da z Berührpunkt von $A_{i_{k-1}+1} = \{x_i \mid i \ge i_{k-1} + 1\}$ ist, existiert ein $i_k > i_{k-1}$ mit $x_{i_k} \in U(z, \frac{1}{k})$. Damit ist eine Teilfolge $(x_{i_k})_{k\in\mathbb{N}}$ von $(x_i)_{i\in\mathbb{N}}$ definiert, die gegen den Punkt $z \in A$ konvergiert. ∎

Mengen eines metrischen Raumes mit der Eigenschaft aus Satz 5.7 nennt man *folgenkompakt*, d.h. wir haben gezeigt, dass in metrischen Räumen Überdeckungskompaktheit Folgenkompaktheit impliziert.

5.8 Folgerung. *Jeder kompakte metrische Raum ist vollständig.*

Beweis. Jede Cauchy-Folge in einem kompakten metrischen Raum besitzt nach Satz 5.7 eine Teilfolge, die gegen einen Punkt z konvergiert. Aus der Cauchy-Eigenschaft der Folge ergibt sich mit der Dreiecksungleichung, dass sie selbst gegen z konvergieren muß. ∎

10 Der euklidische Raum

10.1 Der n-dimensionale euklidische Vektorraum

Unser Arbeitsgebiet in den folgenden Kapiteln wird der n-dimensionale euklidische Raum sein. Diese Begriffsbildung ist aus der Linearen Algebra bekannt, und überhaupt werden Methoden der Linearen Algebra in diesem zweiten Teil der Analysis viel stärker benutzt als im ersten Teil. Wir wollen zunächst an einige Definitionen aus der Vorlesung Lineare Algebra I erinnern und einige für unsere Zwecke praktische Bezeichnungen festlegen. Sodann werden wir sehen, dass der euklidische Raum gegenüber den in Kapitel 9 betrachteten allgemeinen metrischen Räumen noch einige sehr spezielle metrische Eigenschaften hat. Da er zugleich ein endlichdimensionaler Vektorraum ist, ergeben sich aus der Wechselwirkung zwischen Metrik und Vektorraumstruktur zusätzliche Aussagen, die später ständig benutzt werden.

Im folgenden ist n eine natürliche Zahl. Unter dem n-dimensionalen euklidischen Raum werden wir einen reellen Vektorraum der Dimension n mit einem (positiv definiten) Skalarprodukt verstehen. Bekanntlich sind je zwei n-dimensionale reelle Vektorräume isomorph; wir können daher von *dem* n-dimensionalen reellen Vektorraum sprechen und den weiteren Betrachtungen einen bestimmten zugrundelegen. Hierfür wählen wir das n-fache kartesische Produkt

$$\mathbb{R}^n := \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n\text{-mal}},$$

also die Menge aller geordneten n-Tupel reeller Zahlen, zusammen mit den durch

$$\begin{aligned}(x_1, \ldots, x_n) + (y_1, \ldots, y_n) &:= (x_1 + y_1, \ldots, x_n + y_n),\\ \lambda(x_1, \ldots, x_n) &:= (\lambda x_1, \ldots, \lambda x_n) \qquad \text{für } \lambda \in \mathbb{R}\end{aligned}$$

definierten Vektorraumoperationen. Den Nullvektor $(0, \ldots, 0)$ von $\mathbb{R}^n$ bezeichnen wir auch mit 0 (also – wie üblich – mit demselben Symbol wie die reelle Zahl 0). Durch

$$E_1 := (1, 0, \ldots, 0), \quad E_2 := (0, 1, 0, \ldots, 0), \ldots, \quad E_n := (0, \ldots, 0, 1)$$

ist eine Basis $(E_1, \ldots, E_n)$ von $\mathbb{R}^n$ gegeben; wir nennen sie die *Standardbasis* von $\mathbb{R}^n$. Für $X = (x_1, \ldots, x_n) \in \mathbb{R}^n$ ist

$$X = x_1 E_1 + \cdots + x_n E_n,$$

also x_i die i-te Koordinate des Vektors X bezüglich der Standardbasis; wir nennen x_i kurz die *i-te Koordinate* von X. Die Abbildung

$$p_i : \mathbb{R}^n \to \mathbb{R} : (x_1, \ldots, x_n) \mapsto x_i,$$

die also jedem Vektor seine i-te Koordinate zuordnet, heißt i-te *Projektion* $(i = 1, \ldots, n)$.

Bemerkung. In der Sprechweise machen wir keinen Unterschied zwischen „Punkten“ und „Vektoren“. Beides bedeutet hier also dasselbe, nämlich Elemente von $\mathbb{R}^n$. Wir verwenden beide Ausdrücke und sprechen von „Punkten“ meist dann, wenn die Vektorraumstruktur keine Rolle spielt.

Um auf dem Vektorraum $\mathbb{R}^n$ Analysis betreiben zu können, brauchen wir (zum Beispiel) eine Metrik. Sie sollte mit der Vektorraumstruktur gekoppelt sein. Wir führen sie daher ein durch eine spezielle Norm, die durch ein Skalarprodukt induziert ist. An diese Begriffsbildungen soll zunächst in allgemeinerem Rahmen erinnert werden.

1.1 Definition. *Ein* Skalarpodukt *auf dem reellen Vektorraum V ist eine Abbildung*

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R} \colon (u, v) \mapsto \langle u, v \rangle$$

mit folgenden Eigenschaften:

(a) $\langle \lambda u + \mu v, w \rangle = \lambda \langle u, w \rangle + \mu \langle v, w \rangle$, $\langle u, \lambda v + \mu w \rangle = \lambda \langle u, v \rangle + \mu \langle u, w \rangle$,
(b) $\langle u, v \rangle = \langle v, u \rangle$,
(c) $\langle u, u \rangle > 0$, *falls* $u \neq 0$

für alle $u, v, w \in V$ und alle $\lambda, \mu \in \mathbb{R}$.

Ist $\langle \cdot, \cdot \rangle$ ein Skalarprodukt auf V, so wird das Paar $(V, \langle \cdot, \cdot \rangle)$ als euklidischer Vektorraum *bezeichnet.*

Sind $(V, \langle \cdot, \cdot \rangle)$, $(V', \langle \cdot, \cdot \rangle')$ zwei n-dimensionale reelle Vektorräume mit Skalarprodukt, so gibt es, wie man in der Linearen Algebra zeigt, einen Vektorraumisomorphismus $f : V \to V'$ mit $\langle f(u), f(v) \rangle' = \langle u, v \rangle$ für alle $u, v \in V$. In diesem Sinne sind also je zwei euklidische Vektorräume gleicher (endlicher) Dimension isomorph. Alle Skalarprodukte auf einem n-dimensionalen reellen Vektorraum sind also im wesentlichen gleichwertig, und wir können insbesondere bei der Behandlung von $\mathbb{R}^n$ ein spezielles wählen.

1.2 Definition. *Für $X = (x_1, \dots, x_n) \in \mathbb{R}^n$, $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ sei*

$$\langle X, Y\rangle := x_1y_1 + \cdots + x_ny_n.$$

Dann ist $\langle \cdot, \cdot \rangle$ ein Skalarprodukt auf $\mathbb{R}^n$; es heißt das -Skalarprodukt.

dass in der Tat ein Skalarprodukt definiert wird, ist klar. Wir werden es den späteren Betrachtungen zugrundelegen. Unter dem *n-dimensionalen euklidischen Vektorraum* wird dann immer $\mathbb{R}^n$ mit dem Standard-Skalarprodukt verstanden. Zunächst schließen sich aber an die Definition des Skalarprodukts noch einige Begriffsbildungen und Hilfssätze an, die wir ohne Mehraufwand im allgemeineren Rahmen behandeln können.

1.3 Satz. *Ist $\langle \cdot, \cdot \rangle$ ein Skalarprodukt auf dem reellen Vektorraum V, so gilt für alle $u, v \in V$*

$$\langle u, v\rangle^2 \le \langle u, u\rangle\langle v, v\rangle$$

*(*Cauchy-Schwarzsche Ungleichung*) und*

$$\sqrt{\langle u + v, u + v\rangle} \le \sqrt{\langle u, u\rangle} + \sqrt{\langle v, v\rangle}$$

*(*Minkowskische Ungleichung*).*

Beweis. Für $\lambda \in \mathbb{R}$ und $u, v \in V$ gilt nach Definition 1.1

$$\langle u, u\rangle + 2\lambda\langle u, v\rangle + \lambda^2\langle v, v\rangle = \langle u + \lambda v, u + \lambda v\rangle \ge 0.$$

Ist $\langle v, v\rangle \neq 0$, so kann man

$$\lambda := -\frac{\langle u, v\rangle}{\langle v, v\rangle}$$

einsetzen und erhält die erste Ungleichung. Ist $\langle v, v\rangle = 0$, so ist $v = 0$ und daher $\langle u, v\rangle = 0$; die Cauchy-Schwarzsche Ungleichung gilt also trivialerweise. Die Minkowskische Ungleichung folgt aus der Cauchy-Schwarzschen:

$$\begin{aligned}\langle u + v, u + v\rangle &= \langle u, u\rangle + 2\langle u, v\rangle + \langle v, v\rangle \\ &\le \langle u, u\rangle + 2\sqrt{\langle u, u\rangle\langle v, v\rangle} + \langle v, v\rangle \\ &= \left(\sqrt{\langle u, u\rangle} + \sqrt{\langle v, v\rangle}\right)^2.\end{aligned}$$

∎

1.4 Lemma. *Sei $(V, \langle \cdot, \cdot \rangle)$ ein euklidischer Vektorraum. Dann wird durch*

$$\|u\| := \sqrt{\langle u, u\rangle} \qquad \text{für } u \in V$$

auf V eine Norm erklärt. Sie heißt die durch das Skalarprodukt induzierte *Norm.*

Beweis. (Zum Begriff der Norm siehe Beispiele in Abschnitt 9.1.) Nach Definition 1.1 ist $\langle u,u\rangle \geq 0$, also ist $\|u\| = \sqrt{\langle u,u\rangle}$ als reelle Zahl definiert. dass die Axiome für eine Norm erfüllt sind, folgt aus Definition 1.1 und der Minkowskischen Ungleichung, die gerade die Dreiecksungleichung für die induzierte Norm ist. ∎

Gemäß der Beispiele in Abschnitt 9.1 wird damit jetzt durch

$$d(u,v) := \|u-v\| \qquad \text{für } u,v \in V$$

auf V auch eine Metrik d gegeben. Im Falle des n-dimensionalen euklidischen Vektorraums ist

$$d(X,Y) = \sqrt{(x_1-y_1)^2+\cdots+(x_n-y_n)^2}$$

für $X=(x_1,\ldots,x_n)$, $Y=(y_1,\ldots,y_n)\in\mathbb{R}^n$. Wir bezeichnen diese Zahl als den *euklidischen Abstand* der Punkte X,Y und d auch als die *euklidische Metrik* auf $\mathbb{R}^n$.

Mit Hilfe des Skalarprodukts lassen sich weitere, anschaulich vertraute geometrische Grundbegriffe erklären:

1.5 Definition. *Sei $(V,\langle\cdot,\cdot\rangle)$ ein euklidischer Vektorraum. Die Vektoren $u,v\in V$ heißen* orthogonal *(oder* senkrecht*), wenn $\langle u,v\rangle = 0$ ist. Für $u,v\in V\setminus\{0\}$ wird die durch*

$$\frac{\langle u,v\rangle}{\|u\|\|v\|} = \cos\varphi, \qquad 0\leq\varphi\leq\pi,$$

definierte reelle Zahl φ als der Winkel *zwischen u und v bezeichnet.*

dass φ existiert, ist klar wegen

$$|\langle u,v\rangle| \leq \|u\|\|v\|,$$

was gerade die Cauchy-Schwarzsche Ungleichung in neuer Schreibweise ist. Die Eindeutigkeit von φ folgt aus der strengen Monotonie von cos auf $[0,\pi]$.

Bemerkung. Für orthogonale $u,v\in V$ gilt der „Satz des Pythagoras“:

$$\|u+v\|^2 = \|u\|^2+\|v\|^2.$$

Wegen $\langle u,v\rangle = 0$ folgt das aus

$$\|u+v\|^2 = \langle u+v,u+v\rangle = \langle u,u\rangle + 2\langle u,v\rangle + \langle v,v\rangle = \|u\|^2+\|v\|^2.$$

Eine Basis eines euklidischen Vektorraumes, deren Elemente paarweise orthogonal und von der Norm 1 sind, heißt *orthonormiert*. Die Standardbasis des $\mathbb{R}^n$ ist also orthonormiert (bezüglich des Standard-Skalarprodukts).

Von nun an legen wir speziell den n-dimensionalen euklidischen Raum zugrunde, also $\mathbb{R}^n$ mit dem Standard-Skalarprodukt $\langle \cdot, \cdot \rangle$, der daraus abgeleiteten Norm $\| \cdot \|$ und der hierdurch induzierten Metrik. Der euklidische Abstand zweier Punkte $X, Y \in \mathbb{R}^n$ ist durch $\|X - Y\|$ gegeben. Auf diesen metrischen Raum können wir nun alles anwenden, was wir in Kapitel 9 definiert und bewiesen haben. Durch die Koppelung der Metrik mit der Struktur eines endlichdimensionalen Vektorraumes ergeben sich einige Besonderheiten und Vereinfachungen, die wir jetzt zusammenstellen wollen.

1.6 Lemma. *Für $X = (x_1, \ldots, x_n) \in \mathbb{R}^n$ sei*

$$\|X\|_{\max} := \max\{|x_1|, \ldots, |x_n|\}.$$

Dann ist $\| \cdot \|_{\max}$ eine Norm auf $\mathbb{R}^n$, und es gilt

$$\|X\|_{\max} \leq \|X\| \leq \sqrt{n}\|X\|_{\max}.$$

Der Beweis ist trivial.

Diese einfachen Ungleichungen sind ein bequemes Hilfsmittel, um Aussagen über $\mathbb{R}^n$ auf dem Weg über die Koordinaten auf Aussagen über $\mathbb{R}$ zurückzuführen. Die folgenden Sätze werden dies demonstrieren.

1.7 Satz. *Die Menge $\mathbb{Q}^n$ der Punkte mit rationalen Koordinaten ist dicht in $\mathbb{R}^n$.*

Beweis. Sei $X = (x_1, \ldots, x_n) \in \mathbb{R}^n$ gegeben. Sei $\varepsilon \in \mathbb{R}^+$. Da $\mathbb{Q}$ dicht in $\mathbb{R}$ ist, gibt es zu jedem $i \in \{1, \ldots, n\}$ eine Zahl $y_i \in \mathbb{Q}$ mit

$$|x_i - y_i| < \frac{\varepsilon}{\sqrt{n}}.$$

Dann ist $Y := (y_1, \ldots, y_n) \in \mathbb{Q}^n$, und nach Lemma 1.6 ist

$$\|X - Y\| \leq \sqrt{n}\|X - Y\|_{\max} < \varepsilon.$$

In jeder Umgebung eines beliebigen Punktes von $\mathbb{R}^n$ liegen also Punkte aus $\mathbb{Q}^n$. ∎

Der folgende Satz führt die Konvergenz von Punktfolgen im $\mathbb{R}^n$ auf die Konvergenz der Koordinatenfolgen zurück; wir werden ihn häufig zu benutzen haben.

1.8 Satz. *Sei $(X_k)_{k\in\mathbb{N}}$ eine Folge in $\mathbb{R}^n$, sei $X_k = (x_1^{(k)}, \dots, x_n^{(k)})$ für $k \in \mathbb{N}$, sei $X = (x_1, \dots, x_n) \in \mathbb{R}^n$. Dann gilt*

$$\lim_{k\to\infty} X_k = X \quad \Leftrightarrow \quad \lim_{k\to\infty} x_i^{(k)} = x_i \quad \text{für } i = 1, \dots, n.$$

Ferner gilt

$$(X_k)_{k\in\mathbb{N}} \text{ ist Cauchy-Folge } \Leftrightarrow (x_i^{(k)})_{k\in\mathbb{N}} \text{ ist Cauchy-Folge } \text{ für } i = 1, \dots, n.$$

Beweis. Gelte $\lim_{k\to\infty} X_k = X$. Sei $i \in \{1, \dots, n\}$. Sei $\varepsilon \in \mathbb{R}^+$. Es gibt ein $k_0 \in \mathbb{N}$ mit $\|X - X_k\| < \varepsilon$ für $k \geq k_0$. Für alle $k \geq k_0$ gilt also

$$|x_i - x_i^{(k)}| \leq \|X - X_k\|_{\max} \leq \|X - X_k\| < \varepsilon.$$

Damit ist $\lim_{k\to\infty} x_i^{(k)} = x_i$ gezeigt.

Gelte $\lim_{k\to\infty} x_i^{(k)} = x_i$ für $i = 1, \dots, n$. Sei $\varepsilon \in \mathbb{R}^+$. Für $i \in \{1, \dots, n\}$ gibt es ein $k_i \in \mathbb{N}$ mit $|x_i - x_i^{(k)}| < \varepsilon/\sqrt{n}$ für $k \geq k_i$. Für $k \geq k_0 := \max\{k_1, \dots, k_n\}$ gilt dann

$$\|X - X_k\| \leq \sqrt{n}\|X - X_k\|_{\max} < \varepsilon.$$

Damit ist $\lim_{k\to\infty} X_k = X$ gezeigt.

Die zweite Äquivalenz zeigt man völlig analog. ∎

Als Anwendung kann man zeigen, dass die Vektorraumoperationen und das Skalarprodukt stetig sind. Wegen Satz 4.3 aus Kapitel 9 ist das äquivalent mit den folgenden Aussagen.

1.9 Satz. *Für konvergente Folgen $(X_k)_{k\in\mathbb{N}}$, $(Y_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^n$ und $(\lambda_k)_{k\in\mathbb{N}}$ in $\mathbb{R}$ gilt*

$$\begin{aligned}
\lim_{k\to\infty} (X_k + Y_k) &= \lim_{k\to\infty} X_k + \lim_{k\to\infty} Y_k, \\
\lim_{k\to\infty} \lambda_k X_k &= \left(\lim_{k\to\infty} \lambda_k\right)\left(\lim_{k\to\infty} X_k\right), \\
\lim_{k\to\infty} \langle X_k, Y_k\rangle &= \left\langle \lim_{k\to\infty} X_k, \lim_{k\to\infty} Y_k \right\rangle.
\end{aligned}$$

Der Beweis ergibt sich sofort aus Satz 1.8 und bekannten Aussagen über konvergente Folgen reeller Zahlen. Aus dem zweiten Teil von Satz 1.8 erhält man die folgende Aussage.

1.10 Satz. *$\mathbb{R}^n$ ist vollständig.*

Beweis. Sei $(X_k)_{k\in\mathbb{N}}$ mit $X_k = (x_1^{(k)}, \ldots, x_n^{(k)})$ eine Cauchy-Folge in $\mathbb{R}^n$. Für $i \in \{1, \ldots, n\}$ ist $(x_i^{(k)})_{k\in\mathbb{N}}$ nach Satz 1.8 eine Cauchy-Folge in $\mathbb{R}$, wegen der Vollständigkeit von $\mathbb{R}$ also konvergent gegen eine Zahl $x_i \in \mathbb{R}$. Mit $X := (x_1, \ldots, x_n)$ gilt $\lim_{k\to\infty} X_k = X$ nach Satz 1.8. ∎

Ebenso wichtig wie die durch Satz 1.10 ausgedrückte Tatsache, dass in $\mathbb{R}^n$ das Cauchysche Konvergenzkriterium gilt, ist die Tatsache, dass der Satz von Bolzano-Weierstraß sich auf den $\mathbb{R}^n$ übertragen läßt.

1.11 Satz (Bolzano-Weierstraß)**.** *Jede beschränkte Folge in $\mathbb{R}^n$ besitzt eine konvergente Teilfolge.*

Beweis. Sei $(X_k)_{k\in\mathbb{N}}$ mit $X_k = (x_1^{(k)}, \ldots, x_n^{(k)})$ eine beschränkte Folge in $\mathbb{R}^n$. Wegen Lemma 1.6 ist die Folge $(x_1^{(k)})_{k\in\mathbb{N}}$ beschränkt; sie besitzt also nach dem in $\mathbb{R}$ gültigen Satz von Bolzano-Weierstraß eine gegen eine Zahl $x_1 \in \mathbb{R}$ konvergierende Teilfolge $(x_1^{(k_j)})_{j\in\mathbb{N}}$. Die Folge $(x_2^{(k_j)})_{j\in\mathbb{N}}$ ist ebenfalls beschränkt, sie besitzt also eine gegen ein $x_2 \in \mathbb{R}$ konvergierende Teilfolge. So weiter schließend, erhält man nach endlich vielen Schritten eine streng monotone Folge $(m_j)_{j\in\mathbb{N}}$ in $\mathbb{N}$ und Zahlen $x_1, \ldots, x_n \in \mathbb{R}$ mit

$$\lim_{j\to\infty} x_i^{(m_j)} = x_i \qquad \text{für } i = 1, \ldots, n.$$

Mit $X := (x_1, \ldots, x_n)$ gilt also nach Satz 1.8

$$\lim_{j\to\infty} X_{m_j} = X.$$ ∎

Hieraus können wir zum Beispiel eine Verallgemeinerung des aus Analysis I bekannten Intervallschachtelungsprinzips (Kapitel 2, Satz 2.3) herleiten.

1.12 Satz. *Sei $(A_i)_{i\in\mathbb{N}}$ eine Folge abgeschlossener, beschränkter, nichtleerer Teilmengen von $\mathbb{R}^n$ mit $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$. Dann ist $\bigcap_{i\in\mathbb{N}} A_i \neq \emptyset$.*

Beweis. Da $A_i \neq \emptyset$ ist, können wir ein $X_i \in A_i$ auswählen ($i \in \mathbb{N}$). Die Folge $(X_i)_{i\in\mathbb{N}}$ ist (wegen $X_i \in A_1$ und der Voraussetzung) beschränkt, besitzt also nach Satz 1.11 eine konvergente Teilfolge $(X_{i_k})_{k\in\mathbb{N}}$. Sei X ihr Limes. Sei $i \in \mathbb{N}$. Wegen $X_{i_k} \in A_{i_k} \subseteq A_i$ für $i_k \geq i$ ist X Berührpunkt von A_i; also ist $X \in A_i$. ∎

Nun können wir auch zeigen, dass in $\mathbb{R}^n$ der Überdeckungssatz von Heine-Borel gilt:

1.13 Satz. *In $\mathbb{R}^n$ ist jede abgeschlossene, beschränkte Menge kompakt.*

Beweis. Der Beweis verläuft völlig analog zum Beweis von Satz 1.8, Kapitel 4 in Analysis I; dabei sind lediglich Intervalle $[a_i, b_i]$ zu ersetzen durch Würfel $[a_i^{(1)}, b_i^{(1)}] \times \cdots \times [a_i^{(n)}, b_i^{(n)}]$. ∎

Nach diesen wichtigen Beispielen für die Übertragbarkeit von Sätzen von $\mathbb{R}$ auf $\mathbb{R}^n$ kann man sich fragen, welche Rolle dabei die spezielle Wahl der euklidischen Norm $\|\cdot\|$ spielt. Die Antwort lautet, dass man ebenso gut eine beliebige andere Norm hätte wählen können. Dies liegt daran, dass zwei beliebige Normen auf dem Vektorraum $\mathbb{R}^n$ äquivalent sind in folgendem Sinne.

1.14 Satz. *Sind $\|\cdot\|_1$ und $\|\cdot\|_2$ zwei Normen auf $\mathbb{R}^n$, so gibt es Zahlen $k, K \in \mathbb{R}^+$ mit*

$$k\|X\|_1 \le \|X\|_2 \le K\|X\|_1 \qquad \text{für alle } X \in \mathbb{R}^n.$$

Beweis. Es genügt, die Ungleichungen für eine spezielle Norm $\|\cdot\|_1$ zu beweisen; das allgemeine Resultat folgt dann nämlich durch mehrmalige Anwendung des spezielleren. Wir wählen hierzu die Norm $\|\cdot\|_{\max}$ aus Lemma 1.6.

Sei $(E_1, \ldots, E_n)$ die Standardbasis des $\mathbb{R}^n$ und

$$K := \|E_1\|_2 + \cdots + \|E_n\|_2.$$

Für $X = (x_1, \ldots, x_n) \in \mathbb{R}^n$ gilt dann

$$\begin{aligned} \|X\|_2 &= \|x_1E_1 + \cdots + x_nE_n\|_2 \\ &\le |x_1|\|E_1\|_2 + \cdots + |x_n|\|E_n\|_2 \\ &\le K\|X\|_{\max}. \end{aligned}$$

Damit ist schon die Existenz von K gezeigt.

Wir setzen nun $f(X) := \|X\|_2$ für $X \in \mathbb{R}^n$ und zeigen die Stetigkeit der Funktion f. Für $X, Y \in \mathbb{R}^n$ gilt

$$\begin{aligned} |\|X\|_2 - \|Y\|_2| &\le \|X - Y\|_2 && \text{(Dreiecksungleichung)} \\ &\le K\|X - Y\|_{\max} \\ &\le K\|X - Y\| && \text{(nach Lemma 1.6).} \end{aligned}$$

Also ist f stetig. Nun ist die Menge

$$A := \{X \in \mathbb{R}^n : \|X\|_{\max} = 1\}$$

beschränkt (denn für $X \in A$ ist $d(X, 0) = \|X\| \le \sqrt{n}\|X\|_{\max} = \sqrt{n}$) und abgeschlossen (denn aus $X_i \in A$ und $X_i \to X$ folgt $1 = \|X_i\|_{\max} \to \|X\|_{\max}$

nach Lemma 2.5 aus Kapitel 9, also $\|X\|_{\max} = 1$). Nach Satz 1.13 ist A kompakt. Nach Folgerung 5.5 aus Kapitel 9 nimmt f auf A ein Minimum $k \geq 0$ an. Wäre $k = 0$, so gäbe es ein $X \in A$ mit $\|X\|_2 = 0$, also $X = 0$, ein Widerspruch. Also ist $k > 0$. Wir haben also $\|X\|_2 \geq k > 0$ für alle $X \in \mathbb{R}^n$ mit $\|X\|_{\max} = 1$. Für beliebiges $X \in \mathbb{R}^n \setminus \{0\}$ gilt dann mit $\lambda := 1/\|X\|_{\max}$ die Gleichung $\|\lambda X\|_{\max} = 1$, also $|\lambda|\|X\|_2 = \|\lambda X\|_2 \geq k$, folglich $\|X\|_2 \geq k\|X\|_{\max}$. Für $X = 0$ gilt diese Ungleichung trivialerweise. Damit ist auch die Existenz von k gezeigt. ∎

10.2 Abbildungen und Koordinatenfunktionen

Abbildungen zwischen euklidischen Räumen treten in der höherdimensionalen Analysis in verschiedenster Form auf. In diesem Abschnitt sollen einige grundlegende Erläuterungen zu allgemeinen Abbildungen $F : \mathbb{R}^n \to \mathbb{R}^k$ gegeben werden. Wir betrachten zunächst lineare Abbildungen und wiederholen einige einschlägige Begriffe aus der Linearen Algebra.

Eine Abbildung $F : \mathbb{R}^n \to \mathbb{R}^k$ heißt bekanntlich *linear*, wenn

$$F(\lambda X + \mu Y) = \lambda F(X) + \mu F(Y)$$

für alle $X, Y \in \mathbb{R}^n$ und alle $\lambda, \mu \in \mathbb{R}$ gilt. Lineare Abbildungen lassen sich durch Matrizen beschreiben. Wie vereinbart, wollen wir dabei immer die Standardbasen zugrundelegen. Sei also $(E_1, \ldots, E_n)$ die Standardbasis von $\mathbb{R}^n$ und $(E'_1, \ldots, E'_k)$ diejenige von $\mathbb{R}^k$. Für jedes $j \in \{1, \ldots, n\}$ ist der Vektor $F(E_j)$ eindeutig linear kombinierbar aus den Vektoren $E'_1, \ldots, E'_k$, also gilt

$$F(E_j) = \sum_{i=1}^{k} a_{ij} E'_i \qquad \text{für } j = 1, \ldots, n$$

mit reellen Zahlen a_{ij}. Durch die $k \times n$-Matrix

$$(a_{ij})_{\substack{i=1,\ldots,k \\ j=1,\ldots,n}} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & & & \vdots \\ a_{k1} & a_{k2} & \ldots & a_{kn} \end{pmatrix}$$

ist umgekehrt die Abbildung F festgelegt, denn für $X = \sum_{j=1}^{n} x_j E_j \in \mathbb{R}^n$ ist

$$F(X) = \sum_{j=1}^{n} x_j F(E_j) = \sum_{j=1}^{n} x_j \sum_{i=1}^{k} a_{ij} E'_i = \sum_{i=1}^{k} \left(\sum_{j=1}^{n} a_{ij} x_j \right) E'_i.$$

Die i-te Koordinate des Bildvektors ist also gegeben durch

$$\sum_{j=1}^{n} a_{ij} x_j .$$

Die linearen Abbildungen $F : \mathbb{R}^n \to \mathbb{R}^k$ bilden in bekannter Weise einen reellen Vektorraum. Es ist zweckmäßig, hierauf eine Norm einzuführen. Wir nehmen dazu die euklidische Norm des (beliebig geordneten) (kn)-Tupels der Matrixeinträge, durch die F bezüglich der Standardbasen beschrieben wird.

2.1 Lemma. *Sei $F : \mathbb{R}^n \to \mathbb{R}^k$ eine lineare Abbildung. Setze*

$$\|F\| := \sqrt{\sum_{i=1}^{k} \sum_{j=1}^{n} a_{ij}^2},$$

wobei $(a_{ij})_{\substack{i=1,\dots,k \\ j=1,\dots,n}}$ die der Abbildung F wie oben zugeordnete Matrix ist. Dann gilt

$$\|F(X)\| \le \|F\| \|X\| \qquad \textit{für alle } X \in \mathbb{R}^n .$$

Bemerkung. Man beachte, dass hier ungenauerweise dasselbe Symbol $\|\cdot\|$ für drei verschiedene Funktionen benutzt wird: $\|F(X)\|$ ist die Norm von $F(X)$ in $\mathbb{R}^k$, $\|X\|$ ist die Norm von X in $\mathbb{R}^n$, und $\|F\|$ ist eine Norm auf dem Vektorraum $\mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^k)$ der linearen Abbildungen von $\mathbb{R}^n$ in $\mathbb{R}^k$. Da aber keine Gefahr von Mißverständnissen besteht, ist diese unpräzise Bezeichnungsweise durchaus üblich.

Beweis (Lemma 2.1). Für $X = (x_1, \dots, x_n) \in \mathbb{R}^n$ gilt

$$F(X) = \sum_{i=1}^{k} \left(\sum_{j=1}^{n} a_{ij} x_j \right) E_i' ,$$

also

$$\|F(X)\|^2 = \sum_{i=1}^{k} \left(\sum_{j=1}^{n} a_{ij} x_j \right)^2 .$$

Nach der Cauchy-Schwarzschen Ungleichung (in einem beliebigen $\mathbb{R}^m$) $\langle A, B\rangle^2 \le \|A\|^2 \|B\|^2$ ist

$$\left(\sum a_j b_j \right)^2 \le \left(\sum a_j^2 \right) \left(\sum b_j^2 \right),$$

also

$$\|F(X)\|^2 \le \sum_{i=1}^{k} \left(\sum_{j=1}^{n} a_{ij}^2 \right) \left(\sum_{j=1}^{n} x_j^2 \right) = \|F\|^2 \|X\|^2. \qquad \blacksquare$$

2.2 Folgerung. *Ist $F : \mathbb{R}^n \to \mathbb{R}^k$ eine lineare Abbildung, so ist*

$$\|F(X) - F(Y)\| \le \|F\| \|X - Y\| \qquad \text{für } X, Y \in \mathbb{R}^n,$$

also ist F eine Lipschitzabbildung (mit Lipschitz-Konstante $\|F\|$) und somit insbesondere gleichmäßig stetig.

Nun betrachten wir Abbildungen in euklidische Räume, die nicht notwendig linear sind. Der Definitionsbereich darf dann zunächst auch von allgemeinerer Natur sein. Sei zunächst M eine beliebige nichtleere Menge und $F : M \to \mathbb{R}^n$ eine Abbildung. Für jedes $x \in M$ können wir dann den Bildvektor $F(x)$ als Linearkombination der Basisvektoren $E_1, \ldots, E_n$ des $\mathbb{R}^n$ darstellen. Bezeichnen wir die i-te Koordinate von $F(x)$ mit $f_i(x)$, so gilt also

$$F(x) = \sum_{i=1}^{n} f_i(x) E_i \qquad \text{für } x \in M.$$

Dadurch sind reellwertige Funktionen $f_i : M \to \mathbb{R}$, $i = 1, \ldots, n$, definiert. Wir bezeichnen f_i als die i-te *Koordinatenfunktion* der Abbildung F. Sie kann auch als Komposition

$$f_i = p_i \circ F$$

dargestellt werden; dabei ist

$$\begin{aligned} p_i : \quad \mathbb{R}^n \quad &\to \mathbb{R} \\ (x_1, \ldots, x_n) &\mapsto x_i \end{aligned}$$

die i-te Projektion. Allgemein läßt sich also die Untersuchung einer Abbildung in den $\mathbb{R}^n$ zurückführen auf die Untersuchung von n reellwertigen Funktionen, was oft bequem ist. Hierfür ein erstes Beispiel:

2.3 Satz. *Sei M ein metrischer Raum, $F : M \to \mathbb{R}^n$ eine Abbildung und $x \in M$. Dann gilt:*

$$F \text{ ist stetig in } x \;\Leftrightarrow\; p_i \circ F \text{ ist stetig in } x \text{ für } i = 1, \ldots, n.$$

Beweis. Ist F stetig in x, so ist wegen der Stetigkeit der Projektion p_i auch $p_i \circ F$ stetig in x $(i = 1, \ldots, n)$. Seien umgekehrt alle Koordinatenfunktionen $p_i \circ F =: f_i$ stetig in x. Sei $\varepsilon \in \mathbb{R}^+$. Für $i \in \{1, \ldots, n\}$ gibt es eine Umgebung U_i von x in M mit

$$|f_i(x) - f_i(y)| < \frac{\varepsilon}{\sqrt{n}} \qquad \text{für alle } y \in U_i.$$

Dann ist $U := U_1 \cap \cdots \cap U_n$ eine Umgebung von x, und für alle $y \in U$ gilt $|f_i(x) - f_i(y)| < \varepsilon/\sqrt{n}$ für $i = 1, \ldots, n$, also

$$\|F(y) - F(x)\| \le \sqrt{n}\|F(y) - F(x)\|_{\max} < \varepsilon. \qquad \blacksquare$$

Bemerkung. Analoge Aussagen gelten offenbar, wenn „stetig in x“ ersetzt wird durch „stetig“ oder „gleichmäßig stetig“.

Die Stetigkeit einer Abbildung in einem Punkt kann man natürlich ganz analog wie in Analysis I auch unter Verwendung eines Grenzwertbegriffs für Funktionen formulieren. Der folgende Konvergenzbegriff für Abbildungen zwischen euklidischen Räumen wird später häufig benutzt.

2.4 Definition. *Sei $M \subseteq \mathbb{R}^n$, $F : M \to \mathbb{R}^k$ eine Abbildung, X_0 ein Häufungspunkt von M und $Y \in \mathbb{R}^k$. Dann wird definiert*

$$\lim_{X \to X_0} F(X) = Y \quad :\Leftrightarrow$$
$$\forall\, \varepsilon \in \mathbb{R}^+ \;\; \exists \delta \in \mathbb{R}^+ \;\; \forall\, X \in M \setminus \{X_0\} : (\|X - X_0\| < \delta \Rightarrow \|F(x) - Y\| < \varepsilon).$$

Man beachte, dass die Aussage

$$\lim_{X \to X_0} F(X) = Y$$

nicht erfordert, dass F auch an der Stelle X_0 definiert ist.

Die Stetigkeitsdefinition kann jetzt also auch folgendermaßen umformuliert werden:

$$F \text{ stetig in } X_0 \;\Leftrightarrow\; \lim_{X \to X_0} F(X) = F(X_0).$$

Ebenso wie Konvergenz einer Folge in $\mathbb{R}^n$ gleichbedeutend ist mit Konvergenz der entsprechenden Koordinatenfolgen, kann man Limesbeziehungen für Abbildungen zurückführen auf Limesbeziehungen für die Koordinatenfunktionen:

2.5 Satz. *Sei $M \subseteq \mathbb{R}^n$, $F : M \to \mathbb{R}^k$ eine Abbildung, X_0 ein Häufungspunkt von M, $Y = (y_1, \ldots, y_k) \in \mathbb{R}^k$ und $f_i := p_i \circ F$, also $F(X) = (f_1(X), \ldots, f_k(X))$ für $X \in M$. Dann gilt:*

$$\lim_{X \to X_0} F(X) = Y \;\Leftrightarrow\; \lim_{X \to X_0} f_i(X) = y_i \quad \textit{für } i = 1, \ldots, k.$$

Beweis. In Analogie zum Beweis von Satz 1.8 (leichte Übung). $\blacksquare$

Die formalen Eigenschaften des Konvergenzbegriffes aus Definition 2.4 sind sinngemäß dieselben wie früher im eindimensionalen Fall. Wir nennen nur ein später häufig benutztes Beispiel: Aus

$$\lim_{X \to X_0} F(X) = Y \quad \text{und} \quad \lim_{X \to X_0} G(X) = Z$$

folgt

$$\lim_{X \to X_0} (F + G)(X) = Y + Z.$$

11 Differentiation

Wir beginnen mit einigen Vorbetrachtungen. In diesem Kapitel soll die Differentialrechnung für Funktionen von n reellen Veränderlichen behandelt werden. Unter einer reellen Funktion von n reellen Veränderlichen verstehen wir eine Abbildung $f : M \to \mathbb{R}$ mit $M \subseteq \mathbb{R}^n$. Für $X \in \mathbb{R}^n$, $X = (x_1, \ldots, x_n)$ schreibt man üblicherweise

$$f(X) = f(x_1, \ldots, x_n).$$

Man muß sich dabei stets darüber klar sein, dass diese Schreibweise die Festlegung einer Basis des $\mathbb{R}^n$ voraussetzt, auch wenn dies nicht immer explizit gesagt wird. In der Schreibweise $f(X)$ ist dagegen nicht auf eine spezielle Basis Bezug genommen; daher nennt man diese Schreibweise auch „koordinatenunabhängig“ oder „invariant“. Wenn wir in $\mathbb{R}^n$ eine andere Basis $(\tilde{E}_1, \ldots, \tilde{E}_n)$ einführen, so können wir natürlich auch

$$f(X) = f\left(\sum_{i=1}^{n} \tilde{x}_i \tilde{E}_i\right) =: g(\tilde{x}_1, \ldots, \tilde{x}_n)$$

schreiben, aber die Zuordnung $(\tilde{x}_1, \ldots, \tilde{x}_n) \mapsto g(\tilde{x}_1, \ldots, \tilde{x}_n)$ ist natürlich eine andere als die Zuordnung $(x_1, \ldots, x_n) \mapsto f(x_1, \ldots, x_n)$. Aus diesen und anderen Gründen ist es zweckmäßig, so weit wie möglich koordinatenunabhängige Definitionen, Schreibweisen und Schlußweisen zu benutzen. Bei der analytischen Behandlung von Funktionen von mehreren Veränderlichen empfiehlt es sich also, in $f(X)$ immer das Argument X, also einen Punkt des Raumes, als das Wesentliche anzusehen und seine Beschreibung durch Koordinaten nur als Hilfsmittel zu betrachten.

Diese Betrachtungsweise führt uns auch dazu, an den Anfang der Differentialrechnung in höheren Dimensionen nicht die partiellen Ableitungen zu stellen. Was hierunter zu verstehen ist, ist schnell gesagt. Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine Funktion und $\overline{X} = (\overline{x}_1, \ldots, \overline{x}_n) \in \mathbb{R}^n$. Als i-te partielle Ableitung von f an der Stelle $\overline{X}$ wird die Ableitung der Funktion

$$t \mapsto f(\overline{x}_1, \ldots, \overline{x}_{i-1}, t, \overline{x}_{i+1}, \ldots, \overline{x}_n)$$

an der Stelle $t = \overline{x}_i$ bezeichnet, falls sie existiert. Die i-te partielle Ableitung oder partielle Ableitung nach der i-ten Veränderlichen ist also die gewöhnliche Ableitung der Funktion, die sich ergibt, wenn alle Veränderlichen außer der i-ten festgehalten werden. Partielle Ableitungen beziehen sich also immer auf eine Basis. Abgesehen von diesem Schönheitsfehler ist es auch aus anderen Gründen nicht zweckmäßig, den Begriff der Differenzierbarkeit auf die Existenz der partiellen Ableitungen zu gründen. Dies soll jetzt durch Beispiele erläutert werden.

Wir betrachten Funktionen auf $\mathbb{R}^2$. In diesem Fall werden wir die Koordinaten (bezüglich der Standardbasis) eines Vektors X meist mit x, y statt x_1, x_2 bezeichnen. Sei nun $f : \mathbb{R}^2 \to \mathbb{R}$ definiert durch

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

Die partiellen Ableitungen an der Stelle $(0,0)$, bezeichnet mit

$$\frac{\partial f}{\partial x}(0,0) \qquad \text{und} \qquad \frac{\partial f}{\partial y}(0,0),$$

existieren, denn die Funktionen $x \mapsto f(x,0) = 0$ und $y \mapsto f(0,y) = 0$ sind differenzierbar. Die Funktion f ist also an der Stelle (0,0) (und auch an jeder anderen Stelle) partiell differenzierbar. Sie ist aber an der Stelle (0,0) nicht stetig! In der Tat liegen in jeder Umgebung von (0,0) Punkte (a,a) mit $a \neq 0$, und es ist $f(a,a) = 1/2$, während doch $f(0,0) = 0$ ist. Nun sind wir aus Analysis I gewöhnt, dass jede differenzierbare Funktion auch stetig ist. Hier haben wir aber eine partiell differenzierbare Funktion von zwei Veränderlichen, die nicht stetig ist. Dies weist darauf hin, dass man unter der Differenzierbarkeit einer Funktion von mehreren Veränderlichen etwas anderes verstehen sollte als partielle Differenzierbarkeit.

Nun bezieht sich partielle Differenzierbarkeit auf eine spezielle Basis. Man könnte daher von einer Funktion schärfer fordern, dass sie bezüglich jeder Basis partiell differenzierbar ist. Aber auch daraus würde nicht die Stetigkeit folgen. Dies wird belegt durch das Beispiel

$$f(x,y) := \begin{cases} \frac{x^2 y}{x^4+y^2} & \text{für } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y) = (0,0). \end{cases}$$

In jeder Umgebung von $(0,0)$ liegen Punkte (a,a^2) mit $a \neq 0$, und hierfür ist $f(a,a^2) = \frac{1}{2}$. Also ist f nicht stetig in $(0,0)$. Andererseits sind die Funktionen

$$y \mapsto f(0,y) = 0$$

und

$$t \mapsto f(t, \alpha t) = \frac{\alpha t}{t^2 + \alpha^2}, \qquad \alpha \in \mathbb{R}$$

überall differenzierbar. Mit anderen Worten: Für jede Gerade G in $\mathbb{R}^2$ durch den Nullpunkt ist die Einschränkung von f auf G differenzierbar (genauer: für jeden Vektor $E \in \mathbb{R}^2$ ist die Funktion $t \mapsto f(tE)$ differenzierbar). Später werden wir hierfür sagen, dass alle Richtungsableitungen dieser Funktion existieren, insbesondere alle partiellen Ableitungen bei beliebiger Basis. Trotzdem ist die Funktion nicht stetig.

11.1 Differenzierbarkeit

Die Vorbetrachtungen machen klar, dass wir uns gut überlegen müssen, was eigentlich eine sinnvolle Definition der Differenzierbarkeit einer Funktion von mehreren Veränderlichen ist. Wir können uns hier durch den eindimensionalen Fall leiten lassen, müssen ihn aber inhaltlich richtig verstehen und dazu etwas uminterpretieren.

Der Grundgedanke der Differentialrechnung besteht darin, eine gegebene Abbildung von einem euklidischen Raum in einen anderen dadurch zu untersuchen, dass man die Abbildung in der Nähe eines betrachteten Punktes möglichst gut approximiert durch „möglichst einfache“ Abbildungen. Besonders einfach sind konstante Abbildungen und, nach diesem Trivialfall, lineare Abbildungen. Zusammensetzungen aus beiden nennt man affine Abbildungen. Eine Abbildung $F : \mathbb{R}^n \to \mathbb{R}^k$ heißt also *affin*, wenn

$$F(X) = L(X) + Z \qquad \text{für } X \in \mathbb{R}^n$$

ist, wobei $L : \mathbb{R}^n \to \mathbb{R}^k$ eine lineare Abbildung und $Z \in \mathbb{R}^k$ ein fester Vektor ist.

Betrachten wir nun zunächst den Fall einer Funktion $f : \mathbb{R} \to \mathbb{R}$. Eine affine Abbildung $g : \mathbb{R} \to \mathbb{R}$ ist von der Form

$$g(x) = cx + t \qquad \text{für } x \in \mathbb{R}$$

mit Konstanten c und t. Wir wollen nun zu gegebenem $x_0 \in \mathbb{R}$ eine affine Funktion g finden, die f bei x_0 „möglichst gut“ approximiert. Die erste Forderung an g ist natürlich, dass $g(x_0) = f(x_0)$ sein soll; dies wird erfüllt durch

$$g(x) = c(x - x_0) + f(x_0).$$

Schreiben wir wie üblich $x = x_0 + h$, so verschwindet also die Funktion

$$h \mapsto f(x_0+h) - g(x_0+h) = f(x_0+h) - f(x_0) - ch$$

für $h = 0$. „Gute Approximation“ interpretieren wir nun so, dass die Differenz $f(x_0+h) - f(x_0) - ch$ für $h \to 0$ schneller klein werden soll als h. Damit ist gemeint, dass sogar noch

$$\lim_{h\to 0} \frac{f(x_0+h) - f(x_0) - ch}{h} = 0 \tag{1.1}$$

sein soll. Hiermit sind wir genau bei der Definition der Differenzierbarkeit aus Analysis I angelangt: (1.1) ist äquivalent damit, dass f bei x_0 differenzierbar und dass $f'(x_0) = c$ ist. Die obigen Überlegungen legen es aber jetzt nahe, nicht die Zahl c, sondern die durch sie definierte lineare Abbildung $L : h \mapsto ch$ in den Mittelpunkt zu stellen, also zu sagen: Die Funktion f heißt differenzierbar in x_0, wenn es eine lineare Funktion $L : \mathbb{R} \to \mathbb{R}$ gibt mit

$$\lim_{h\to 0} \frac{f(x_0+h) - f(x_0) - L(h)}{h} = 0.$$

In dieser Form läßt sich die Definition sofort auf Abbildungen zwischen höherdimensionalen Räumen übertragen.

1.2 Definition. *Sei $M \subseteq \mathbb{R}^n$ offen, sei $F : M \to \mathbb{R}^k$ eine Abbildung, sei $X_0 \in M$. Die Abbildung F heißt* differenzierbar in X_0, *wenn es eine lineare Abbildung $L : \mathbb{R}^n \to \mathbb{R}^k$ gibt mit*

$$\lim_{H\to 0} \frac{F(X_0+H) - F(X_0) - L(H)}{\|H\|} = 0.$$

Gibt es eine solche lineare Abbildung, so ist sie eindeutig bestimmt; sie heißt dann das Differential von F in X_0 *und wird mit DF_{X_0} bezeichnet. F heißt* differenzierbar, *wenn F differenzierbar in X ist für alle $X \in M$.*

Wohldefiniertheit von Definition 1.2. Zu zeigen ist, dass es höchstens eine lineare Abbildung mit der genannten Eigenschaft gibt. Seien also $L_1, L_2 : \mathbb{R}^n \to \mathbb{R}^k$ zwei lineare Abbildungen dieser Art. Dann gilt

$$\lim_{H\to 0} \frac{L_1(H) - L_2(H)}{\|H\|} = 0.$$

Für jeden Vektor $X \in \mathbb{R}^n \setminus \{0\}$ folgt insbesondere

$$0 = \lim_{t\searrow 0} \frac{L_1(tX) - L_2(tX)}{\|tX\|} = \frac{L_1(X) - L_2(X)}{\|X\|},$$

also $L_1(X) = L_2(X)$; somit ist $L_1 = L_2$. ∎

Bemerkung. In der älteren Literatur wird Differenzierbarkeit im Sinne von Definition 1.2 oft als „vollständige“ oder „totale“ Differenzierbarkeit bezeichnet.

In Definition 1.2 dürfen n und k beliebige natürliche Zahlen sein. Ist eine dieser Zahlen gleich 1, so kann man sich die Bedeutung des Differentials gut veranschaulichen.

Sei zunächst $k = 1$ und, der Einfachheit halber, $n = 2$. In diesem Fall empfiehlt es sich, von der Funktion $f : M \to \mathbb{R}$ (mit $M \subseteq \mathbb{R}^2$) den Graphen

$$\text{Graph } f := \{(x, y, z) \in \mathbb{R}^3 \mid (x, y) \in M,\ z = f(x, y)\}$$

zu betrachten. Sei $X_0 \in M$ und Df_{X_0} das Differential von f in X_0. Dann ist also $Df_{X_0} : \mathbb{R}^2 \to \mathbb{R}$ eine lineare Funktion. Differenzierbarkeit von f in X_0 kann auch so formuliert werden, dass

$$f(X) = f(X_0) + Df_{X_0}(X - X_0) + R(X_0, X)$$

mit

$$\lim_{X \to X_0} \frac{R(X_0, X)}{\|X - X_0\|} = 0 \tag{1.3}$$

gilt. Die durch

$$g(X) := f(X_0) + Df_{X_0}(X - X_0), \quad X \in \mathbb{R}^2,$$

definierte Funktion g ist affin, und ihr Graph

$$\text{Graph } g = \{(x, y, z) \in \mathbb{R}^3 \mid (x, y) \in \mathbb{R}^2,\ z = g(x, y)\}$$

ist eine Ebene durch den Punkt $(X_0, f(X_0))$. Die Beziehung (1.3), also

$$\lim_{X \to X_0} \frac{f(X) - g(X)}{\|X - X_0\|} = 0$$

besagt, dass der Graph von f bei Annäherung von X an X_0 sich dem Graphen von g besonders gut anschmiegt. Der Graph von g wird daher auch als *Tangentialebene* des Graphen von f im Punkt $(X_0, f(X_0))$ bezeichnet. Differenzierbarkeit einer reellen Funktion kann also auch als Existenz einer Tangentialebene an den Graphen interpretiert werden. Das Differential beschreibt die Stellung dieser Tangentialebene.

Betrachten wir jetzt den Fall $n = 1$ und $k \geq 2$. Die Menge $M \subseteq \mathbb{R}$ wollen wir etwa als Intervall annehmen. Die Abbildung $F : M \to \mathbb{R}^k$ bezeichnet man als eine *parametrisierte Kurve*. (Eine physikalische Interpretation im Fall $k = 3$ wäre etwa, dass F die Bahn eines Massepunktes in Abhängigkeit von der Zeit beschreibt.) Sei $t_0 \in M$ und F differenzierbar in t_0. Das Differential $DF_{t_0} : \mathbb{R} \to \mathbb{R}^k$ ist eine lineare Abbildung von $\mathbb{R}$ in $\mathbb{R}^k$, also gilt $DF_{t_0}(h) = hDF_{t_0}(1)$. Wir setzen

$$DF_{t_0}(1) =: F'(t_0);$$

das ist also ein Vektor in $\mathbb{R}^k$. Die Limesgleichung in der Definition 1.2 der Differenzierbarkeit lautet dann

$$\lim_{h\to 0} \frac{F(t_0+h) - F(t_0) - F'(t_0)h}{|h|} = 0,$$

was äquivalent ist mit

$$\lim_{h\to 0} \frac{F(t_0+h) - F(t_0)}{h} = F'(t_0).$$

Damit ist eine anschauliche Deutung des Vektors $F'(t_0)$ gefunden. Wir bezeichnen ihn als *Ableitung* der Abbildung F oder als *Tangentenvektor* der parametrisierten Kurve F in t_0. Ist f_i die i-te Koordinatenfunktion von F, also

$$F(t) = (f_1(t), \ldots, f_k(t)) \qquad \text{für } t \in M,$$

so ist wegen Satz 2.5 aus Kapitel 10

$$F'(t_0) = (f_1'(t_0), \ldots, f_k'(t_0)).$$

Man erhält den Ableitungs- oder Tangentenvektor also durch koordinatenweises Differenzieren.

Wir kehren zum allgemeinen Fall zurück und wollen zunächst einige grundlegende Aussagen über differenzierbare Abbildungen zusammenstellen. Danach untersuchen wir insbesondere reellwertige Funktionen, die wir anschliessend in Gestalt von Koordinatenfunktionen wieder für die Behandlung allgemeiner differenzierbarer Abbildungen nutzbar machen.

Die folgenden Sätze zeigen, dass der durch Definition 1.2 eingeführte Differenzierbarkeitsbegriff, anders als die Forderung der partiellen Differenzierbarkeit, die aus Analysis I bekannten Implikationen von Differenzierbarkeit auszudehnen gestattet.

1.4 Satz. *Ist $F : M \to \mathbb{R}^k$ in X_0 differenzierbar, so ist F in X_0 stetig.*

Beweis. Sei F in X_0 differenzierbar. Zu gegebenem $\varepsilon \in \mathbb{R}^+$ existiert dann ein $\delta_1 \in \mathbb{R}^+$ mit

$$\frac{\|F(X_0+H) - F(X_0) - DF_{X_0}(H)\|}{\|H\|} < \frac{\varepsilon}{2}$$

für alle H mit $X_0 + H \in M$ und $0 < \|H\| < \delta_1$. Da lineare Abbildungen gleichmäßig stetig sind, existiert ein $\delta_2 \in \mathbb{R}^+$ mit $\|DF_{X_0}(H)\| < \varepsilon/2$ für alle $H \in \mathbb{R}^n$ mit $\|H\| < \delta_2$. Für alle $X_0 + H \in M$ mit $\|H\| < \delta := \min\{\delta_1, \delta_2, 1\}$ gilt dann

$$\begin{aligned}
&\|F(X_0+H)-F(X_0)\| \\
&\qquad \le \|F(X_0+H)-F(X_0)-DF_{X_0}(H)\| + \|DF_{X_0}(H)\| \\
&\qquad \le \frac{\varepsilon}{2}\|H\| + \frac{\varepsilon}{2} \\
&\qquad \le \varepsilon.
\end{aligned}$$

∎

1.5 Satz. *Sind die Abbildungen $F : M \to \mathbb{R}^k$ und $G : M \to \mathbb{R}^k$ differenzierbar in X_0, so auch $F+G$, und es gilt*

$$D(F+G)_{X_0} = DF_{X_0} + DG_{X_0}.$$

Beweis. Die Behauptung ergibt sich aus

$$\begin{aligned}
&(F+G)(X_0+H)-(F+G)(X_0)-(DF_{X_0}+DG_{X_0})(H) \\
&= [F(X_0+H)-F(X_0)-DF_{X_0}(H)] + [G(X_0+H)-G(X_0)-DG_{X_0}(H)]
\end{aligned}$$

nach Division durch $\|H\|$ und Limesbildung $H \to 0$. ∎

Die wichtige Kettenregel besagt, dass die Komposition differenzierbarer Abbildungen differenzierbar ist und dass das Differential der Komposition die Komposition der Differentiale ist.

1.6 Satz (Kettenregel)**.** *Seien $M \subseteq \mathbb{R}^n$ und $N \subseteq \mathbb{R}^k$ offen, seien $F : M \to N$ und $G : N \to \mathbb{R}^m$ Abbildungen. Sei $X_0 \in M$, sei F differenzierbar in X_0 und G differenzierbar in $F(X_0)$. Dann ist $G \circ F$ differenzierbar in X_0, und es gilt*

$$D(G \circ F)_{X_0} = DG_{F(X_0)} \circ DF_{X_0}.$$

Beweis. Wir setzen $F(X_0) =: Y_0$ und $F(X_0+H)-F(X_0) = Z_H$ für $H \in \mathbb{R}^n \setminus \{0\}$ und $X_0+H \in M$. Für diese H gilt

$$\begin{aligned}
&\frac{1}{\|H\|}\left[(G\circ F)(X_0+H)-(G\circ F)(X_0)-DG_{F(X_0)}\circ DF_{X_0}(H)\right] \\
&\qquad = \frac{1}{\|H\|}\left[G(Y_0+Z_H)-G(Y_0)-DG_{Y_0}(Z_H)\right] \\
&\qquad\quad + \frac{1}{\|H\|}\left[DG_{Y_0}(F(X_0+H)-F(X_0)) - DG_{Y_0}(DF_{X_0}(H))\right] \\
&\qquad = \frac{1}{\|H\|}\left[G(Y_0+Z_H)-G(Y_0)-DG_{Y_0}(Z_H)\right] \\
&\qquad\quad + DG_{Y_0}\left(\frac{1}{\|H\|}\left(F(X_0+H)-F(X_0)-DF_{X_0}(H)\right)\right),
\end{aligned}$$

wobei zuletzt benutzt wurde, dass das Differential eine lineare Abbildung ist. Wegen der Differenzierbarkeit von F in X_0 und der Stetigkeit der linearen Abbildung DG_{Y_0} gilt

$$\lim_{H\to 0} DG_{Y_0}\left(\frac{1}{\|H\|}\left(F(X_0+H)-F(X_0)-DF_{X_0}(H)\right)\right)=0.$$

Für den ersten Summanden haben wir

$$\begin{aligned}&\frac{1}{\|H\|}\left[G(Y_0+Z_H)-G(Y_0)-DG_{Y_0}(Z_H)\right]\\&\quad=\begin{cases}0, & \text{falls } Z_H=0,\\ \frac{\|Z_H\|}{\|H\|}\frac{1}{\|Z_H\|}\left[G(Y_0+Z_H)-G(Y_0)-DG_{Y_0}(Z_H)\right] & \text{sonst.}\end{cases}\end{aligned}$$

Aus der Differenzierbarkeit von G in Y_0, der Stetigkeit von F in X_0 und der Beschränktheit von $\|Z_H\|/\|H\|$ in einer Umgebung von 0 (die aus der Differenzierbarkeit von F in X_0 folgt) ergibt sich jetzt

$$\lim_{H\to 0}\frac{1}{\|H\|}\left[G(Y_0+Z_H)-G(Y_0)-DG_{Y_0}(Z_H)\right]=0.$$

Daraus folgt die Behauptung. ∎

Bei konkreten Berechnungen wird man im allgemeinen die Kettenregel in Koordinatenschreibweise benutzen; hierauf kommen wir später zurück.

Den Fall $n=1$ der Kettenregel (Satz 1.6), der häufig vorkommen wird, schreiben wir noch mit Verwendung des Ableitungsvektors.

1.7 Folgerung. *Seien $I\subseteq\mathbb{R}$ und $N\subseteq\mathbb{R}^k$ offen, seien $F: I\to N$ und $G: N\to\mathbb{R}^m$ Abbildungen. Ist F differenzierbar in $t_0\in I$ und G differenzierbar in $F(t_0)$, so gilt*

$$(G\circ F)'(t_0)=DG_{F(t_0)}(F'(t_0)).$$

Beweis. Nach Satz 1.6 ist

$$D(G\circ F)_{t_0}=DG_{F(t_0)}\circ DF_{t_0};$$

für $h\in\mathbb{R}$ folgt also

$$\begin{aligned}h(G\circ F)'(t_0)&=D(G\circ F)_{t_0}(h)=DG_{F(t_0)}(DF_{t_0}(h))\\&=DG_{F(t_0)}(hF'(t_0))=hDG_{F(t_0)}(F'(t_0)).\end{aligned}$$ ∎

Hiermit läßt sich auch die Bedeutung des Differentials noch etwas einsichtiger machen. Mit den Bezeichnungen aus Folgerung 1.7 ist $F: I\to N$ (wobei I ein Intervall sei) eine parametrisierte Kurve durch $F(t_0)=:X_0$, und $F'(t_0)$ ist ihr Tangentenvektor an dieser Stelle. Umgekehrt ist jeder Vektor $T\in\mathbb{R}^k$ Tangentenvektor einer passenden parametrisierten Kurve durch X_0;

zum Beispiel kann man $F(t) := X_0 + tT$ für $t \in \mathbb{R}$ setzen. Nun ist die Komposition $G \circ F$ eine parametrisierte Kurve in $\mathbb{R}^m$ durch $F(X_0)$. Nach Folgerung 1.7 gilt

$$(G \circ F)'(t_0) = DG_{X_0}(F'(t_0)).$$

Man erhält also den Tangentenvektor der Bildkurve $G \circ F$, indem man auf den Tangentenvektor der ursprünglichen parametrisierten Kurve F das Differential der Abbildung G an der Stelle X_0 anwendet.

Einer der nützlichsten Sätze aus Analysis I über differenzierbare Funktionen ist der Mittelwertsatz (Kapitel 6, Satz 2.2). Er besagt, dass man für eine differenzierbare Funktion $f : [x, y] \to \mathbb{R}$ die Differenz $f(y) - f(x)$ darstellen kann durch

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

wobei $z \in (x, y)$ eine passende Zwischenstelle ist. Ausgenutzt wird dies meist in folgender Weise. Weiß man, dass für alle $z \in (x, y)$ die Abschätzung $|f'(z)| \leq c$ mit einer Konstanten c gilt, so folgt

$$|f(y) - f(x)| \leq c|y - x|.$$

Mit anderen Worten, die Differenz der Funktionswerte bei x und y ist klein, wenn im ganzen Intervall der Betrag der Ableitung klein ist. In dieser Form läßt sich der Mittelwertsatz auf differenzierbare Abbildungen F von $\mathbb{R}^n$ in den $\mathbb{R}^k$ übertragen, nicht jedoch in der vorherigen Form einer Gleichung. Qualitativ besagt diese Verallgemeinerung des Mittelwertsatzes, dass der Abstand $\|F(Y) - F(X)\|$ nicht groß ist, wenn das Differential von F längs der Verbindungsstrecke von X und Y nicht groß ist. Die Größe des Differentials wird dabei im Sinne der in Lemma 2.1 aus Kapitel 10 eingeführten Norm linearer Abbildungen gemessen. Mit

$$[X, Y] := \{(1 - \lambda)X + \lambda Y \mid 0 \leq \lambda \leq 1\}$$

bezeichnen wir die Verbindungsstrecke der Punkte $X, Y \in \mathbb{R}^n$.

1.8 Satz. *Sei $M \subseteq \mathbb{R}^n$ offen, seien $X, Y \in M$ Punkte mit $[X, Y] \subseteq M$. Die Abbildung $F : M \to \mathbb{R}^k$ sei stetig in M und differenzierbar in $(1 - \lambda)X + \lambda Y$ für alle $\lambda \in (0, 1)$. Gilt*

$$\|DF_Z\| \leq c \quad \text{für } Z = (1 - \lambda)X + \lambda Y \text{ mit } \lambda \in (0, 1),$$

mit einer Konstanten c, so gilt

$$\|F(Y) - F(X)\| \leq c\|Y - X\|.$$

Beweis. Zunächst behandeln wir einen Spezialfall. Sei $G : [0,1] \to \mathbb{R}^k$ eine stetige Abbildung derart, dass die Einschränkung $G|_{(0,1)}$ differenzierbar ist und dass

$$\|G'(t)\| \leq c \qquad \text{für alle } t \in (0,1)$$

gilt. Wähle $\varepsilon \in \mathbb{R}^+$ und setze

$$A := \{t \in [0,1] \mid \|G(t) - G(0)\| \leq (c+\varepsilon)t + \varepsilon\}.$$

Da G in 0 stetig ist, enthält A jedenfalls ein Intervall $[0,\tau]$ mit $\tau > 0$. Setze $s := \sup A$; dann ist also $0 < s \leq 1$. Da G in s stetig ist, gilt

$$\|G(s) - G(0)\| \leq (c+\varepsilon)s + \varepsilon,$$

also $s \in A$. Angenommen, es wäre $s < 1$. Da G in s differenzierbar ist, gibt es ein $h > 0$ mit $s + h \leq 1$ und

$$\left\| \frac{G(s+h) - G(s)}{h} - G'(s) \right\| \leq \varepsilon,$$

also mit

$$\left\| \frac{G(s+h) - G(s)}{h} \right\| \leq \|G'(s)\| + \varepsilon \leq c + \varepsilon.$$

Es folgt

$$\begin{aligned} \|G(s+h) - G(0)\| &\leq \|G(s+h) - G(s)\| + \|G(s) - G(0)\| \\ &\leq (c+\varepsilon)h + (c+\varepsilon)s + \varepsilon \\ &= (c+\varepsilon)(s+h) + \varepsilon, \end{aligned}$$

also $s+h \in A$, im Widerspruch zur Definition von s. Damit ist $s = 1$ bewiesen; es gilt also

$$\|G(1) - G(0)\| \leq c + 2\varepsilon.$$

Da $\varepsilon \in \mathbb{R}^+$ beliebig war, folgt

$$\|G(1) - G(0)\| \leq c.$$

Nun sei F wie in Satz 1.8 vorausgesetzt. Wir definieren $K : [0,1] \to \mathbb{R}^n$ durch

$$K(t) := (1-t)X + tY \qquad \text{für } t \in [0,1].$$

Trivialerweise ist K differenzierbar und $K'(t) = Y - X$. Nach der Kettenregel (Satz 1.6) ist $F \circ K$ differenzierbar in $t \in (0,1)$, und es gilt

$$D(F \circ K)_t = DF_{K(t)} \circ DK_t,$$

also

$$(F \circ K)'(t) = DF_{K(t)}(K'(t)) = DF_{K(t)}(Y - X).$$

Nach Lemma 2.1 aus Kapitel 10 folgt

$$\|(F \circ K)'(t)\| = \|DF_{K(t)}(Y - X)\| \leq \|DF_{K(t)}\|\|Y - X\| \leq c\|Y - X\|.$$

Nach dem bereits Bewiesenen (mit $G := F \circ K$ und $c\|Y - X\|$ statt c) folgt

$$\|F(Y) - F(X)\| = \|(F \circ K)(1) - (F \circ K)(0)\| \leq c\|Y - X\|. \qquad \blacksquare$$

1.9 Folgerung. *Sei $M \subseteq \mathbb{R}^n$ offen und zusammenhängend, sei $F : M \to \mathbb{R}^k$ eine differenzierbare Abbildung mit $DF_X = 0$ für alle $X \in M$. Dann ist F konstant.*

Beweis. Seien $X, Y \in M$. Da M zusammenhängend und offen ist, gibt es Punkte $X_1, \ldots, X_k$ mit $X_1 = X$, $X_k = Y$ und $[X_i, X_{i+1}] \subseteq M$ für $i = 1, \ldots, k-1$. (Denn für festes $X \in M$ ist die Menge A_1 aller $Y \in M$, die derart mit X durch einen Streckenzug in M verbindbar sind, offen und nicht leer. Die Menge $A_2 := M \setminus A_1$ ist ebenfalls offen. Da $A_1 \cup M_2 = M$, $A_1 \cap A_2 = \emptyset$ und M zusammenhängend ist, muß $A_2 = \emptyset$ sein, also $M = A_1$.) Für $i \in \{1, \ldots, k-1\}$ gilt nach Satz 1.8 die Gleichung $F(X_{i+1}) - F(X_i) = 0$. Hieraus folgt $F(Y) = F(X)$. Da $X, Y \in M$ beliebig waren, folgt die Behauptung. $\blacksquare$

11.2 Partielle Ableitungen und Koordinatenfunktionen

Bei der Behandlung konkreter differenzierbarer Abbildungen wird man zweckmäßigerweise ihre Koordinatenfunktionen heranziehen. Differentiale und Kettenregel treten dann in Matrizenschreibweise auf. Diese Umformulierungen wollen wir in diesem Abschnitt vornehmen. Dazu müssen wir zunächst partielle Ableitungen reellwertiger Funktionen betrachten.

Im ersten Teil des Folgenden sei stets $M \subset \mathbb{R}^n$ eine offene Menge, $X_0 \in M$ und $f : M \to \mathbb{R}$ eine reelle Funktion.

Für reellwertige Funktionen hängt der Begriff des Differentials eng zusammen mit dem der Richtungsableitung.

2.1 Definition. *Sei $E \in \mathbb{R}^n \setminus \{0\}$ ein Vektor. Wenn die Funktion*

$$t \mapsto f(X_0 + tE)$$

in 0 differenzierbar ist, so wird ihre Ableitung mit $f'(X_0; E)$ bezeichnet, also

$$f'(X_0; E) := \lim_{h \to 0} \frac{f(X_0 + hE) - f(X_0)}{h}.$$

$f'(X_0; E)$ *heißt* Richtungsableitung *von f bezüglich E im Punkt X_0.*

2.2 Satz. *Ist f in X_0 differenzierbar, so existieren alle Richtungsableitungen von f in X_0, und es gilt*

$$f'(X_0; E) = Df_{X_0}(E) \qquad \text{für } E \in \mathbb{R}^n.$$

Beweis. Definieren wir $g(t) := f(X_0 + tE)$ für alle $t \in \mathbb{R}$ mit $X_0 + tE \in M$, so ist g nach der Kettenregel (Folgerung 1.7) in 0 differenzierbar, und es gilt $g'(0) = Df_{X_0}(E)$. ∎

Behandelt man Aufgaben über differenzierbare Funktionen unter Zuhilfenahme von Koordinaten, so treten insbesondere die Richtungsableitungen in denjenigen Richtungen auf, die durch die Vektoren $E_1, \ldots, E_n$ der Standardbasis des $\mathbb{R}^n$ gegeben sind. Hierfür hat man daher besondere Bezeichnungen:

2.3 Definition. *Für $i \in \{1, \ldots, n\}$ schreibt man*

$$f'(X_0; E_i) =: \partial_i f(X_0),$$

falls diese Ableitung existiert. $\partial_i f(X_0)$ heißt i-te partielle Ableitung von f in X_0.

Eine häufig verwendete Schreibweise ist auch

$$\partial_i f = \frac{\partial f}{\partial x_i}.$$

Wie partielle Ableitungen zu berechnen sind, ist klar: Für $X_0 = (x_1^0, \ldots, x_n^0)$ gilt

$$\begin{aligned} \partial_i f(X_0) = f'(X_0; E_i) &= \lim_{h \to 0} \frac{f(X_0 + hE_i) - f(X_0)}{h} \\ &= \lim_{h \to 0} \frac{f(x_1^0, \ldots, x_{i-1}^0, x_i^0 + h, x_{i+1}^0, \ldots, x_n^0) - f(x_1^0, \ldots, x_n^0)}{h}. \end{aligned}$$

Das ist die gewöhnliche Ableitung der Funktion

$$t \mapsto f(x_1^0, \ldots, x_{i-1}^0, t, x_{i+1}^0, \ldots, x_n^0)$$

an der Stelle x_i^0. Hieraus erklärt sich die Bezeichnung „partielle Ableitung“, und es ergibt sich eine einfache Berechnungsvorschrift.

Durch die partiellen Ableitungen läßt sich nun auch die Richtungsableitung allgemein und damit das Differential einer differenzierbaren Funktion

leicht ausdrücken. Für $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ gilt wegen der Linearität des Differentials

$$\begin{aligned} f'(X_0; Y) &= Df_{X_0}(y_1 E_1 + \cdots + y_n E_n) \\ &= y_1 Df_{X_0}(E_1) + \cdots + y_n Df_{X_0}(E_n) \\ &= y_1 \partial_1 f(X_0) + \cdots + y_n \partial_n f(X_0) \\ &= \langle Y, (\partial_1 f(X_0), \ldots, \partial_n f(X_0)) \rangle. \end{aligned}$$

dass man, wie hier, ein lineares Funktional auf dem $\mathbb{R}^n$ als Skalarprodukt mit einem festen Vektor darstellen kann, ist aus der Linearen Algebra wohlbekannt. Für den in diesem Fall auftretenden Vektor hat man eine besondere Bezeichnung.

2.4 Definition. *Existieren für $f : M \to \mathbb{R}$ (mit $M \subseteq \mathbb{R}^n$) die partiellen Ableitungen in X_0, so heißt der Vektor*

$$\nabla f(X_0) := (\partial_1 f(X_0), \ldots, \partial_n f(X_0))$$

der Gradient von f in X_0. *Ist f in X_0 differenzierbar, so gilt*

$$f'(X_0; E) = Df_{X_0}(E) = \langle \nabla f(X_0), E \rangle \quad \textit{für alle } E \in \mathbb{R}^n.$$

Man beachte, dass die Darstellung

$$\nabla f(X_0) = (\partial_1 f(X_0), \ldots, \partial_n f(X_0))$$

von einer speziellen Basis des $\mathbb{R}^n$ Gebrauch macht, dass der Gradient aber, wenn f in X_0 differenzierbar ist, nicht von der gewählten Basis abhängt. Der Gradient $\nabla f(X_0)$ ist ja durch die basisunabhängige Gleichung

$$\langle \nabla f(X_0), E \rangle = f'(X_0; E) \qquad \text{für } E \in \mathbb{R}^n$$

festgelegt.

Der Gradient hat eine einfache anschauliche Bedeutung: Nach der Cauchy-Schwarzschen Ungleichung (Satz 1.3 aus Kapitel 10) gilt für $\|E\| = 1$

$$f'(X_0; E) = \langle \nabla f(X_0), E \rangle \leq \|\nabla f(X_0)\|,$$

und das Gleichheitszeichen gilt (wie am Beweis von Satz 1.3 aus Kapitel 10 abzulesen ist) genau dann, wenn

$$\nabla f(X_0) = \lambda E \qquad \text{mit einem } \lambda \geq 0$$

ist. Wir halten dies als Satz fest.

2.5 Satz. *Sei f in X_0 differenzierbar und $\nabla f(X_0) \neq 0$. Dann nimmt die Richtungsableitung von f in X_0 bezüglich Einheitsvektoren E genau für*

$$E = \frac{\nabla f(X_0)}{\|\nabla f(X_0)\|}$$

ihr Maximum an; dieses Maximum ist gleich $\|\nabla f(X_0)\|$.

Mit anderen Worten: Der Gradient weist in die Richtung stärksten Anstiegs der Funktion; seine Länge gibt die Größe dieses Anstiegs an.

Für differenzierbare reellwertige Funktionen gilt der Mittelwertsatz in Form einer Gleichung; er läßt sich bequem mit Hilfe des Gradienten formulieren.

2.6 Satz (Mittelwertsatz)**.** *Seien $X, Y \in M$ Punkte mit $[X, Y] \subseteq M$. Sei $f : M \to \mathbb{R}$ stetig in M und differenzierbar in $(1-t)X + tY$ für $t \in (0,1)$. Dann gibt es eine Zahl $\vartheta \in (0,1)$ mit*

$$f(Y) - f(X) = \langle \nabla f((1-\vartheta)X + \vartheta Y), Y - X \rangle.$$

Beweis. Setze $g(t) := f((1-t)X + tY)$ für $0 \leq t \leq 1$. Dann ist g stetig und nach der Kettenregel differenzierbar in $(0,1)$. Nach dem Mittelwertsatz aus Analysis I (Kapitel 6, Satz 2.2) existiert daher eine Zahl $\vartheta \in (0,1)$ mit

$$\begin{aligned} f(Y) - f(X) &= g(1) - g(0) = g'(\vartheta) \\ &= \langle \nabla f((1-\vartheta)X + \vartheta Y), Y - X \rangle. \end{aligned}$$

∎

Wir wollen nun auf den grundsätzlichen Zusammenhang zwischen Differenzierbarkeit und partieller Differenzierbarkeit eingehen. Dabei heißt eine reellwertige Funktion *partiell differenzierbar*, wenn ihre partiellen Ableitungen existieren. Für eine partiell differenzierbare Funktion $f : M \to \mathbb{R}$ ist also für jedes $i \in \{1, \ldots, n\}$ die Funktion

$$\partial_i f : M \to \mathbb{R} \colon X \mapsto \partial_i f(X)$$

erklärt, die man als *i-te partielle Ableitung* von f bezeichnet. Aus Differenzierbarkeit folgt partielle Differenzierbarkeit, nach Satz 2.2 sogar die Existenz aller Richtungsableitungen. Umgekehrt folgt aus der Existenz aller Richtungsableitungen nach einem Beispiel vom Anfang dieses Kapitels nicht einmal die Stetigkeit, also gewiß nicht die Differenzierbarkeit. Andererseits ist die partielle Differenzierbarkeit oft leicht nachprüfbar, und partielle Ableitungen sind in konkreten Fällen gut zu berechnen. Es sind daher Aussagen von Nutzen, die aus partieller Differenzierbarkeit und zusätzlichen Aussagen auf Differenzierbarkeit schließen lassen.

2.7 Satz. *Ist $f\colon M \to \mathbb{R}$ in einer Umgebung von X_0 partiell differenzierbar und sind die partiellen Ableitungen in X_0 stetig, so ist f in X_0 differenzierbar.*

Beweis. Sei U eine offene Kugel mit Mittelpunkt $X_0 = (x_1, \dots, x_n)$, in der f partiell differenzierbar ist. Für $H = (h_1, \dots, h_n) \in \mathbb{R}^n$ mit $X_0 + H \in U$ gilt nach dem Mittelwertsatz der Differentialrechnung

$$\begin{aligned} &f(X_0 + H) - f(X_0) \\ &\quad = \sum_{i=1}^{n} [f(x_1, \dots, x_{i-1}, x_i + h_i, x_{i+1} + h_{i+1}, \dots, x_n + h_n) \\ &\qquad - f(x_1, \dots, x_{i-1}, x_i, x_{i+1} + h_{i+1}, \dots, x_n + h_n)] \\ &\quad = \sum_{i=1}^{n} h_i \partial_i f(x_1, \dots, x_{i-1}, x_i + c_i h_i, x_{i+1} + h_{i+1}, \dots, x_n + h_n) \end{aligned}$$

mit geeigneten $c_i \in (0,1)$, also

$$\begin{aligned} &\frac{1}{\|H\|} \left| f(X_0 + H) - f(X_0) - \langle \nabla f(X_0), H \rangle \right| \\ &\quad = \frac{1}{\|H\|} \left| \sum_{i=1}^{n} h_i [\partial_i f(x_1, \dots, x_{i-1}, x_i + c_i h_i, \dots, x_n + h_n) - \partial_i f(X_0)] \right| \\ &\quad \le \sum_{i=1}^{n} \left| \partial_i f(x_1, \dots, x_{i-1}, x_i + c_i h_i, x_{i+1} + h_{i+1}, \dots, x_n + h_n) - \partial_i f(X_0) \right|. \end{aligned}$$

Da nach Voraussetzung jede partielle Ableitung $\partial_i f$ in X_0 stetig ist, ist der Limes der rechten Seite für $\|H\| \to 0$ gleich 0. Daraus folgt die Behauptung. ∎

Nach dieser Behandlung reellwertiger Funktionen kehren wir nun zu allgemeinen differenzierbaren Abbildungen in den $\mathbb{R}^k$ zurück. Wir verwenden jetzt Koordinatenfunktionen und deren partielle Ableitungen.

Gegeben seien im folgenden, wenn nichts anderes gesagt ist, eine offene Menge $M \subseteq \mathbb{R}^n$, ein Punkt $X_0 \in M$ und eine Abbildung $F : M \to \mathbb{R}^k$. Für $i = 1, \dots, k$ sei $f_i : M \to \mathbb{R}$ die i-te Koordinatenfunktion von F bezüglich der Standardbasis $E_1', \dots, E_k'$ des $\mathbb{R}^k$, also

$$F(X) = (f_1(X), \dots, f_k(X)) = \sum_{i=1}^{k} f_i(X) E_i' \qquad \text{für } X \in M.$$

Mit $X = (x_1, \dots, x_n)$, $Y = (y_1, \dots, y_k)$ lautet also die Gleichung $Y = F(X)$ in Koordinatenschreibweise

$$\begin{aligned} y_1 &= f_1(x_1, \dots, x_n), \\ &\vdots \quad \vdots \\ y_k &= f_k(x_1, \dots, x_n). \end{aligned}$$

Differenzierbarkeit einer Abbildung ist, analog wie bei der Stetigkeit, gleichwertig mit Differenzierbarkeit der Koordinatenfunktionen.

2.8 Satz. *Die Abbildung F ist genau dann differenzierbar in X_0, wenn alle Koordinatenfunktionen f_i in X_0 differenzierbar sind $(i = 1, \dots, k)$. Ist das der Fall, so gilt*

$$DF_{X_0}(H) = \sum_{i=1}^{k} (Df_i)_{X_0}(H) E_i' \qquad \text{für } H \in \mathbb{R}^n.$$

Beweis. Sei $L : \mathbb{R}^n \to \mathbb{R}^k$ eine lineare Abbildung. Nach Satz 2.5 aus Kapitel 10 gilt

$$\begin{aligned} & \lim_{H \to 0} \frac{F(X_0 + H) - F(X_0) - L(H)}{\|H\|} = 0 \\ \Leftrightarrow \quad & \lim_{H \to 0} \frac{f_i(X_0 + H) - f_i(X_0) - (p_i \circ L)(H)}{\|H\|} = 0 \quad \text{für } i = 1, \dots, k. \end{aligned}$$

Daraus folgen die Behauptungen unmittelbar. ∎

An der Gleichung in Satz 2.8 können wir nun sofort ablesen, durch welche Matrix das Differential von F in X_0 bezüglich der Standardbasen beschrieben wird. Sei $E_1, \dots, E_n$ die Standardbasis von $\mathbb{R}^n$. Ist $L : \mathbb{R}^n \to \mathbb{R}^k$ eine lineare Abbildung, so ist die ihr bezüglich der Standardbasen zugeordnete Matrix $(a_{ij})_{\substack{i=1,\dots,k \\ j=1,\dots,n}}$ definiert durch

$$L(E_j) = \sum_{i=1}^{k} a_{ij} E_i', \qquad j = 1, \dots, n.$$

Setzen wir nun in der Gleichung in Satz 2.8 speziell $H = E_j$ ein, so erhalten wir wegen $Df_{X_0}(E_j) = \partial_j f(X_0)$ die Gleichungen

$$DF_{X_0}(E_j) = \sum_{i=1}^{k} \partial_j f_i(X_0) E_i',$$

also $a_{ij} = \partial_j f_i(X_0)$. Wir halten das fest und definieren:

2.9 Definition. *Sei $F : M \to \mathbb{R}^k$ mit $M \subseteq \mathbb{R}^n$ differenzierbar in $X_0 \in M$. Dann wird das Differential DF_{X_0} bezüglich der Standardbasen in $\mathbb{R}^n$ und $\mathbb{R}^k$ beschrieben durch die $k \times n$-Matrix*

$$JF(X_0) := \begin{pmatrix} \partial_1 f_1(X_0) & \cdots & \partial_n f_1(X_0) \\ \vdots & & \vdots \\ \partial_1 f_k(X_0) & \cdots & \partial_n f_k(X_0) \end{pmatrix}.$$

Sie heißt die Funktionalmatrix *oder* Jacobische Matrix *der Abbildung F an der Stelle X_0. Im Fall $k = n$ wird die Determinante* $\det JF(X_0)$ *der Funktionalmatrix als* Funktionaldeterminante *von F in X_0 bezeichnet.*

Bemerkung. Für die Funktionaldeterminante von F (im Fall $n = k$) findet man auch oft die Bezeichnung

$$\frac{\partial(f_1, \ldots, f_n)}{\partial(x_1, \ldots, x_n)}.$$

Beschreibt man lineare Abbildungen (nach Einführung von Basen in den beteiligten Räumen) durch Matrizen, so wird einer Komposition von Abbildungen als Matrix bekanntlich das Matrizenprodukt der Matrizen der einzelnen Abbildungen zugeordnet. Mit der Bezeichnung aus Definition 2.9 lautet die Kettenregel aus Satz 1.6 daher jetzt folgendermaßen: Die Funktionalmatrix einer Komposition ist gleich dem Matrizenprodukt (in der richtigen Reihenfolge) der Funktionalmatrizen der einzelnen Abbildungen.

2.10 Satz (Kettenregel in Koordinatenschreibweise)**.** *Seien $M \subseteq \mathbb{R}^n$ und $N \subseteq \mathbb{R}^k$ offen, seien $F : M \to N$ und $G : N \to \mathbb{R}^m$ Abbildungen, sei F differenzierbar in X_0 und G differenzierbar in $F(X_0)$. Dann gilt*

$$J(G \circ F)(X_0) = JG(F(X_0)) \cdot JF(X_0)$$

(Matrizenprodukt), also mit $G \circ F =: H$

$$\partial_j h_i(X_0) = \sum_{r=1}^{k} \partial_r g_i(F(X_0)) \partial_j f_r(X_0)$$

für $i = 1, \ldots, m$ und $j = 1, \ldots, n$.

Etwas übersichtlicher ist vielleicht die Schreibweise

$$\frac{\partial h_i}{\partial x_j}(X_0) = \sum_{r=1}^{k} \frac{\partial g_i}{\partial y_r}(Y_0) \frac{\partial f_r}{\partial x_j}(X_0), \qquad Y_0 = F(X_0).$$

Im Fall $m = n = 1$ reduziert sich die Kettenregel in Koordinatenschreibweise auf eine einzige Gleichung, die wir für

$$h(t) = g(f_1(t), \ldots, f_k(t))$$

in der Form

$$\frac{dh}{dt} = \frac{\partial g}{\partial x_1}\frac{df_1}{dt} + \cdots + \frac{\partial g}{\partial x_k}\frac{df_k}{dt}$$

schreiben können. Hat hier df_i/dt (und damit auch dh/dt) das Argument t, so muß $\partial g/\partial x_i$ das Argument $(f_1(t), \ldots, f_k(t))$ haben.

11.3 Höhere Ableitungen, Taylorformel, lokale Extrema

Im folgenden seien stets eine offene Menge $M \subseteq \mathbb{R}^n$, ein Punkt $X_0 \in M$ und eine reellwertige Funktion $f : M \to \mathbb{R}$ gegeben.

In diesem Abschnitt betrachten wir höhere partielle Ableitungen reellwertiger Funktionen. Es liegt auf der Hand, wie sie zu definieren sind. Existiert die i-te partielle Ableitung $\partial_i f$ von f in einer Umgebung von X_0 und existiert die k-te partielle Ableitung von $\partial_i f$ in X_0, so wird sie mit

$$\partial_k \partial_i f(X_0) \quad \text{oder} \quad \frac{\partial^2 f}{\partial x_k \partial x_i}(X_0)$$

bezeichnet. Entsprechend wird allgemein die partielle Ableitung

$$\partial_{i_r} \ldots \partial_{i_1} f = \frac{\partial^r f}{\partial x_{i_r} \ldots \partial x_{i_1}}$$

rekursiv definiert. Man bezeichnet sie als *partielle Ableitung r-ter Ordnung*. Existieren alle partiellen Ableitungen r-ter Ordnung in X_0, so heißt f *r-mal partiell differenzierbar* in X_0, und wenn dies für alle $X_0 \in M$ gilt, heißt f *r-mal partiell differenzierbar*. Ist f r-mal partiell differenzierbar und sind alle partiellen Ableitungen r-ter Ordnung der Funktion f stetig, so heißt f *r-mal stetig differenzierbar* (im Fall $r = 1$ kurz *stetig differenzierbar*). Die Menge der auf M r-mal stetig differenzierbaren reellen Funktionen wird mit $C^r(M)$ bezeichnet. Unter $C^0(M)$ wird die Menge der stetigen reellen Funktionen auf M verstanden. Ein Element von $C^r(M)$ heißt auch *Funktion der Klasse C^r* auf M. Es gilt also

$$C^0(M) \supset C^1(M) \supset C^2(M) \supset \ldots,$$

und zwar jeweils mit strikter Inklusion.

Bemerkung. Wird mehrmals nach derselben Veränderlichen differenziert, so verwendet man abkürzende Schreibweisen wie z.B.

$$\frac{\partial^2 f}{\partial x \partial x} =: \frac{\partial^2 f}{\partial x^2},$$

$$\underbrace{\partial_i \cdots \partial_i}_{s\text{-mal}} \underbrace{\partial_k \cdots \partial_k}_{r\text{-mal}} f = \frac{\partial^{r+s} f}{\partial x_i^s \partial x_k^r}.$$

Das folgende Beispiel zeigt, dass es ohne zusätzliche Voraussetzungen nicht gleichgültig ist, in welcher Reihenfolge die partiellen Differentiationen ausgeführt werden.

Beispiel. Sei $f : \mathbb{R}^2 \to \mathbb{R}$ erklärt durch

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

Die partiellen Ableitungen zweiter Ordnung existieren, und man erhält

$$\frac{\partial^2 f}{\partial y \partial x}(0,0) = -1, \qquad \frac{\partial^2 f}{\partial x \partial y}(0,0) = 1.$$

Derartiges kann jedoch nicht passieren, wenn die partiellen Ableitungen zweiter Ordnung noch stetig sind, wie der folgende Satz zeigt.

3.1 Satz (Vertauschbarkeit der Differentiationsreihenfolge)**.** *Ist $r \geq 2$ und $f \in C^r(M)$, so sind die partiellen Ableitungen von f bis zur r-ten Ordnung unabhängig von der Reihenfolge der Differentiationen, d.h. es gilt*

$$\partial_{i_1} \cdots \partial_{i_r} f = \partial_{i_{\sigma(1)}} \cdots \partial_{i_{\sigma(r)}} f$$

für jede Permutation σ der Zahlen $1, \ldots, r$.

Beweis. Da jede Permutation der Zahlen $1, \ldots, r$ Produkt von solchen Permutationen ist, bei denen nur zwei benachbarte Zahlen vertauscht werden, genügt es, den Beweis für den Fall $r = 2$ zu führen. Außerdem bedeutet es keine Einschränkung der Allgemeinheit, $n = 2$ anzunehmen, weil bei den in Betracht kommenden Differentiationen alle bis auf zwei Veränderliche festgehalten werden.

Sei $X = (x_1, x_2) \in M$ und $U \subseteq M$ eine offene Kugel mit Mittelpunkt X. Im folgenden sei $H = (h_1, h_2)$ so klein, dass $X + H \in U$ ist. Dann gilt

$$
\begin{aligned}
&f(x_1+h_1,x_2+h_2)-f(x_1+h_1,x_2)-f(x_1,x_2+h_2)+f(x_1,x_2)\\
&\quad=\varphi(x_1+h_1)-\varphi(x_1) && \text{mit } \varphi(t):=f(t,x_2+h_2)-f(t,x_2)\\
&\quad=\varphi'(x_1+c_1h_1)h_1 && \text{mit } c_1\in(0,1) \text{ (Mittelwertsatz)}\\
&\quad=[\partial_1 f(x_1+c_1h_1,x_2+h_2)-\partial_1 f(x_1+c_1h_1,x_2)]h_1\\
&\quad=[\psi(x_2+h_2)-\psi(x_2)]h_1 && \text{mit } \psi(t):=\partial_1 f(x_1+c_1h_1,t)\\
&\quad=\psi'(x_2+c_2h_2)h_1h_2 && \text{mit } c_2\in(0,1) \text{ (Mittelwertsatz)}\\
&\quad=\partial_2\partial_1 f(x_1+c_1h_1,x_2+c_2h_2)h_1h_2.
\end{aligned}
$$

Analog findet man

$$
\begin{aligned}
&f(x_1+h_1,x_2+h_2)-f(x_1+h_1,x_2)-f(x_1,x_2+h_2)+f(x_1,x_2)\\
&\qquad=\partial_1\partial_2 f(x_1+d_1h_1,x_2+d_2h_2)h_1h_2
\end{aligned}
$$

mit passenden $d_1,d_2\in(0,1)$. Für $h_1h_2\neq 0$ folgt also

$$\partial_2\partial_1 f(x_1+c_1h_1,x_2+c_2h_2)=\partial_1\partial_2 f(x_1+d_1h_1,x_2+d_2h_2).$$

Man beachte, dass c_i und d_i hier von h_i abhängen ($i=1,2$), da der Mittelwertsatz auf dem Intervall (x_1,x_1+h_1) bzw. (x_2,x_2+h_2) angewandt wurde. Es gilt jedoch $c_i,d_i\in(0,1)$ und daher mit $h_i\to 0$ auch $c_ih_i\to 0$ und $d_ih_i\to 0$. Wegen der vorausgesetzten Stetigkeit der partiellen Ableitungen zweiter Ordnung ergibt sich mit $(h_1,h_2)\to(0,0)$ die Behauptung

$$\partial_2\partial_1 f(x_1,x_2)=\partial_1\partial_2 f(x_1,x_2).$$ ∎

Einer der wichtigsten Sätze aus Analysis I über mehrfach differenzierbare Funktionen war die Taylorformel (Kapitel 6, Satz 3.3). Wir wollen nun untersuchen, was sie für Funktionen von n Veränderlichen liefert, wenn man diese auf Geraden durch einen betrachteten Punkt einschränkt. Zuerst erinnern wir uns an die Taylorformel aus Analysis I; dabei können wir uns auf einen Spezialfall beschränken. Sei $g:[-\varepsilon,h]\to\mathbb{R}$ mit $\varepsilon,h>0$ eine $(k+1)$-mal differenzierbare Funktion ($k\in\mathbb{N}$). Dann gilt

$$g(h)=\sum_{j=0}^{k}\frac{1}{j!}g^{(j)}(0)h^j+\frac{1}{(k+1)!}g^{(k+1)}(c)h^{k+1}$$

mit einer passenden Zwischenstelle $c\in(0,h)$.

Wir betrachten nun eine Funktion $f:M\to\mathbb{R}$ mit offenem Definitionsbereich $M\subseteq\mathbb{R}^n$ und einen Punkt $X_0\in M$ und machen die folgende Voraussetzung.

Voraussetzung. Sei $k\in\mathbb{N}$, $f\in C^k(M)$, die partiellen Ableitungen k-ter Ordnung von f sind in M differenzierbar, $H\in\mathbb{R}^n$ ist ein Vektor mit $[X_0,X_0+H]\subseteq M$.

Nun setzen wir

$$g(t) := f(X_0 + tH) \qquad \text{für } t \in [-\varepsilon, 1]$$

wobei $\varepsilon > 0$ so gewählt sei, dass $[X_0, X_0 - \varepsilon H] \subseteq M$ ist. Da f differenzierbar ist (denn die partiellen Ableitungen sind noch differenzierbar, also stetig, somit ist f nach Satz 2.7 differenzierbar), ist g nach der Kettenregel differenzierbar, und nach Satz 2.10 gilt

$$g'(t) = \sum_{i=1}^{n} \partial_i f(X_0 + tH) h_i,$$

wenn $H = (h_1, \ldots, h_n)$ ist. Nach Voraussetzung sind die partiellen Ableitungen $\partial_i f$ $(i = 1, \ldots, n)$ noch differenzierbare Funktionen, daher können wir abermals die Kettenregel anwenden und erhalten

$$g''(t) = \sum_{i=1}^{n} \sum_{j=1}^{n} \partial_j \partial_i f(X_0 + tH) h_i h_j.$$

So fortfahrend, erhalten wir schließlich

$$g^{(r)}(t) = \sum_{i_1, \ldots, i_r = 1}^{n} \partial_{i_1} \cdots \partial_{i_r} f(X_0 + tH) h_{i_1} \cdots h_{i_r}$$

für $r = 1, \ldots, k+1$. Auf g können wir die oben zitierte eindimensionale Taylorformel anwenden (mit $h = 1$). Es gibt also eine Zahl $c \in (0, 1)$ mit

$$g(1) = \sum_{r=0}^{k} \frac{1}{r!} g^{(r)}(0) + \frac{1}{(k+1)!} g^{(k+1)}(c).$$

Setzen wir hier die oben berechneten Ableitungen von g ein, so erhalten wir den folgenden Satz.

3.2 Satz (Taylorformel)**.** *Sei $M \subseteq \mathbb{R}^n$ offen, $X_0 \in M$, $H \in \mathbb{R}^n$ mit $[X_0, X_0 + H] \subseteq M$, $k \in \mathbb{N}$, $f \in C^k(M)$, und die partiellen Ableitungen k-ter Ordnung von f seien in M differenzierbar. Dann gibt es eine Zahl $c \in (0, 1)$ mit*

$$f(X_0 + H) = f(X_0) + \sum_{r=1}^{k} \frac{1}{r!} \sum_{i_1, \ldots, i_r = 1}^{n} \partial_{i_1} \cdots \partial_{i_r} f(X_0) h_{i_1} \cdots h_{i_r}$$
$$+ \frac{1}{(k+1)!} \sum_{i_1, \ldots, i_{k+1} = 1}^{n} \partial_{i_1} \cdots \partial_{i_{k+1}} f(X_0 + cH) h_{i_1} \cdots h_{i_{k+1}}.$$

Speziell für $k = 2$ ist also

$$f(X_0 + H) = f(X_0) + \langle \nabla f(X_0), H \rangle + \frac{1}{2} \sum_{i,j=1}^{n} \partial_i \partial_j f(X_0) h_i h_j + R(X_0; H)$$

mit

$$R(X_0; H) = \frac{1}{6} \sum_{i,j,m=1}^{n} \partial_i \partial_j \partial_m f(Y) h_i h_j h_m$$

und passendem $Y \in [X_0, X_0 + H]$. Wegen $|h_i| \leq \|H\|$ folgt

$$\lim_{H \to 0} \frac{R(X_0; H)}{\|H\|^2} = 0,$$

falls $\partial_i \partial_j \partial_m f$ auf M beschränkt ist. Analog zu Analysis I, Kapitel 6, Satz 3.4 (und daraus herleitbar) gilt diese Schlußfolgerung auch schon unter der Voraussetzung, dass $f \in C^1(M)$ ist und die partiellen Ableitungen von f in X_0 differenzierbar sind.

Man kann den durch

$$f(X_0+H) = f(X_0) + \langle \nabla f(X_0), H\rangle + \frac{1}{2}\sum_{i,j=1}^{n} \partial_i\partial_j f(X_0) h_i h_j + R(X_0;H),$$
$$\lim_{H\to 0} \frac{R(X_0;H)}{\|H\|^2} = 0$$

ausgedrückten Sachverhalt auch folgendermaßen beschreiben: In zweiter Näherung verhält sich die Funktion f an der Stelle X_0 wie ein Polynom zweiten Grades. Anschaulich läßt sich folgendes sagen. Betrachten wir den Graphen von f, so gibt der erste Term $f(X_0)$ der Taylorformel die Höhe des Graphen über X_0 an, der zweite Term, der durch den Gradienten von f in X_0 bestimmt ist, legt die Tangentialebene an den Graphen über X_0 fest, und der nächste Term,

$$\frac{1}{2}\sum_{i,j=1}^{n} \partial_i\partial_j f(X_0) h_i h_j,$$

gibt eine erste Information über die Gestalt des Graphen bei X_0. Dies zeigt sich insbesondere bei der Untersuchung lokaler Extremwerte. Zuvor wollen wir für diesen Term eine besondere Bezeichnung einführen.

3.3 Definition. *Sei $f : M \to \mathbb{R}$ zweimal partiell differenzierbar in X_0. Dann heißt die durch*

$$Q(f, X_0; H) := \sum_{i,j=1}^{n} \partial_i\partial_j f(X_0) h_i h_j \qquad \text{für } H = (h_1, \ldots, h_n) \in \mathbb{R}^n$$

definierte Funktion $Q(f, X_0; \cdot)$ die Hesse-Form *von f in X_0, und die Matrix*

$$\mathrm{Hess}(f)_{X_0} := \begin{pmatrix} \partial_1\partial_1 f(X_0) & \cdots & \partial_1\partial_n f(X_0) \\ \vdots & & \vdots \\ \partial_n\partial_1 f(X_0) & \cdots & \partial_n\partial_n f(X_0) \end{pmatrix}$$

heißt die Hessesche Matrix *von f in X_0.*

Falls f in einer Umgebung von X_0 zweimal stetig differenzierbar ist, dann ist nach Satz 3.1 die Hessesche Matrix von f symmetrisch. Die Hesse-Form ist also eine quadratische Form, die Hessesche Matrix ist die ihr (bezüglich der Standardbasis) zugeordnete Matrix. Über quadratische Formen, deren Behandlung in der Linearen Algebra erfolgt, wollen wir hier nur das für die folgenden Betrachtungen Notwendige bereitstellen.

3.4 Definition. *Eine Funktion* $q : \mathbb{R}^n \to \mathbb{R}$ *mit*

$$q(X) = \sum_{i,j=1}^{n} a_{ij} x_i x_j \qquad \text{für } X = (x_1, \ldots, x_n) \in \mathbb{R}^n$$

($a_{ij} \in \mathbb{R}$ *für* $i, j = 1, \ldots, n$*) heißt* quadratische Form *über* $\mathbb{R}^n$*. Die quadratische Form* q *heißt* positiv definit*, wenn* $q(X) > 0$ *für* $X \in \mathbb{R}^n \setminus \{0\}$ *gilt,* positiv semidefinit*, wenn* $q(X) \geq 0$ *für* $X \in \mathbb{R}^n$ *gilt,* negativ definit *(*negativ semidefinit*), wenn* $-q$ *positiv definit (bzw. positiv semidefinit) ist, und* indefinit*, wenn* q *weder positiv noch negativ semidefinit ist.*

Nach diesen Vorbereitungen können wir nun lokale Extremwerte differenzierbarer Funktionen untersuchen.

3.5 Definition. *Die Funktion* $f : M \to \mathbb{R}$ hat in X_0 ein lokales Maximum (lokales Minimum)*, wenn es eine Umgebung* $U \subseteq M$ *von* X_0 *gibt mit* $f(X_0) \geq f(Y)$ *(bzw.* $f(X_0) \leq f(Y)$*) für alle* $Y \in U$*. Gilt* $f(X_0) > f(Y)$ *(bzw.* $f(X_0) < f(Y)$*) für* $Y \in U \setminus \{X_0\}$*, so hat* f *in* X_0 *ein* striktes lokales Maximum (striktes lokales Minimum)*. Die Funktion* f *hat in* X_0 *ein lokales Extremum, wenn sie dort ein lokales Maximum oder ein lokales Minimum hat.*

Zunächst stellen wir, analog wie in Analysis I, für das Vorliegen eines lokalen Extremums eine notwendige Bedingung auf.

3.6 Satz. *Sei* $M \subseteq \mathbb{R}^n$ *offen und* $f : M \to \mathbb{R}$ *differenzierbar in* X_0*. Hat* f *in* X_0 *ein lokales Extremum, so ist* $\nabla f(X_0) = 0$*.*

Beweis. Hat f in X_0 ein lokales Extremum, so hat für jeden Vektor $H \in \mathbb{R}^n$ die durch

$$g(t) := f(X_0 + tH) \qquad \text{für } X_0 + tH \in M$$

erklärte Funktion g in 0 ein lokales Extremum. Nach dem Kriterium aus Analysis I (Kapitel 6, Satz 3.5) gilt daher

$$\langle \nabla f(X_0), H \rangle = g'(0) = 0.$$

Da $H \in \mathbb{R}^n$ beliebig war, ist $\nabla f(X_0) = 0$. ∎

3.7 Definition. *Ist* f *differenzierbar in einer Umgebung von* X_0 *und ist* $\nabla f(X_0) = 0$*, so heißt* X_0 *ein* kritischer Punkt *von* f*.*

Wir wollen nun hinreichende Bedingungen angeben für das Vorliegen eines strikten lokalen Extremums.

3.8 Satz. *Sei $M \subseteq \mathbb{R}^n$ offen und $f \in C^2(M)$. Sei X_0 kritischer Punkt von f. Ist die Hesse-Form $Q(f, X_0; \cdot)$ von f in X_0 positiv definit (negativ definit), so hat f in X_0 ein striktes lokales Minimum (striktes lokales Maximum).*

Beweis. Sei etwa $Q(f, X_0; \cdot)$ positiv definit. Wir zeigen zunächst die Existenz einer Zahl $\delta \in \mathbb{R}^+$ mit $U(X_0, \delta) \subseteq M$ und der Eigenschaft, dass für alle $H \in \mathbb{R}^n$ mit $\|H\| < \delta$ auch die Hesse-Form $Q(f, X_0 + H; \cdot)$ positiv definit ist.

Da die Funktion $Q(f, X_0; \cdot)$ stetig ist, nimmt ihre Einschränkung auf die wegen Satz 1.13 aus Kapitel 10 kompakte Menge $\{E \in \mathbb{R}^n \mid \|E\| = 1\}$ nach Folgerung 5.5 aus Kapitel 9 ein Minimum an. Da $Q(f, X_0; \cdot)$ positiv definit ist, ist dieses Minimum positiv. Es gibt also eine Zahl $a \in \mathbb{R}^+$ mit

$$Q(f, X_0; E) > 2a \quad \text{für alle } E \in \mathbb{R}^n \text{ mit } \|E\| = 1.$$

Da nach Voraussetzung die partiellen Ableitungen zweiter Ordnung der Funktion f in X_0 stetig sind (und M offen ist), gibt es ein $\delta \in \mathbb{R}^+$ mit $U(X_0, \delta) \subseteq M$ und

$$|\partial_i \partial_j f(X_0 + H) - \partial_i \partial_j f(X_0)| < \frac{a}{n^2} \quad \text{für alle} \quad H \in \mathbb{R}^n \quad \text{mit} \quad \|H\| < \delta$$

$(i, j = 1, \ldots, n)$. Für beliebiges $E = (e_1, \ldots, e_n) \in \mathbb{R}^n$ mit $\|E\| = 1$ und für $\|H\| < \delta$ folgt

$$\begin{aligned} &|Q(f, X_0 + H; E) - Q(f, X_0; E)| \\ &\quad \leq \sum_{i,j=1}^{n} |\partial_i \partial_j f(X_0 + H) - \partial_i \partial_j f(X_0)|\, |e_i|\, |e_j| \\ &\quad \leq \frac{a}{n^2} \sum_{i,j=1}^{n} 1 = a. \end{aligned}$$

Also ist $Q(f, X_0 + H; E) > a > 0$ für $\|E\| = 1$ und $\|H\| < \delta$; daraus folgt $Q(f, X_0 + H; Y) > 0$ für alle $Y \in \mathbb{R}^n \setminus \{0\}$. Für alle $H \in \mathbb{R}^n$ mit $\|H\| < \delta$ ist also $Q(f, X_0 + H; \cdot)$ positiv definit.

Nun gibt es nach Satz 3.2 (für $k = 1$) und wegen der Voraussetzung $\nabla(X_0) = 0$ ein $c \in (0, 1)$ mit

$$f(X_0 + H) = f(X_0) + \frac{1}{2} \sum_{i,j=1}^{n} \partial_i \partial_j f(X_0 + cH) h_i h_j.$$

Für $0 < \|H\| < \delta$ folgt

$$f(X_0 + H) > f(X_0).$$

Die Funktion f hat also in X_0 ein striktes lokales Minimum. ∎

Umgekehrt läßt sich aus dem Vorliegen eines lokalen Extremums lediglich folgern, dass die Hesse-Form in diesem Punkt semidefinit ist:

3.9 Satz. *Sei M offen und $f \in C^2(M)$. In X_0 habe f ein lokales Minimum (lokales Maximum). Dann ist die Hesse-Form $Q(f, X_0, \cdot)$ positiv (negativ) semidefinit.*

Beweis. Die Funktion f habe etwa in X_0 ein lokales Minimum. Angenommen, $Q(f, X_0; \cdot)$ wäre nicht positiv semidefinit. Dann gibt es ein $E = (e_1, \ldots, e_n) \in \mathbb{R}^n$ mit $Q(f, X_0; E) < 0$. Analog wie im Beweis von Satz 3.8 zeigt man die Existenz eines $\delta \in \mathbb{R}^+$ mit $U(X_0, \delta) \subseteq M$ und $Q(f, X_0 + H; E) < 0$ für alle $H \in \mathbb{R}^n$ mit $\|H\| < \delta$. Wählen wir $H := \lambda E$ mit hinreichend kleinem $\lambda \in \mathbb{R}^+$, so folgt $Q(f, X_0 + cH; H) < 0$ für $c \in (0, 1)$. Aus Satz 3.2 ergibt sich dann $f(X_0 + H) < f(X_0)$. Da hier $\|H\| > 0$ beliebig klein gewählt werden kann, hat f in X_0 kein lokales Minimum, ein Widerspruch. ∎

3.10 Folgerung. *Sei M offen und $f \in C^2(M)$. Ist die Hesse-Form $Q(f, X_0; \cdot)$ von f in X_0 indefinit, so hat f in X_0 kein lokales Extremum.*

Aus den vorstehenden Sätzen ergibt sich, dass man zur Untersuchung von lokalen Extrema Kriterien haben muß für den Definitheitscharakter quadratischer Formen. Wir geben ein solches Kriterium an für den Fall $n = 2$.

3.11 Satz. *Die durch*

$$q(x, y) = a_{11}x^2 + 2a_{12}xy + a_{22}y^2 \qquad \text{für } (x, y) \in \mathbb{R}^2$$

definierte quadratische Form q ist genau dann positiv definit, wenn

$$a_{11} > 0 \qquad \text{und} \qquad a_{11}a_{22} - a_{12}^2 > 0$$

ist. Im Fall $a_{11}a_{22} - a_{12}^2 < 0$ ist sie indefinit.

Beweis. Sei q positiv definit. Dann ist $a_{11} = q(1, 0) > 0$. Durch Ausrechnen bestätigt man die Identität

$$q(x, y) = \frac{1}{a_{11}} \left[(a_{11}x + a_{12}y)^2 + (a_{11}a_{22} - a_{12}^2)y^2 \right].$$

Daraus folgt

$$a_{11}a_{22} - a_{12}^2 = a_{11}\, q\left(-\frac{a_{12}}{a_{11}}, 1 \right) > 0.$$

Umgekehrt folgt aus $a_{11} > 0$ und $a_{11}a_{22} - a_{12}^2 > 0$ wegen der obigen Identität sofort $q(x, y) > 0$ für $(x, y) \neq (0, 0)$. ∎

11.4 Differenzierbare Abbildungen

Im ersten Teil dieses Abschnitts befassen wir uns mit differenzierbaren Abbildungen zwischen Mengen gleicher Dimension. Solche Abbildungen treten in den Anwendungen unter anderem als sogenannte Koordinatentransformationen auf. Wir erläutern dies zunächst an einem speziellen Beispiel.

Gegeben sei eine offene Menge $M \subseteq \mathbb{R}^2$ und eine Funktion $f : M \to \mathbb{R}$. Bei konkreten Fragestellungen ist es manchmal nicht zweckmäßig, einen Punkt aus M durch seine Koordinaten bezüglich der Standardbasis zu beschreiben, sondern man kann oft mit Vorteil andere Zahlenpaare wählen, durch die die Punkte ebenso festgelegt werden, die aber der Problemstellung besser angepaßt sind („krummlinige Koordinaten“). Nehmen wir etwa an, die Funktion f sei rotationssymmetrisch, das heißt $f(X)$ hänge nur vom „Radius“ $\|X\|$ ab. Dann empfiehlt es sich, in $M \setminus \{0\}$ Polarkoordinaten einzuführen: Jeder Punkt X ist eindeutig festgelegt durch seinen Abstand r vom Nullpunkt und (im Fall $X \neq 0$) durch den Winkel φ, den der Vektor X mit E_1 bildet. Für $X = (x, y)$ gilt dann

$$\begin{aligned} x &= r\cos\varphi, \\ y &= r\sin\varphi, \end{aligned}$$

und man schreibt etwa

$$f(x, y) = f(r\cos\varphi, r\sin\varphi) = \tilde{f}(r, \varphi)$$

und sagt, man habe damit f auf Polarkoordinaten bezogen. Dies ist folgendermaßen zu präzisieren. Wir setzen

$$U := \{(r, \varphi) \in \mathbb{R}^2 \,\big|\, r > 0,\ \ 0 \leq \varphi < 2\pi\}$$

und definieren eine Abbildung $\tau : U \to \mathbb{R}^2$ durch

$$\tau(r, \varphi) := (r\cos\varphi, r\sin\varphi).$$

Offenbar ist $\tau(U) = \mathbb{R}^2 \setminus \{0\}$, und τ ist injektiv. Die Abbildung τ ist differenzierbar, ihre Umkehrabbildung τ^{-1} ebenfalls, denn sie ist explizit gegeben durch

$$\begin{aligned} r &= \sqrt{x^2 + y^2}, \\ \varphi &= \operatorname{arccot}\big(\frac{x}{y} + k\pi\big) \qquad \text{für } y \neq 0, \\ \varphi &= \arctan\big(\frac{y}{x}\big) + k\pi \qquad \text{für } x \neq 0, \end{aligned}$$

wo $k \in \{0, 1, 2\}$ jeweils so zu wählen ist, dass $\varphi(1, 0) = 0$ und φ für $(x, y) \neq (a, 0)$, $a > 0$, stetig ist. Die oben mit $\tilde{f}$ bezeichnete Funktion ist

die Komposition $\tilde{f} = f \circ \tau$. Aus ihr gewinnt man f in der Form $f = \tilde{f} \circ \tau^{-1}$. Dabei sind jeweils noch die Definitionsbereiche zu beachten sowie die Tatsache, dass $(0,0)$ nicht im Bild von τ liegt.

Allgemein bezeichnet man als *Koordinatentransformation* eine Abbildung $F : M \to \mathbb{R}^n$ mit $M \subseteq \mathbb{R}^n$ (M offen) mit den Eigenschaften: F ist injektiv, F und die Umkehrabbildung F^{-1} (definiert auf $F(M)$) sind differenzierbar. Sei F eine Abbildung mit diesen Eigenschaften. Setzen wir $X = (x_1, \ldots, x_n)$, $F(X) = Y = (y_1, \ldots, y_n)$, so ist also

$$\left.\begin{aligned} y_1 &= f_1(x_1, \ldots, x_n), \\ &\vdots \\ y_n &= f_n(x_1, \ldots, x_n), \end{aligned}\right\} \tag{4.1}$$

wo $f_1, \ldots, f_n$ die Koordinatenfunktionen von F sind. Sind $g_1, \ldots, g_n$ die Koordinatenfunktionen der Umkehrabbildung $F^{-1} : F(M) \to M$, so ist also (4.1) äquivalent mit

$$\left.\begin{aligned} x_1 &= g_1(y_1, \ldots, y_n), \\ &\vdots \\ x_n &= g_n(y_1, \ldots, y_n). \end{aligned}\right\} \tag{4.2}$$

Im allgemeinen wird es nicht möglich sein, bei explizit gegebenen Funktionen $f_1, \ldots, f_n$ die Funktionen $g_1, \ldots, g_n$ explizit zu berechnen (also das Gleichungssystem (4.1) „aufzulösen nach $x_1, \ldots, x_n$“). Häufig benötigt man jedoch nur die partiellen Ableitungen der Funktionen $g_1, \ldots, g_n$, und diese kann man stets explizit berechnen aus den partiellen Ableitungen der Funktionen $f_1, \ldots, f_n$. Dies ergibt sich in folgender Weise aus bekannten Resultaten.

Wir setzen voraus, dass die Abbildung $F : M \to \mathbb{R}$ differenzierbar ist in einem Punkt $X \in M$, dass die Umkehrabbildung F^{-1} existiert und differenzierbar ist im Punkt $Y := F(X)$. Da die Abbildung $F^{-1} \circ F$ auf M die Identität ist, gilt nach der Kettenregel

$$D(F^{-1})_Y \circ DF_X = \text{Identität},$$

oder in Koordinatenschreibweise

$$JF^{-1}(Y)JF(X) = \text{Einheitsmatrix}.$$

Es folgt, dass das Differential DF_X und damit die Funktionalmatrix $JF(X)$ regulär sind und dass

$$JF^{-1}(Y) = JF(X)^{-1}$$

ist. Ausgeschrieben lautet diese Gleichung

$$\begin{pmatrix} \partial_1 g_1(Y) & \cdots & \partial_n g_1(Y) \\ \vdots & & \vdots \\ \partial_1 g_n(Y) & \cdots & \partial_n g_n(Y) \end{pmatrix} = \begin{pmatrix} \partial_1 f_1(X) & \cdots & \partial_n f_1(X) \\ \vdots & & \vdots \\ \partial_1 f_n(X) & \cdots & \partial_n f_n(X) \end{pmatrix}^{-1}.$$

Nach bekannten Regeln für die Berechnung einer inversen Matrix ergibt sich hieraus für jede Funktion $\partial_j g_i$ eine Darstellung als Quotient, wo Zähler und Nenner Polynome in den Funktionen $\partial_k f_m \circ F^{-1}$ $(k, m = 1, \ldots, n)$ sind und der Nenner nicht verschwindet. Durch weitere Differentiationen (nach der Kettenregel) ergibt sich dann: Sind $f_1, \ldots, f_n$ Funktionen der Klasse C^r, so sind $g_1, \ldots, g_n$ von der Klasse C^r.

Von besonderer Wichtigkeit ist nun naheliegenderweise die Frage, ob und gegebenenfalls wie man einer gegebenen differenzierbaren Abbildung $F : M \to \mathbb{R}^n$ (mit $M \subseteq \mathbb{R}^n$) ansehen kann, ob sie eine differenzierbare Umkehrabbildung besitzt, also eine Koordinatentransformation definiert. Notwendig ist, wie wir eben gesehen haben, jedenfalls die Regularität (Umkehrbarkeit) des Differentials, also das Nichtverschwinden der Funktionaldeterminante. Für eine differenzierbare Funktion $f : (a, b) \to \mathbb{R}$ ist die Bedingung $f'(t) \neq 0$ für $t \in (a, b)$ bekanntlich auch hinreichend für die Existenz und Differenzierbarkeit der Umkehrfunktion, aber im Fall $n > 1$ ist die Sachlage nicht so einfach.

Beispiel. Sei $M = \mathbb{R}^2 \setminus \{0\}$ und $F : M \to \mathbb{R}^2$ definiert durch

$$F(x, y) := (x^2 - y^2, 2xy) \qquad \text{für } (x, y) \in \mathbb{R}^2.$$

Die Funktionalmatrix

$$JF(x, y) = \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix}$$

hat die Determinante $4(x^2 + y^2) \neq 0$, ist also regulär. Wegen $F(X) = F(-X)$ ist die Abbildung F aber nicht umkehrbar.

Die Regularität des Differentials an einer gegebenen Stelle reicht jedoch aus, um zumindest in einer hinreichend kleinen Umgebung dieser Stelle die Existenz und Differenzierbarkeit der Umkehrabbildung zu sichern. Diesen wichtigen Satz wollen wir nun beweisen.

Die Abbildung $F : M \to \mathbb{R}^n$ heißt *von der Klasse* C^r, wenn alle Koordinatenfunktionen von F r-mal stetig differenzierbar sind. Wir stellen zunächst einen Hilfssatz bereit.

4.3 Lemma. *Sei $M \subseteq \mathbb{R}^n$ offen, sei $F : M \to \mathbb{R}^n$ eine Abbildung der Klasse C^1, sei $L : \mathbb{R}^n \to \mathbb{R}^n$ eine lineare Abbildung. Sind $X, Y \in M$ Punkte mit $[X, Y] \subseteq M$, so gilt*

$$\|F(X) - F(Y) - L(X - Y)\| \leq \|X - Y\| \max_{Z \in [X,Y]} \|DF_Z - L\|.$$

Beweis. Setze $G(X) := F(X) - L(X)$ für $X \in M$; dann ist $DG_Z = DF_Z - L$ für $Z \in M$. Da F von der Klasse C^1 ist, ist die Funktion $Z \mapsto \|DF_Z - L\|$ stetig, daher existiert

$$\max_{Z \in [X,Y]} \|DF_Z - L\| =: c.$$

Aus dem Mittelwertsatz (Satz 1.8) folgt jetzt

$$\|G(X) - G(Y)\| \leq c\|X - Y\|,$$

also die Behauptung. ∎

4.4 Satz (über die Umkehrabbildung)**.** *Sei $M \subseteq \mathbb{R}^n$ offen, $X_0 \in M$, sei $F : M \to \mathbb{R}^n$ eine Abbildung der Klasse C^r (für ein $r \in \mathbb{N}$). Das Differential DF_{X_0} von F in X_0 sei regulär (d.h. die Funktionaldeterminante von F in X_0 sei $\neq 0$). Dann gibt es eine offene Umgebung $U \subseteq M$ von X_0 mit folgenden Eigenschaften:*

(a) die Einschränkung $F|_U$ ist injektiv,
(b) die Bildmenge $V := F(U)$ ist offen,
(c) die Umkehrabbildung $(F|_U)^{-1} : V \to U$ ist von der Klasse C^r.

Beweis. Im folgenden bezeichnet I die identische Abbildung von $\mathbb{R}^n$ auf sich. Zur Abkürzung wird für $\alpha > 0$

$$U_\alpha := U(0, \alpha) = \{X \in \mathbb{R}^n \mid \|X\| < \alpha\}$$

gesetzt. Wir machen zunächst eine spezielle Voraussetzung.

Voraussetzung. $X_0 = 0$, $F(0) = 0$, $DF_0 = I$.

Nach Voraussetzung ist die Funktion $X \mapsto \|DF_X - I\|$ in M stetig, und es ist $\|DF_0 - I\| = 0$. Daher existiert zu vorgegebenem $\varepsilon > 0$ ein $\alpha > 0$ mit $U_\alpha \subseteq M$ und

$$\|DF_X - I\| \leq \varepsilon \qquad \text{für alle } X \in U_\alpha.$$

Aus Lemma 4.3 folgt dann für alle $X, Y \in U_\alpha$

$$\|F(X) - F(Y) - (X - Y)\| \leq \varepsilon \|X - Y\| \tag{4.5}$$

und hieraus durch Anwendung der Dreiecksungleichung

$$(1 - \varepsilon)\|X - Y\| \leq \|F(X) - F(Y)\|. \tag{4.6}$$

Wir wählen nun ein positives $\varepsilon < 1$ und dazu $\alpha > 0$ wie oben und so, dass $\overline{U_\alpha} \subseteq M$ ist.

Behauptung (1). Es gilt $U_{(1-\varepsilon)\alpha} \subseteq F(U_\alpha)$.

Beweis. Sei $Y \in U_{(1-\varepsilon)\alpha}$. Es ist ein $X \in U_\alpha$ zu finden mit $Y = F(X)$. Hierzu benutzen wir den Banachschen Fixpunktsatz. Definiere $\Phi : \overline{U_\alpha} \to \mathbb{R}^n$ durch

$$\Phi(X) := Y - F(X) + X \qquad \text{für } X \in \overline{U_\alpha}.$$

Für $X \in \overline{U_\alpha}$ gilt wegen $F(0) = 0$ nach (4.5)

$$\|\Phi(X)\| \leq \|Y\| + \|F(X) - X\| \leq \|Y\| + \varepsilon\|X\| < (1 - \varepsilon)\alpha + \varepsilon\alpha = \alpha,$$

also $\Phi(X) \in U_\alpha$. Somit bildet Φ die Menge $\overline{U_\alpha}$ in sich ab. Für $X, Z \in \overline{U_\alpha}$ gilt nach (4.5)

$$\|\Phi(X) - \Phi(Z)\| = \|F(X) - F(Z) - (X - Z)\| \leq \varepsilon\|X - Z\|,$$

wegen $\varepsilon < 1$ ist also Φ kontrahierend. Da $\overline{U_\alpha}$ nach Satz 1.10 aus Kapitel 10 und Satz 2.9 aus Kapitel 9 vollständig ist, gibt es nach dem Banachschen Fixpunktsatz (Satz 3.2 aus Kapitel 9) einen Punkt $X \in \overline{U_\alpha}$ mit $\Phi(X) = X$. Wie oben gezeigt, ist $X = \Phi(X) \in U_\alpha$. Es ist also ein Punkt $X \in U_\alpha$ gefunden mit $Y = F(X)$. Damit ist die 1. Behauptung bewiesen. ∎

Nun setzen wir $V := U_{(1-\varepsilon)\alpha}$ und $U := F^{-1}(V)$. Dann ist U eine offene Umgebung von 0 und $F(U) = V$. Aus (4.6) folgt, dass $F|_U$ injektiv ist. Für den Moment bezeichne $G : V \to U$ die Umkehrabbildung von $F|_U$.

Behauptung (2). G ist in 0 differenzierbar.

Beweis. Sei $\varepsilon' \in \mathbb{R}^+$ gegeben. Wie anfangs gezeigt ((4.5) mit $Y = 0$) existiert ein $\alpha' \in \mathbb{R}^+$ mit $U_{\alpha'} \subseteq M$ und

$$\|F(X) - X\| \leq \frac{\varepsilon'}{1 + \varepsilon'}\|X\| \qquad \text{für } \|X\| < \alpha'$$

und daher

$$\|X\| \leq (1+\varepsilon')\|F(X)\| \qquad \text{für } \|X\| < \alpha'.$$

Sei $H \in V$ ein Vektor mit $\|H\| < \alpha'(1-\varepsilon)$. Für $X := G(H)$ ist dann $X \in U$, also $\|X\| < \alpha$ und daher nach (4.6)

$$\|X\| \leq \frac{1}{1-\varepsilon}\|F(X)\| = \frac{1}{1-\varepsilon}\|H\| < \alpha',$$

also

$$\|G(H) - H\| = \|X - F(X)\| \leq \frac{\varepsilon'}{1+\varepsilon'}\|X\| \leq \varepsilon'\|F(X)\| = \varepsilon'\|H\|,$$

somit

$$\frac{\|G(H) - H\|}{\|H\|} \leq \varepsilon' \quad \text{für } 0 < \|H\| < \alpha'(1-\varepsilon).$$

Damit ist

$$\lim_{H \to 0} \frac{G(H) - H}{\|H\|} = 0$$

gezeigt, also (wegen $G(0) = 0$) die Differenzierbarkeit von G in 0 sowie $DG_0 = I$. ∎

Wir lassen nun die spezielle Voraussetzung fallen, dass $X_0 = 0$, $F(0) = 0$ und $DF_0 = I$ sein soll.

Es ist also jetzt vorausgesetzt, dass $X_0 \in M$ ein Punkt ist derart, dass DF_{X_0} regulär ist. Durch passende Transformationen gewinnen wir eine Abbildung $\tilde{F}$, die die spezielle Voraussetzung erfüllt. Hierzu bezeichne $T_Z : \mathbb{R}^n \to \mathbb{R}^n$ die Translation um den Vektor Z, also $T_Z(X) := X + Z$ für $X \in \mathbb{R}^n$. Wir setzen

$$\begin{aligned} L &:= DF_{X_0}, \\ \tilde{M} &:= (L \circ T_{-X_0})(M) \\ \tilde{F}(X) &:= T_{-F(X_0)} \circ F \circ T_{X_0} \circ L^{-1}(X) \qquad \text{für } X \in \tilde{M}. \end{aligned}$$

Wie man leicht nachprüft, gilt $0 \in \tilde{M}$, $\tilde{F}(0) = 0$ und $D\tilde{F}_0 = I$. Anwendung des bereits Bewiesenen auf $\tilde{F}$ ergibt dann die lokale Umkehrbarkeit und Differenzierbarkeit von F. Hierzu beachte man, dass

$$F = T_{F(X_0)} \circ \tilde{F} \circ L \circ T_{-X_0}$$

ist. Die Differenzierbarkeit der Umkehrabbildung ergibt sich zunächst nur im Punkt $F(X_0)$. Nun ist aber in X_0 die Funktionaldeterminante von F verschieden von Null, und das gilt, da sie stetig ist, noch in einer ganzen Umgebung

von X_0. Auf jeden Punkt dieser Umgebung kann man das bereits Bewiesene anwenden, erhält also die Differenzierbarkeit der Umkehrabbildung. dass mit F auch die lokale Umkehrung von der Klasse C^r ist, haben wir bereits vor Lemma 4.3 begründet. ∎

Eine injektive Abbildung $F : M \to \mathbb{R}^n$ (mit offenem $M \subseteq \mathbb{R}^n$) der Klasse C^r (für ein $r \in \mathbb{N}$) mit überall regulärem Differential nennt man auch einen *Diffeomorphismus* der Klasse C^r von M auf $F(M)$. Ist nun $F : M \to \mathbb{R}^n$ eine Abbildung der Klasse C^r und DF_{X_0} regulär, so gibt es nach Satz 4.4 eine offene Umgebung U von X_0 derart, dass $F|_U$ ein Diffeomorphismus der Klasse C^r von U auf $F(U)$ ist. Für diesen Sachverhalt sagt man auch, F sei bei X_0 lokal ein C^r-Diffeomorphismus.

Als eine Folgerung aus dem Satz über die Umkehrabbildung wollen wir nun eine Aussage über sogenannte „implizite Funktionen“ gewinnen. Dieses für manche Anwendungen wichtige Thema wollen wir zunächst anschaulich erläutern.

Verschiedenartige Aufgabenstellungen erfordern die Untersuchung von Gebilden im euklidischen Raum, wie „Kurven“ oder „Flächen“, die beschrieben werden durch Gleichungen, die zwischen den Koordinaten ihrer Punkte bestehen. Als einfaches Beispiel betrachten wir im $\mathbb{R}^2$ die „Einheitssphäre“

$$S^2 := \{X \in \mathbb{R}^2 \mid x_1^2 + x_2^2 - 1 = 0\}.$$

Die genauere Untersuchung dieser Punktmenge kann dadurch erleichtert werden, dass man die definierende Gleichung $x_1^2 + x_2^2 = 1$ nach einer Variablen auflöst, indem man etwa

$$x_2 = \sqrt{1 - x_1^2}$$

schreibt. Man betrachtet dann die Punktmenge

$$S_+^1 := \{X \in \mathbb{R}^2 \mid x_2 = \sqrt{1 - x_1^2},\ x_1^2 < 1\}.$$

Allerdings ist S_+^1 nur ein echter Teil von S^1, die „obere Halbsphäre“. Offenbar kann die ganze Menge S^1 nicht einheitlich dadurch dargestellt werden, dass man eine Koordinate als Funktion der beiden anderen schreibt. Aber zur lokalen Untersuchung von S^1 ist eine derartige Darstellung gut geeignet. Es stört nicht, dass man zur Untersuchung von S^1 nicht mit nur einer solchen Darstellung auskommt. Will man S^1 in einer Umgebung des Punktes $(0, -1)$ untersuchen, so benutzt man die „Auflösung“ $x_2 = -\sqrt{1 - x_1^2}$, in einer Umgebung von $(1, 0)$ benutzt man die Auflösung $x_1 = \sqrt{1 - x_2^2}$, usw.

Analog kann man Punktmengen untersuchen, die durch mehrere Gleichungen zwischen den Koordinaten definiert sind. Als Beispiel betrachten wir die Punktmenge

$$K := \left\{ X \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 - 1 = 0 \text{ und } \left(x_1 - \frac{1}{2}\right)^2 + x_2^2 - \frac{1}{4} = 0 \right\},$$

also den Durchschnitt der Sphäre $S^2 \subseteq \mathbb{R}^3$ mit einem gewissen Kreiszylinder mit Erzeugenden parallel zur x_3-Achse. Eine „lokale Auflösung“ lautet

$$x_1 = 1 - x_3^2, \;\; x_2 = x_3\sqrt{1 - x_3^2}, \;\; |x_3| < 1,$$

eine andere

$$x_1 = 1 - x_3^2, \;\; x_2 = -x_3\sqrt{1 - x_3^2}, \;\; |x_3| < 1.$$

Diese Darstellungen können zur lokalen Untersuchung der Schnittkurve K benutzt werden.

Im allgemeinen wird es nicht möglich sein, eine „lokale Auflösung“ explizit (formelmäßig) zu bewerkstelligen. Hier sind dann Aussagen von Interesse, die die grundsätzliche Möglichkeit einer solchen Auflösung (und die Differenzierbarkeit der darstellenden Funktionen) gewährleisten. Zunächst soll aber die Fragestellung allgemeiner und genauer gefaßt werden.

Eine Punktmenge N des $\mathbb{R}^n$ sei definiert und beschrieben durch ein Gleichungssystem

$$\left.\begin{aligned} f_1(x_1, \ldots, x_n) &= 0, \\ f_2(x_1, \ldots, x_n) &= 0, \\ &\vdots \\ f_k(x_1, \ldots, x_n) &= 0. \end{aligned}\right\} \tag{4.7}$$

Dabei sei $M \subseteq \mathbb{R}^n$ eine offene Teilmenge, und

$$f_i : M \to \mathbb{R}, \qquad i = 1, \ldots, k < n,$$

seien reellwertige Funktionen der Klasse C^1. Sodann sei

$$N := \{X \in M \mid X = (x_1, \ldots, x_n) \;\; \text{erfüllt (4.7)}\}.$$

(Der Buchstabe N soll an „Nullstellenmenge“ erinnern.) Das Problem besteht nun darin, die Punktmenge N in einer Umgebung eines gegebenen Punktes $X_0 \in N$ zu beschreiben, indem k Koordinaten als Funktionen der übrigen $n - k$ Koordinaten dargestellt werden, etwa

$$\begin{aligned} x_{n-k+1} &= g_1(x_1, \ldots, x_{n-k}), \\ &\vdots \\ x_n &= g_k(x_1, \ldots, x_{n-k}) \end{aligned}$$

für $(x_1, \ldots, x_{n-k})$ aus einer passenden Menge $V \subseteq \mathbb{R}^{n-k}$. Das ist so zu verstehen: Setzt man die so berechneten Koordinaten in (4.7) ein, so soll dieses System erfüllt sein, also

$$f_i(x_1, \ldots, x_{n-k}, g_1(x_1, \ldots, x_{n-k}), \ldots, g_k(x_1, \ldots, x_{n-k})) = 0$$

für $(x_1, \ldots, x_{n-k}) \in V$ und $i = 1, \ldots, k$. Genauer gesagt: Jeder Punkt

$$(x_1, \ldots, x_{n-k}, g_1(x_1, \ldots, x_{n-k}), \ldots, g_k(x_1, \ldots, x_{n-k}))$$

mit $(x_1, \ldots, x_{n-k}) \in V$ soll zu N gehören; umgekehrt soll aber auch der Durchschnitt von N mit einer ganzen Umgebung von X_0 auf diese Weise erhalten werden. Man sagt dann, die Funktionen $g_1, \ldots, g_k$ seien „implizit definiert“ durch die Gleichungen

$$f_1(X) = 0, \ldots, f_k(X) = 0.$$

Nützlich ist die hierdurch gegebene lokale Auflösung im Algemeinen nur, wenn die Funktionen $g_1, \ldots, g_k$ ihrerseits so oft differenzierbar sind wie die Funktionen $f_1, \ldots, f_k$. Für die Möglichkeit einer solchen Auflösung gibt Satz 4.8 eine hinreichende Bedingung.

Es ist im folgenden zweckmäßig, die Funktionen $f_1, \ldots, f_k$ als Koordinatenfunktionen einer Abbildung $F : M \to \mathbb{R}^k$ und ebenso die Funktionen $g_1, \ldots, g_k$ als Koordinatenfunktionen einer Abbildung $G : V \to \mathbb{R}^k$ aufzufassen. Unter dem *Graphen* von G versteht man die Menge

$$\text{Graph } G := \{(X, G(X)) \in \mathbb{R}^{n-k} \times \mathbb{R}^k \mid X \in V\}.$$

Der zu beweisende Satz läßt sich dann so formulieren, dass man eine Nullstellenmenge unter passenden Voraussetzungen lokal als Graph darstellen kann.

4.8 Satz (über implizit definierte Funktionen)**.** *Sei $k < n$, $M \subseteq \mathbb{R}^n$ offen, $F : M \to \mathbb{R}^k$ eine Abbildung der Klasse C^r (für ein $r \in \mathbb{N}$), sei*

$$N := \{X \in M \mid F(X) = 0\}.$$

Sei $X_0 \in N$ und DF_{X_0} vom Rang k. Dann gibt es (nach passender Identifizierung von $\mathbb{R}^n$ mit $\mathbb{R}^{n-k} \times \mathbb{R}^k$) eine offene Umgebung U von X_0 in M, eine offene Menge V in $\mathbb{R}^{n-k}$ und eine Abbildung $G : V \to \mathbb{R}^k$ der Klasse C^r mit

$$N \cap U = \text{ Graph } G.$$

Beweis. Das Differential DF_{X_0} ist nach Voraussetzung vom Rang k. Die Funktionalmatrix von F in X_0 hat also k linear unabhängige Spalten. Wir dürfen o.B.d.A. (d.h. nach passender Umnumerierung der Basisvektoren) annehmen, dass dies die letzten k Spalten sind. Wir identifizieren dann $\mathbb{R}^{n-k}$

mit dem von den ersten $n-k$ Basisvektoren und $\mathbb{R}^k$ mit dem von den letzten k Basisvektoren des $\mathbb{R}^n$ aufgespannten Unterraum, und wir identifizieren $\mathbb{R}^n$ mit $\mathbb{R}^{n-k} \times \mathbb{R}^k$. Jeder Punkt X aus $\mathbb{R}^n$ läßt sich also eindeutig in der Form (X', X'') mit $X' \in \mathbb{R}^{n-k}$ und $X'' \in \mathbb{R}^k$ schreiben. Wir definieren die Abbildung

$$\Phi : M \to \mathbb{R}^{n-k} \times \mathbb{R}^k : (X', X'') \mapsto (X', F(X)).$$

Dann ist Φ eine Abbildung der Klasse C^r. Wegen

$$J\Phi = \begin{pmatrix} 1 & 0 \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 \cdots & 1 & 0 & \cdots & 0 \\ \partial_1 f_1 & \cdots & \partial_{n-k} f_1 & \partial_{n-k+1} f_1 & \cdots & \partial_n f_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \partial_1 f_k & \cdots & \partial_{n-k} f_k & \partial_{n-k+1} f_k & \cdots & \partial_n f_k \end{pmatrix}$$

ist $J\Phi$ an der Stelle X_0 vom Rang n. Nach Satz 4.4 ist daher Φ in einer Umgebung U_0 von X_0 ein Diffeomorphismus der Klasse C^r. Sei Ψ seine Umkehrabbildung. Ψ ist definiert auf einer offenen Umgebung W von $\Phi(X_0) = (X_0', F(X_0)) = (X_0', 0)$. Wir setzen

$$V := \{X' \in \mathbb{R}^{n-k} \mid (X', 0) \in W\},$$

dann ist $V \subseteq \mathbb{R}^{n-k}$ eine offene Umgebung von X_0'. Wegen $\Phi(X', X'') = (X', F(X', X''))$ ist

$$\Psi(X', Z) = (X', \Psi_2(X', Z)),$$

wodurch eine Abbildung $\Psi_2 : W \to \mathbb{R}^k$ definiert wird. Wir setzen nun

$$G(X') := \Psi_2(X', 0) \qquad \text{für } X' \in V.$$

Dann ist $G : V \subseteq \mathbb{R}^{n-k} \to \mathbb{R}^k$ eine Abbildung der Klasse C^r. Anwendung von Ψ auf die Gleichung

$$\Phi(X_0', X_0'') = (X_0', F(X_0)) = (X_0', 0)$$

ergibt

$$(X_0', X_0'') = \Psi(X_0', 0) = (X_0', \Psi_2(X', 0)) = (X_0', G(X_0')),$$

also $G(X_0') = X_0''$. Für $X' \in V$ gilt $\Psi(X', 0) = (X', G(X'))$ (nach Definition von Ψ_2 und G). Anwendung von Φ ergibt

$$(X', 0) = \Phi(X', G(X')) = (X', F(X', G(X'))),$$

also $F(X', G(X')) = 0$ und damit

$$(X', G(X')) \in N.$$

Wir können o.B.d.A. $V \times G(V) \subseteq U_0$ annehmen, denn dies läßt sich durch Verkleinerung von V erreichen, da G in X_0' stetig und $G(X_0') = X_0''$ ist. Setze $U := (V \times \mathbb{R}^k) \cap U_0$. Dann gilt

$$N \cap U = \text{ Graph } G.$$

Zum Beweis sei $(X', Y) \in$ Graph G, also $X' \in V$ und $Y = G(X')$. Dann ist $(X', Y) = (X', G(X')) \in V \times G(V) \subseteq U$ und, wie eben gezeigt, $(X', G(X')) \in N$.

Ist umgekehrt $(X', Y) \in N \cap U$, so ist

$$(X', F(X', Y)) = (X', 0) = (X', F(X', G(X'))),$$

also nach Definition von Φ

$$\Phi(X', Y) = \Phi(X', G(X')).$$

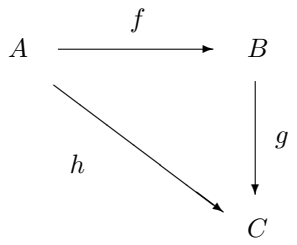
Da Φ injektiv ist, folgt $Y = G(X')$, also $(X', Y) \in$ Graph G. ∎

Satz 4.4 gibt die Möglichkeit, auch für $k \neq n$ differenzierbare Abbildungen $F : M \to \mathbb{R}^k$ mit nichtausgeartetem Differential in einer Weise zu beschreiben, die zeigt, dass solche Abbildungen sich lokal qualitativ so wie ihr Differential verhalten. Für $k < n$ ist diese Beschreibung im wesentlichen schon in Satz 4.8 enthalten; für $k > n$ wird sie auf analoge Weise erhalten.

4.9 Definition. *Sei $F : M \subseteq \mathbb{R}^n \to \mathbb{R}^k$ differenzierbar in $X_0 \in M$. Dann wird der Rang des Differentials DF_{X_0} (= Rang der Funktionalmatrix $JF(X_0)$) als* Rang *der Abbildung F in X_0 bezeichnet.*

Sei nun $M \subseteq \mathbb{R}^n$ offen und $F : M \to \mathbb{R}^k$ eine Abbildung der Klasse C^1. Der Rang von F in $X \in M$ kann dann höchstens gleich $\min\{n, k\}$ sein. Ist er gleich dieser Zahl, so sagt man, F sei in X *von maximalem Rang.* Eine Teilaussage von Satz 4.4 kann man auch so formulieren: Im Fall $k = n$ ist jede C^1-Abbildung, die in einem Punkt von maximalem Rang ist, in einer Umgebung dieses Punktes ein Diffeomorphismus. Wir wollen analoge Aussagen für $k \neq n$ gewinnen. Sie besagen, dass man differenzierbare Abbildungen maximalen Ranges lokal in sehr übersichtlicher Weise durch Diffeomorphismen und lineare Abbildungen beschreiben kann.

Bemerkung. Sind $f : A \to B$, $g : B \to C$, $h : A \to B$ (A, B, C Mengen) Abbildungen mit $g \circ f = h$, so drückt man dies auch aus durch die Sprechweise: „Das Diagramm



ist kommutativ.“ [Anschaulich: Man kommt, ausgehend von einem $a \in A$, zu demselben Element von C, gleichgültig auf welchem Wege man den Pfeilen folgt.]

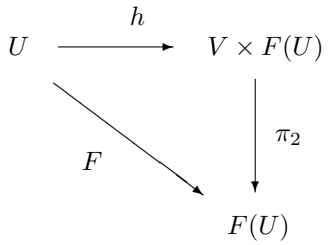
Bemerkung. Sind A, B Mengen, so sind auf dem kartesischen Produkt $A \times B$ die *Projektionen* π_1, π_2 definiert durch

$$\pi_1 : A \times B \to A : (a, b) \mapsto a$$
$$\pi_2 : A \times B \to B : (a, b) \mapsto b.$$

Sind A, B Vektorräume, so wird die *kanonische Injektion* i von A in $A \times B$ definiert durch

$$i : A \to A \times B$$
$$a \mapsto (a, 0).$$

4.10 Satz (über lokal surjektive Abbildungen)**.** *Sei $k < n$. Sei $M \subseteq \mathbb{R}^n$ offen, sei $F : M \to \mathbb{R}^k$ eine Abbildung der Klasse C^r ($r \geq 1$). Sei $X_0 \in M$ und F in X_0 vom Rang k (also DF_{X_0} surjektiv). Dann gibt es eine offene Umgebung U von X_0 in M, eine offene Menge V in $\mathbb{R}^{n-k}$ und einen C^r-Diffeomorphismus $h : U \to V \times F(U)$, so dass das Diagramm*



kommutativ ist.

Bemerkung. Die Abbildung F ist also in einer Umgebung des Punktes X_0, in dem das Differential surjektiv ist, dargestellt als Komposition eines C^r-Diffeomorphismus und einer surjektiven linearen Abbildung.

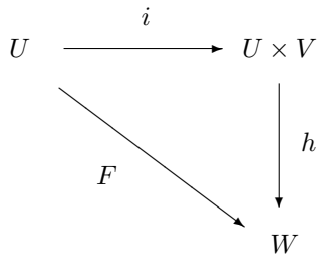
von Satz 4.10. Nach Voraussetzung hat die Funktionaldeterminante $JF(X_0)$ k linear unabhängige Spalten. Wir können o.B.d.A. annehmen, dass dies die letzten k Spalten sind. Wir fassen jetzt wieder $\mathbb{R}^n$ als das kartesische Produkt $\mathbb{R}^{n-k} \times \mathbb{R}^k$ auf und bezeichnen mit π_1, π_2 die zugehörigen Projektionen. Aus Satz dem Beweis von Satz 4.8 folgt, dass die Abbildung

$$\varphi : M \to \mathbb{R}^{n-k} \times \mathbb{R}^k : X \mapsto (\pi_1(X), F(X))$$

auf einer offenen Umgebung U' von X_0 ein C^r-Diffeomorphismus ist. Das Bild $\varphi(U')$ enthält eine offene Umgebung von $\varphi(X_0) = (\pi_1(X_0), F(X_0))$ der Form $V \times W$, dabei ist also V offen in $\mathbb{R}^{n-k}$. Setze $U := \varphi^{-1}(V \times W)$ und $h := \varphi|U$. Dann ist $F(U) = W$, und für $X \in U$ gilt $h(X) = (\pi_1(X), F(X))$, also $\pi_2 \circ h = F$. ∎

Der folgende Satz ist das Gegenstück zu Satz 4.10 für den Fall $k > n$.

4.11 Satz (über lokal injektive Abbildungen)**.** *Sei $k > n$. Sei $M \subseteq \mathbb{R}^n$ offen, sei $F : M \to \mathbb{R}^k$ eine Abbildung der Klasse C^r $(r \geq 1)$. Sei $X_0 \in M$ und F in X_0 vom Rang n (also DF_{X_0} injektiv). Dann gibt es eine offene Umgebung U von X_0 in M, eine offene Umgebung V von 0 in $\mathbb{R}^{k-n}$, eine offene Umgebung W von $F(X_0)$ in $\mathbb{R}^k$ und einen C^r-Diffeomorphismus $h : U \times V \to W$, so dass das Diagramm*



(wo i die kanonische Injektion bezeichnet) kommutativ ist.

Bemerkung. Die Abbildung F ist also in einer Umgebung des Punktes X_0, in dem das Differential injektiv ist, dargestellt als Komposition einer injektiven linearen Abbildung und eines C^r-Diffeomorphismus.

Beweis (Satz 4.11). Nach Voraussetzung hat die Funktionalmatrix $JF(X_0)$ n linear unabhängige Zeilen. Wir können o.B.d.A. annehmen, dass dies die ersten n Zeilen sind. Wir fassen den Raum $\mathbb{R}^k$ als das kartesische Produkt $\mathbb{R}^n \times \mathbb{R}^{k-n}$ auf und definieren

$$\psi : M \times \mathbb{R}^{k-n} \to \mathbb{R}^k : (X, Y) \mapsto F(X) + (0, Y).$$

Dann ist ψ von der Klasse C^r, und die Funktionalmatrix von ψ ist von der Form

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

wo A aus den ersten n Zeilen der Funktionalmatrix von F gebildet ist und E die $(k-n)$-reihige Einheitsmatrix ist. Im Punkt $(X_0, 0)$ ist also die Funktionaldeterminante von ψ ungleich Null. Nach Satz 4.4 gibt es daher offene Umgebungen $U \times V$ (o.B.d.A. in dieser Produktform) von $(X_0, 0)$ und W von $\psi(X_0, 0) = F(X_0)$, so dass $h := \psi|_{U \times V}$ ein C^r-Diffeomorphismus von $U \times V$ auf W ist. Für $X \in U$ gilt $h \circ i(X) = h(X, 0) = F(X)$. ∎

Es ist nun Gelegenheit, die Untersuchung von Extremwerten reeller Funktionen fortzuführen. Wir wollen eine notwendige Bedingung für das Vorliegen eines lokalen Extremums unter Nebenbedingungen aufstellen, die sogenannte *Multiplikatorenregel von Lagrange*. Zunächst betrachten wir als Beispiel eine typische Aufgabe dieser Art. Es seien die Extrema der Funktion $f(x, y) = x + y$ zu finden unter der Nebenbedingung $5x^2 + 6xy + 2y^2 = 1$. Mit anderen Worten, es sei

$$N := \{(x, y) \in \mathbb{R}^2 \mid 5x^2 + 6xy + 2y^2 = 1\},$$

und es sollen die Extrema der Einschränkung $f|_N$ gefunden werden. In diesem einfachen Fall kann man natürlich so vorgehen, dass man die Gleichung $5x^2 + 6xy + 2y^2 = 1$ nach x oder y auflöst, zum Beispiel

$$y = -\frac{3}{2}x \pm \sqrt{\frac{1}{2} - \frac{1}{4}x^2}.$$

Wir setzen daher

$$N^{\pm} := \left\{(x, y) \in \mathbb{R}^2 \mid |x| < \sqrt{2},\ y = -\frac{3}{2}x \pm \sqrt{\frac{1}{2} - \frac{1}{4}x^2}\right\},$$

$$g^{\pm}(x) := x + \left(-\frac{3}{2}x \pm \sqrt{\frac{1}{2} - \frac{1}{4}x^2}\right) \quad \text{für } |x| < \sqrt{2}.$$

Für $(x, y) \in N^+$ gilt dann $f(x, y) = g^+(x)$. Für $(x, y) \in N^+$ hat also f in (x, y) genau dann ein lokales Extremum, wenn g^+ in x ein lokales Extremum

hat. Nun ist $(g^+)'(x) = 0$ nur für $x = -1$. Hat also f in N^+ ein lokales Extremum, so im Punkt $(-1, 2)$. Analog gilt: Hat f in N^- ein lokales Extremum, so im Punkt $(1, -2)$. Man muß allerdings noch nachweisen, z.B. durch Auflösung nach x, dass f in den nicht erfaßten Punkten mit $|x| = \sqrt{2}$ keine lokalen Extrema hat. Da f ein Maximum und ein Minimum annimmt, hat f also genau in den beiden Punkten $(-1, 2)$ und $(1, -2)$ lokale Extrema.

Im allgemeinen wird eine solche explizite Auflösung nicht möglich sein. Sie ist aber auch nicht erforderlich, da man folgendermaßen argumentieren kann. Wir können (wie später allgemeiner gezeigt und verwendet wird), die Menge N lokal parametrisieren, das heißt die Punkte $(x, y) \in N$ darstellen in der Form $x = \varphi(t)$, $y = \psi(t)$ mit stetig differenzierbaren Funktionen φ, ψ. Dann setzen wir $h(t) = f(\varphi(t), \psi(t))$ und bestimmen die kritischen Stellen von h. Eine notwendige Bedingung ist $h'(t) = 0$, also

$$\partial_1 f(\varphi(t), \psi(t))\varphi'(t) + \partial_2 f(\varphi(t), \psi(t))\psi'(t) = 0.$$

Ist $g(x, y) = 0$ die Nebenbedingung, so gilt $g(\varphi(t), \psi(t)) = 0$, also

$$\partial_1 g(\varphi(t), \psi(t))\varphi'(t) + \partial_2 g(\varphi(t), \psi(t))\psi'(t) = 0.$$

An einer kritischen Stelle von h sind also, wenn dort nicht gerade $\varphi' = \psi' = 0$ ist, die Spaltenvektoren $(\partial_1 f, \partial_1 g)$ und $(\partial_2 f, \partial_2 g)$ linear abhängig. Somit sind also auch die Zeilenvektoren $(\partial_1 f, \partial_2 f)$ und $(\partial_1 g, \partial_2 g)$ linear abhängig. Es gibt also, wenn an der kritischen Stelle $(\partial_1 g, \partial_2 g) \neq (0, 0)$ ist, eine Zahl λ mit

$$\begin{aligned} \partial_1 f - \lambda \partial_1 g &= 0, \\ \partial_2 f - \lambda \partial_2 g &= 0. \end{aligned}$$

Zusammen mit $g = 0$ sind das drei Gleichungen, aus denen man im Prinzip die Koordinaten des kritischen Punktes und den Wert von λ bestimmen kann. Dabei ist die Zahl λ, der Multiplikator von Lagrange, in der Regel nicht von Interesse, sondern hat nur eine Hilfsfunktion.

Im obigen Beispiel ist $f(x, y) = x + y$, $g(x, y) = 5x^2 + 6xy + 2y^2 - 1$. Man erhält also die Gleichungen

$$\begin{aligned} 1 - \lambda(10x + 6y) &= 0, \\ 1 - \lambda(6x + 4y) &= 0, \\ 5x^2 + 6xy + 2y^2 - 1 &= 0 \end{aligned}$$

mit den Lösungen $(x, y, \lambda) = (-1, 2, \frac{1}{2})$ und $(1, -2, -\frac{1}{2})$.

Wir wollen dieses Verfahren nun auf mehrere Nebenbedingungen ausdehnen und exakt begründen.

4.12 Satz (Multiplikatorenregel von Lagrange)**.** *Sei $M \subseteq \mathbb{R}^n$ offen, seien $f : M \to \mathbb{R}$ und $F : M \to \mathbb{R}^k$ mit $1 \leq k < n$ Abbildungen der Klasse C^1, sei*

$$N := \{X \in M \mid F(X) = 0\}.$$

Die Einschränkung $f|_N$ habe an der Stelle $X_0 \in N$ ein lokales Extremum, und die Abbildung F mit Koordinatenfunktionen $f_1, \ldots, f_k$ sei an der Stelle X_0 vom Rang k. Dann gibt es reelle Zahlen $\lambda_1, \ldots, \lambda_k$ mit

$$\partial_i \left(f - \sum_{j=1}^{k} \lambda_j f_j \right) (X_0) = 0 \qquad \text{für } i = 1, \ldots, n.$$

Beweis. Nach Satz 4.8 gibt es eine Umgebung $U \subseteq \mathbb{R}^n$ von X_0, eine offene Menge $V \subseteq \mathbb{R}^{n-k}$ und eine injektive Abbildung $\varphi : V \subseteq \mathbb{R}^{n-k} \to \mathbb{R}^n$ der Klasse C^1 vom Rang $n-k$ mit $U \cap N = \varphi(V)$ (nämlich $\varphi(Z) = (Z, G(Z))$ mit G wie in Satz 4.8). Sei Z_0 das Urbild von X_0 unter der Funktion φ. Die Funktion $h = f \circ \varphi$ hat an der Stelle Z_0 ein lokales Extremum. Daher gilt $Dh_{Z_0} = 0$, nach der Kettenregel also

$$Df_{X_0}(D\varphi_{Z_0}(Y)) = 0 \qquad \text{für alle } Y \in \mathbb{R}^{n-k}.$$

Wegen $F \circ \varphi = 0$ gilt auch $DF_{X_0}(D\varphi_{Z_0}(Y)) = 0$, also

$$(Df_i)_{X_0}(D\varphi_{Z_0}(Y)) = 0 \qquad \text{für alle } Y \in \mathbb{R}^{n-k}.$$

Die im folgenden auftretenden Differentiale von f und f_i sind sämtlich an der Stelle X_0 genommen. Wir haben gezeigt, dass

$$Df(H) = 0 \quad \text{und} \quad Df_i(H) = 0 \quad \text{für } i = 1, \ldots, k \text{ und } H \in D\varphi_{Z_0}(\mathbb{R}^{n-k})$$

gilt. Sei $L := D\varphi_{Z_0}(\mathbb{R}^{n-k})$ und $L^\perp$ der Orthogonalraum in $\mathbb{R}^n$. Dann ist $\dim L^\perp = k$. Sei $(W_1, \ldots, W_k)$, $W_i \in \mathbb{R}^n$, eine Basis von $L^\perp$. Die Zeilen der Matrix

$$\begin{pmatrix} Df_1(W_1) & \cdots & Df_k(W_1) \\ \vdots & & \vdots \\ Df_1(W_k) & \cdots & Df_k(W_k) \end{pmatrix}$$

sind die Koordinatenvektoren der Bildvektoren $DF_{X_0}(W_1), \ldots, DF_{X_0}(W_k)$, sind also linear unabhängig. Daher sind auch die Spalten der Matrix linear unabhängig, und es gibt reelle Zahlen $\lambda_1, \ldots, \lambda_k$ mit

$$\begin{pmatrix} Df(W_1) \\ \vdots \\ Df(W_k) \end{pmatrix} = \lambda_1 \begin{pmatrix} Df_1(W_1) \\ \vdots \\ Df_1(W_k) \end{pmatrix} + \cdots + \lambda_k \begin{pmatrix} Df_k(W_1) \\ \vdots \\ Df_k(W_k) \end{pmatrix}.$$

Da jeder Vektor $E \in \mathbb{R}^n$ in der Form $E = H + c_1 W_1 + \cdots + c_k W_k$ mit $H \in L$ und $c_1, \ldots, c_k \in \mathbb{R}$ dargestellt werden kann, folgt

$$Df(E) = \lambda_1 Df_1(E) + \cdots + \lambda_k Df_k(E).$$

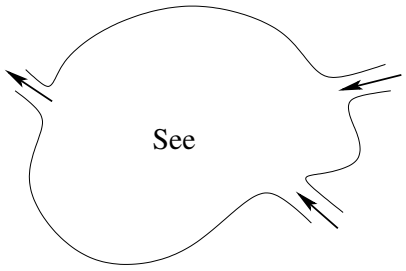
Mit $E = E_i$, $i = 1, \ldots, n$, folgt die Behauptung. ∎

12 Gewöhnliche Differentialgleichungen

12.1 Motivation

Gewöhnliche Differentialgleichungen (ODE's) beschreiben Prozesse, die zeitabhängig sind. Sie liefern Einsichten in diese Prozesse, die oft anders nicht zu erhalten sind. Allerdings sind ODE's nur Modelle für die Realität. Selbst wenn man ODE's exakt lösen kann, muss dies nicht exakt die Realität sein.

Beispiel (Fischpopulation in einem See).



Wir betrachten einen See voller Fische. Dabei bezeichnen wir die "Anzahl" der Fische (in Tonnen) zum Zeitpunkt t mit $y(t)$. Wie viele Fische kann die Fischindustrie fangen, ohne dass die Fischpopulation ausstirbt?

Annahme: Der See ist isoliert, d.h. entweder ist die Anzahl der Fische, die durch Flüsse in den See kommen gleich der Anzahl der Fische, die den See durch Flüsse verlassen oder es existieren keine zu- bzw. abfließenden Flüsse.

Wir modellieren die Fischpopulation über ein *Erhaltungsgesetz*

$$\text{Änderung der Population} = \text{"Zuwachs"} - \text{"Abgang"}.$$

Dabei identifizieren wir

$$\begin{aligned} \text{Änderung der Population} &\cong y'(t), \\ \text{Zuwachs} &\cong \text{Geburtenrate } G(t), \\ \text{Abgang} &\cong \text{Todesrate } T(t) + \text{ Fischfangrate } H(t), \end{aligned}$$

und nehmen an, dass die Geburtenrate und die Todesrate proportional zur Population sind, d.h. $G(t) = b\,y(t)$ und $T(t) = m\,y(t)$ für $b, m > 0$. Damit erhalten wir die Gleichung

$$y'(t) = (b - m)\,y(t) - H(t).$$

Weiter nehmen wir an, dass $b - m =: a > 0$ gilt. Dabei kann die Größe a über Beobachtungen gemessen werden und $H(t)$ kann kontrolliert werden. Zu einen gegebenen Anfangszustand $y(t_0) = y_0$ zum Zeitpunkt t_0 können wir nun berechnen, ob sich das Fischen lohnt. Betrachten wir die einfache Situation $H(t) = H$, d.h. die Fischfangrate ist konstant in der Zeit, und wählen $t_0 = 0$, so erhalten wir das erste Modell

$$\begin{aligned} y' &= a\,y - H, \\ y(0) &= y_0. \end{aligned} \tag{1.1}$$

Dies ist eine ähnliche Gleichung wie bei der Zinsrechnung, wir erwarten also eine Exponentialfunktion als Lösung. Diese leiten wir im folgenden heuristisch her, eine rigorose Herleitung folgt später. Multiplizieren wir die erste Zeile von (1.1) mit dem positiven Wert $e^{-a\,t}$ so folgt

$$\begin{aligned} -e^{-a\,t}H &= e^{-a\,t}\big(y'(t) - a\,y(t)\big) \\ &= e^{-a\,t}(-a)\,y(t) + e^{a\,t}y'(t) = (e^{-a\,t}y(t))'. \end{aligned}$$

Integrieren wir diese Gleichung bzgl. t (bzw. bilden wir die Stammfunktion), so erhalten wir

$$e^{-a\,t}\,y(t) = \frac{H}{a}\,e^{-a\,t} + C.$$

Multiplizieren wir noch mit $e^{a\,t}$, so folgt

$$y(t) = \frac{H}{a} + C\,e^{a\,t},$$

d.h. falls y eine Lösung von (1.1) ist, hat sie diese Form. Da in der Rechnung alle Operationen umkehrbar sind, löst das so definierte $y(t)$ die erste Gleichung in (1.1). Setzen wir $t = 0$ in die Gleichung für $y(t)$ ein und verwenden die Bedingung $y(0) = y_0$ für den Anfangszustand, so folgt

$$\frac{H}{a} + C = y_0 \qquad \text{bzw.} \qquad C = y_0 - \frac{H}{a}.$$

Also ist

$$y(t) = \frac{H}{a} + \Big(y_0 - \frac{H}{a}\Big) e^{a\,t}$$

für $t \geq 0$ Lösung von (1.1). Betrachten wir diese Lösung, so wächst die Population bei einer Fischfangrate echt größer der Anfangspopulation exponentiell, bei einer Fischfangrate gleich der Anfangspopulation bleibt die Population stabil und sonst sterben die Fische aus (vergleiche Abbildung 1.1). Dies ist

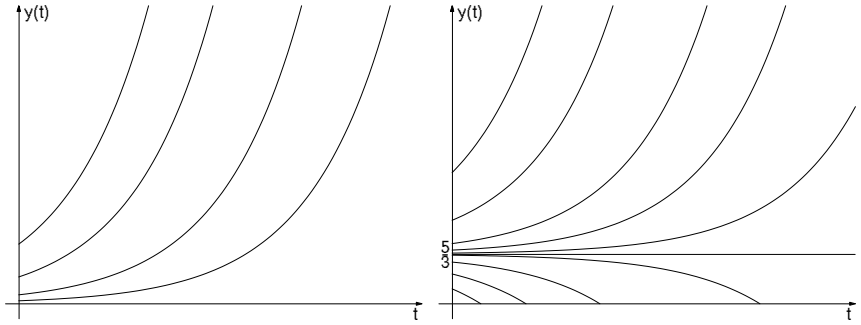


Abb. 1.1. Lösungen von (1.1) zu verschiedenen Anfangswerten. Die Parameter wurden im linken Bild mit $H = 0$, $a = 1$, im rechten Bild mit $H = \frac{5}{3}$, $a = 1$ gewählt.

offensichtlich unrealistisch und somit unser Modell zu einfach gewählt. Insbesondere die Annahme der Proportionalität zwischen der Todesrate und der Population ist problematisch. In abgeschlossenen Seen ist die Todesrate höher als linear abhängig zur Population, z.B. $T(t) = c\,y^2(t)$. Dies führt zu einem neuen Modell

$$\begin{aligned} y'(t) &= b\,y(t) - c\,y^2(t) - H(t), \\ y(0) &= y_0. \end{aligned} \tag{1.2}$$

Wählen wir speziell $b = 1$, $c = \frac{1}{12}$ und $H = \frac{5}{3}$, so gilt

$$\begin{aligned} y'(t) &= y(t) - \frac{1}{12}\,y^2(t) - \frac{5}{3} \\ &= -\frac{1}{12}\big(y(t) - 10\big)\big(y(t) - 2\big) =: f(y(t)). \end{aligned}$$

Wir betrachten zuerst die Gleichung

$$y' = f(y)$$

heuristisch. Für $y \in (2, 10)$ ist $f(y) > 0$, d.h. $y' > 0$ und somit y wachsend, für $y \in (0, 2)$ oder $y > 10$ ist $f(y) < 0$, d.h. $y' < 0$ und somit y fallend. Für

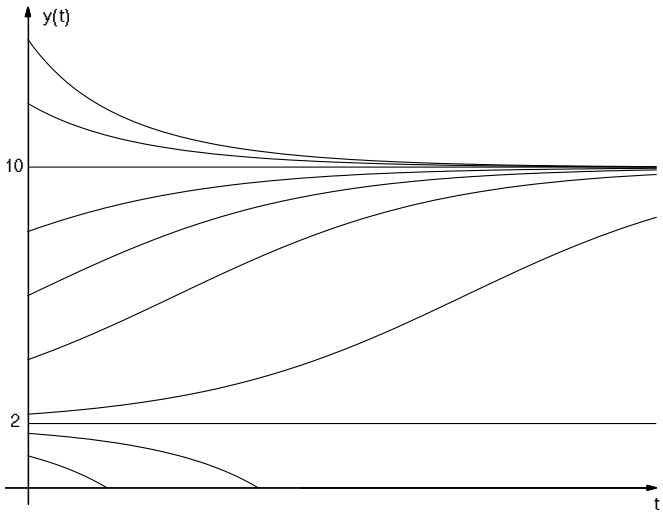


Abb. 1.2. Lösungen von (1.2) zu verschiedenen Anfangswerten. Die Parameter wurden als $b = 1$, $c = \frac{1}{12}$ und $H = \frac{5}{3}$ gewählt.

$y = 2$ und $y = 10$ ist $f(y) = 0$ und somit y konstant. Dies ist in Abbildung 1.2 graphisch dargestellt.

Wir verwenden nun die Methode der *Separation der Variablen*, um diese Heuristik zu untermauern. So folgt aus

$$\frac{dy}{dt} = -\frac{1}{12}\big(y-10\big)\big(y-2\big)$$

für $y \neq 2, 10$ die Gleichung

$$\frac{dy}{(y-10)(y-2)} = -\frac{dt}{12}$$

bzw.

$$\frac{1}{8}\Big(\frac{1}{y-10} - \frac{1}{y-2}\Big)\,dy = -\frac{dt}{12}.$$

Da auf jeder Seite der Gleichung nur eine Variable vorkommt, können wir auf beiden Seiten die Stammfunktion (bezüglich der jeweiligen Variablen) bilden und erhalten

$$\frac{1}{8}\big(\ln|y-10| - \ln|y-2|\big) = -\frac{t}{12} + c.$$

Verwenden wir nun die Rechenregeln für den Logarithmus folgt

$$\ln\left|\frac{y-10}{y-2}\right| = 8\,c - \frac{2}{3}\,t.$$

Nun wenden wir die Exponentialfunktion auf diese Gleichung an und erhalten

$$\left|\frac{y-10}{y-2}\right| = e^{8\,c}\,e^{-\frac{2}{3}\,t}.$$

Setzen wir $c_0 := \exp(8\,c)$, impliziert diese Gleichung $c_0 > 0$. Lösen wir nun den Betrag per Fallunterscheidung auf, so müssen wir $c_0 \in \mathbb{R}$ zulassen. Somit können wir die Gleichung nach y auflösen und erhalten

$$y(t) = \frac{10 - 2\,c_0\,e^{-\frac{2}{3}t}}{1 - c_0\,e^{-\frac{2}{3}t}}.$$

Aus der Forderung $y(0) = y_0$ folgt

$$c_0 = \frac{y_0 - 10}{y_0 - 2}.$$

Setzen wir also

$$y(t) := \frac{10 - 2\,\frac{y_0-10}{y_0-2}\,e^{-\frac{2}{3}t}}{1 - \frac{y_0-10}{y_0-2}\,e^{-\frac{2}{3}t}}$$

für $y_0 \neq 2, 10$ (und somit $y(t) \neq 2, 10$), so löst y die Gleichung (1.2). Weiter sind die konstanten Funktionen $y(t) := 2$ bzw. $y(t) := 10$ Lösungen der Gleichung (1.2) zu den Anfangswerten $y_0 = 2$ bzw. $y_0 = 10$ und stellen somit ein Gleichgewicht zwischen Fischfang und Populationswachstum dar. Dieses Modell ist realistischer als das erste und erlaubt Vorhersagen: So stabilisiert sich der Fischbestand trotz fischen bei einer Anfangspopulation größer gleich 2 Tonnen, bei einer geringeren Anfangspopulation stirbt diese aus (vergleiche Abbildung 1.2).

Für allgemeine, aber konstante b, c und H kann man ebenfalls Lösungsformeln für (1.2) herleiten, bei zeitabhängigen Fischfangraten sind diese nicht bekannt.

12.2 Existenztheorie

Für $f : \mathbb{R}^{n+1} \to \mathbb{R}$ heißt die Gleichung

$$y^{(n)}(t) = f\big(t, y(t), y'(t), \ldots, y^{n-1}(t)\big) \tag{2.1}$$

explizite *Differentialgleichung n-ter Ordnung*. Fordert man zusätzlich

$$y(t_0) = y_0,\ y'(t_0) = y_1,\ \ldots,\ y^{(n-1)}(t_0) = y_{n-1}$$

für $(t_0,\, y_1,\, \ldots,\, y_{n-1}) \in \mathbb{R}^{n+1}$, so spricht man von einem *Anfangswertproblem*.

2.2 Definition. *Sei I ein Intervall. Eine Funktion $y : I \to \mathbb{R}$ heißt* Lösung *von (2.1) im Intervall I, falls y in I n-mal differenzierbar ist und*

$$y^n(t) = f\big(t, y(t), y'(t), \dots, y^{n-1}(t)\big)$$

für alle $t \in I$ identisch erfüllt ist.

Man beachte, dass die Mengen $I = (a, b)$, $I = [a, b)$ oder $I = [a, b]$, mit $a, b \in \mathbb{R}$ und $a < b$, alle zulässige Intervalle sind.

Sei $F : \Omega \subseteq \mathbb{R}^{n+1} \to \mathbb{R}^n$. Das Gleichungssystem

$$Y'(t) = F(t, Y(t)) \tag{2.3}$$

heißt *System von Differentialgleichungen 1. Ordnung*. Falls $F = (f_1, \dots, f_n)$ und $Y = (y_1, \dots, y_n)$, so können wir (2.3) komponentenweise schreiben

$$\begin{aligned} y_1'(t) &= f_1(t, y_1(t), \dots, y_n(t)) \\ &\vdots \\ y_n'(t) &= f_n(t, y_1(t), \dots, y_n(t)). \end{aligned} \tag{2.4}$$

Ergänzt man (2.3) durch die Bedingung

$$Y(t_0) = Y_0$$

für ein $(t_0, Y_0) \in \Omega$ bzw. (2.4) durch

$$y_1(t_0) = y_1^0, \ \dots, \ y_n(t_0) = y_n^0$$

für ein $(t_0, y_1^0, \dots, y_n^0) \in \Omega$, so spricht man wieder von einem *Anfangswertproblem*.

2.5 Definition. *Sei I ein Intervall. Eine vektorwertige Funktion $Y : I \to \mathbb{R}^n$ heißt Lösung des Systems (2.3) in dem Intervall I, falls Y in I differenzierbar ist und*

$$Y'(t) = F\big(t, Y(t)\big)$$

für alle $t \in I$ identisch erfüllt ist.

Bemerkung. (i) Für eine Lösung von (2.3) gilt insbesondere $(t, Y(t)) \in \Omega$ für alle $t \in I$.

(ii) Die gesamte Existenztheorie wird direkt für Systeme von Differentialgleichungen bewiesen, da kein Mehraufwand im Vergleich zur Behandlung einer Differentialgleichung entsteht. Man sollte sich die Aussagen trotzdem für den anschaulich einfacheren Fall von einer Differentialgleichung klar machen.

(iii) Die explizite Differentialgleichung n-ter Ordnung (2.1) lässt sich in ein System von n Differentialgleichung 1. Ordnung umschreiben:

$$\begin{aligned} y_1'(t) &= y_2(t), \\ y_2'(t) &= y_3(t), \\ &\vdots \\ y_{n-1}'(t) &= y_n(t), \\ y_n'(t) &= f\big(t, y_1(t), \dots, y_n(t)\big). \end{aligned} \tag{2.6}$$

Denn sei y Lösung von (2.1). Dann ist $Y := (y, y', \dots, y^{n-1})$ eine Lösung von (2.6). Sei umgekehrt $Y = (y_1, \dots, y_n)$ Lösung von (2.6), so ist $y := y_1$ Lösung von (2.1). Analog transformieren sich die entsprechenden Anfangswertprobleme.

Wir formulieren nun eine Voraussetzung an F, um die Lösbarkeit eines Systems von Differentialgleichungen 1. Ordnung zu sichern.

Grundvoraussetzung (S). *Es existiert eine offene Menge $\Omega \subseteq \mathbb{R}^{n+1}$ so, dass $F : \Omega \subseteq \mathbb{R}^{n+1} \to \mathbb{R}^n$ stetig ist.*

2.7 Lemma. *Sei die Voraussetzung (S) erfüllt und sei $(t_0, Y_0) \in \Omega$, I ein Intervall und $t_0 \in I$. Dann ist Y genau dann eine Lösung von (2.3) mit $Y(t_0) = Y_0$ auf dem Intervall I, wenn Y eine stetige Lösung der Integralgleichung*

$$Y(t) = Y_0 + \int_{t_0}^{t} F\big(s, Y(s)\big)\, ds, \qquad t \in I \tag{2.8}$$

ist.

Das Integral für vektorwertige Funktionen ist komponentenweise definiert, d.h. für $Y(t) = (y_1(t), \dots, y_n(t))$ gilt

$$\int_a^b Y(s)\, ds := \Big(\int_a^b y_1(s)\, ds, \dots, \int_a^b y_n(s)\, ds \Big),$$

falls die Integrale $\int_a^b y_i(s)\, ds$, $i = 1, \dots n$ existieren.

Beweis (Lemma 2.7). " $\Rightarrow$ " Sei Y Lösung von (2.3) mit $Y(t_0) = Y_0$. Somit ist Y differenzierbar in I und insbesondere stetig. Damit ist auch die Abbildung $t \mapsto F\big(t, Y(t)\big)$ stetig. Aufgrund der Gleichung $Y'(t) = F\big(t, Y(t)\big)$ für $t \in I$ und dem Hauptsatz der Differential- und Integralrechnung (Analysis I, Kapitel 7, Satz 3.1) folgt für alle $t \in I$

$$Y(t) - Y_0 = Y(t) - Y(t_0) = \int_{t_0}^{t} Y'(s)\, ds = \int_{t_0}^{t} F\big(s, Y(s)\big)\, ds.$$

Also ist Y eine stetige Lösung der Integralgleichung (2.8).

" $\Leftarrow$ " Sei Y stetige Lösung von (2.8). Nach dem Hauptsatz der Differential- und Integralrechnung ist Y stetig differenzierbar in I und es gilt

$$Y'(t) = \frac{d}{dt} \int_{t_0}^{t} F\big(s, Y(s)\big)\, ds = F\big(t, Y(t)\big),$$

d.h. Y ist stetig differenzierbar. Weiter gilt für den Anfangswert

$$Y(t_0) = Y_0 + \int_{t_0}^{t_0} F\big(s, Y(s)\big)\, ds = Y_0.$$

Somit ist Y eine Lösung von (2.3) auf dem Intervall I mit $Y(t_0) = Y_0$. ∎

Wir wollen im Folgenden zeigen, dass die Bedingung (S) für die Existenz einer Lösung von (2.3) ausreicht. Diese Voraussetzung ist schwächer als die Lipschitz-Stetigkeit, die wir im Existenzsatz von Picard-Lindelöf (Satz 3.5 aus Kapitel 9) benötigt haben. Doch zunächst benötigen wir einige Hilfsmittel.

2.9 Definition. *Eine Menge $M = \{f_\alpha\}_{\alpha \in A}$ von stetigen Funktionen auf dem Intervall $[a, b]$, d.h. $M \subseteq C([a, b])$, heißt* gleichgradig stetig, *wenn zu jedem $\varepsilon > 0$ ein $\delta > 0$ existiert, so dass für alle $s, t \in [a, b]$ mit $|s - t| < \delta$ und alle $\alpha \in A$ gilt: $|f_\alpha(s) - f_\alpha(t)| < \varepsilon$.*

Eine stetige Funktion auf einem kompakten Intervall $[a, b]$ ist gleichmäßig stetig (Satz 3.5 im Kapitel 4), d.h.

$$\forall \varepsilon > 0\ \exists \delta > 0 : \forall s, t \in [a, b] \text{ mit } |s - t| < \delta : |f(s) - f(t)| < \varepsilon.$$

Bei der gleichgradigen Stetigkeit ist neu, dass δ unabhängig von f_α ist.

2.10 Lemma. *Sei $(f_n)_{n \in \mathbb{N}} \subseteq C([a, b])$ eine Folge gleichgradig stetiger Funktionen, die auf einer dichten Menge $M \subseteq [a, b]$ gegen eine Funktion $f : M \to \mathbb{R}$ konvergiert. Dann existiert eine Fortsetzung $\widetilde{f} : [a, b] \to \mathbb{R}$ der Funktion f (d.h. es gilt $\widetilde{f}(x) = f(x)$ für alle $x \in M$), so dass (f_n) gleichmäßig auf $[a, b]$ gegen $\widetilde{f}$ konvergiert. Insbesondere ist $\widetilde{f}$ stetig auf $[a, b]$.*

Zur Erinnerung: (i) M dicht in $[a, b]$ bedeutet, dass für alle $s \in [a, b]$ eine Folge $(s_n)_{n \in \mathbb{N}} \subseteq M$ mit $s_n \to s$ existiert (z.B. $\mathbb{Q}$ ist dicht in $\mathbb{R}$). (ii) Die Funktionenfolge $(f_n)_{n \in \mathbb{N}}$ konvergiert gleichmäßig gegen eine Funktion f auf der Menge D, falls für alle $\varepsilon > 0$ ein $N \in \mathbb{N}$ existiert, so dass für alle $n \geq N$ und für alle $x \in D$ gilt: $|f_n(x) - f(x)| < \varepsilon$.

Beweis (Lemma 2.10). Sei $\varepsilon > 0$ und $(f_n)_{n\in\mathbb{N}}$ gleichgradig stetig. Somit existiert ein $\delta > 0$, so dass

$$\forall n \in \mathbb{N}\, \forall s,t \in [a,b] \text{ mit } |s-t| < \delta : |f_n(s) - f_n(t)| < \varepsilon.$$

Wir wählen nun ein $N \in \mathbb{N}$ mit $\frac{|b-a|}{N} < \frac{\delta}{2}$ und zerlegen das Intervall $[a,b]$ in die Intervalle $[t_i, t_{i+1}]$, $i = 0,\dots,N-1$, wobei $t_0 := a$ und $t_i := a + i\,\frac{|b-a|}{N}$ gewählt werden (vergleiche Abbildung 2.1). Somit gilt $t_i < t_{i+1}$, $t_N = b$ sowie

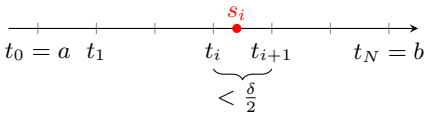


Abb. 2.1. Zerlegung des Intervalls $[a,b]$ in N gleichgroße Teilintervalle.

$|t_i - t_{i+1}| < \frac{\delta}{2}$. Da M dicht in $[a,b]$ liegt, existieren zu den Mittelpunkten der Intervalle $[t_i, t_{i+1}]$ eine Folge aus Elementen von M, die gegen diese konvergieren. Da die Mittelpunkte einen positiven Abstand zum Rand der Intervalle haben, existieren insbesondere $s_i \in [t_i, t_{i+1}] \cap M$ für $i = 0,\dots,N-1$. Weiter konvergiert nach Voraussetzung $(f_n)_{n\in\mathbb{N}}$ auf M gegen f und somit ist $(f_n)_{n\in\mathbb{N}}$ eine Cauchyfolge auf M. Da wir nur N viele s_i haben, folgern wir

$$\exists n_0 \in \mathbb{N}\, \forall m,n \geq n_0\, \forall i = 0,\dots,N-1 : |f_n(s_i) - f_m(s_i)| < \varepsilon.$$

Wählen wir nun ein beliebiges $s \in [a,b]$. Nach Konstruktion existiert ein $i \in \{0,\dots,N-1\}$ mit $|s_i - s| < \delta$. Für $n,m \geq n_0$ gilt dann

$$\begin{aligned} |f_n(s) - f_m(s)| &\leq |f_n(s) - f_n(s_i)| + |f_n(s_i) - f_m(s_i)| + |f_m(s_i) - f_m(s)| \\ &< \varepsilon + \varepsilon + \varepsilon = 3\varepsilon. \end{aligned}$$

Bilden wir das Supremum über alle $s \in [a,b]$ folgt, dass $(f_n)_{n\in\mathbb{N}}$ eine Cauchyfolge bezüglich der Maximumsnorm ist. Somit konvergiert $(f_n)_{n\in\mathbb{N}}$ auf $[a,b]$ gleichmäßig gegen ein $\widetilde{f} \in C([a,b])$ (siehe Analysis I, Kapitel 8, Satz 1.1 und Satz 1.2). Da gleichmäßige Konvergenz die punktweise Konvergenz impliziert und der Grenzwert eindeutig ist, folgt $\widetilde{f}|_M = f$. ∎

2.11 Satz (Arzela-Ascoli)**.** *Sei $(f_n)_{n\in\mathbb{N}} \subseteq C([a,b])$ eine Folge gleichgradig stetiger, gleichmäßig beschränkter Funktionen, d.h. es existiert eine Konstante M, so dass für alle $n \in \mathbb{N}$ und $s \in [a,b]$ gilt $|f_n(s)| \leq M$. Dann existiert eine Teilfolge, die gleichmäßig gegen eine Funktion $f \in C([a,b])$ konvergiert, d.h.*

$$\exists\, (f_{n_k})_{k\in\mathbb{N}} \subseteq (f_n)_{n\in\mathbb{N}},\ \exists\, f \in C([a,b]) \text{ mit } f_{n_k} \rightrightarrows f \text{ für } k \to \infty.$$

Beweis. Sei $M = \{x_i\}_{i\in\mathbb{N}}$ eine Aufzählung der rationalen Zahlen aus $[a, b]$. Da nach Voraussetzung die Folge $(f_n(x_1))_{n\in\mathbb{N}}$ beschränkt ist, existiert nach dem Satz von Bolzano–Weierstraß (Analysis I, Kapitel 4, Satz 1.3) eine Teilfolge $(f_{n_k^1})_{k\in\mathbb{N}} \subseteq (f_n)_{n\in\mathbb{N}}$ und ein $f(x_1) \in \mathbb{R}$ mit

$$f_{n_k^1}(x_1) \to f(x_1) \text{ für } k \to \infty.$$

Nun ist die Folge $(f_{n_k^1}(x_2))_{k\in\mathbb{N}}$ beschränkt und es existiert wiederum nach dem Satz von Bolzano-Weierstraß eine Teilfolge $(f_{n_k^2})_{k\in\mathbb{N}} \subseteq (f_{n_k^1})_{k\in\mathbb{N}}$ und ein $f(x_2) \in \mathbb{R}$ mit

$$f_{n_k^2}(x_2) \to f(x_2) \text{ für } k \to \infty.$$

Führen wir dies für alle $i \in \mathbb{N}$ fort, so erhalten wir ein System von Teilfolgen, die Aufgrund der Schachtelung $(f_{n_k^i})_{k\in\mathbb{N}} \subseteq (f_{n_k^{i-1}})_{k\in\mathbb{N}} \subseteq \ldots \subseteq (f_n)_{n\in\mathbb{N}}$ für immer weitere Punkte konvergieren:

$$\begin{aligned}
&f_{n_1^1}, f_{n_2^1}, f_{n_3^1}, \ldots && \text{konvergiert für } x = x_1\\
&f_{n_1^2}, f_{n_2^2}, f_{n_3^2}, \ldots && \text{konvergiert für } x = x_1, x_2\\
&\quad\vdots\\
&f_{n_1^i}, f_{n_2^i}, f_{n_3^i}, \ldots && \text{konvergiert für } x = x_1, x_2, \ldots, x_i\\
&\quad\vdots
\end{aligned}$$

Wählen wir nun die Diagonalfolge $(f_{n_k^k})_{k\in\mathbb{N}}$, so konvergiert diese für alle x_i gegen $f(x_i)$, denn ab dem i-ten Folgenglied ist $(f_{n_k^k})_{k\in\mathbb{N}}$ eine Teilfolge von $(f_{n_k^i})_{k\in\mathbb{N}}$ und somit konvergent. Wir haben also eine Teilfolge gefunden, die auf einer dichten Menge gegen die Funktion f (gegeben durch $f(x_i)$ am Punkt x_i) konvergiert. Wenden wir nun Lemma 2.10 an, so ist der Satz bewiesen. ∎

2.12 Folgerung. *Sei $F_k : [a, b] \to \mathbb{R}^n$, $k \in \mathbb{N}$, eine Folge gleichgradig stetiger vektorwertiger Funktionen, d.h.*

$$\forall\varepsilon > 0\ \exists\delta > 0 : \forall k \in \mathbb{N},\ \forall s, t \in [a, b] \text{ mit } |s - t| < \delta : \|F_k(s) - F_k(t)\| < \varepsilon,$$

die gleichmäßig beschränkt sind. Dann existiert eine Teilfolge $(F_{k_\ell})_{\ell\in\mathbb{N}}$, die gleichmäßig konvergiert.

Beweis. Schreiben wir $F_k = (f_1^k, \ldots, f_n^k)$, $F = (f_1, \ldots, f_n)$, so ist die gleichmäßige Konvergenz von F_k gegen F äquivalent zur gleichmäßigen Konvergenz der f_i^k gegen f_i, $i = 1, \ldots n$. Da nach Voraussetzung die f_i^k, $i = 1, \ldots n$, ebenfalls gleichgradig stetig und gleichmäßig beschränkt sind, folgt die Behauptung mit Satz 2.11, sukzessive angewendet auf die Koordinatenfunktionen (f_i^k) und die sich ergebenden Teilfolgen $(f_i^{k_\ell})_{\ell\in\mathbb{N}}$, $i = 1, \ldots, n$. ∎

2.13 Satz (Peano)**.** *Erfülle F die Voraussetzung (S), sei $(t_0, Y_0) \in \Omega$ und seien $a, b > 0$ so, dass*

$$Q := \big\{(t, Y) \in \mathbb{R}^{n+1} \;\big|\; |t - t_0| \le a, \|Y - Y_0\| \le b\big\}$$

eine Teilmenge von Ω ist, d.h. $Q \subseteq \Omega$. Sei $K > 0$ derart, dass $\|F(t, Y)\| \le K$ für alle $(t, Y) \in Q$ gilt. Dann besitzt das System gewöhnlicher Differentialgleichungen (2.3) mit $Y(t_0) = Y_0$ eine stetig differenzierbare Lösung $Y(\cdot)$ auf dem Intervall $[t_0 - c, t_0 + c] =: I$ wobei die Konstante $c = \min(a, \frac{b}{K})$ gewählt werden kann.

Beweis. Nach Lemma 2.7 genügt es, eine stetige Funktion Y mit

$$Y(t) = Y_0 + \int_{t_0}^{t} F\big(s, Y(s)\big)\, ds, \quad t \in I$$

zu finden. Wir wollen eine Lösung der Integralgleichung per Approximation finden: Dazu definieren wir für $t \in I$ die Funktion $Y_0(t) := Y_0$ und für $n \in \mathbb{N}$ die Funktionen

$$Y_{n+1}(t) := Y_0 + \int_{t_0}^{t} F\big(s, Y_n(s)\big)\, ds \qquad t \in I.$$

Für die Wohldefiniertheit von $Y_{n+1} : I \to \mathbb{R}^n$ reicht es die Stetigkeit von Y_n und $(t, Y_n(t)) \in Q \subseteq \Omega$ für $t \in I$ zu zeigen. Dies machen wir per vollständiger Induktion:

Induktionsanfang: Offenbar ist Y_0 stetig und es gilt $(t, Y_0) \in Q$ für $t \in I$ nach der Wahl von c und der Definition von Q.
Induktionshypothese: Gelte Y_n ist stetig und $(t, Y_n(t)) \in Q$ für $t \in I$.
Induktionsschritt: Für $t \in I$ folgt einerseits $|t - t_0| \le a$ wegen $c \le a$ und andererseits

$$\begin{aligned}
\big\|Y_{n+1}(t) - Y_0\big\| &= \Big\| \int_{t_0}^{t} F\big(s, Y_n(s)\big)\, ds \Big\| \\
&\le \int_{t_0}^{t} \big\|F\big(s, Y_n(s)\big)\big\|\, ds \\
&\le K\,|t - t_0| \le K\,c \le K\,\frac{b}{K} = b.
\end{aligned}$$

Somit gilt $(t, Y_{n+1}(t)) \in Q$. Da nach Induktionshypothese Y_n stetig ist und F nach Bedingung (S) auf Ω stetig ist, folgt mit dem Hauptsatz der Differential- und Integralrechnung (Analysis I, Kapitel 7, Satz 3.1) die Stetigkeit von Y_{n+1} und der Induktionsschritt ist bewiesen. Mit der gleichen Rechnung wie im Beweis der Wohldefiniertheit folgt für alle $n \in \mathbb{N}$

$$\sup_{t\in I}\|Y_n(t)\| \le \sup_{t\in I}\|Y_n(t)-Y_0\| + \|Y_0\| \le b + \|Y_0\|$$

bzw. für $n=0$

$$\sup_{t\in I}\|Y_0(t)\| \le \sup_{t\in I}\|Y_0\| \le b + \|Y_0\|.$$

Also ist die Folge $(Y_n)_{n\in\mathbb{N}_0}$ gleichmäßig beschränkt. Wir zeigen nun, dass diese Folge auch gleichgradig stetig ist: Sei dazu $\varepsilon > 0$ und $\delta := \frac{\varepsilon}{K}$. Dann gilt für alle $q,t \in I$ mit $|q-t| < \delta$ und alle $n \in \mathbb{N}$

$$\begin{aligned}\Big\|Y_n(q) - Y_n(t)\Big\| &= \Big\|Y_0 + \int_{t_0}^{q} F\big(s, Y_{n-1}(s)\big)\,ds - Y_0 + \int_{t_0}^{t} F\big(s, Y_{n-1}(s)\big)\,ds\Big\| \\ &\le \Big|\int_{q}^{t} \big\|F\big(s, Y_{n-1}(s)\big)\big\|\,ds\Big| \\ &\le K\,|t-q| < K\,\delta = K\,\frac{\varepsilon}{K} = \varepsilon.\end{aligned}$$

Für $n=0$ folgt trivialerweise

$$\|Y_0(q) - Y_0(t)\| = \|Y_0 - Y_0\| = 0 < \varepsilon.$$

Wir können also den Satz von Arzela-Ascoli (bzw. Folgerung 2.12) anwenden und erhalten eine Teilfolge $(Y_{n_k})_{k\in\mathbb{N}} \subseteq (Y_n)_{n\in\mathbb{N}}$ und eine stetige Funktion $Y : I \to R^n$ mit

$$Y_{n_k} \rightrightarrows Y \qquad \text{für } k \to \infty.$$

Wir bemerken zunächst, dass nach Definition Q abgeschlossen ist und wegen $(s, Y_n(s)) \in Q$ für alle $s \in I$, $n \in \mathbb{N}$ sowie der Konvergenz der Y_n gegen Y damit $(s, Y(s)) \in Q$ für alle $s \in I$ folgt. Da weiter F stetig auf der kompakten Menge Q ist, folgt die gleichmäßige Stetigkeit von F auf Q, d.h. für alle $\varepsilon > 0$ existiert ein $\delta > 0$, so dass für alle $(t,Y),(s,Z) \in Q$ mit $\|(t,Y)-(s,Z)\| < \delta$ folgt: $\|F(s,Z) - F(t,Y)\| < \varepsilon$. Wählen wir nun zu $\varepsilon > 0$ das $\delta > 0$ aus der gleichmäßigen Stetigkeit und das k_0 aus der gleichmäßigen Konvergenz von Y_{n_k} gegen Y mit

$$\|Y_{n_k}(s) - Y(s)\| < \delta \qquad \forall s \in I,\ k \ge k_0,$$

so haben wir die gleichmäßige Konvergenz von $F(\cdot, Y_{n_k}(\cdot))$ gegen $F(\cdot, Y(\cdot))$ gezeigt. Nach Analysis I, Kapitel 7, Satz 2.4 folgt die Konvergenz des Integrals, d.h. mit der Definition der Folge Y_n gilt

$$\begin{aligned}Y(t) &= \lim_{k\to\infty} Y_{n_{k+1}}(t) \\ &= \lim_{k\to\infty} Y_0 + \int_{t_0}^{t} F\big(s, Y_{n_k}(s)\big)\,ds \\ &= Y_0 + \int_{t_0}^{t} F\big(s, Y(s)\big)\,ds.\end{aligned}$$

Somit löst Y die Integralgleichung und nach Lemma 2.7 haben wir die Existenz einer Lösung von (2.3) bewiesen. ∎

Der Satz von Peano ist ein lokaler Existenzsatz, d.h. er liefert nur die Existenz einer Lösung auf einem kleinen Intervall um den Anfangswert. Wir stellen uns nun die Frage, ob wir diese Lösung fortsetzen können, in dem wir den Satz von Peano am Ende des Intervalls erneut anwenden, also für den Anfangswert die Lösung am Ende des Intervalls auswerten.

2.14 Lemma. *Erfülle F die Voraussetzung (S) und sei $Y : (a, b) \to \mathbb{R}^n$ eine Lösung von (2.3). Weiter existiere der linksseitige Grenzwert*

$$\lim_{t \nearrow b} Y(t) =: Y_1$$

und es gelte $(b, Y_1) \in \Omega$. Dann existiert ein $\delta > 0$, so dass man die Lösung Y zu einer Lösung auf dem Intervall $(a, b + \delta]$ fortsetzen kann.

Beweis. Da $(b, Y_1) \in \Omega$ gilt und Ω offen ist, existiert eine offene Umgebung von (b, Y_1), die vollständig in Ω liegt. Insbesondere existieren also $\alpha, \beta > 0$, so dass

$$Q := \left\{(t, Y) \in \mathbb{R}^{n+1} \;\middle|\; |t - b| \le \alpha,\ \|Y - Y_1\| \le \beta\right\} \subseteq \Omega$$

gilt (vergleiche Abbildung 2.2). Setzen wir weiter $K := \sup_{(t,Y)\in Q} \|F(t, Y)\|$,

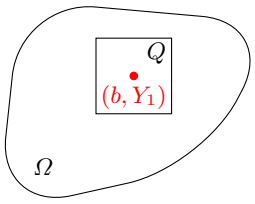


Abb. 2.2. Schachtelung der Mengen um den Punkt (b, Y_1).

so folgt mit dem Satz von Peano die Existenz einer Lösung $\widetilde{Y}$ von

$$\widetilde{Y}'(t) = F(t, \widetilde{Y}(t)),$$
$$\widetilde{Y}(b) = Y_1$$

auf dem Intervall $I = [b - \delta, b + \delta]$, wobei δ durch $\delta := \min(\alpha, \frac{\beta}{K})$ gegeben ist (siehe Abbildung 2.3). Wir definieren nun für $t \in (a, b + \delta]$

$$Z(t) := \begin{cases} Y(t) & t \in (a, b) \\ \widetilde{Y}(t) & t \in [b, b + \delta]. \end{cases}$$

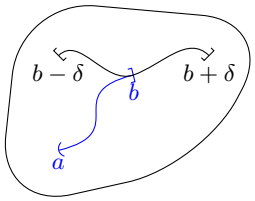


Abb. 2.3. Lösungen $Y, \widetilde{Y}$ auf den Intervallen $(a, b]$ bzw. $[b-\delta, b+\delta]$

Offensichtlich ist Z stetig in $(a, b+\delta]$ und stetig differenzierbar in (a, b) sowie $(b, b+\delta)$. Weiter erfüllt Z die Gleichung (2.3) in den Intervallen (a, b) sowie $(b, b+\delta]$. Wir müssen also nur die Existenz von $Z'(b)$ und die Identität $Z'(b) = F(b, Z(b))$ zeigen. Dazu bestimmen wir die Rechts- und Linksseitige Ableitung von Z: Nach Lemma 2.7 ist die Gleichung (2.3) äquivalent zur Integralgleichung

$$Y(t) = Y(t_0) + \int_{t_0}^{t} F\big(s, Y(s)\big)\, ds \qquad t_0, t \in (a, b).$$

Da man nach Voraussetzung Y durch $Y(b) := Y_1$ stetig auf das Intervall $(a, b]$ fortsetzen kann, folgt die Stetigkeit von $F(\cdot, Y(\cdot))$ auf dem Intervall $(a, b]$. Somit gilt die obige Integralgleichung für alle $t \in (a, b]$. Nach dem Hauptsatz der Differential- und Integralrechnung (Analysis I, Kapitel 7, Satz 3.1) gilt

$$Y'(t) = F\big(t, Y(t)\big) \qquad \forall t \in (a, b].$$

Wegen $Y(b) = Y_1 = \widetilde{Y}(b)$ folgt $Z = Y$ auf $(a, b]$ und somit die Existenz der linksseitigen Ableitung im Punkt b mit

$$Z'_-(b) = F\big(b, Y_1\big) = F\big(b, Z(b)\big).$$

Wenden wir analog Lemma 2.7 auf $\widetilde{Y}$ und das Intervall $[b, b+\delta]$ an, so folgt

$$\widetilde{Y}(t) = Y_1 + \int_{b}^{t} F\big(s, \widetilde{Y}(s)\big)\, ds \qquad t \in [b, b+\delta]$$

Wie oben erhalten wir damit

$$\widetilde{Y}'(t) = F\big(t, \widetilde{Y}(t)\big) \qquad t \in [b, b+\delta]$$

und somit die Existenz der rechtsseitigen Ableitung im Punkt b mit

$$Z'_+(b) = F\big(b, \widetilde{Y}(b)\big) = F\big(b, Z(b)\big).$$

Somit existiert $Z'(b)$ und es gilt $Z'(b) = F\big(b, Z(b)$. Wir haben also eine passende Fortsetzung der Lösung gefunden. ∎

Bemerkung. Eine analoge Aussage gilt für das linksseitige Fortsetzen: Erfülle F die Voraussetzung (S) und sei $Y : (a, b) \to \mathbb{R}^n$ eine Lösung von (2.3). Weiter existiere der linksseitige Grenzwert $\lim_{t \searrow a} Y(t) =: Y_1$ und es gelte $(a, Y_1) \in \Omega$. Dann existiert ein $\delta > 0$, so dass man die Lösung Y zu einer Lösung auf dem Intervall $[a - \delta, b)$ fortsetzen kann.

Nun stellt sich die Frage, wie weit man diese Lösungen fortsetzen kann. Es könnte passieren, dass bei der Anwendung von Peano die Intervalle immer kleiner werden und der Prozess stoppt. Dies werden wir im Folgenden untersuchen.

2.15 Lemma. *Erfülle F die Voraussetzung (S), sei $Y : (a, b) \to \mathbb{R}^n$ eine Lösung von (2.3) und sei K eine kompakte Teilmenge von Ω. Dann lässt sich Y über K hinaus fortsetzen, d.h. es existiert eine Funktion $\widetilde{Y} : (c, d) \to \mathbb{R}^n$ mit $c \le a < b \le d$, so dass $\widetilde{Y}$ eine Lösung von (2.3) auf dem Intervall (c, d) ist, $\widetilde{Y} = Y$ auf (a, b) gilt und $c < \tau_1 < \tau_2 < d$ existieren, so dass für alle $t \in (c, \tau_1) \cup (\tau_2, d)$ gilt: $(t, \widetilde{Y}(t)) \notin K$. Insbesondere ist* $\operatorname{graph}(\widetilde{Y}) \not\subseteq K$ *(vergleiche Abbildung 2.4).*

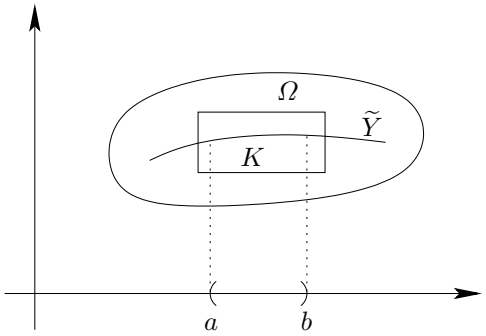


Abb. 2.4. Fortsetzung der Lösung über eine kompakte Menge K hinaus.

Beweis. Für zwei Mengen $A, B \subseteq \mathbb{R}^{n+1}$ ist der Abstand voneinander durch

$$\operatorname{dist}(A, B) := \inf_{X \in A,\, Y \in B} \|X - Y\|$$

definiert. Im Fall $\operatorname{dist}(A, B) = \delta > 0$ gilt folglich für alle $X \in A$ und alle $Y \in B$ die Ungleichung $\|X - Y\| \ge \delta > 0$ (vergleiche Abbildung 2.5).

Da K kompakt in Ω liegt, existiert ein $\delta > 0$ mit $\operatorname{dist}(K, \partial\Omega) = 3\delta$, denn sonst wäre $\operatorname{dist}(K, \partial\Omega) = 0$. Somit würden Folgen $(x_n)_{n\in\mathbb{N}} \subseteq K$ und $(y_n)_{n\in\mathbb{N}} \subseteq \partial\Omega$ mit $\|x_n - y_n\| \to 0$ für $n \to \infty$ existieren. Da K kompakt ist, gäbe es ein

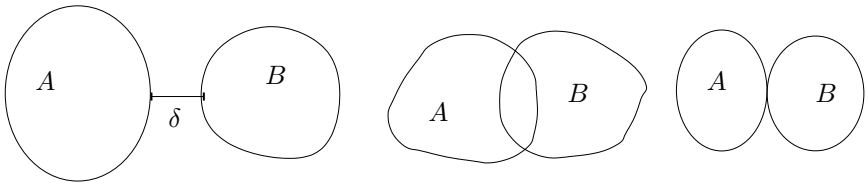


Abb. 2.5. Abstand zwischen zwei Mengen A, B. Links: $\operatorname{dist}(A,B) = \delta$. Mitte und Rechts: $\operatorname{dist}(A,B) = 0$.

$x_0 \in K$ und eine Teilfolge mit $x_{n_k} \to x_0$ für $k \to \infty$. Damit würde auch folgen, dass $y_{n_k} \to x_0$ für $k \to \infty$ und somit wäre $x_0 \in \partial\Omega$ aufgrund der Abgeschlossenheit des Randes. Da Ω offen ist, ist die Menge Ω^c abgeschlossen und insbesondere gilt $\partial\Omega \subseteq \Omega^c$. Nun haben wir einen Widerspruch, denn $K \subseteq \Omega$ impliziert $K \cap \Omega^c = \emptyset$.

Definieren wir nun

$$K_{2\delta} := \big\{ X \in \Omega \;\big|\; \operatorname{dist}(X,K) \le 2\delta \big\},$$

so folgt $K \subseteq K_{2\delta} \subseteq \Omega$ nach der Wahl von δ und man sieht leicht, dass $K_{2\delta}$ beschränkt und abgeschlossen ist. Damit ist $K_{2\delta}$ kompakt und

$$\sup_{(t,Y)\in K_{2\delta}} \|F(t,Y)\| =: C$$

ist endlich. Wir definieren

$$Q(t_0,Y_0) := \big\{ (t,Y) \in \mathbb{R}^{n+1} \;\big|\; |t_0 - t| \le \delta,\ \|Y - Y_0\| \le \delta \big\}.$$

Für $(t_0,Y_0) \in K$ folgt für alle $(t,Y) \in Q(t_0,Y_0)$

$$\|(t_0,Y_0) - (t,Y)\|^2 = |t_0 - t|^2 + \|Y - Y_0\|^2 \le 2\delta^2$$

und somit $\operatorname{dist}(Q(t_0,Y_0),K) \le \sqrt{2}\delta \le 2\delta$, d.h. $Q(t_0,Y_0) \subseteq K_{2\delta}$.

Wir wählen nun ein $t_0 \in (a,b)$ und betrachten die Lösung $Y : (a,b) \to \mathbb{R}^n$ von (2.3) eingeschränkt auf die Intervalle $[t_0,b)$ und $(a,t_0]$ separat. Wir beginnen mit $Y : [t_0,b) \to \mathbb{R}^n$ und zeigen, dass eine "rechtsseitige" Fortsetzung $\widetilde{Y}$ der Lösung mit $\operatorname{graph}(\widetilde{Y}) \not\subseteq K$ existiert. Sei $\operatorname{graph}(Y) \subseteq K$, da sonst $\widetilde{Y} = Y$ gewählt werden kann. Damit gilt: $(t,Y(t)) \in \operatorname{graph}(Y) \subseteq K \subseteq K_{2\delta}$ für $t \in [t_0,b)$ und es folgt mit der Lösungseigenschaft

$$\|Y'(t)\| = \|F\big(t,Y(t)\big)\| \le C \qquad t \in [t_0,b).$$

Wählen wir eine Folge $t_n \nearrow b$, so folgt mit dem Mittelwertsatz

$$\|Y(t_n) - Y(t_m)\| \le \|Y'(x_{n,m})(t_n - t_m)\| \le C\,|t_n - t_m|.$$

Somit ist $(Y(t_n))_{n\in\mathbb{N}}$ eine Cauchyfolge in $\mathbb{R}^n$ und es existiert ein $Y_1 \in \mathbb{R}^n$ mit $Y(t_n) \to Y_1$ für $n \to \infty$. Für eine weitere Folge $\widetilde{t}_n \nearrow b$ mit $Y(\widetilde{t}_n) \to \widetilde{Y}_1$ für $n \to \infty$ folgt wieder mit dem Mittelwertsatz

$$\begin{aligned}\|Y_1 - \widetilde{Y}_1\| &\le \|Y_1 - Y(t_n)\| + \|Y(t_n) - Y(\widetilde{t}_n)\| + \|Y(\widetilde{t}_n) - \widetilde{Y}_1\| \\ &\le \|Y_1 - Y(t_n)\| + C\left(|t_n - b| + |b - \widetilde{t}_n|\right) + \|Y(\widetilde{t}_n) - \widetilde{Y}_1\| \xrightarrow{n\to\infty} 0.\end{aligned}$$

Somit ist der Grenzwert Y_1 unabhängig von der gewählten Folge und

$$\lim_{t\nearrow b} Y(t) = Y_1$$

existiert. Da K abgeschlossen ist, folgt mit der Konvergenz $(t_n, Y(t_n)) \to (b, Y_1)$, dass $(b, Y_1) \in K$ und somit $(b, Y_1) \in \Omega$ gilt. Verwenden wir nun Lemma 2.14, so existiert eine Fortsetzung $\widetilde{Y}$ der Lösung auf einem Intervall $[t_0, b + \rho]$. Betrachten wir den Beweis von Lemma 2.14 genauer, so kann in unserer Situation $\rho = \min(\delta, \frac{\delta}{C})$ gewählt werden. Gilt nun $\operatorname{graph}(\widetilde{Y}) \not\subseteq K$ sind wir fertig, ansonsten wiederholen wir den Prozess. Da jede weitere Fortsetzung wieder um das feste ρ verlängert wird und K beschränkt ist, muss der Prozess nach endlich vielen Schritten abbrechen, d.h. $\operatorname{graph}(\widetilde{Y}) \not\subseteq K$. Für das Fortsetzen über den linken Rand betrachten wir $Y : (a, t_0] \to \mathbb{R}^n$ und argumentieren analog. ∎

2.16 Lemma. *Erfülle F die Voraussetzung (S) und sei $(t_0, Y_0) \in \Omega$. Sei $(Y_n)_{n\in\mathbb{N}}$ eine Folge von Lösungen von (2.3), wobei Y_n auf dem Intervall I_n definiert ist. Gelte weiter $t_0 \in I_n$ und $Y_n(t_0) = Y_0$ sowie*

$$Y_n(t) = Y_m(t) \qquad \forall t \in I_n \cap I_m$$

und alle $n, m \in \mathbb{N}$. Dann existiert im Intervall $I := \bigcup_{n\in\mathbb{N}} I_n$ eine Lösung Y von (2.3) mit $Y(t_0) = Y_0$, für die $Y(t) = Y_n(t)$ für alle $t \in I_n$ und alle $n \in \mathbb{N}$ gilt.

Beweis. Sei $t \in I$. Dann existiert ein $n \in \mathbb{N}$ mit $t \in I_n$. Wir definieren $Y(t) := Y_n(t)$. Diese Definition ist wohldefiniert, d.h. unabhängig von der Wahl von n, denn für ein $m \in \mathbb{N}$ mit $t \in I_m$ folgt nach Voraussetzung $Y_n(t) = Y_m(t)$. Weiter ist Y auch eine Lösung von (2.3), denn für $s \in I$ existiert ein $n \in \mathbb{N}$ mit $s \in I_n$ und insbesondere $[t_0, s] \subseteq I_n$ (bzw. $[s, t_0] \subseteq I_n$). Da Y_n eine Lösung von (2.3) auf dem Intervall I_n ist und $Y(t) = Y_n(t)$ für alle $t \in [t_0, s]$ (bzw. $t \in [s, t_0]$)) gilt, ist also auch Y eine Lösung auf diesem Intervall. Da s beliebig war, ist Y also eine gewünschte Lösung. ∎

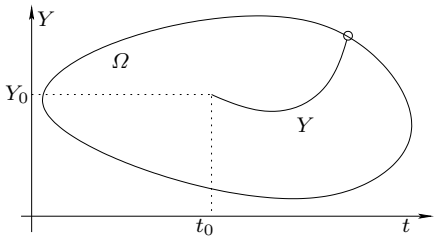
2.17 Definition. *Sei Y eine Lösung von (2.3) mit $Y(t_0) = Y_0$ und sei G der abgeschlossene Graph von Y, d.h. $G := \overline{\operatorname{graph}(Y)}$. Wir sagen, dass Y nach rechts dem Rand beliebig nahe kommt, wenn*

$$G^+ := \big\{(t,Z) \in G \;\big|\; t \geq t_0\big\}$$

keine kompakte Teilmenge von Ω ist. Analog ist nach links dem Rand beliebig nahe kommen definiert.

Bemerkung. Man kann zeigen, dass für eine Lösung $Y : [t_0, b) \to \mathbb{R}^n$ der "rechtsseitige" Graph G^+ keine kompakte Teilmenge von Ω ist, falls einer folgenden Fälle eintritt: a) $\lim_{t \nearrow b} \operatorname{dist}(\partial\Omega, (t, Y(t))) = 0$, oder b) $b = \infty$, oder c) $\lim_{t \nearrow b} \|Y(t)\| = \infty$. Diese drei Fälle kann man sich wie folgt vorstellen, wobei wir nur die Zusammenhangskomponente von Ω, die (t_0, Y_0) enthält, betrachten:

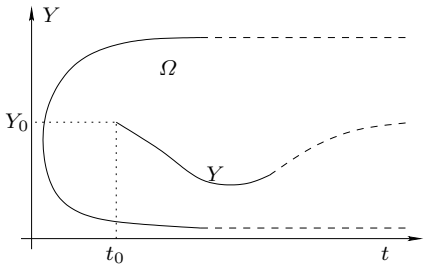
a)



$Y : [t_0, b) \to \mathbb{R}^n$ mit

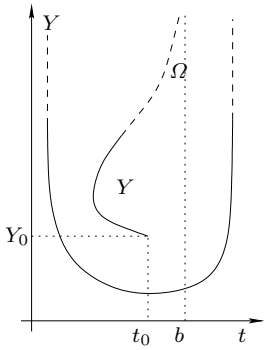
$$\lim_{t \nearrow b} \operatorname{dist}(\partial\Omega, (t, Y(t))) = 0.$$

b)



$Y : [t_0, b) \to \mathbb{R}^n$ mit $b = \infty$

c)



$Y : [t_0, b) \to \mathbb{R}^n$ mit

$$\lim_{t \nearrow b} \|Y(t)\| = \infty.$$

2.18 Satz. *Erfülle F die Voraussetzung (S) und sei $(t_0, Y_0) \in \Omega$. Dann existiert eine Lösung Y von (2.3) mit $Y(t_0) = Y_0$, die sich so fortsetzen lässt,*

dass Y links und rechts dem Rand $\partial\Omega$ beliebig nahe kommt. Eine derartig fortgesetzte Lösung nennen wir maximale Lösung.

Beweis. Da Ω offen ist und $(t_0, Y_0) \in \Omega$ gilt, existieren $a, b > 0$ mit

$$Q := \big\{(t, Y) \in \mathbb{R}^{n+1} \;\big|\; |t - t_0| \le a,\ \|Y - Y_0\| \le b\big\} \subseteq \Omega.$$

Offenbar ist Q kompakt und somit $\sup_{(t,Y)\in Q} \|F(t, Y)\|$ endlich. Nach dem Satz von Peano existiert also eine lokale Lösung Y von (2.3) mit $Y(t_0) = Y_0$. Definieren wir für $n \in \mathbb{N}$ die Mengen

$$M_n := \big\{(t, Y) \in \Omega \;\big|\; \operatorname{dist}((t, Y), \partial\Omega) \ge \frac{1}{n},\ \|(t, Y)\| \le n\big\},$$

so sind diese kompakt und es gilt $M_n \subseteq M_{n+1}$ sowie $\bigcup_{n\in\mathbb{N}} M_n = \Omega$. Falls Y rechts dem Rand $\partial\Omega$ nicht beliebig nahe kommt, also falls

$$G^+ = \big\{(t, Z) \in \overline{\operatorname{graph}(Y)} \;\big|\; t \ge t_0\big\}$$

eine kompakte Teilmenge von Ω ist, folgt wie im Beweis von Lemma 2.15 $\operatorname{dist}(G^+, \partial\Omega) > 0$. Somit existiert ein $n \in \mathbb{N}$ mit $G^+ \subseteq M_n$. Da M_n kompakt ist, existiert nach Lemma 2.15 eine Fortsetzung $Y_n : [t_0, b_n] \to \mathbb{R}^n$ von Y über M_n hinaus. Nun gibt es zwei Möglichkeiten: Entweder $\operatorname{graph}(Y_n) \not\subseteq M_{n+1}$, dann setzen wir $Y_{n+1} := Y_n$. Oder $\operatorname{graph}(Y_n) \subseteq M_{n+1}$. Da M_{n+1} kompakt ist, erhalten wir wieder nach Lemma 2.15 eine Fortsetzung $Y_{n+1} : [t_0, b_{n+1}] \to \mathbb{R}^n$ über M_{n+1} hinaus. Insgesamt haben wir eine Folge $(Y_k)_{k\in\mathbb{N}}$ von Lösungen auf den Intervallen $I_k := [t_0, b_k]$ konstruiert, die $b_{k+1} \ge b_k$ und $Y_k = Y_\ell$ auf I_k für $\ell \ge k$ erfüllen. Nach Lemma 2.16 existiert eine Lösung Y auf $[t_0, b)$ mit $b = \lim_{k\to\infty} b_k$[1] und $Y|_{I_k} = Y_k$. Nach Konstruktion ist $\operatorname{graph}(Y)$ in keiner der Mengen M_n und somit in keiner kompakten Teilmenge von Ω enthalten. Also kommt Y recht dem Rand $\partial\Omega$ beliebig nahe. Analog zeigt man die Existenz einer Fortsetzung, die dem Rand $\partial\Omega$ nach links beliebig nahe kommt. ∎

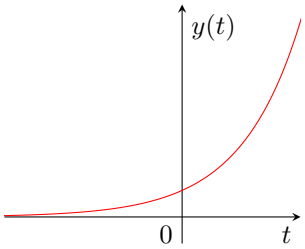
Beispiele. Die folgenden Lösungen von (2.3) kommen dem Rand $\partial\Omega$ beliebig nahe.

1. Betrachte $\Omega = \mathbb{R}^2$, $f(t, y) = y$ und $y(0) = 1$, d.h. (2.3) entspricht

   $$y'(t) = y(t)$$
   $$y(0) = 1.$$

   Dann ist $y(t) := e^t$ eine Lösung auf dem Intervall $I = (-\infty, \infty)$, d.h.

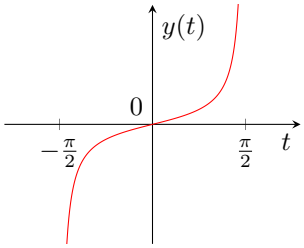
   $$\lim_{t\to\infty} y(t) = \infty.$$



[1] Hier ist $b = \infty$ zugelassen.

2. Betrachte $\Omega = \mathbb{R}^2$, $f(t,y) = 1 + y^2$ und $y(0) = 0$, d.h. (2.3) entspricht

$$\begin{aligned} y'(t) &= 1 + y^2(t) \\ y(0) &= 0. \end{aligned}$$



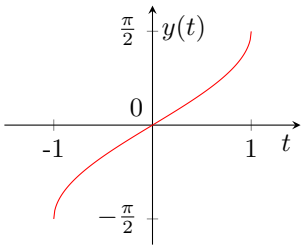
Dann ist $y : (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$, $y(t) = \tan(t)$ eine Lösung, denn es gilt

$$y'(t) = \frac{1}{\cos^2(t)} = \frac{\sin^2(t) + \cos^2(t)}{\cos^2(t)} = 1 + \tan^2(t) = 1 + y(t)^2$$

und offenbar ist $y(0) = 0$. Hier haben wir ein endliches Zeitintervall, aber es gilt $\lim_{t \to \pm\frac{\pi}{2}} y(t) = \pm\infty$.

3. Betrachte $\Omega = \mathbb{R} \times (-\frac{\pi}{2}, \frac{\pi}{2})$, $f(t,y) = \frac{1}{\cos(y)}$ und $y(0) = 0$, d.h. (2.3) entspricht

$$\begin{aligned} y'(t) &= \frac{1}{\cos(y(t))} \\ y(0) &= 0. \end{aligned}$$



Dann ist $y : (-1,1) \to \mathbb{R}$, $y(t) = \arcsin(t)$ eine Lösung, denn es gilt unter Verwendung von $t = \sin(y(t))$

$$y'(t) = \frac{1}{\sqrt{1-t^2}} = \frac{1}{\sqrt{1 - \sin^2(y(t))}} = \frac{1}{\cos(y(t))}$$

und offenbar ist $y(0) = 0$. Hier haben wir ein endliches Zeitintervall und es gilt $\lim_{t \to \pm 1} y(t) = \pm\frac{\pi}{2}$, aber $(\pm 1, \pm\frac{\pi}{2}) \in \partial\Omega$.

Als nächstes rechtfertigen wir den Namen "maximale Lösung".

2.19 Folgerung. *Erfülle F die Voraussetzung (S) und sei $(t_0, Y_0) \in \Omega$. Eine maximale Lösung von (2.3) mit $Y(t_0) = Y_0$ kann nicht fortgesetzt werden.*

Beweis. Sei $Y : (a,b) \to \mathbb{R}^n$ eine maximale Lösung. Wir nehmen an, dass für $\delta > 0$ die Funktion $\widetilde{Y} : (a, b+\delta) \to \mathbb{R}^n$ eine Fortsetzung der Lösung Y ist. Da $\widetilde{Y}$ stetig ist und $\widetilde{Y}|_{(a,b)} = Y$ gilt, folgt

$$\lim_{t \nearrow b} Y(t) = \lim_{t \nearrow b} \widetilde{Y}(t) = \widetilde{Y}(b).$$

Da $\widetilde{Y}$ eine Lösungs ist, gilt $(b, \widetilde{Y}(b)) \in \Omega$. Also folgt

$$G^+ = \left\{(t, Z) \in \overline{\operatorname{graph}(Y)} \;\middle|\; t \geq t_0\right\} \subseteq [t_0, b] \times \overline{\operatorname{Bild}(\widetilde{Y}|_{[t_0,b]})}.$$

Aufgrund der Stetigkeit von $\widetilde{Y}$ ist die rechte Seite kompakt und somit G^+ beschränkt. Da weiter G^+ abgeschlossen ist, folgt die Kompaktheit von G^+. Da Y eine Lösung ist, folgt $(t, Y(t)) \in \Omega$, $t \in [t_0, b)$, und mit dem schon gezeigten $(b, \widetilde{Y}(b)) \in \Omega$ von oben folgt weiter $G^+ \subseteq \Omega$. Dies ist aber ein Widerspruch, da Y maximale Lösung ist. Analog zeigt man, dass die Lösung nicht nach links fortgesetzt werden kann. ∎

Als nächstes stellen wir die Frage nach der Eindeutigkeit der Lösungen. Zunächst betrachten wir dies an einem Beispiel.

Beispiel. Sei $\Omega = \mathbb{R}^2$ und $f(t, y) = |y|^{\frac{1}{2}}$, d.h. (2.3) entspricht

$$y'(t) = |y(t)|^{\frac{1}{2}}.$$

Sei $y : I \to \mathbb{R}$ eine Lösung auf dem Intervall I, so ist $z(t) := -y(-t)$ eine Lösung auf dem gespiegelten Intervall, denn

$$z'(t) = (-y(-t))' = -y'(-t)\,(-1) = y'(-t) = |y(-t)|^{\frac{1}{2}} = |z(t)|^{\frac{1}{2}}.$$

Es genügt also, positive Lösungen zu betrachten - negative Lösungen erhält man dann durch Substitution. Wir verwenden die Methode der Separation der Variablen:

1. Für positive Lösungen betrachten wir

   $$\frac{dy}{dt} = y^{\frac{1}{2}}.$$

   Separieren wir die Variablen, erhalten wir

   $$\frac{dy}{y^{\frac{1}{2}}} = dt.$$

   Bilden wir nun die Stammfunktion, so folgt

   $$2\,y^{\frac{1}{2}} = t + c.$$

   Da wir eine positive Lösung angenommen hatten, muss weiter $t + c \geq 0$ und somit $t \geq -c$ gelten. Somit ist

   $$y(t) := \frac{(t+c)^2}{4}, \qquad t \geq -c$$

   eine Lösung.

2. Offenbar ist $y(t) = 0$ eine Lösung

3. Durch Substitution erhalten wir die negative Lösung

$$y(t) := -\frac{(c-t)^2}{4} \qquad t \le c$$

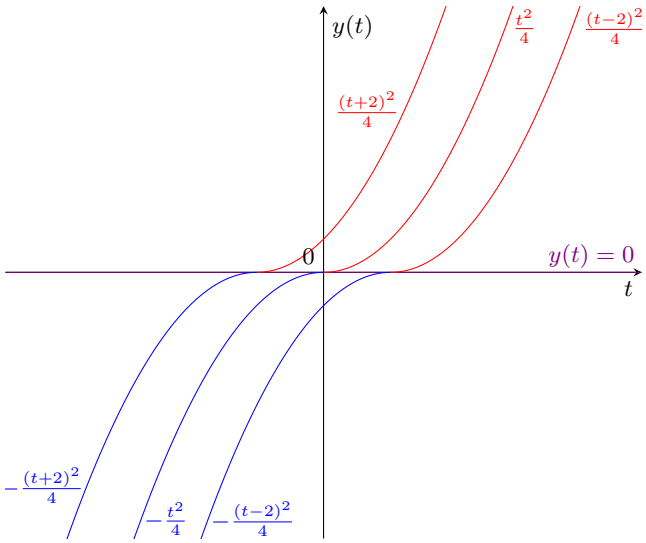


Abb. 2.6. Lösungen von $y'(t) = \sqrt{|y(t)|}$ zu verschiedenen Konstanten c.

Betrachten wir also das Anfangswertproblem $y'(t) = |y(t)|^{\frac{1}{2}}$ mit $y(2) = 1$, so sind sowohl

$$y(t) := \begin{cases} \frac{t^2}{4} & t \ge 0 \\ 0 & t < 0 \end{cases}$$

als auch

$$y(t) := \begin{cases} \frac{t^2}{4} & t \ge 0 \\ 0 & -2 < t < 0 \\ -\frac{(t+2)^2}{4} & t \le -2 \end{cases}$$

Lösungen. Für "nur" stetige f ist also das Anfangswertproblem (2.3) nicht eindeutig.

2.20 Satz. *Erfülle F die Voraussetzung (S) und sei $F : \Omega \to \mathbb{R}^n$ lokal Lipschitz stetig bezüglich Y, d.h. für alle $(t_0, Y_0) \in \Omega$ existiert eine Umgebung U des Punktes (t_0, Y_0) und eine Konstante $L > 0$ so, dass für alle $(t, Z), (t, Y) \in U$ gilt*

$$\|F(t,Y) - F(t,Z)\| \le L\,\|Y - Z\|.$$

Dann ist die maximale Lösung aus Satz 2.18 eindeutig.

Beweis. Sei $(t_0, Y_0) \in \Omega$ und seien $Y_1 : I_1 \to \mathbb{R}^n$, $Y_2 : I_2 \to \mathbb{R}^n$ maximale Lösungen von (2.3) mit

$$Y_1(t_0) = Y_2(t_0) = Y_0.$$

Gelte $Y_1 \neq Y_2$, d.h. oBdA existiert ein $\tau > t_0$ mit $Y_1(\tau) \neq Y_2(\tau)$. Wir definieren die Menge

$$A := \big\{t \in I_1 \cap I_2 \;\big|\; t \ge t_0,\, Y_1(t) \neq Y_2(t)\big\}.$$

Wegen $\tau \in A$ folgt $A \neq \emptyset$ und offensichtlich ist A durch t_0 nach unten beschränkt. Somit gilt $t_0 \le \inf A =: \widetilde{t} \le \tau$ und für alle $t_0 \le t < \widetilde{t}$ folgt $Y_1(t) = Y_2(t)$. Aufgrund der Stetigkeit von Y_1, Y_2 gilt also auch $Y_1(\widetilde{t}) = Y_2(\widetilde{t})$. Insbesondere folgt $(\widetilde{t}, Y_1(\widetilde{t})) \in \Omega$. Aufgrund der lokalen Lipschitz Stetigkeit von F existiert eine Umgebung U von $(\widetilde{t}, Y_1(\widetilde{t}))$ und ein $L > 0$ mit

$$\|F(t,Y) - F(t,Z)\| \le L\,\|Y - Z\| \qquad \forall (t,Y),(t,Z) \in U.$$

Da U offen ist, existieren $r, \widetilde{\delta} > 0$ mit $(\widetilde{t} - 2\widetilde{\delta}, \widetilde{t} + 2\widetilde{\delta}) \times B_{2r}(Y_1(\widetilde{t})) \subseteq U$. Aufgrund der Stetigkeit der Y_1, Y_2 im Punkt $\widetilde{t}$ existieren weiter $\delta_i > 0$, $i = 1, 2$, so dass für alle $t \in \mathbb{R}$ mit $|\widetilde{t} - t| < \delta_i$ die Abschätzung $|Y_i(\widetilde{t}) - Y_i(t)| < r$ folgt. Setzen wir also $\delta := \min\{\delta_1, \delta_2, \widetilde{\delta}, \frac{1}{2L}\}$, so gilt $(t, Y_i(t)) \in U$ für $t \in [\widetilde{t}, \widetilde{t} + \delta]$. Nach Lemma 2.7 lösen die Y_i, $i = 1, 2$, die Integralgleichungen

$$Y_i(t) = Y_0 + \int_{t_0}^{t} F\big(s, Y_i(s)\big)\, ds.$$

Verwenden wir $Y_1(s) = Y_2(s)$ für $t_0 \le s < \widetilde{t}$, so folgt für $t \in [\widetilde{t}, \widetilde{t} + \delta]$

$$\begin{aligned}
\|Y_1(t) - Y_2(t)\| &= \Big\| \int_{t_0}^{t} F\big(s, Y_1(s)\big) - F\big(s, Y_2(s)\big)\, ds \Big\| \\
&= \Big\| \int_{\widetilde{t}}^{t} F\big(s, Y_1(s)\big) - F\big(s, Y_2(s)\big)\, ds \Big\| \\
&\le \int_{\widetilde{t}}^{t} \|F\big(s, Y_1(s)\big) - F\big(s, Y_2(s)\big)\|\, ds \\
&\le L \int_{\widetilde{t}}^{t} \|Y_1(s) - Y_2(s)\|\, ds \\
&\le L \sup_{s \in [\widetilde{t}, \widetilde{t}+\delta]} \|Y_1(s) - Y_2(s)\|\, |t - \widetilde{t}| \\
&\le L\,\delta \sup_{s \in [\widetilde{t}, \widetilde{t}+\delta]} \|Y_1(s) - Y_2(s)\| \\
&\le \frac{1}{2} \sup_{s \in [\widetilde{t}, \widetilde{t}+\delta]} \|Y_1(s) - Y_2(s)\|.
\end{aligned}$$

Bilden wir nun das Supremum über alle $t \in [\widetilde{t}, \widetilde{t}+\delta]$, so folgt

$$\sup_{s\in[\widetilde{t},\widetilde{t}+\delta]} \|Y_1(s) - Y_2(s)\| \leq \frac{1}{2} \sup_{s\in[\widetilde{t},\widetilde{t}+\delta]} \|Y_1(s) - Y_2(s)\|.$$

Da Y_1, Y_2 auf dem Intervall $[\widetilde{t}, \widetilde{t}+\delta]$ stetig sind, sind die Suprema endlich und nach Konstruktion ungleich 0. Somit haben wir einen Widerspruch und die Aussage ist bewiesen. ∎

12.3 Spezialfälle für Gleichungen 1. Ordnung

In diesem Abschnitt betrachten wir spezielle Funktionen $f : \Omega \subseteq \mathbb{R}^2 \to \mathbb{R}$, welche die Voraussetzung (S) erfüllen und suchen Lösungen des Anfangswertproblemes

$$\begin{aligned} y'(t) &= f(t, y(t)) \\ y(t_0) &= y_0. \end{aligned} \tag{3.1}$$

12.3.1 Ortsunabhängige rechte Seiten $f = f(t)$

Falls die rechte Seite f unabhängig von y ist, so können wir die Lösung leicht berechnen. Sei I ein offenes Intervall mit $t_0 \in I$. Erfülle $f : I \times \mathbb{R} \to \mathbb{R}$ die Voraussetzung (S) und gelte $f(t, y) = f(t)$. Wir suchen also nur eine Stammfunktion von f. Da für $t \in I$ die Funktion f stetig auf dem Intervall $[t_0, t]$ ist, ist sie dort integrierbar und

$$\varphi(t) := \int_{t_0}^{t} f(s)\, ds, \qquad t \in I$$

ist wohldefiniert. Mit dem Hauptsatz der Differential und Integralrechnung folgt

$$\varphi'(t) = f(t).$$

Somit sind die Stammfunktionen von f durch

$$y(t) := \varphi(t) + c, \qquad t \in I$$

mit $c \in \mathbb{R}$ gegeben. Für den Anfangswert ergibt sich weiter $y(t_0) = c$. Somit hat (3.1) für alle $y_0 \in \mathbb{R}$ eine Lösung. Wir fassen dies im folgendem Satz zusammen.

3.2 Satz. *Sei I ein offenes Intervall, erfülle $f : I \times \mathbb{R} \to \mathbb{R}$ die Voraussetzung (S) und gelte $f(t, y) = f(t)$ für alle $(t, y) \in I \times \mathbb{R}$. Dann existiert für alle $(t_0, y_0) \in I \times \mathbb{R}$ eine eindeutige, maximale Lösung $y : I \to \mathbb{R}$ von (3.1) mit $y(t_0) = y_0$.*

Beweis. Wir haben oben gezeigt, dass eine Lösung y von (3.1) mit $y(t_0) = y_0$ auf dem Intervall I existiert. Wegen $\Omega = I \times \mathbb{R}$ ist dies bereits eine maximale Lösung. Da $f(t, y) = f(t)$ Lipschitzytetig bezüglich y ist, folgt mit Satz 2.20 die Eindeutigkeit der Lösung und somit die Behauptung. ∎

12.3.2 Zeitunabhängige rechte Seiten $f = f(y)$

Nun vertauschen wir die Rolle von t und y, wir betrachten also ein Intervall $I = (a, b)$ und eine Funktion $f : \mathbb{R} \times I \to \mathbb{R}$, welche die Voraussetzung (S) sowie $f(t, y) = f(y)$ erfüllt. Nehmen wir weiter an, dass $f(y) \neq 0$ für alle $y \in (a, b)$ gilt. Falls $\varphi : (\alpha, \beta) \to I$ eine Lösung von (3.1) ist, dann folgt

$$\varphi'(t) = f(\varphi(t)) \neq 0, \qquad t \in (\alpha, \beta)$$

bzw.

$$\frac{\varphi'(t)}{f(\varphi(t))} = 1, \qquad t \in (\alpha, \beta).$$

Aufgrund unserer Voraussetzungen an f ist $\frac{1}{f}$ stetig und es existiert somit eine Stammfunktion F von $\frac{1}{f}$ auf dem Intervall (α, β). Es folgt für $t \in (\alpha, \beta)$

$$\frac{d}{dt} F(\varphi(t)) = F'(\varphi(t))\, \varphi'(t) = \frac{\varphi'(t)}{f(\varphi(t)} = 1.$$

Nach dem Hauptsatz der Differential- und Integralrechnung existiert also ein $c \in \mathbb{R}$ mit

$$F(\varphi(t)) = t + c, \qquad t \in (\alpha, \beta).$$

Aufgrund der Stetigkeit von f und

$$F'(y) = \frac{1}{f(y)} \neq 0, \qquad y \in (\alpha, \beta)$$

ist F strikt monoton. Insbesondere existiert die Umkehrfunktion F^{-1}. Somit erhalten wir als Gleichung für φ

$$\varphi(t) = F^{-1}(t + c), \qquad t \in (\alpha, \beta).$$

Wir fassen dieses Ergebnis im folgenden Lemma zusammen.

3.3 Lemma. *Sei I ein offenes Intervall und sei $f : \mathbb{R} \times I \to \mathbb{R}$ durch $f(t, y) = f(y)$ gegeben, wobei die Funktion $f : I \to \mathbb{R}$ stetig ist und $f \neq 0$ auf I erfüllt. Falls $\varphi : (\alpha, \beta) \to I$ eine Lösung von (3.1) ist, dann existiert ein $c \in \mathbb{R}$ so, dass für alle $t \in (\alpha, \beta)$ gilt:*

$$\varphi(t) = F^{-1}(t + c),$$

wobei F eine Stammfunktion von $\frac{1}{f}$ ist.

Beweis. Siehe Rechnungen vorher. ∎

Falls für ein $\hat{y} \in (a, b)$ gilt $f(\hat{y}) = 0$, ist $y(t) := \hat{y}$, $t \in \mathbb{R}$, eine maximale Lösung von (3.1) im Falle von zeitunabhängigen rechten Seiten. Es stellt sich die Frage, ob man in der Situation von Lemma 3.3 immer Lösungen in obiger Form finden kann.

3.4 Satz. *Erfülle f die Voraussetzungen von Lemma 3.3. Dann existiert für alle $(t_0, y_0) \in \mathbb{R} \times I$ eine eindeutige maximale Lösung $y : J_0 \to \mathbb{R}$ von (3.1) mit $y(t_0) = y_0$. Diese Lösung ist von der Form*

$$y(t) = F^{-1}(t - t_0), \qquad t \in J_0,$$

wobei F durch

$$F(y) := \int_{y_0}^{y} \frac{1}{f(x)}\, dx, \qquad y \in I$$

gegeben ist.

Bemerkung. Das Intervall J_0 aus Satz 3.4 ist durch $J_0 := \mathrm{Bild}(F) + t_0$ gegeben.

Beweis von Satz 3.4. Sei $I = (a, b)$. Wir definieren $F : I \to \mathbb{R}$ durch

$$F(y) := \int_{y_0}^{y} \frac{1}{f(x)}\, dx, \qquad y \in I.$$

Aufgrund der Eigenschaften von f ist F wohldefiniert. Nach dem Hauptsatz der Differential- und Integralrechnung gilt

$$F'(y) = \frac{1}{f(y)} \neq 0, \qquad y \in I.$$

Wieder mit der Stetigkeit von $\frac{1}{f}$ folgt die strikte Monotonie von F. Setzen wir $J := \mathrm{Bild}(F)$, so ist J aufgrund der strikten Monotonie ein offenes Intervall und die Umkehrfunktion $F^{-1} : J \to I$ existiert. Wegen $F(y_0) = 0$ folgt $0 \in J$.

Verschieben wir das Intervall J um t_0, so existieren $\alpha, \beta \in \mathbb{R}$ mit $\alpha < \beta$ und $(\alpha, \beta) = J + t_0$, d.h. es gilt $t \in (\alpha, \beta)$ genau dann, wenn $t - t_0 \in J$ gilt. Wir definieren nun

$$y(t) := F^{-1}(t - t_0), \qquad t \in (\alpha, \beta).$$

Dann gilt nach dem Satz über Ableitungen inverser Funktionen (Analysis I, Kapitel 6, Satz 1.4) für $t \in (\alpha, \beta)$

$$y'(t) = (F^{-1})'(t - t_0) = \frac{1}{F'(F^{-1}(t - t_0))} = \frac{1}{F'(y(t))} = f(y(t)).$$

Wegen $0 \in J$ und somit $t_0 \in (\alpha, \beta)$, sowie $F(0) = y_0$ folgt weiter

$$y(t_0) = F^{-1}(0) = y_0.$$

Somit ist y eine Lösung der gesuchten Form von (3.1) mit $y(t_0) = y_0$ auf dem Intervall (α, β). Nach Satz 2.18 existiert eine maximale Lösung $\varphi : (\widetilde{\alpha}, \widetilde{\beta}) \to I$ von (3.1) mit $t_0 \in (\widetilde{\alpha}, \widetilde{\beta})$ und $\varphi(t_0) = y_0$. Nach Lemma 3.3 folgt (mit der spezieller Wahl der Stammfunktion als unser konkretes F) die Existenz eines $c \in \mathbb{R}$ mit

$$\varphi(t) = F^{-1}(t + c), \qquad t \in (\widetilde{\alpha}, \widetilde{\beta}).$$

Wegen $t_0 \in (\alpha, \beta) \cap (\widetilde{\alpha}, \widetilde{\beta})$ folgt insbesondere

$$F^{-1}(t_0 + c) = \varphi(t_0) = y_0 = y(t_0) = F^{-1}(0).$$

Aufgrund der Injektivität von F^{-1} folgt $0 = t_0 + c$ bzw. $c = -t_0$. Es gilt also $y = \varphi$ auf $(\alpha, \beta) \cap (\widetilde{\alpha}, \widetilde{\beta})$ und nach Konstruktion der α, β, d.h. $(\alpha, \beta) = \mathrm{Bild}(F) + t_0$, bzw. den Eigenschaften einer maximalen Lösung folgt $\alpha = \widetilde{\alpha}$ sowie $\beta = \widetilde{\beta}$. Damit ist y eine maximale Lösung. Analog zeigt man die Eindeutigkeit dieser Lösung. ∎

Beispiele. Wir verwenden nun Satz 3.4, um Lösungen von (3.1) zu finden.

a) Wählen wir $f(t, y) = 1 + y^2$ und $(t_0, y_0) \in \mathbb{R} \times \mathbb{R}$. Wir suchen also eine Lösung von

$$\begin{aligned} y'(t) &= 1 + y^2, \\ y(t_0) &= y_0. \end{aligned}$$

Da arctan die Stammfunktion von $(1 + y^2)^{-1}$ ist und $\mathrm{Bild}(\tan) = \mathbb{R}$ gilt, existiert für alle $y_0 \in \mathbb{R}$ ein $c \in (-\frac{\pi}{2}, \frac{\pi}{2})$ mit $y_0 = \tan(c)$. Wir berechnen für $y \in \mathbb{R}$

$$F(y) = \int_{y_0}^{y} \frac{1}{1+x^2}\, dx = \arctan(y) - \arctan(y_0) = \arctan(y) - c.$$

Wegen $\text{Bild}(\arctan) = (-\frac{\pi}{2}, \frac{\pi}{2})$ folgt

$$\text{Bild}(F) = \big(-\frac{\pi}{2} - c, \frac{\pi}{2} - c\big)$$

und F^{-1} ist auf $\text{Bild}(F)$ durch

$$F^{-1}(t) = \tan(t + c)$$

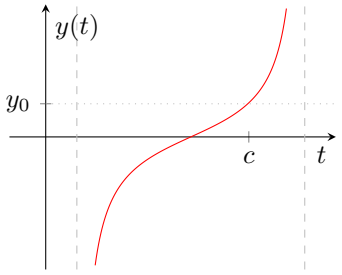
gegeben. Somit ist die maximale Lösung nach Satz 3.4 auf dem Intervall

$$J_0 := (-\frac{\pi}{2} + t_0 - c, \frac{\pi}{2} + t_0 - c)$$

durch

$$y(t) = \tan(t - t_0 + c)$$

gegeben. Die maximale Lösung ist folglich nur auf einem beschränkten Zeitintervall definiert. Obwohl f also auf ganz $\mathbb{R}^2$ definiert ist, kann hier keine Lösung für alle Zeiten $t \in \mathbb{R}$ gefunden werden.



b) Wählen wir $f(t, y) = 3\, y^{\frac{2}{3}}$, so gilt $f(y) = 0$ genau dann, wenn $y = 0$ gilt. Somit erfüllt f nicht die Voraussetzungen von Satz 3.4 auf dem Intervall $I = \mathbb{R}$. Wenden wir für $(t_0, y_0) \in \mathbb{R} \times (0, \infty)$ stattdessen Satz 3.4 auf $f|_{(0,\infty)}$ an, so ist F auf $(0, \infty)$ durch

$$F(y) = \int_{y_0}^{y} \frac{1}{3\, x^{\frac{2}{3}}}\, dx = y^{\frac{1}{3}} - y_0^{\frac{1}{3}}.$$

gegeben. Damit ist für $t \in (-y_0^{\frac{1}{3}}, \infty)$

$$F^{-1}(t) = (t + y_0^{\frac{1}{3}})^3$$

und die (zu $f|_{(0,\infty)} : \mathbb{R} \times (0, \infty) \to \mathbb{R}$ maximale) Lösung gemäß Satz 3.4 durch

$$y(t) = (t - t_0 + y_0^{\frac{1}{3}})^3, \qquad t \in (-y_0^{\frac{1}{3}} + t_0, \infty)$$

gegeben. Allerdings ist dies keine maximale Lösung des ursprünglichen Problems, denn

$$y_0(t) := \begin{cases} 0 & t \le -y_0^{\frac{1}{3}} + t_0, \\ (t - t_0 + y_0^{\frac{1}{3}})^3 & t > -y_0^{\frac{1}{3}} + t_0 \end{cases}$$

ist eine (maximale) Lösung auf $\mathbb{R}$. Betrachten wir analog $f|_{(-\infty,0)}$, so ist

$$y_1(t) = (t - t_1 + y_1^{\frac{1}{3}})^3, \qquad t \in (-\infty, -y_1^{\frac{1}{3}} + t_1)$$

eine Lösung für $(t_1, y_1) \in \mathbb{R} \times (-\infty, 0)$. Weiter ist für alle $(t_0, y_0) \in \mathbb{R} \times (0, \infty)$ und $(t_1, y_1) \in \mathbb{R} \times (-\infty, 0)$ mit $y_0^{\frac{1}{3}} + t_0 > -y_1^{\frac{1}{3}} + t_1$

$$y_2(t) := \begin{cases} (t - t_1 + y_1^{\frac{1}{3}})^3 & t < -y_1^{\frac{1}{3}} + t_1 \\ 0 & -y_1^{\frac{1}{3}} + t_1 \le t \le -y_0^{\frac{1}{3}} + t_0, \\ (t - t_0 + y_0^{\frac{1}{3}})^3 & t > -y_0^{\frac{1}{3}} + t_0 \end{cases}$$

eine Lösung auf dem Intervall $\mathbb{R}$. Es gibt folglich zu dem Anfangswert (t_0, y_0) unendlich viele maximale Lösungen.

c) Das Fischpopulationsmodell aus der Einführung hatte die Form $f(y) = y(b - c\,y) - H$ für $b, c, H > 0$ und kann folglich mit dem Satz 3.4 gelöst werden. Die Lösungen haben wir schon dort mit derselben Methode berechnet.

12.3.3 Separierbare rechte Seiten $f = h(t)g(y)$

Eine weitere Klasse von Gleichungen sind separierbare Differentialgleichungen erster Ordnung. Hierbei besitzt $f : J \times I \to \mathbb{R}$ die Form $f(t, y) = h(t)\, g(y)$, wobei $I = (a, b)$ und $J = (c, d)$ offene Intervalle sind und die Funktionen g, h dort stetig sind. Weiter gelte $g(y) \neq 0$ für alle $y \in I$. Wir können dabei ähnlich wie in Abschnitt 12.3.2 vorgehen. Falls $\varphi : (\alpha, \beta) \subseteq (c, d) \to (a, b)$ eine Lösung von (3.1) ist, d.h.

$$\varphi'(t) = h(t)\, g(\varphi(t)), \qquad t \in (\alpha, \beta)$$

gilt, so folgt

$$\frac{\varphi'(t)}{g(\varphi(t))} = h(t), \qquad t \in (\alpha, \beta).$$

Aufgrund der Stetigkeit der Funktionen g, h auf I bzw. J, sowie $g \neq 0$ auf I existieren Stammfunktionen G von $\frac{1}{g}$ und H von h. Damit folgt für $t \in (\alpha, \beta)$

$$\frac{d}{dt}\, G(\varphi(t)) = G'(\varphi(t))\varphi'(t) = \frac{\varphi'(t)}{g(\varphi(t))} = h(t) = \frac{d}{dt}\, H(t).$$

Nach dem Hauptsatz der Differential- und Integralrechnung existiert eine Konstante $c \in \mathbb{R}$ mit

$$G(\varphi(t)) = H(t) + c, \qquad t \in (\alpha, \beta).$$

Schließlich ist G aufgrund der Stetigkeit von g und $G' = \frac{1}{g} \neq 0$ strikt monoton und die Umkehrfunktion G^{-1} existiert. Somit folgt

$$\varphi(t) = G^{-1}(H(t) + c), \qquad t \in (\alpha, \beta).$$

3.5 Lemma. *Seien I, J offene Intervalle und sei $f : J \times I \to \mathbb{R}$ durch $f(t, y) = h(t)\, g(y)$ gegeben, wobei die Funktionen $g : I \to \mathbb{R}$, $g \neq 0$ und $h : J \to \mathbb{R}$ stetig sind. Falls $\varphi : (\alpha, \beta) \subseteq J \to I$ eine Lösung von (3.1) ist, dann existiert ein $c \in \mathbb{R}$ so, dass für alle $t \in (\alpha, \beta)$*

$$\varphi(t) = G^{-1}(H(t) + c)$$

gilt, wobei G eine Stammfunktion von $\frac{1}{g}$ und H eine Stammfunktion von h ist.

Beweis. Siehe obige Rechnung. ∎

3.6 Satz. *Erfülle f die Voraussetzungen von Lemma 3.5. Dann existiert für alle $(t_0, y_0) \in J \times I$ eine eindeutige maximale Lösung $y : J_0 \to I$ von (3.1) mit $y(t_0) = y_0$. Diese Lösung ist von der Form*

$$y(t) = G^{-1}\big(H(t)\big),$$

wobei $G(y) = \int_{y_0}^{y} \frac{1}{g(x)}\, dx$ für $y \in I$ und $H(t) = \int_{t_0}^{t} h(s)\, ds$ für $t \in J$ ist.

Bemerkung. Das Intervall J_0 aus Satz 3.6 ist dabei das größte offene Intervall, welches in $H^{-1}(\mathrm{Bild}(G))$ liegt und t_0 beinhaltet.

Beweis von Satz 3.6. Sei $I = (a, b)$, $J = (c, d)$ und $(t_0, y_0) \in J \times I$ sowie

$$G(y) := \int_{y_0}^{y} \frac{1}{g(x)}\, dx, \qquad y \in I$$
$$H(t) := \int_{t_0}^{t} h(s)\, ds, \qquad t \in J.$$

Wir definieren $K := \mathrm{Bild}(G)$. Es gilt $G' = \frac{1}{g} \neq 0$ auf I und folglich ist $G : I \to K$ strikt monoton sowie K ein offenes Intervall. Insbesondere existiert die Umkehrfunktion $G^{-1} : K \to I$ und wegen $G(y_0) = 0$ folgt $0 \in K$. Definieren wir

$$M := \big\{ t \in J \;\big|\; H(t) \in K \big\} = H^{-1}(K),$$

so ist aufgrund der Stetigkeit von H die Menge M offen. Weiter folgt wegen $H(t_0) = 0$, dass $t_0 \in M$ gilt. Sei nun (α, β) das größte offene Intervall, welches in M liegt und t_0 beinhaltet. Damit ist

$$y(t) := G^{-1}(H(t)), \qquad t \in (\alpha, \beta)$$

wohldefiniert und es folgt mit dem Satz über die Ableitung einer inversen Funktion für $t \in (\alpha, \beta)$

$$y'(t) = (G^{-1})'(H(t))H'(t) = \frac{h(t)}{G'(G^{-1}(H(t)))} = \frac{h(t)}{G'(y(t))} = g(y(t))h(t).$$

Weiter gilt

$$y(t_0) = G^{-1}(H(t_0)) = G^{-1}(0) = y_0.$$

Somit ist y eine Lösung von (3.1) mit $y(t_0) = y_0$ auf dem Intervall (α, β). Nach Satz 2.18 existiert eine maximale Lösung $\varphi : (\widetilde{\alpha}, \widetilde{\beta}) \to I$ von (3.1) mit $y(t_0) = y_0$. Nach Lemma 3.5 existiert ein $c \in \mathbb{R}$ mit

$$\varphi(t) = G^{-1}(H(t) + c), \qquad t \in (\widetilde{\alpha}, \widetilde{\beta}),$$

wobei G und H wie oben definiert sind. Insbesondere gilt $t_0 \in (\widetilde{\alpha}, \widetilde{\beta}) \cap (\alpha, \beta)$ und somit

$$G^{-1}(H(t_0) + c) = \varphi(t_0) = y_0 = y(t_0) = G^{-1}(H(t_0)) = G^{-1}(0).$$

Aufgrund der Injektivität von G^{-1} folgt

$$0 = H(t_0) + c = c$$

und somit $c = 0$. Damit gilt $\varphi = y$ auf $(\widetilde{\alpha}, \widetilde{\beta}) \cap (\alpha, \beta)$. Nach der Definition des Intervalls (α, β) und $(\widetilde{\alpha}, \widetilde{\beta}) \subseteq M$ folgt $(\widetilde{\alpha}, \widetilde{\beta}) \subseteq (\alpha, \beta)$ und somit $\alpha = \widetilde{\alpha}$ bzw. $\beta = \widetilde{\beta}$. Wir haben also gezeigt, dass y eine maximale Lösung ist. Analog zeigt man die Eindeutigkeit dieser Lösung. ∎

Falls für ein $\hat{y} \in I$ gilt $g(\hat{y}) = 0$, ist $y(t) := \hat{y}$, $t \in J$, eine maximale Lösung von (3.1) im Falle von separierbaren rechten Seiten.

Beispiel. Wir betrachten das Anfangswertproblem

$$\begin{aligned} y'(t) &= 2\,t\,(1 + y^2), \\ y(t_0) &= y_0 \end{aligned}$$

mit $(t_0, y_0) \in \mathbb{R}^2$. Mit der Notation von Satz 3.6 gilt $g : \mathbb{R} \to \mathbb{R} : y \mapsto 1 + y^2$ sowie $h : \mathbb{R} \to \mathbb{R} : t \mapsto 2t$. Damit folgt

$$G(y) = \int_{y_0}^{y} \frac{1}{1+x^2}\, dx = \arctan(y) - \arctan(y_0) =: \arctan(y) - c_0,$$
$$H(t) = \int_{t_0}^{t} 2s\, ds = t^2 - t_0^2.$$

Weiter gilt $\text{Bild}(G) = (-\frac{\pi}{2} - c_0, \frac{\pi}{2} - c_0)$ mit $c_0 \in (-\frac{\pi}{2}, \frac{\pi}{2})$. Zu der Bestimmung von $H^{-1}(\text{Bild}(G))$ benötigen wir nun eine Fallunterscheidung:

1. **Fall $-t_0^2 > -\frac{\pi}{2} - c_0$:** Die Urbildmenge ist in diesem Fall durch ein Intervall gegeben (vergleiche Abbildung 3.1). Setzen wir $\alpha = \sqrt{\frac{\pi}{2} - c_0 + t_0^2}$, so gilt $H^{-1}(\text{Bild}(G)) = (-\alpha, \alpha)$. Die maximale Lösung gemäß Satz 3.6 ist also durch

$$y(t) = \tan(t^2 - t_0^2 + \arctan(y_0)) \qquad t \in (-\alpha, \alpha)$$

gegeben. Offenbar gilt

$$\lim_{t \nearrow \alpha} y(t) = \infty \qquad \text{und} \qquad \lim_{t \searrow -\alpha} y(t) = \infty$$

(vergleiche Abbildung 3.1).

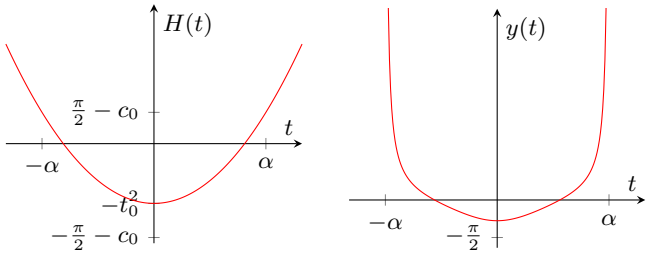


Abb. 3.1. Die Graphen von H und y mit $c_0 = \frac{\pi}{4}$ und $t_0^2 = 1.5$.

2. **Fall $-t_0^2 \leq -\frac{\pi}{2} - c_0$:** Die Urbildmenge besteht aus zwei Intervallen (vergleiche Abbildung 3.2). Setzen wir

$$\alpha := \sqrt{-\frac{\pi}{2} - c_0 + t_0^2} \quad \text{und} \quad \beta := \sqrt{\frac{\pi}{2} - c_0 + t_0^2},$$

so ist $H^{-1}(\text{Bild}(G)) = (-\beta, -\alpha) \cup (\alpha, \beta)$. Sei oBdA $t_0 \in (\alpha, \beta)$, so ist die maximale Lösung gemäß Satz 3.6 durch

$$y(t) = \tan(t^2 - t_0^2 + \arctan(y_0)) \qquad t \in (\alpha, \beta)$$

gegeben. Offenbar gilt

$$\lim_{t \nearrow \beta} y(t) = \infty \qquad \text{und} \qquad \lim_{t \searrow \alpha} y(t) = -\infty$$

(vergleiche Abbildung 3.2).

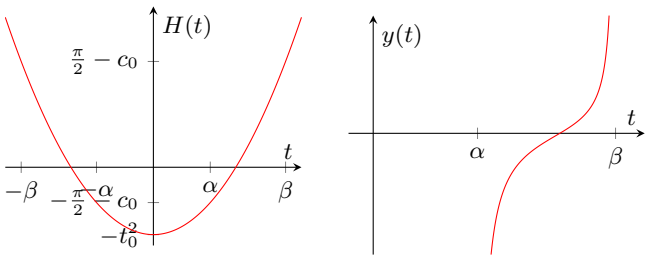


Abb. 3.2. Die Graphen von H und y mit $c_0 = -\frac{\pi}{4}$ und $t_0 = \sqrt{1.5}$.

12.3.4 Lineare Gleichungen

Wir betrachten rechte Seiten der Form $f(t, y) = h(t)\, y + p(t)$, wobei die Funktionen $p, h : I = (a, b) \to \mathbb{R}$ stetig sind und somit $f : I \times \mathbb{R} \to \mathbb{R}$ die Voraussetzung (S) erfüllt. Dabei unterscheiden wir zwei Typen von Gleichungen: Gilt

i) $p(t) \equiv 0$, so nennen wir (3.1) eine *homogene Gleichung,*

ii) $p(t) \not\equiv 0$, so nennen wir (3.1) eine *inhomogene Gleichung.*

Der Name *lineare Gleichung* entspringt dabei folgende Anschauung: Schreiben wir (3.1) zu

$$y'(t) - h(t)\, y(t) = p(t)$$

um, so erhalten wir Lösungen dieser Gleichung, wenn wir das Urbild der rechten Seite $p \in C^0(I)$ besüglich des Operators

$$L : C^1(I) \to C^0(I) : y \mapsto y' - h\, y,$$

d.h. $L(y)(t) := y'(t) - h(t)\, y(t)$, $t \in I$, $y \in C^1(I)$, suchen. Der Operator L ist linear, es gilt also

$$L(\alpha\, y + \beta\, z) = \alpha\, L(y) + \beta\, L(z) \qquad \alpha, \beta \in \mathbb{R},\ y, z \in C^1(I).$$

Suchen wir nun eine Lösung der homogenen Gleichung, so ist dies ein Spezialfall der separierbaren Differentialgleichungen. Mit der Notation aus Abschnitt 12.3.3 gilt "$h(t) = h(t)$" sowie $g(y) = y$, allerdings können wir aufgrund von $g(0) = 0$ den Satz 3.6 nicht anwenden. Wir argumentieren dennoch ähnlich, um einen Kandidaten für die Lösung zu erhalten. Eine Stammfunktion von g ist auf $\mathbb{R} \setminus \{0\}$ durch $G(y) := \ln|y|$ gegeben und es folgt $G^{-1}(t) = e^t$. Weiter sei H eine Stammfunktion von h. Für $y \neq 0$ können wir (3.1) wieder zu

$$\frac{dy}{y} = h\,dt$$

umschreiben. Bilden wir die Stammfunktionen auf beiden Seiten, so existiert ein $c \in \mathbb{R}$ mit

$$\ln|y(t)| = H(t) + c$$

bzw.

$$|y(t)| = e^c e^{H(t)}.$$

Wir können wir den Betrag vernachlässigen, falls wir e^c durch eine Konstante $\widetilde{c} \in \mathbb{R} \setminus \{0\}$ ersetzen. Wir suchen also eine Lösung der Form

$$y(t) = \widetilde{c}\, e^{H(t)}.$$

Daher definieren wir zu einen Anfangswert $(t_0, y_0) \in I \times (\mathbb{R} \setminus \{0\})$ die Funktion

$$y(t) := y_0\, e^{\widetilde{H}(t)}, \qquad t \in I$$

mit der speziellen Stammfunktion

$$\widetilde{H}(t) := \int_{t_0}^{t} h(s)\,ds, \qquad t \in I$$

von h. Da y auf ganz I definiert ist, sieht man sofort, dass die Funktion $y : I \to \mathbb{R}$ eine maximale Lösung von (3.1) mit $y(t_0) = y_0$ und $y \neq 0$ ist. Für den Anfangswert $(t_0, 0) \in I \times \{0\}$ ist offensichtlich die Funktion

$$y : I \to \mathbb{R}, \quad y(t) := 0$$

maximale Lösung von (3.1) mit $y(t_0) = 0 = y_0$. Tatsächlich wird dieser Fall ebenfalls von obiger Formel abgedeckt. Schließlich bemerken wir, dass die Funktion g Lipschitz-Stetig in y ist und nach Satz 2.20 die gefundenen Lösungen eindeutig sind. Bevor wir uns den inhomogenen Gleichungen zuwenden betrachten wir folgendes Lemma.

3.7 Lemma. *Sei $I = (a, b)$ und für $f : I \times \mathbb{R} \to \mathbb{R}$ gelte $f(t, y) = h(t)\,y + p(t)$, wobei p und h auf I stetig sind. Seien y_1, y_2 Lösungen der inhomogenen Gleichung (3.1) und sei y_0 eine Lösung der homogenen Gleichung (3.1). Dann gilt:*

i) $y_1 - y_2$ *ist eine Lösung der homogenen Gleichung.*

ii) $y_1 + y_0$ *ist eine Lösung der inhomogenen Gleichung.*

Beweis. Dies folgt durch triviales Einsetzen. ∎

Anders formuliert bedeutet dies

$$\text{Lösung inhomogene Gleichung} = \text{Lösung homogene Gleichung} + \text{Lösung inhomogene Gleichung},$$

man kann also eine *spezielle* Lösung der inhomogenen Gleichung suchen und dazu alle Lösungen der homogenen Gleichung addieren.

Um die inhomogene Gleichung zu lösen, verwenden wir folgende, auf Lagrange zurückzuführende Idee der "Variation der Konstanten": Wir benutzen die Formel für die homogene Gleichung, wobei wir die Konstante $\widetilde{c}$ durch eine Funktion $c(t)$ ersetzen. Wir betrachten also folgenden Ansatz

$$y(t) = c(t)\, e^{H(t)}.$$

Damit gilt

$$y'(t) = c'(t)\, e^{H(t)} + c(t)\, e^{H(t)}\, h(t).$$

Um die Gleichung (3.1) zu erfüllen, muss also

$$p(t) + h(t)\, y = y'(t) = c'(t)\, e^{H(t)} + c(t)\, e^{H(t)}\, h(t)$$

gelten. Wir müssen also $c'(t) = p(t)\, e^{-H(t)}$ wählen, d.h. wir wählen c als Stammfunktion von $p(t)\, e^{-H(t)}$. Dies fassen wir in folgendem Satz zusammen.

3.8 Satz. *Sei* $I = (a, b)$ *und für* $f : I \times \mathbb{R} \to \mathbb{R}$ *gelte* $f(t, y) = h(t)\, y + p(t)$, *wobei* p *und* h *auf* I *stetig sind. Dann existiert für alle* $(t_0, y_0) \in I \times \mathbb{R}$ *eine eindeutige maximale Lösung* $y : I \to \mathbb{R}$ *von (3.1) mit* $y(t_0) = y_0$. *Diese Lösung ist von der Form*

$$y(t) = e^{H(t)} \big(y_0 + \int_{t_0}^{t} p(s)\, e^{-H(s)}\, ds\big),$$

wobei $H : I \to \mathbb{R}$ *durch*

$$H(t) = \int_{t_0}^{t} h(s)\, ds$$

gegeben ist.

Beweis. Es gilt

$$y(t_0) = e^{H(t_0)}(y_0 + 0) = e^0\, y_0 = y_0.$$

Zusammen mit den obigen Rechnungen ist $y : I \to \mathbb{R}$ folglich eine maximale Lösung von (3.1). Seien weiter y_1, y_2 Lösungen von (3.1) mit $y_i(t_0) = y_0$, $i = 1, 2$, so ist $y_0 := y_1 - y_2$ eine Lösung der homogenen Gleichung mit $y_0(t_0) = 0$. Wir hatten bereits gesehen, dass $y_0(t) = 0$ die eindeutige Lösung der homogenen Gleichung mit $y_0(t_0) = 0$ ist. Damit folgt $y_1 = y_2$ und die Lösung y ist eindeutig. ∎

13 Systeme linearer Differentialgleichungen

13.1 Grundlagen

Wir betrachten Systeme von Differentialgleichungen 1. Ordnung der Form

$$Y'(t) = \mathbf{A}(t)\,Y(t) + B(t) \tag{1.1}$$

mit $\mathbf{A}(t) \in \mathbb{R}^{n\times n}$, $B(t) \in \mathbb{R}^n$ sowie $Y(t) \in \mathbb{R}^n$ auf einem Intervall $I = (a, b)$[1]. Definieren wir $F : I \times \mathbb{R}^n \to \mathbb{R}^n$ durch $F(t, Y) := \mathbf{A}(t)\,Y + B(t)$, so ist dies ein Spezialfall der Gleichung (2.3) aus Kapitel 12. Fordern wir also die Stetigkeit der Funktionen $\mathbf{A} : I \to \mathbb{R}^{n\times n}$ und $B : I \to \mathbb{R}^n$, so erfüllt F die Voraussetzung (S) auf $I \times \mathbb{R}^n$ und ist insbesondere lokal Lipschitz stetig bezüglich Y. Nach Satz 2.20 aus Kapitel 12 existiert also für alle $t_0 \in I$ und $Y_0 \in \mathbb{R}^n$ genau eine maximale Lösung von (1.1) mit $Y(t_0) = Y_0$. Wir wollen nun zeigen, dass diese Lösung auf ganz I definiert ist.

1.2 Lemma (Gronwall)**.** *Sei J ein Intervall, $t_0 \in J$ und $\alpha, \beta \in [0, \infty)$. Ferner sei $x : J \to [0, \infty)$ stetig und erfülle*

$$x(t) \le \alpha + \beta \left| \int_{t_0}^{t} x(s)\, ds \right| \qquad t \in J.$$

Dann gilt

$$x(t) \le \alpha\, e^{\beta\,|t-t_0|} \qquad t \in J.$$

Beweis. Wir betrachten zunächst den Fall $t \ge t_0$ für $t \in J$. Dazu definieren wir

$$h(s) := \beta\, e^{\beta\,(t_0-s)} \int_{t_0}^{s} x(\tau)\, d\tau \qquad s \in [t_0, t].$$

Dann gilt mit dem Hauptsatz der Differential- und Integralrechnung

[1] Wir lassen hier $a = -\infty$ sowie $b = \infty$ zu.

$$\begin{aligned} h'(s) &= \beta\, e^{\beta\,(t_0-s)}\,(-1)\,\beta \int_{t_0}^{s} x(\tau)\,d\tau + \beta\, e^{\beta\,(t_0-s)} x(s) \\ &= -\beta\, h(s) + \beta\, e^{\beta\,(t_0-s)} x(s) \\ &\le -\beta\, h(s) + \alpha\,\beta\, e^{\beta\,(t_0-s)} + \beta\,\beta\, e^{\beta\,(t_0-s)} \Big| \int_{t_0}^{s} x(\tau)\,d\tau \Big| \\ &= -\beta\, h(s) + \alpha\,\beta\, e^{\beta\,(t_0-s)} + \beta\, h(s) \\ &= \frac{d}{ds}\big(-\alpha\, e^{\beta\,(t_0-s)}\big). \end{aligned}$$

Integrieren wir nun über das Intervall $[t_0, t]$ so folgt

$$\int_{t_0}^{t} h'(s)\,ds = h(t) - h(t_0) = h(t) = \beta\, e^{\beta\,(t_0-t)} \int_{t_0}^{t} x(\tau)\,d\tau$$

bzw.

$$\int_{t_0}^{t} \frac{d}{ds}\big(-\alpha\, e^{\beta\,(t_0-s)}\big)\,ds = -\alpha\, e^{\beta\,(t_0-t)} + \alpha\, e^{\beta\,(t_0-t_0)} = \alpha - \alpha\, e^{\beta\,(t_0-t)}.$$

Damit folgt mit der Monotonie des Integrals

$$\beta\, e^{\beta\,(t_0-t)} \int_{t_0}^{t} x(\tau)\,d\tau \le \alpha - \alpha\, e^{\beta\,(t_0-t)}$$

und somit

$$\beta \int_{t_0}^{t} x(\tau)\,d\tau \le \alpha\, e^{\beta\,(t-t_0)} - \alpha.$$

Verwenden wir noch die Voraussetzung, so haben wir

$$x(t) \le \alpha + \beta \int_{t_0}^{t} x(\tau)\,d\tau \le \alpha\, e^{\beta\,(t-t_0)}$$

gezeigt. Analog folgt der Fall $t < t_0$ für $t \in J$. ∎

1.3 Satz. *Seien* $\mathbf{A}$, B *stetige Funktionen auf dem Intervall* $I = (a, b)$. *Dann ist jede maximale Lösung von (1.1) auf ganz* (a, b) *definiert.*

Beweis. Sei $Y : (\alpha, \beta) \to \mathbb{R}^n$ eine maximale Lösung von (1.1) mit $Y(t_0) = Y_0$ für ein $t_0 \in (\alpha, \beta)$ und ein $Y_0 \in \mathbb{R}^n$. Da Y eine maximale Lösung ist kann $\lim_{t \nearrow \beta} Y(t)$ nicht existieren. Anderenfalls gäbe es nach Lemma 2.14 aus Kapitel 12 eine Fortsetzung von Y über β hinaus. Dies ist aber nach Folgerung 2.19 aus Kapitel 12 ein Widerspruch zur Maximalität der Lösung. Gelte nun $(\alpha, \beta) \subsetneq (a, b)$, wobei wir oBdA $\beta < b$ und somit insbesondere $\beta < \infty$ annehmen können. Zu $n \in \mathbb{N}$ definieren wir

$$K_n := \big\{(t,Y) \in \mathbb{R}^{n+1} \;\big|\; t \in [t_0,\beta],\, \|Y\| \le n\big\}.$$

Offenbar ist K_n eine kompakte Teilmenge von $(a,b)\times\mathbb{R}^n$. Nach der Definition der maximalen Lösung (siehe Satz 2.18 aus Kapitel 12) ist

$$G^+ = \big\{(t,Z) \in \overline{\operatorname{graph}(Y)} \;\big|\; t \ge t_0\big\}$$

keine kompakte Teilmenge von $\Omega = (a,b)\times\mathbb{R}^n$. Es gilt also $G^+ \not\subseteq K_n$. Da $\lim_{t\nearrow\beta} Y(t)$ nicht existiert, muss $\tau_n \in [t_0,\beta)$ existieren mit $\|Y(\tau_n)\| = n$. Somit existiert der Grenzwert

$$\lim_{n\to\infty} \|Y(\tau_n)\| = \infty.$$

Dies wollen wir zum Widerspruch führen. Dafür definieren wir

$$\begin{aligned} n(t) &:= \|Y(t)\|, \qquad t \in (\alpha,\beta),\\ \delta &:= \max_{t\in[t_0,\beta]} \|\mathbf{A}(t)\|_{\mathbb{R}^{n\times n}},\\ \gamma &:= \max_{t\in[t_0,\beta]} \|B(t)\|_{\mathbb{R}^n}. \end{aligned}$$

Aufgrund der Stetigkeit von Y, $\mathbf{A}$ und B ist n stetig sowie δ,γ wohldefiniert und endlich. Da Y Lösung von (1.1) ist, löst Y nach Lemma 2.7 aus Kapitel 12 die Integralgleichung

$$Y(t) = Y(t_0) + \int_{t_0}^{t} \mathbf{A}(s)\,Y(s) + B(s)\,ds, \qquad t \in [t_0,\beta).$$

Mit der Cauchy-Schwarzschen Ungleichung und der Monotonie des Integrals folgt

$$\|Y(t)\| \le \|Y(t_0)\| + \int_{t_0}^{t} \|\mathbf{A}(s)\|\,\|Y(s)\| + \|B(s)\|\,ds, \qquad t \in [t_0,\beta)$$

und somit

$$n(t) \le \|Y(t_0)\| + (\beta - t_0)\gamma + \delta \int_{t_0}^{t} n(s)\,ds, \qquad t \in [t_0,\beta).$$

Wenden wir nun das Gronwallsche Lemma mit $x(t) = n(t)$, $\beta = \delta$ und $\alpha = \|Y(t_0)\| + (\beta - t_0)\gamma$ an, so folgt

$$\|Y(t)\| = n(t) \le \Big(\|Y(t_0)\| + (\beta - t_0)\gamma\Big) e^{\delta|t-t_0|}, \qquad t \in [t_0,\beta).$$

Dabei ist die rechte Seite stetig auf dem Intervall $[t_0,\beta]$ und somit gleichmäßig beschränkt, ein Widerspruch zu $\lim_{n\to\infty}\|Y(\tau_n)\| = \infty$. Die Annahme $(\alpha,\beta) \subsetneq (a,b)$ war somit falsch und die maximale Lösung ist folglich auf dem ganzen Intervall (a,b) definiert. ∎

1.4 Definition. *Das System (1.1) heißt* homogen*, falls $B(t) \equiv 0$ gilt. Das homogene System (1.1) reduziert sich also zu*

$$Y'(t) = \mathbf{A}(t)\, Y(t). \tag{1.5}$$

Ist $B(t) \not\equiv 0$, so heißt das System inhomogen*.*

Wir wenden uns zunächst den homogenen Systemen zu.

13.2 Homogene Systeme

Die (maximalen) Lösungen des homogenen Systems (1.5) bilden einen Vektorraum. In der Tat, seien Y_1, Y_2 Lösungen von (1.5), so folgt für $\alpha \in \mathbb{R}$

$$\begin{aligned}\frac{d}{dt}\big(Y_1(t) + Y_2(t)\big) &= Y_1'(t) + Y_2'(t)\\ &= \mathbf{A}(t)\, Y_1(t) + \mathbf{A}(t)\, Y_2(t)\\ &= \mathbf{A}(t)\, (Y_1(t) + Y_2(t))\end{aligned}$$

sowie

$$\frac{d}{dt}\big(\alpha\, Y_1(t)\big) = \alpha Y_1'(t) = \alpha\, \mathbf{A}(t)\, Y_1(t) = \mathbf{A}(t)\, \big(\alpha\, Y_1(t)\big).$$

Dieser Vektorraum ist ein Teilraum von $C^1(I)$.

2.1 Lemma. *Sei $I = (a, b)$ und $\mathbf{A} : I \to \mathbb{R}^{n\times n}$ eine stetige Funktion. Dann hat der Vektorraum der maximalen Lösungen des homogenen Systems (1.5) die Dimension n.*

Beweis. Sei $I = (a, b)$ und $t_0 \in I$ fest und sei E_i, $i = 1, \dots, n$ die Standardbasis von $\mathbb{R}^n$. Dann gilt:

1. Es existieren n linear unabhängige Lösungen:
   Seien $Z_i : I \to \mathbb{R}^{n\times n}$ die maximalen Lösungen von

   $$\begin{aligned}Z_i'(t) &= \mathbf{A}(t)\, Z_i(t)\\ Z_i(t_0) &= E_i\end{aligned}$$

   für $i = 1, \dots, n$. Dann sind die Z_i als Funktionen auf I linear unabhängig, da diese in t_0 linear unabhängige Vektoren sind.

2. Die Z_i, $i = 1, \dots, n$ bilden eine Basis:
   Sei Y eine Lösung des homogenen Systems mit

$$Y(t_0) = (y_1(t_0), \dots, y_n(t_0)).$$

   Dann ist

$$Z(t) := \sum_{i=1}^{n} y_i(t_0)\, Z_i(t) \qquad t \in I$$

   Lösung des homogenen Systems und es gilt

$$Z(t_0) = \sum_{i=1}^{n} y_i(t_0)\, E_i = Y(t_0).$$

   Aufgrund der Eindeutigkeit der Lösung mit Anfangswert $(t_0, Y(t_0))$ folgt $Y = Z$. Die Z_i, $i = 1, \dots, n$ bilden also eine Basis. ∎

2.2 Definition. *Sei $I = (a, b)$ und $\mathbf{A} : I \to \mathbb{R}^{n\times n}$ eine stetige Funktion. Jede Basis des Lösungsraumes des homogenen Systemes (1.5) nennen wir* Fundamentalsystem *von (1.5). Sei $Y^{(i)}$, $i = 1, \dots, n$ ein Fundamentalsystem von (1.5), dann heißt die Matrix*

$$\mathbf{Y} := \left(Y^{(1)}, \dots, Y^{(n)}\right) = \begin{pmatrix} Y_1^{(1)} & \cdots & Y_1^{(n)} \\ \vdots & & \vdots \\ Y_n^{(1)} & \cdots & Y_n^{(n)} \end{pmatrix}$$

Fundamentalmatrix *von (1.5).*

Bemerkungen. 1. Sei $\mathbf{Y} = (Y_{i,j})_{i,j=1,\dots,n}$ eine Fundamentalmatrix von (1.5). Setzen wir $\mathbf{Y}' := (Y'_{ij})_{i,j=1,\dots,n}$, so folgt $\mathbf{Y}' = \mathbf{A}\,\mathbf{Y}$.

2. Sei $t_0 \in I$ und $\mathbf{Y}$ eine Fundamentalmatrix von (1.5). Gilt $\mathbf{Y}(t_0) = \mathbf{E}$, wobei $\mathbf{E}$ die Einheitsmatrix bezeichnet, so ist die Lösung von

$$Y' = \mathbf{A}\,Y, \qquad Y(t_0) = Y_0$$

   durch $Y(t) := \mathbf{Y}(t)\,Y_0$ gegeben (vergleiche Beweis von Lemma 2.1).

3. Sei $t_0 \in I$ und $\mathbf{Y}$ eine Fundamentalmatrix von (1.5). Für alle t_0 ist die Matrix $\mathbf{Y}(t_0)$ invertierbar. Anderenfalls gäbe es eine nichttriviale Linearkombination der Spaltenvektoren $Y^{(i)}(t_0)$, d.h. $\sum_{i=1}^{n} \lambda_i Y^{(i)}(t_0) = 0$. Dann würde $Y(t) := \sum_{i=1}^{n} \lambda_i Y^{(i)}(t)$ das System $Y' = \mathbf{A}\,Y$ mit $Y(t_0) = 0$ lösen. Da die Lösungen von (1.5) eindeutig sind, folgt $Y(t) = 0$, was ein Widerspruch zur Definition des Fundamentalsystems wäre. Definieren wir

$$\mathbf{Z}(t) := \mathbf{Y}(t)\,\mathbf{Y}^{-1}(t_0),$$

so gilt

$$\mathbf{Z}' = \mathbf{Y}'\,\mathbf{Y}^{-1}(t_0) = \mathbf{A}\,\mathbf{Y}\,\mathbf{Y}^{-1}(t_0) = \mathbf{A}\,\mathbf{Z}.$$

Insbesondere gilt $\mathbf{Z}(t_0) = \mathbf{E}$. Damit ist $\mathbf{Z}$ wieder eine Fundamentalmatrix von (1.5). Die Lösung von

$$Y' = \mathbf{A}\,Y, \qquad Y(t_0) = Y_0$$

ist also durch

$$Y(t) = \mathbf{Y}(t)\,\mathbf{Y}^{-1}(t_0)\,Y_0$$

gegeben.

4. Im Allgemeinen existiert keine Methode die Fundamentalmatrix zu bestimmen, falls $n > 1$ gilt und $\mathbf{A}(t)$ nicht konstant ist.

2.3 Definition. *Sei $I = (a, b)$ und $\mathbf{A} : I \to \mathbb{R}^{n\times n}$ eine stetige Funktion. Sei $\mathbf{Y}$ eine Fundamentalmatrix des homogenen Systems (1.5). Wir nennen*

$$W(t) := \det \mathbf{Y}(t), \qquad t \in I$$

die Wronski-Determinante.

2.4 Satz. *Sei $I = (a, b)$, $t_0 \in I$ und $\mathbf{A} : I \to \mathbb{R}^{n\times n}$ eine stetige Funktion. Sei $\mathbf{Y}$ eine Fundamentalmatrix des homogenen Systems (1.5). Dann gilt für die zugehörige Wronski-Determinante*

$$W(t) = W(t_0)\, e^{\int_{t_0}^{t} \operatorname{tr} \mathbf{A}(s)\, ds}, \qquad t \in I,$$

wobei für $\mathbf{A} = (a_{ij})_{i,j=1\ldots,n}$ die Spur der Matrix durch $\operatorname{tr}\mathbf{A} = \sum_{i=1}^{n} a_{ii}$ definiert ist.

Beweis. Es gilt $\operatorname{tr}\mathbf{A} \in \mathbb{R}$ sowie $W = \det \mathbf{Y} \in \mathbb{R}$, wir betrachten also eine skalare Gleichung. Nach der Theorie der linearen Gleichungen (siehe Satz 3.8 aus Kapitel 12) gilt die gewünschte Formel (als Lösungsformel von (3.1) aus Kapitel 12 mit Anfangswert $W(t_0)$) genau dann, wenn

$$W'(t) = \operatorname{tr}\mathbf{A}(t)\, W(t), \qquad t \in I$$

gilt. Wir benötigen also eine Formel für die Ableitung der Determinante. Es gilt für $\mathbf{B} = (B_1, \ldots, B_n) = (b_{ij})_{i,j=1,\ldots,n}$

$$\det \mathbf{B} = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma)\, b_{1\sigma(1)} \cdot \ldots \cdot b_{n\sigma(n)},$$

wobei S_n die symmetrische Gruppe vom Grad n ist und das Signum der Permutation σ durch $\operatorname{sgn}(\sigma) := \det(E_{\sigma(1)}, \dots, E_{\sigma(n)}) \in \{-1, 1\}$ gegeben ist. Somit gilt

$$\begin{aligned}(\det \mathbf{B})' &= \sum_{i=1}^{n} \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, b_{1\sigma(1)} \cdot \ldots \cdot b'_{i\sigma(i)} \cdot \ldots \cdot b_{n\sigma(n)} \\ &= \sum_{i=1}^{n} \det(B_1, \dots, B'_i, \dots, B_n).\end{aligned}$$

Wählen wir nun ein beliebiges $t_1 \in I$, so können wir wie in den obigen Bemerkungen eine Fundamentalmatrix $\mathbf{Z}(t) = (Z_1(t), \dots, Z_n(t))$ des homogenen Systems (1.5) mit $\mathbf{Z}(t_1) = \mathbf{E}$, also $Z_i(t_1) = E_i$ konstruieren. Wie wir gesehen hatten gilt $Z'_i(t_1) = \mathbf{A}(t_1)\, Z_i(t_1) = \mathbf{A}(t_1)\, E_i$ sowie

$$\mathbf{Y}(t) = \mathbf{Z}(t)\, \mathbf{Y}(t_1), \qquad t \in I.$$

Für die Wronski-Determinante folgt

$$W(t) = \det \mathbf{Y}(t) = \det \mathbf{Z}(t) \det \mathbf{Y}(t_1) = \det \mathbf{Z}(t)\, W(t_1), \qquad t \in I$$

und somit

$$W'(t) = \big(\det \mathbf{Z}(t)\big)'\, W(t_1), \qquad t \in I.$$

Mit der Formel für die Ableitung der Determinante erhalten wir

$$\begin{aligned}(\det \mathbf{Z}(t_1))' &= \sum_{i=1}^{n} \det(Z_1(t_1), \dots, Z'_i(t_1), \dots Z_n(t_1)) \\ &= \sum_{i=1}^{n} \det(E_1, \dots, \mathbf{A}(t_1)\, E_i, \dots, E_n) \\ &= \sum_{i=1}^{n} a_{ii}(t_1) \\ &= \operatorname{tr} \mathbf{A}(t_1).\end{aligned}$$

Es gilt folglich

$$W'(t_1) = \operatorname{tr} \mathbf{A}(t_1)\, W(t_1).$$

Da $t_1 \in I$ beliebig war, folgt also mit Satz 3.8 die Behauptung. ∎

2.5 Folgerung. *Die Wronski-Determinante einer Fundamentalmatrix ist überall ungleich Null.*

13.3 Inhomogene Systeme

Man sieht leicht, dass eine analoge Version von Lemma 3.7 aus Kapitel 12 auch für Systeme von linearen Differentialgleichungen 1. Ordnung gilt: Falls Y_1, Y_2 Lösungen des inhomogenen Systemes sind, so ist $Y_1 - Y_2$ eine Lösung des homogenen Systemes. Falls weiter Y_0 eine Lösung des homogenen Systemes ist, so ist $Y_0 + Y_1$ eine Lösung des inhomogenen Systemes. Anders ausgedrückt:

$$\begin{aligned}&\text{Lösungen des inhomogenen Systems}\\ &\quad = \text{ spezielle Lösung des inhomogenen Systems}\\ &\qquad + \text{ Lösungen des homogenen Systems.}\end{aligned}$$

Wir können also wie im Fall $n = 1$ die Methode der Variation der Konstanten verwenden. Ersetzen wir die Konstante Y_0 mit $C(t)$ in der Lösungsformel, betrachten wir also

$$Z(t) = \mathbf{Y}(t)\,\mathbf{Y}^{-1}(t_0)\,C(t),$$

so kommen wir auf die Lösungsformel im folgenden Satz.

3.1 Satz. *Sei $I = (a, b)$ und $\mathbf{A} : I \to \mathbb{R}^{n\times n}$, $B : I \to \mathbb{R}^n$ stetige Funktionen. Sei $\mathbf{Y}$ eine Fundamentalmatrix des homogenen Systems (1.5). Dann gibt es für alle $(t_0, Y_0) \in I \times \mathbb{R}^n$ eine eindeutige maximale Lösung von (1.1) mit $Y(t_0) = Y_0$. Diese Lösung hat die Form*

$$Y(t) = \mathbf{Y}(t)\,\mathbf{Y}^{-1}(t_0)\,Y_0 + \mathbf{Y}(t)\int_{t_0}^{t} \mathbf{Y}^{-1}(s)\,B(s)\,ds, \qquad t \in I.$$

Beweis. Die Existenz einer eindeutigen maximalen Lösung wurde schon bewiesen (vergleiche die Einleitung des Kapitels bzw. Satz 2.20 aus Kapitel 12). Ist Y wie im Satz definiert, so gilt

$$Y(t_0) = \mathbf{Y}(t_0)\,\mathbf{Y}^{-1}(t_0)\,Y_0 + \mathbf{Y}(t_0)\int_{t_0}^{t_0} \mathbf{Y}^{-1}(s)\,B(s)\,ds = Y_0.$$

Weiter ist aufgrund der Eigenschaften der Fundamentalmatrix die Funktion $Y(t) = \mathbf{Y}(t)\,\mathbf{Y}^{-1}(t_0)\,Y_0$ eine Lösung des homogenen Systems (1.5) mit $Y(t_0) = Y_0$. Wir müssen also nur zeigen, dass

$$Z(t) := \mathbf{Y}(t)\int_{t_0}^{t} \mathbf{Y}^{-1}(s)\,B(s)\,ds, \qquad t \in I$$

eine Lösung des inhomogenen Systems (1.1) ist. Dazu berechnen wir

$$\begin{aligned} Z'(t) &= \mathbf{Y}'(t) \int_{t_0}^{t} \mathbf{Y}^{-1}(s)\, B(s)\, ds + \mathbf{Y}(t)\, \mathbf{Y}^{-1}(t)\, B(t) \\ &= \mathbf{A}(t)\, \mathbf{Y}(t) \int_{t_0}^{t} \mathbf{Y}^{-1}(s)\, B(s)\, ds + B(t) \\ &= \mathbf{A}(t)\, Z(t) + B(t). \end{aligned}$$

Damit ist Z eine spezielle Lösung von (1.1) mit $Z(t_0) = 0$ und folglich Y die maximale Lösung von (1.1) mit $Y(t_0) = Y_0$. ∎

13.4 Systeme mit konstantem A

Wir betrachten Systeme von Differentialgleichungen 1. Ordnung der Form

$$Y'(t) = \mathbf{A}\, Y(t) \tag{4.1}$$

mit konstanter Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. Betrachten wir die lineare Abbildung

$$A : \mathbb{R}^n \to \mathbb{R}^n : Y \mapsto \mathbf{A}\, Y,$$

so gilt

$$\mathbf{A} = \mathbf{M}_E^E(A),$$

wobei $\mathbf{M}_E^E(A)$ die Darstellungsmatrix von A bezüglich der Standardbasis ist. Dabei ist für zwei Basen $B = \{B_1, \ldots, B_n\}$, $C = \{C_1, \ldots, C_n\}$ mit B_i, $C_i \in \mathbb{R}^n$, $i = 1, \ldots, n$, die Matrix $\mathbf{M}_C^B(A)$ durch die Koordinatenspaltenvektoren bezüglich C, die sich aus den Bildern $A(B_1), \ldots, A(B_n)$ der Basis B ergeben, gegeben. Bezüglich der Basis B hat die Abbildung A also eine andere Darstellungsmatrix $\mathbf{M}_B^B(A)$. Wir suchen nun eine solche Basis, so dass $\mathbf{M}_B^B(A)$ möglichst einfach ist. Aus der Linearen Algebra ist bekannt, dass

$$\mathbf{M}_B^B(A) = \mathbf{M}_B^E(id)\, \mathbf{M}_E^E(A)\, \mathbf{M}_E^B(id)$$

gilt. Weiter ist die Matrix $\mathbf{M}_E^B(id) =: \mathbf{B}$ invertierbar und es gilt $\mathbf{M}_B^E(id) = (\mathbf{M}_E^B(id))^{-1}$. Es folgt also

$$\mathbf{M}_B^B(A) = \mathbf{B}^{-1}\, \mathbf{A}\, \mathbf{B} =: \mathbf{D}.$$

Setzen wir

$$Z(t) := \mathbf{B}^{-1}\, Y(t) \qquad \text{und somit} \qquad Y(t) = \mathbf{B}\, Z(t),$$

so folgt aus (4.1) die Gleichung

$$Z'(t) = \mathbf{B}^{-1}\, Y'(t) = \mathbf{B}^{-1}\, \mathbf{A}\, Y(t) = \mathbf{B}^{-1}\, \mathbf{A}\, \mathbf{B}\, Z(t) = \mathbf{D}\, Z(t). \tag{4.2}$$

13.4.1 Symmetrische Matrizen A

Betrachten wir nun den speziellen Fall, dass $\mathbf{A}$ symmetrisch ist. Dann existiert, wie in der Linearen Algebra gezeigt wurde, eine Matrix $\mathbf{B}$, so dass $\mathbf{D} = diag(\lambda_1, \ldots, \lambda_n)$ gilt. Die Gleichung

$$Z'(t) = diag(\lambda_1, \ldots, \lambda_n)\, Z(t)$$

lässt sich also als System

$$\begin{aligned} z_1'(t) &= \lambda_1\, z_1(t), \\ &\vdots \\ z_n'(t) &= \lambda_n\, z_n(t) \end{aligned}$$

von n linearen Differentialgleichungen 1. Ordnung schreiben. Diese haben nach Satz 3.8 aus Kapitel 12 die Lösungen $z_i(t) = c_i\, e^{\lambda_i\, t}$, $i = 1, \ldots, n$. Mit der Wahl von $Z(t) := (z_1(t), \ldots, z_n(t))$ haben wir somit eine Lösung der Gleichung (4.2) gefunden. Mit der Notation $Z_i(t) := e^{\lambda_i t} E_i$ ist

$$\mathbf{Z}(t) = (Z_1(t), \ldots, Z_n(t)) = \begin{pmatrix} e^{\lambda_1\, t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2\, t} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & e^{\lambda_n\, t} \end{pmatrix}, \qquad t \in \mathbb{R} \tag{4.3}$$

eine Fundamentalmatrix von $Z' = \mathbf{D}\, Z$. Um zurück zur Gleichung (4.1) zu kommen, setzen wir mit der Notation $\mathbf{B} = (B_1, \ldots, B_n)$

$$\mathbf{Y} = \mathbf{B}\, \mathbf{Z} = (\mathbf{B}\, Z_1, \ldots, \mathbf{B}\, Z_n) = (B_1\, e^{\lambda_1\, t}, \ldots, B_n\, e^{\lambda_n\, t}).$$

Offenbar ist $\mathbf{Y}$ eine Fundamentalmatrix von (4.1) und wir können die Lösungsformel aus Satz 3.1 für inhomogene Systeme anwenden.

13.4.2 Matrizen A mit nur reellen Eigenwerten

Betrachten wir nun den allgemeineren Fall einer Matrix $\mathbf{A}$, welche nur reelle Eigenwerte besitzt. In diesem Fall liefert die Lineare Algebra die Existenz einer Matrix $\mathbf{B}$, so dass

$$\mathbf{D} = \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{J}_m \end{pmatrix}, \qquad \text{wobei} \qquad \mathbf{J}_i = \begin{pmatrix} \lambda_i & 0 & \cdots & \cdots & 0 \\ 1 & \lambda_i & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & \lambda_i \end{pmatrix}$$

für $i = 1, \ldots, m$ gilt. Dabei wird $\mathbf{J}_i \in \mathbb{R}^{k \times k}$ *Jordan Block der Ordnung* k genannt. Insbesondere sind mehrfache Eigenwerte, also $\lambda_i = \lambda_j$ für $i \neq j$ möglich. Analog zum Fall der symmetrischen Matrizen reicht es, sich die einzelnen Jordan Blöcke anzusehen: Für einen Jordan Block müssen wir also die Gleichungen

$$\begin{aligned} z_1'(t) &= \lambda\, z_1(t), \\ z_2'(t) &= z_1(t) + \lambda\, z_2(t), \\ &\vdots \\ z_k'(t) &= z_{k-1}(t) + \lambda\, z_k(t) \end{aligned}$$

lösen. Wir hatten schon gesehen, dass $z_1(t) = c_1\, e^{\lambda\,(t)}$ eine Lösung der ersten Gleichung ist. Für die zweite Gleichung verwenden wir wieder die Methode der Variation der Konstanten: Betrachten wir

$$z_2(t) = c(t)\, e^{\lambda\, t},$$

so folgt

$$z_2'(t) = c'(t)\, e^{\lambda\, t} + \lambda\, c(t)\, e^{\lambda\, t} = c'(t)\, e^{\lambda\, t} + \lambda\, z_2(t).$$

Wählen wir also $c'(t) = c_1$ bzw. $c(t) = c_1\, t + c_2$, so ist $z_2(t) = (c_1\, t + c_2)\, e^{\lambda\, t}$ eine Lösung der zweiten Gleichung. Analog erhalten wir, dass

$$\begin{aligned} z_1(t) &= c_1\, e^{\lambda\, t}, \\ z_2(t) &= (c_1\, t + c_2)\, e^{\lambda\, t}, \\ &\vdots \\ z_k(t) &= \Big(c_1\, \frac{t^{k-1}}{(k-1)!} + \ldots + c_k\Big)\, e^{\lambda\, t} \end{aligned}$$

eine Lösung des obigen Gleichungssystemes ist. Eine (Teil-) Fundamentalmatrix von $Z' = \mathbf{D}\, Z$ bezüglich dieses Jordan Blocks $\mathbf{J}_i$ ist also durch

$$\mathbf{Z}_i = (Z_1^i, \ldots, Z_k^i) = \begin{pmatrix} e^{\lambda_i\, t} & 0 & \cdots & 0 \\ t\, e^{\lambda_i\, t} & e^{\lambda_i\, t} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ \frac{t^{k-1}}{(k-1)!} e^{\lambda_i\, t} & \frac{t^{k-2}}{(k-2)!} e^{\lambda_i\, t} & \cdots & e^{\lambda_i\, t} \end{pmatrix} \tag{4.4}$$

gegeben. Setzt man nun die Jordan Blöcke zusammen, so ist

$$\mathbf{Z} = (Z_1, \ldots, Z_n) = \begin{pmatrix} \mathbf{Z}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{Z}_m \end{pmatrix}$$

eine Fundamentalmatrix von $Z' = \mathbf{D}\,Z$. Beispielsweise erhalten wir aus der unten gegebenen Matrix $\mathbf{D}$ die Fundamentalmatrix $\mathbf{Z}$

$$\mathbf{D} = \left(\begin{array}{c|cc} \mu & 0 & 0 \\ \hline 0 & \lambda & 0 \\ 0 & 1 & \lambda \end{array}\right) \qquad \mathbf{Z} = \left(\begin{array}{c|cc} e^{\mu\,t} & 0 & 0 \\ \hline 0 & e^{\lambda\,t} & 0 \\ 0 & t\,e^{\lambda\,t} & e^{\lambda\,t} \end{array}\right).$$

Die Rücktransformation

$$\mathbf{Y} = \mathbf{B}\,\mathbf{Z}$$

ist im Allgemeinen schwer aufzuschreiben, kann aber in jedem konkreten Fall ausgeführt werden.

Beispiele. (a) Sei die Matrix $\mathbf{A}$ durch

$$\mathbf{A} = \begin{pmatrix} 1 & 4 \\ -1 & -3 \end{pmatrix}$$

gegeben. Wir müssen zunächst die Matrix $\mathbf{A}$ in Jordan Normalform bringen. Dazu berechnen wir das charakteristische Polynom

$$\det(\mathbf{A} - \lambda\,\mathbf{E}) = \begin{vmatrix} 1-\lambda & 4 \\ -1 & -3-\lambda \end{vmatrix} = -(1-\lambda)\,(3+\lambda) + 4 = (\lambda+1)^2$$

und erhalten $\lambda = -1$ als doppelten Eigenwert von $\mathbf{A}$. Um eine Basis des Eigenraumes $E(-1) = \mathrm{Ker}(\mathbf{A} + \mathbf{E})$ zu bestimmen, betrachten wir

$$\begin{pmatrix} 2 & 4 \\ -1 & -2 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Dieses System wird z.B. von $u = -2$ sowie $v = 1$ erfüllt und wir erhalten den Eigenvektor $V_1 = (-2, 1)^\top$, der eine Basis von $E(-1)$ ist. Insbesondere ist $\dim E(-1) = 1$. Zur Bestimmung der Nilpotenzordnung des Eigenwertes $\lambda = -1$ betrachten wir

$$(\mathbf{A} + \mathbf{E})^2 = \begin{pmatrix} 2 & 4 \\ -1 & -2 \end{pmatrix} \begin{pmatrix} 2 & 4 \\ -1 & -2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Die Nilpotenzordnung ist also gleich 2 und der verallgemeinerte Eigenraum $F(-1) = \mathrm{Ker}(\mathbf{A} + \mathbf{E})^2$. Dieser enthält den Eigenraum $E(-1)$. Eine Basis von $F(-1)$ ist z.B. durch (V_2, V_1) mit $V_2 = E_1$ gegeben. Bezüglich dieser Basis muss die Darstellungsmatrix von $F(-1)$ allerdings nicht die Gestalt eines Jordan Blockes haben. Um dies sicherzustellen modifizieren wir den Basisvektor V_1 wie folgt:

$$\widetilde{V}_1 = (\mathbf{A} + \mathbf{E})\,V_2 = \begin{pmatrix} 2 & 4 \\ -1 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

Somit ist $(V_2, \widetilde{V}_1)$ eine Basis von $F(-1)$ bezüglich derer die Darstellungsmatrix von $F(-1)$ die Gestalt eines Jordan Blockes hat. Wir erhalten somit die Basiswechselmatrix $\mathbf{B}$ und die Darstellungsmatrix $\mathbf{D}$ von $\mathbf{A}$ bezüglich dieser Basis als

$$\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} \qquad \mathbf{D} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \end{pmatrix}.$$

Die Fundamentalmatrix von $Z' = \mathbf{D}\,Z$ ist also durch

$$\mathbf{Z}(t) = \begin{pmatrix} e^{-t} & 0 \\ t\,e^{-t} & e^{-t} \end{pmatrix}$$

gegeben. Die Rücktransformation ergibt

$$\mathbf{Y}(t) = \mathbf{B}\,\mathbf{Z}(t) = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} e^{-t} & 0 \\ t\,e^{-t} & e^{-t} \end{pmatrix} = \begin{pmatrix} e^{-t} + 2\,t\,e^{-t} & 2\,e^{-t} \\ -t\,e^{-t} & -e^{-t} \end{pmatrix}$$

als Fundamentalmatrix von $Y' = \mathbf{A}\,Y$. Um die Lösungsformel mit $t_0 = 0$ anwenden zu können, müssen wir noch $\mathbf{Y}(0)^{-1}$ berechnen. Offenbar gilt

$$\mathbf{Y}(0) = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} = \mathbf{Y}(0)^{-1}.$$

Damit ist für alle $Y_0 \in \mathbb{R}^2$

$$Y(t) := \mathbf{Y}(t)\mathbf{Y}(0)^{-1}\,Y_0$$

Lösung von (4.1) mit $Y(0) = Y_0$.

(b) Sei die Matrix $\mathbf{A}$ durch

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & -3 \\ 1 & 4 & 4 \\ 0 & 0 & -1 \end{pmatrix}$$

gegeben. Wir berechnen wieder das charakteristische Polynom

$$\begin{aligned} \det(\mathbf{A} - \lambda\mathbf{E}) &= \begin{vmatrix} 2-\lambda & -1 & -3 \\ 1 & 4-\lambda & 4 \\ 0 & 0 & -1-\lambda \end{vmatrix} \\ &= (-1-\lambda)\begin{vmatrix} 2-\lambda & -1 \\ 1 & 4-\lambda \end{vmatrix} \\ &= -(1+\lambda)\Big((2-\lambda)(4-\lambda)+1\Big) \\ &= -(1+\lambda)(\lambda-3)^2 \end{aligned}$$

und erhalten $\lambda_1 = -1$ als einfachen Eigenwert sowie $\lambda_2 = 3$ als doppelten Eigenwert. Zur Bestimmung des Eigenraumes $E(3)$ und des verallgemeinerten Eigenraumes $F(3)$ betrachten wir

$$\big(\mathbf{A} - 3\mathbf{E}\big) = \begin{pmatrix} -1 & -1 & -3 \\ 1 & 1 & 4 \\ 0 & 0 & -4 \end{pmatrix}, \quad \big(\mathbf{A} - 3\mathbf{E}\big)^2 = \begin{pmatrix} 0 & 0 & 11 \\ 0 & 0 & -15 \\ 0 & 0 & 16 \end{pmatrix},$$

sowie

$$\big(\mathbf{A} - 3\mathbf{E}\big)^3 = \begin{pmatrix} 0 & 0 & -44 \\ 0 & 0 & 60 \\ 0 & 0 & -64 \end{pmatrix}.$$

Die Nilpotenzordnung ist also 2. Weiterhin gilt: $E(3) = \mathrm{Ker}(\mathbf{A} - 3\mathbf{E})$, $F(3) = \mathrm{Ker}(\mathbf{A} - 3\mathbf{E})^2$ sowie $\dim E(3) = 1, \dim F(3) = 2$. Eine Basis von $E(3)$ ist z.B. durch

$$V_1 := \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

gegeben. Diese ergänzen wir durch

$$V_2 := \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

zu einer Basis von $F(3)$. Um sicherzustellen, dass die Darstellungsmatrix ein Jordan Block ist, modifizieren wir V_1 wie folgt:

$$\widetilde{V}_1 = \big(\mathbf{A} - 3\mathbf{E}\big)\, V_2 = \begin{pmatrix} -1 & -1 & -3 \\ 1 & 1 & 4 \\ 0 & 0 & -4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -2 \\ 2 \\ 0 \end{pmatrix}.$$

Weiter benötigen wir eine Basis von $E(-1)$. Mit

$$\mathbf{A} - (-1)\mathbf{E} = \begin{pmatrix} 3 & -1 & -3 \\ 1 & 5 & 4 \\ 0 & 0 & 0 \end{pmatrix}$$

sieht man leicht, dass

$$V_3 = \begin{pmatrix} 11 \\ -15 \\ 16 \end{pmatrix}$$

eine Basis von $\mathrm{Ker}(\mathbf{A} + \mathbf{E}) = E(-1)$ ist. Bezüglich der Basis $(V_3, V_2, \widetilde{V}_1)$ ist nun die Darstellungsmatrix $\mathbf{D}$ von $\mathbf{A}$ sowie die Basiswechselmatrix $\mathbf{B}$ durch

$$\mathbf{B} = \begin{pmatrix} 11 & 1 & -2 \\ -15 & 1 & 2 \\ 16 & 0 & 0 \end{pmatrix} \qquad \mathbf{D} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 1 & 3 \end{pmatrix}$$

gegeben. Damit erhalten wir die Fundamentalmatrix von $Z' = \mathbf{D}\,Z$ als

$$\mathbf{Z}(t) = \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & e^{3\,t} & 0 \\ 0 & t\,e^{3\,t} & e^{3\,t} \end{pmatrix}.$$

Die Rücktransformation ergibt

$$\mathbf{Y}(t) = \begin{pmatrix} 11 & 1 & -2 \\ -15 & 1 & 2 \\ 16 & 0 & 0 \end{pmatrix} \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & e^{3\,t} & 0 \\ 0 & t\,e^{3\,t} & e^{3\,t} \end{pmatrix} = \begin{pmatrix} 11\,e^{t} & (e^{3\,t} - 2\,t\,e^{3\,t}) & -2\,e^{3\,t} \\ -15\,e^{t} & (e^{3\,t} + 2\,t\,e^{3\,t}) & 2\,e^{3\,t} \\ 16\,e^{t} & 0 & 0 \end{pmatrix}.$$

Für eine Lösung von (4.1) können wir also wieder die Lösungsformel anwenden.

13.4.3 Matrizen A mit komplexen Eigenwerten

Falls $\lambda = a + ib$ ein (komplexer) Eigenwert der (reellen) Matrix **A** ist, so ist auch $\overline{\lambda} = a - ib$ ein Eigenwert der Matrix **A**. Falls die Dimension des Systems $n \leq 3$ ist, kann es folglich höchstens die einfachen komplexen Eigenwerte $\lambda = a + ib$ und $\overline{\lambda} = a - ib$ geben. Wir werden im weiteren nur den speziellen Fall $n = 2$ betrachten. Gehen wir nun wie im Fall symmetrischer Matrizen vor (und erlauben dabei komplexe Koeffizienten), so transformieren wir die Matrix **A** mithilfe der komplexen Matrix $\widetilde{\mathbf{B}}$[2] in die *komplexe Jordanmatrix*

$$\widetilde{\mathbf{A}} := \begin{pmatrix} \lambda & 0 \\ 0 & \overline{\lambda} \end{pmatrix}$$

und erhalten die *komplexe Fundamentalmatrix*

$$\widetilde{\mathbf{Z}}(t) = \begin{pmatrix} e^{\lambda\,t} & 0 \\ 0 & e^{\overline{\lambda}\,t} \end{pmatrix}$$

von $Z' = \widetilde{\mathbf{A}}\,Z$. Wir suchen nun eine reelle Darstellung dieser Fundamentalmatrix bzw. der Fundamentalmatrix des ursprünglichen Systems $Y' = \mathbf{A}\,Y$. Sei dafür $V \in \mathbb{C}^2$ ein (komplexer) Eigenvektor von **A** zum Eigenwert λ, d.h. $\mathbf{A}V = \lambda V$. Für den (komponentenweise) komplex konjugierten Vektor $\overline{V}$ gilt die Gleichung $\mathbf{A}\overline{V} = \overline{\lambda V}$. Somit ist $\overline{v}$ ein Eigenvektor von **A** zum Eigenwert $\overline{\lambda}$. Wir betrachten nun die Darstellung des komplexen Eigenvektors

[2] Die Matrix $\widetilde{\mathbf{B}}$ besteht aus den Spaltenvektoren V und $\overline{V}$, wobei V ein komplexer Eigenvektor von **A** ist.

und Eigenwertes $V = U + iW$ mit $U, W \in \mathbb{R}^2$ und $\lambda = a + i\,b$ für $b > 0$. Damit zerlegen wir die Gleichung $\mathbf{A}V = \lambda V$ in den Imaginär- und Realteil:

$$\begin{aligned}\mathbf{A}U &= a\,U - b\,W,\\ \mathbf{A}W &= a\,W + b\,U.\end{aligned}$$

Dies können wir auch als Matrixmultiplikation schreiben:

$$\mathbf{A}\begin{pmatrix}U & W\end{pmatrix} = \begin{pmatrix}U & W\end{pmatrix}\begin{pmatrix}a & b\\ -b & a\end{pmatrix}.$$

Da V und $\overline{V}$ linear unabhängig sind, folgt auch die lineare Unabhängigkeit von U und W. Somit ist $\mathbf{B} := \begin{pmatrix}U & W\end{pmatrix}$ regulär und es gilt

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{B} = \begin{pmatrix}a & b\\ -b & a\end{pmatrix} =: \hat{\mathbf{A}}. \tag{4.5}$$

Die Matrix $\mathbf{B}$ transformiert also $\mathbf{A}$ in die *reelle Normalform* $\hat{\mathbf{A}}$. Wir könnten jetzt direkt die Fundamentalmatrix des Systems $Z' = \hat{\mathbf{A}}\,Z$ ausrechnen. Aufgrund folgender Überlegung kann man diese aber direkt aus der komplexen Fundamentalmatrix des Systems $Z' = \widetilde{\mathbf{A}}\,Z$ ablesen. Wenn man das Transformationsverhalten der Matrix $\mathbf{A}$ beachtet, d.h.

$$\begin{pmatrix}a + i\,b & 0\\ 0 & a - i\,b\end{pmatrix} = \widetilde{\mathbf{A}} \xleftarrow[\text{mit } \widetilde{\mathbf{B}}]{\text{Transformation}} \mathbf{A} \xrightarrow[\text{mit } \mathbf{B}]{\text{Transformation}} \hat{\mathbf{A}} = \begin{pmatrix}a & b\\ -b & a\end{pmatrix},$$

und vermutet, dass sich dieses Transformationsverhalten auf die Fundamentalmatrix $\mathbf{Y}(t)$ des Systems $Y' = \mathbf{A}\,Y$ überträgt, so erwarten wir

$$e^{a\,t}\begin{pmatrix}\cos(b\,t) + i\,\sin(b\,t) & 0\\ 0 & \cos(b\,t) - i\,\sin(b\,t)\end{pmatrix} = \widetilde{\mathbf{Z}}(t)$$

$$\xleftarrow[\text{mit } \widetilde{\mathbf{B}}]{\text{Transformation}} \text{ Fundamentalmatrix } \mathbf{Y}(t) \xrightarrow[\text{mit } \mathbf{B}]{\text{Transformation}}$$

$$\mathbf{Z}(t) := e^{a\,t}\begin{pmatrix}\cos(b\,t) & \sin(b\,t)\\ -\sin(b\,t) & \cos(b\,t)\end{pmatrix}.$$

Durch Nachrechnen kann man leicht überprüfen, dass

$$\mathbf{Z}(t) = e^{a\,t}\begin{pmatrix}\cos(b\,t) & \sin(b\,t)\\ -\sin(b\,t) & \cos(b\,t)\end{pmatrix}$$

in der Tat die reelle Fundamentalmatrix von $Z' = \hat{\mathbf{A}}\,Z$ ist. Um zum ursprünglichen Problem zurückzukommen, beachten wir die Identität $\mathbf{B}^{-1}\mathbf{A}\mathbf{B} = \hat{\mathbf{A}}$. Setzten wir also

$$\mathbf{Y}(t) := \mathbf{B}\,\mathbf{Z}(t), \qquad t \in I,$$

so ist $\mathbf{Y}$ eine *reelle Fundamentalmatrix* des Systems $Y' = \mathbf{A}\,Y$. Da die Matrix $\mathbf{B} = (U, W)$ aus dem Realteil sowie Imaginärteil des Eigenvektors V aufgebaut ist, haben wir also eine reelle Darstellung einer Fundamentalmatrix von $Y' = \mathbf{A}\,Y$ gefunden.

Beispiel. Sei die Matrix $\mathbf{A}$ durch

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix}$$

gegeben. Für das charakteristische Polynom berechnen wir

$$\begin{aligned} \det(\mathbf{A} - \lambda\mathbf{E}) &= \det \begin{pmatrix} 1-\lambda & 2 \\ -1 & 3-\lambda \end{pmatrix} \\ &= (1-\lambda)(3-\lambda) + 2 = \lambda^2 - 4\,\lambda + 5. \end{aligned}$$

Die Eigenwerte sind also durch $\lambda_{1,2} = (2 \pm \sqrt{4-5}\,=)2 \pm i$ gegeben. Wir setzen $\lambda := 2 + i$ und $\overline{\lambda} = 2 - i$. Wir suchen nun Eigenvektoren zu diesen Eigenwerten. Dazu betrachten wir $(\mathbf{A} - \lambda\mathbf{E})\,V = 0$:

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1-2-i & 2 \\ -1 & 3-2-i \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} -1-i & 2 \\ -1 & 1-i \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

Diese Gleichung wird offenbar von $V = \begin{pmatrix} 2 \\ 1+i \end{pmatrix}$ erfüllt. Es folgt

$$U = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad W = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad \text{und somit} \qquad \mathbf{B} = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}.$$

Die Fundamentalmatrix von $Z' = \hat{\mathbf{A}}\,Z$

$$\mathbf{Z}(t) = e^{2\,t} \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}$$

transformiert sich also zur Fundamentalmatrix

$$\mathbf{Y}(t) = \mathbf{B}\,\mathbf{Z}(t) = e^{2\,t} \begin{pmatrix} 2\cos(t) & 2\sin(t) \\ \cos(t) - \sin(t) & \cos(t) + \sin(t) \end{pmatrix}$$

von $Y' = \mathbf{A}\,Y$. Für den Anfangswert zum Zeitpunkt $t_0 = 0$ benötigen wir eine Fundamentalmatrix $\widetilde{\mathbf{Y}}(t)$ mit $\widetilde{\mathbf{Y}}(0) = \mathbf{E}$. Dazu bestimmen wir $\mathbf{Y}(0)^{-1}$:

$$\mathbf{Y}(0) = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix} \qquad \text{und somit} \qquad \mathbf{Y}(0)^{-1} = \begin{pmatrix} \frac{1}{2} & 0 \\ -\frac{1}{2} & 1 \end{pmatrix}.$$

Damit erhalten wir als gewünschte Fundamentalmatrix

$$\widetilde{\mathbf{Y}}(t) = \mathbf{Y}(t)\,\mathbf{Y}(0)^{-1} = e^{2\,t} \begin{pmatrix} \cos(t) - \sin(t) & 2\sin(t) \\ -\sin(t) & \cos(t) + \sin(t) \end{pmatrix}.$$

13.4.4 Reelle Systeme für $n = 2$

Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ mit $\det \mathbf{A} \neq 0$. Für die Jordan'sche Normalform gibt es folgende Möglichkeiten:

1. $\begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}$ mit $\lambda, \mu \in \mathbb{R}$, $\lambda, \mu \neq 0$ ($\lambda = \mu$ möglich)

2. $\begin{pmatrix} \lambda & 0 \\ 1 & \lambda \end{pmatrix}$ mit $\lambda \in \mathbb{R} \setminus \{0\}$

3. $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$ mit $\lambda \in \mathbb{C}$, $\lambda = a + i\,b$

Wir hatten bereits gesehen, dass für alle $t \in \mathbb{R}$ und alle Anfangswerte $Y_0 \in \mathbb{R}^2$ maximale Lösungen existieren. Es stellt sich nun die Frage nach dem asymptotischen Verhalten der Lösungen für $t \to \infty$. Wir werden dabei die Fälle 1. - 3. einzeln betrachten.

Zu Fall 1: Die Lösung der auf eine Basis von Eigenvektoren transformierten Gleichung (4.1) ist durch

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} c_1\, e^{\lambda\, t} \\ c_2\, e^{\mu\, t} \end{pmatrix} \tag{4.6}$$

gegeben. Um uns einen besseren Überblick über das qualitative Lösungsverhalten zu verschaffen, betrachten wir das *Phasenportrait* der Lösung. Dazu interpretieren wir die Lösungsformel (4.6) als Parameterdarstellung zu verschiedenen c_1, c_2 einer Kurve im $\mathbb{R}^2$, den sogenannten *Phasenkurven* oder *Trajektorien* der Gleichung. Fassen wir nun mehrere Trajektorien in einem Diagramm zusammen und orientieren diese mit der üblichen Orientierung von $\mathbb{R}$, so sprechen wir von einem *Phasenportrait*. Dazu unterscheiden wir die folgenden vier Fälle.

(a) Gelte $\lambda, \mu < 0$. Direkt aus der Lösungsformel sehen wir

$$\lim_{t\to\infty} x(t) = 0, \qquad \lim_{t\to\infty} y(t) = 0. \tag{4.7}$$

Für den Fall $c_1, c_2 \neq 0$ erhalten wir aus der Lösungsformel die Trajektorien

$$y(t) = c_2\, e^{\mu\, t} = c_2 \left(e^{\lambda\, t}\right)^{\frac{\mu}{\lambda}} = c_2 \left(\frac{x(t)}{c_1}\right)^{\frac{\mu}{\lambda}}.$$

Man beachte, dass $\frac{\mu}{\lambda} > 0$. Für $c_1 = 0$ ist die Trajektorie $x(t) = 0$, für $c_2 = 0$ erhalten wir $y(t) = 0$ und für $c_1 = c_2 = 0$ ist die Trajektorie durch $x(t) = y(t) = 0$ gegeben. Damit erhalten wir das Phasenportrait in Abbildung 4.1. Die Richtung der Pfeile ist durch (4.7) gegeben. Insbesondere laufen alle Trajektorien in den Ursprung $(0, 0)$. Deshalb nennen wir $(0, 0)$ einen *stabilen Knoten*. Das Phasenportrait (Abbildung 4.1) ist bezüglich der Basis aus Eigenwerten gezeichnet.

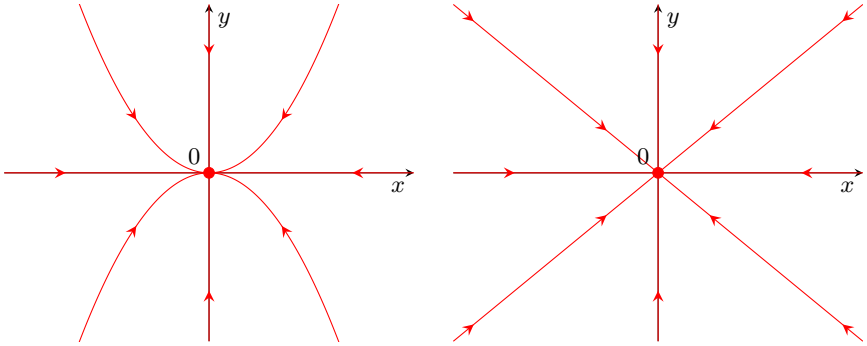


Abb. 4.1. Phasenportrait im Fall 1 und $\lambda \leq \mu < 0$. Links $\frac{\mu}{\lambda} = 2$, Rechts $\frac{\mu}{\lambda} = 1$.

Betrachten wir die Standardbasis (und somit das ursprüngliche Problem (4.1)), so ist das Phasenportrait durch eine affine Deformation des obigen gegeben (siehe Abbildung 4.2).

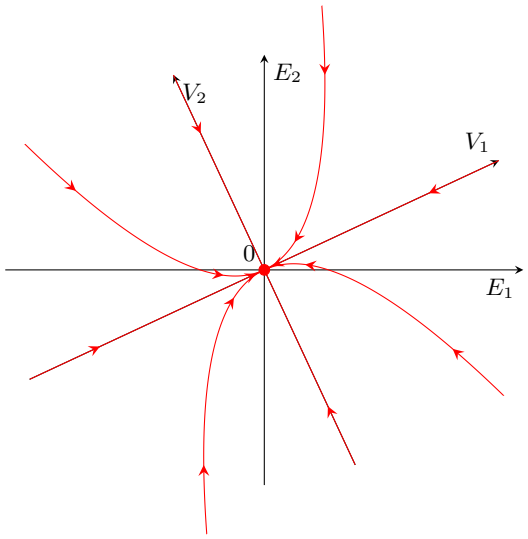


Abb. 4.2. Phasenportrait im Fall 1 und $\lambda \leq \mu < 0$ bezüglich der Standardbasis ($\frac{\mu}{\lambda} = 2$, V_1, V_2 Eigenvektoren).

(b) Gelte $\lambda, \mu > 0$. Dann gilt für $c_1, c_2 \neq 0$

$$\lim_{t\to\infty} |x(t)| = \infty, \qquad \lim_{t\to\infty} |y(t)| = \infty. \tag{4.8}$$

Ansonsten sind die Rechnungen aus dem vorherigen Fall auch hier gültig. Insbesondere ist $\frac{\mu}{\lambda} > 0$. Durch (4.8) drehen sich allerdings die Pfeile um. Das Phasenportrait ist in Abbildung 4.3 veranschaulicht. Da jetzt alle Trajektorien aus $(0,0)$ heraus laufen, nennen wir $(0,0)$ einen *instabilen Knoten*.

(c) Gelte $\lambda < 0 < \mu$. Dann gilt für $c_1, c_2 \neq 0$

$$\lim_{t\to\infty} x(t) = 0, \qquad \lim_{t\to\infty} |y(t)| = \infty.$$

Man beachte, dass in der Darstellung

$$y(t) = c_2 \left(\frac{x(t)}{c_1}\right)^{\frac{\mu}{\lambda}} \tag{4.9}$$

der Exponent $\frac{\mu}{\lambda}$ jetzt negativ ist. Dadurch ergibt sich ein völlig anderes Phasenportrait. Die Richtung der Pfeile ist durch (4.9) gegeben. Das Phasenportrait ist dann in Abbildung 4.4 veranschaulicht. Der Ursprung wird als *instabiler Sattelpunkt* bezeichnet, da die Trajektorie auf der "x"-Achse in ihn hineinläuft und die Trajektorie auf der "y"-Achse aus ihm heraus läuft.

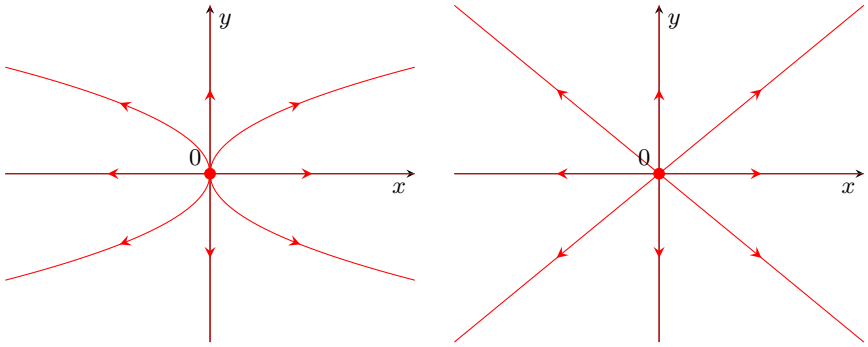


Abb. 4.3. Phasenportrait im Fall 1 und $\lambda \geq \mu > 0$ mit $\frac{\mu}{\lambda} = 0.5$ (links) und $\frac{\mu}{\lambda} = 1$ (rechts).

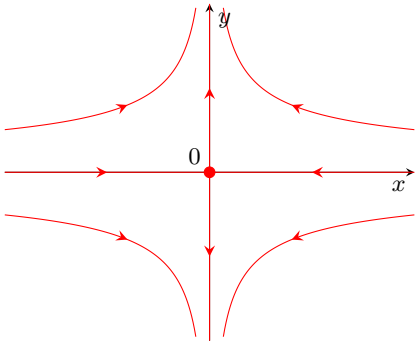


Abb. 4.4. Phasenportrait im Fall 1 und $\lambda < 0 < \mu$ mit $\frac{\mu}{\lambda} = -0.5$.

(d) Gelte $\mu < 0 < \lambda$. Dann gilt für $c_1, c_2 \neq 0$

$$\lim_{t\to\infty} |x(t)| = \infty, \qquad \lim_{t\to\infty} y(t) = 0.$$

In diesem Fall vertauschen sich gerade die Rollen von x und y. Das Phasenportrait ist dann in Abbildung 4.5 dargestellt. Auch hier wird der Ursprung als ein *instabiler Sattelpunkt* bezeichnet.

Zu Fall 2: Wie wir im Abschnitt 13.4.2 gesehen haben, ist die Lösung der auf eine Basis von Eigenvektoren transformierten Gleichung (4.1) durch

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} c_1 \\ c_1\, t + c_2 \end{pmatrix} e^{\lambda\, t} \qquad t \in \mathbb{R} \tag{4.10}$$

gegeben. Auch hier betrachten wir unterschiedliche Fälle.

(a) Gelte $\lambda < 0$. Dann folgt aus (4.10)

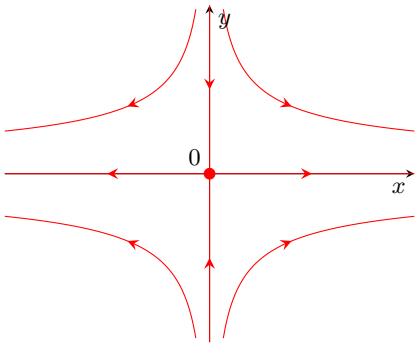


Abb. 4.5. Phasenportrait im Fall 1 und $\mu < 0 < \lambda$ mit $\frac{\mu}{\lambda} = -0.5$.

$$\lim_{t\to\infty} x(t) = 0, \qquad \lim_{t\to\infty} y(t) = 0. \tag{4.11}$$

Weiter erhalten wir für $c_1 \neq 0$

$$\frac{y(t)}{x(t)} = \frac{c_1\, t\, e^{\lambda t} + c_2\, e^{\lambda t}}{c_1\, e^{\lambda t}} = t + \frac{c_2}{c_1}$$

und somit

$$y(t) = t\, x(t) + \frac{c_2}{c_1}\, x(t).$$

Weiter folgt aus $\frac{x(t)}{c_1} = e^{\lambda t} > 0$ die Gleichung $\ln(\frac{x(t)}{c_1}) = \lambda t$. Somit können wir die explizite t-Abhängigleit eliminieren und erhalten für die Trajektorien

$$y(t) = \frac{1}{\lambda}\, x(t)\, \ln\left(\frac{x(t)}{c_1}\right) + \frac{c_2}{c_1}\, x(t).$$

Das Phasenportrait ist in Abbildung 4.6 veranschaulicht. Die Richtung der Pfeile ist durch (4.11) gegeben. Insbesondere ist der Ursprung ein stabiler Knoten.

(b) Gelte $\lambda > 0$. Dann folgt für $c_1 \neq 0$ aus (4.10)

$$\lim_{t\to\infty} |x(t)| = \infty, \qquad \lim_{t\to\infty} |y(t)| = \infty. \tag{4.12}$$

Das Phasenportrait ist in Abbildung 4.7 dargestellt. Die Richtung der Pfeile dreht sich aufgrund von (4.12) im Vergleich zum vorherigen Fall um. Insbesondere ist der Ursprung jetzt ein instabiler Knoten.

Zu Fall 3: Wie wir in Abschnitt 13.4.3 gesehen haben, ist die reelle Fundamentalmatrix für das transformierte System durch

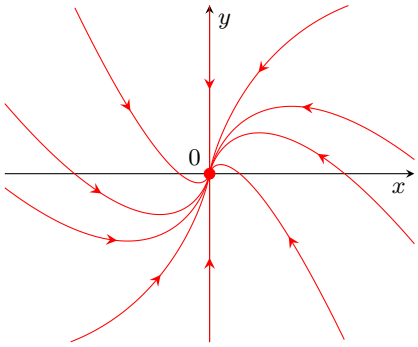


Abb. 4.6. Phasenportrait im Fall 2 und $\lambda = -1$.

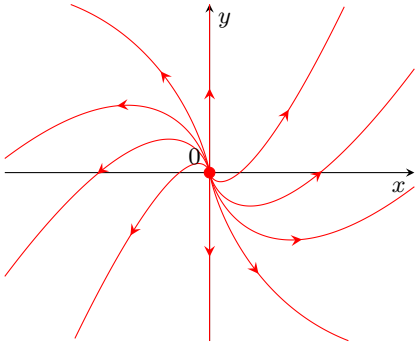


Abb. 4.7. Phasenportrait im Fall 2 und $\lambda = 1$.

$$\mathbf{Z}(t) = e^{a\,t} \begin{pmatrix} \cos(b\,t) & \sin(b\,t) \\ -\sin(b\,t) & \cos(b\,t) \end{pmatrix}$$

gegeben. Zur Veranschaulichung der Lösungen ist es sinnvoll $\mathbb{R}^2$ und $\mathbb{C}$ zu identifizieren, d.h. $(x, y) \in \mathbb{R}^2$ wird mit $x + iy \in \mathbb{C}$ identifiziert. Unter Berücksichtigung von

$$\cos(b\,t) - i\,\sin(b\,t) = e^{-i\,b\,t}$$

erhalten wir die komplexen Basislösungen

$$\begin{aligned} z_1(t) &= e^{a\,t}\,e^{-i\,b\,t}, \\ z_2(t) &= i\,z_1(t). \end{aligned}$$

Es reicht also die Basislösung z_1 zu veranschaulichen, da z_2 durch eine Drehung aus z_1 hervorgeht. Es können nun die folgenden Fälle auftreten.

(a) Gilt $a = 0$, so folgt

$$|z_1(t)| = const.$$

Das Phasenportrait ist dann durch Abbildung 4.8 gegeben und wir nennen den Ursprung ein *stabiles Zentrum*.

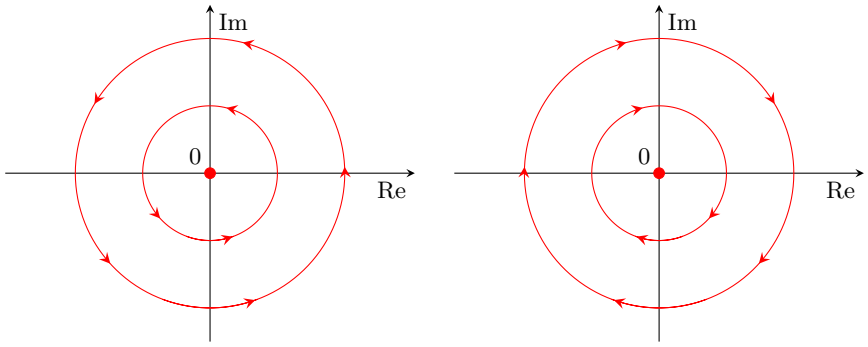


Abb. 4.8. Phasenportrait im Fall 3 für $a = 0$ und links $b < 0$, rechts $b > 0$.

(b) Gilt $a > 0$, so folgt für Anfangswerte ungleich Null

$$\lim_{t\to\infty} |z_1(t)| = \infty.$$

Das Phasenportrait ist dann in Abbildung 4.9 veranschaulicht. Der Ursprung wird *instabiler Strudel* genannt.

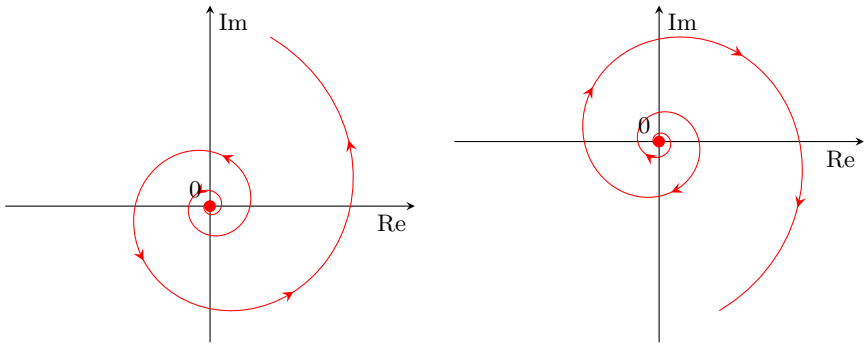


Abb. 4.9. Phasenportrait im Fall 3 für $a > 0$ und links $b < 0$, rechts $b > 0$.

(c) Gilt $a < 0$, so folgt

$$\lim_{t\to\infty} z_1(t) = 0.$$

Das Phasenportrait ist dann in Abbildung 4.10 dargestellt. Der Ursprung wird *stabiler Strudel* genannt.

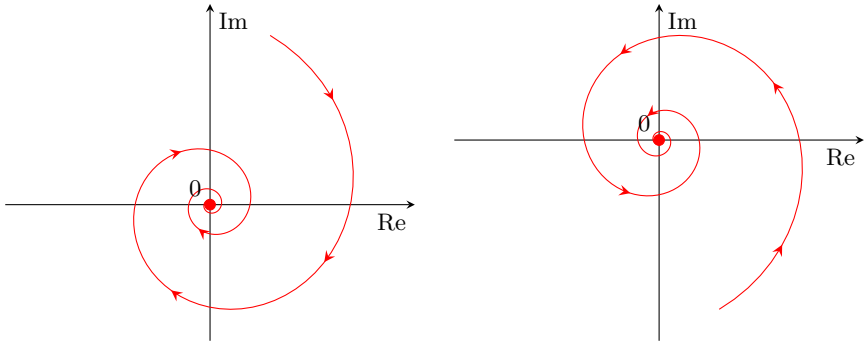


Abb. 4.10. Phasenportrait im Fall 3 für $a < 0$ und links $b < 0$, rechts $b > 0$.

13.5 Exponentialfunktion für Matrizen

Für eine lineare homogene Differentialgleichung mit konstanten Koeffizienten $y' = a\,y$ haben wir eine explizite Lösungsformel mithilfe der Exponentialfunktion hergeleitet, nämlich $y(t) = c\,e^{at}$. Um diese Formal auf lineare Systeme mit konstanten Koeffizienten $Y' = \mathbf{A}\,Y$ zu verallgemeinern benötigen wir die Exponentialfunktion für Matrizen.

5.1 Definition. *Sei* $\mathbf{A} \in \mathbb{R}^{n\times n}$. *Wir definieren die* Exponentialfunktion *für* $\mathbf{A}$ *durch*

$$e^{\mathbf{A}} := \sum_{n=0}^{\infty} \frac{\mathbf{A}^n}{n!}. \tag{5.2}$$

Bemerkungen. 1. Für $\mathbf{A} \in \mathbb{R}^{n\times n}$ gilt

$$\|\mathbf{A}^n\| = \|\mathbf{A}\mathbf{A}\ldots\mathbf{A}\| \le \|\mathbf{A}\| \cdot \ldots \cdot \|\mathbf{A}\| = \|\mathbf{A}\|^n.$$

Somit folgt

$$\|\sum_{n=0}^{N} \frac{\mathbf{A}^n}{n!}\| \le \sum_{n=0}^{N} \frac{\|\mathbf{A}\|^n}{n!}.$$

Also konvergiert die Reihe in (5.2) absolut, da die reelle Exponentialreihe $\sum_{n=0}^{\infty} \frac{\|\mathbf{A}\|^n}{n!}$ eine Majorante ist.

2. Analog kann man Funktionen von Matrizen für beliebige Potenzreihen definieren, z.B. für $\sin(\mathbf{A})$, $\tan(\mathbf{A})$ oder $\ln(\mathbf{A})$. Diese konvergieren absolut innerhalb des Konvergenzradius der Reihe.

5.3 Lemma. *Es gilt:*

(i) $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}} \cdot e^{\mathbf{B}} = e^{\mathbf{B}} \cdot e^{\mathbf{A}}$, *falls* $\mathbf{A}\,\mathbf{B} = \mathbf{B}\,\mathbf{A}$.

(ii) $e^{\mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{B}} = \mathbf{B}^{-1}\,e^{\mathbf{A}}\,\mathbf{B}$, *falls* $\det \mathbf{B} \neq 0$.

(iii) $e^{diag\,(\lambda_1,\ldots,\lambda_n)} = diag\,(e^{\lambda_1},\ldots,e^{\lambda_n})$.

Beweis. (i): Aufgrund unserer Voraussetzung folgt

$$(\mathbf{A}+\mathbf{B})(\mathbf{A}+\mathbf{B}) = \mathbf{A}^2 + \mathbf{A}\,\mathbf{B} + \mathbf{B}\,\mathbf{A} + \mathbf{B}^2 = \mathbf{A} + 2\,\mathbf{A}\,\mathbf{B} + \mathbf{B}^2.$$

Da die Reihen absolut konvergieren ist weiter insbesondere das Cauchyprodukt dieser Reihen wohldefiniert. Somit folgt analog zum Beweis für die reelle Exponentialreihe

$$\begin{aligned} e^{\mathbf{A}+\mathbf{B}} = \sum_{n=0}^{\infty} \frac{(\mathbf{A}+\mathbf{B})^n}{n!} &= \sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{\mathbf{B}^k\,\mathbf{A}^{n-k}}{k!(n-k)!} \\ &= \Big(\sum_{n=0}^{\infty} \frac{\mathbf{B}^k}{k!}\Big)\Big(\sum_{\ell=0}^{\infty} \frac{\mathbf{A}^\ell}{\ell!}\Big) = e^{\mathbf{B}} \cdot e^{\mathbf{A}}. \end{aligned}$$

Vertauschen wir die Rolle von $\mathbf{A}$ und $\mathbf{B}$, so folgt $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{B}} \cdot e^{\mathbf{A}}$.

(ii): Es gilt

$$(\mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{B})^2 = (\mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{B})(\mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{B}) = \mathbf{B}^{-1}\,\mathbf{A}^2\,\mathbf{B}$$

uns somit induktiv

$$(\mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{B})^k = \mathbf{B}^{-1}\,\mathbf{A}^k\,\mathbf{B}.$$

Also folgt

$$\begin{aligned} e^{\mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{B}} &= \sum_{k=0}^{\infty} \frac{(\mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{B})^k}{k!} \\ &= \sum_{k=0}^{\infty} \mathbf{B}^{-1}\frac{\mathbf{A}^k}{k!}\mathbf{B} \\ &= \mathbf{B}^{-1}\Big(\sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!}\Big)\mathbf{B} = \mathbf{B}^{-1}\,e^{\mathbf{A}}\,\mathbf{B}. \end{aligned}$$

(iii): Es gilt offenbar $\operatorname{diag}(\lambda_1,\dots,\lambda_n)^2 = \operatorname{diag}(\lambda_1^2,\dots,\lambda_n^2)$. Somit folgt analog zum Beweis von (ii) die Behauptung. ∎

5.4 Satz. *Für $\mathbf{A} \in \mathbb{R}^{n\times n}$ setzen wir*

$$\mathbf{S}(t) := e^{\mathbf{A}\,t}, \qquad t \in \mathbb{R}. \tag{5.5}$$

Dann gilt:

1. $\mathbf{S}(0) = \mathbf{E}$,
2. $\mathbf{S}(s+t) = \mathbf{S}(s)\,\mathbf{S}(t) = \mathbf{S}(t)\,\mathbf{S}(s)$,
3. $\frac{d}{dt}\mathbf{S}(t) = \mathbf{A}\,\mathbf{S}(t) = \mathbf{S}(t)\,\mathbf{A}$.

Mit anderen Worten: $\mathbf{S}$ *ist eine Fundamentalmatrix von* $Y' = \mathbf{A}\,Y$.

Beweis. Nach Definition gilt

$$\mathbf{S}(t) = \sum_{n=0}^{\infty} \frac{t^n\,\mathbf{A}^n}{n!}.$$

Somit folgt

$$\mathbf{S}(0) = \sum_{n=0}^{\infty} \frac{0^n\,\mathbf{A}^n}{n!} = \frac{0^0\,\mathbf{A}^0}{0!} + \mathbf{0} + \ldots = \frac{1\,\mathbf{E}}{1} = \mathbf{E}.$$

Es gilt also 1. Weiter folgt mit der Binomialformel und dem Cauchyprodukt

$$\begin{aligned}
\mathbf{S}(s+t) &= \sum_{n=0}^{\infty} \frac{(s+t)^n\,\mathbf{A}^n}{n!} \\
&= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{t^k\,s^{n-k}\,\mathbf{A}^{n-k}\,\mathbf{A}^k}{k!(n-k)!} \\
&= \sum_{k=0}^{\infty} \frac{t^k\,\mathbf{A}^k}{k!} \sum_{\ell=0}^{\infty} \frac{s^\ell\,\mathbf{A}^\ell}{\ell!} \\
&= \mathbf{S}(t)\cdot\mathbf{S}(s).
\end{aligned}$$

Vertauschen wir die Rolle von s und t, so haben wir 2. gezeigt. Verwenden wir diese Aussage, so erhalten wir

$$\begin{aligned}
\frac{d}{dt}\mathbf{S}(t)\Big|_{t=t_0} &= \lim_{h\to 0}\frac{\mathbf{S}(t_0+h)-\mathbf{S}(t_0)}{h}\\
&= \mathbf{S}(t_0)\,\lim_{h\to 0}\frac{\mathbf{S}(h)-\mathbf{E}}{h}\\
&= \mathbf{S}(t_0)\,\lim_{h\to 0}\frac{1}{h}\Big(\sum_{k=0}^{\infty}\frac{h^k\,\mathbf{A}^k}{k!}-\mathbf{E}\Big)\\
&= \mathbf{S}(t_0)\,\lim_{h\to 0}\Big(\mathbf{A}+\sum_{k=2}^{\infty}\frac{h^{k-1}\,\mathbf{A}^k}{k!}\Big)\\
&= \mathbf{S}(t_0)\,\mathbf{A},
\end{aligned}$$

wobei wir die absolute Konvergenz der Reihe verwendet haben. Im letzten Schritt haben wir auch benutzt, dass

$$\begin{aligned}
\Big\|\sum_{k=2}^{\infty}\frac{h^{k-1}\,\mathbf{A}^k}{k!}\Big\| &= \Big\|\lim_{N\to\infty}\sum_{k=2}^{N}\frac{h^{k-1}\,\mathbf{A}^k}{k!}\Big\|\\
&= \lim_{N\to\infty}\Big\|\sum_{k=2}^{N}\frac{h^{k-1}\,\mathbf{A}^k}{k!}\Big\|\\
&\le \lim_{N\to\infty}\sum_{k=2}^{N}|h|^{k-1}\frac{\|\mathbf{A}\|^k}{k!}\\
&\le \lim_{N\to\infty}|h|\|\mathbf{A}\|^2\sum_{k=0}^{N-2}\frac{h^k\|\mathbf{A}\|^k}{(k+2)!}\\
&\le \lim_{N\to\infty}|h|\|\mathbf{A}\|^2\sum_{k=0}^{N-2}\frac{h^k\|\mathbf{A}\|^k}{k!}\\
&\le |h|\|\mathbf{A}\|^2\,e^{h\,\|\mathbf{A}\|}\xrightarrow{h\to 0}0.
\end{aligned}$$

Wir haben also 3. gezeigt und somit den Satz bewiesen. ∎

Wir haben somit die Analogie zum Fall $n=1$ gerechtfertigt. In diesem Fall wussten wir schon, dass die Gleichung

$$y'(t)=a\,y(t)$$

die Lösung

$$y(t)=c\,e^{a\,t}$$

besitzt. Jetzt haben wir gezeigt, dass

$$\mathbf{Y}(t):=e^{\mathbf{A}\,t},$$

eine Fundamentalmatrix des Systems $Y' = \mathbf{A}\,Y$ ist und $\mathbf{Y}(0) = \mathbf{E}$ erfüllt. Somit hat das Anfangswertproblem

$$Y'(t) = \mathbf{A}\,Y(t)\,, \quad Y(0) = Y_0$$

die Lösung

$$Y(t) = e^{\mathbf{A}\,t}\,Y_0.$$

Wir haben auch eine neue Berechtigung für (4.4) gefunden. Wir können einen Jordan Block der Ordnung k wie folgt zerlegen

$$\mathbf{J} = \begin{pmatrix} \lambda & 0 & \cdots & \cdots & 0 \\ 1 & \lambda & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & \lambda \end{pmatrix} = \lambda\,\mathbf{E} + \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} =: \lambda\,\mathbf{E} + \mathbf{F}.$$

Somit folgt

$$\mathbf{F}^2 = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ 1 & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 & 0 \end{pmatrix}$$

und insbesondere $\mathbf{F}^k = \mathbf{0}$. Für die Exponentialfunktion erhalten wir also

$$\begin{aligned} e^{\mathbf{F}\,t} = \sum_{n=0}^{\infty} \frac{\mathbf{F}^n\,t^n}{n!} &= \mathbf{E} + \frac{\mathbf{F}^1\,t^1}{1} + \ldots + \frac{\mathbf{F}^{k-1}\,t^{k-1}}{(k-1)!} + \mathbf{0} + \ldots \\ &= \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ t & 1 & \ddots & \ddots & \vdots \\ \frac{1}{2}\,t^2 & t & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \frac{t^{k-1}}{(k-1)!} & \cdots & \frac{1}{2}\,t^2 & t & 1 \end{pmatrix}. \end{aligned}$$

Da $\mathbf{EF} = \mathbf{FE}$, gilt mithilfe von Lemma 5.3

$$e^{\mathbf{J}\,t} = e^{\lambda\,\mathbf{E}\,t}\,e^{\mathbf{F}\,t} = e^{\lambda\,t}\,\mathbf{E}\,e^{\mathbf{F}\,t} = \begin{pmatrix} e^{\lambda\,t} & 0 & \cdots & 0 \\ t\,e^{\lambda\,t} & e^{\lambda\,t} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ \frac{t^{k-1}}{(k-1)!}e^{\lambda\,t} & \frac{t^{k-2}}{(k-2)!}e^{\lambda\,t} & \ldots & e^{\lambda\,t} \end{pmatrix},$$

also die Formel (4.4). Gilt $\mathbf{J} = \mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{B}$, so folgt $\mathbf{A} = \mathbf{B}\,\mathbf{J}\,\mathbf{B}^{-1}$ und wir erhalten mit Lemma 5.3

$$e^{\mathbf{A}\,t} = \mathbf{B}\,e^{\mathbf{J}\,t}\,\mathbf{B}^{-1}.$$

A Anhang

A.1 Die Jordan'sche Normalform

In Kapitel 13 haben wir mehrfach die Jordan'sche Normalform verwendet und berechnet. Diese wurde bereits in der Linearen Algebra behandelt, der Vollständigkeit halber werden wir sie an dieser Stelle ebenfalls einführen. Die grundlegende Zielsetzung ist dabei, einen nilpotenten Endomorphismus, d.h. eine lineare Abbildung $f : K^n \to K^n$ eines Vektorraumes in sich selbst mit

$$f^k \equiv 0$$

für ein $k \in \mathbb{N}$ bzw. deren Darstellungsmatrix in eine möglichst einfache Gestalt zu transformieren.

1.1 Definition. *Sei K eine Körper und $n \in \mathbb{N}$. Die Matrix*

$$J_n := \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \in K^{n\times n}$$

heißt Jordanmatrix der Größe n zum Eigenwert 0.

1.2 Folgerung. (a) *Für das charakteristische Polynom gilt $\chi_{J_n} = T^n$. Insbesondere ist 0 der einzige Eigenwert von J_n.*

(b) *Sei $\mathbf{A} = J_n$. Der zugehörige Endomorphismus $f_{\mathbf{A}}$ ist durch*

$$f_{\mathbf{A}} : K^n \to K^n, \quad e_i \mapsto \begin{cases} e_{i+1} & i < n, \\ 0 & i = n \end{cases}$$

gegeben. Es folgt für $\ell \in \mathbb{N}$

$$f_{\mathbf{A}}^{\ell}(e_i) = \begin{cases} e_{i+\ell} & i+\ell \leq n, \\ 0 & i+\ell > n. \end{cases}$$

Insbesondere gilt $f_{\mathbf{A}}^{n} = 0$ und damit $\mathbf{A}^n = 0$. Also ist $f_{\mathbf{A}}$ bzw. $\mathbf{A}$ nilpotent. Insbesondere entsteht der i-te Basisvektor durch $(i-1)$-malige Anwendung von $f_{\mathbf{A}}$ auf den ersten Basisvektor, also

$$e_i = f_{\mathbf{A}}^{i-1}(e_1).$$

Es folgt

$$\operatorname{Ker}(f_{\mathbf{A}}^{n-1}) = \bigoplus_{i=2}^{n} K \cdot e_i = \Big\{ \big(0, \alpha_1, \ldots, \alpha_n\big)^T \;\Big|\; \alpha_i \in K,\, i = 2, \ldots, n \Big\}.$$

Der erste Basisvektor e_1 zeichnet sich also dadurch aus, dass er jede Basis von $\operatorname{Ker}(f_{\mathbf{A}}^{n-1})$ zu einer Basis von K^n ergänzt.

Die einfachste Gestalt nilpotenter Endomorphismen ist durch die Jordanmatrizen gegeben. Genauer gilt:

1.3 Satz (Struktur nilpotenter Matrizen)**.** *Sei $g : V \to V$ ein nilpotenter Endomorphismus eines n-dimensionalen K-Vektorraumes. Sei $d := \min\{i \in \mathbb{N} \,|\, g^i \equiv 0\}$. Dann existiert eine (nicht eindeutig bestimmte) Basis $\mathcal{B}$ von V, so dass*

$$M_{\mathcal{B}}^{\mathcal{B}}(g) = \begin{pmatrix} J_d & & & & & & 0 \\ & \ddots & & & & & \\ & & J_d & & & & \\ & & & J_{d-1} & & & \\ & & & & \ddots & & \\ & & & & & J_1 & \\ & & & & & & \ddots \\ 0 & & & & & & J_1 \end{pmatrix} \begin{matrix} \} r_d - mal \\ \} r_{d-1} - mal \\ \} r_1 - mal \end{matrix}$$

eine Blockdiagonalmatrix aus Jordanmatrizen ist. Dabei gilt

$$r_\ell = 2\dim(\operatorname{Ker}(g^\ell)) - \dim(\operatorname{Ker}(g^{\ell-1})) - \dim(\operatorname{Ker}(g^{\ell+1}))$$

für alle $1 \leq \ell \leq d$.

Beweis. Wir zeigen zunächst die Existenz von $\mathcal{B}$.

i) **Idee:** Finde einen Basisvektor $w_1 \in \mathrm{Ker}(g^d) \setminus \mathrm{Ker}(g^{d-1})$. Setzen wir $w_i := g^{i-1}(w_1)$, $1 \leq i \leq d$ so erhalten wir J_d.
Finde nun $w_{d+1} \in \mathrm{Ker}(g^d) \setminus \mathrm{Ker}(g^{d-1})$, welches linear unabhängig von $(w_1, \ldots, w_d)$ ist und setze $w_{d+i} := g^{i-1}(w_{d+1})$, $1 \leq i \leq d$. Damit erhalten wir die nächste Jordanmatrix J_d. Führe dies solange fort, bis alle Jordanmatrizen gefunden sind.

ii) Betrachte $d := \min\{i \in \mathbb{N} \mid g^i \equiv 0\}$. Aufgrund der Nilpotenz von g ist d wohldefiniert und endlich. Weiter gilt

$$\{0\} \subset \mathrm{Ker}(g) \subset \mathrm{Ker}(g^2) \subset \ldots \subset \mathrm{Ker}(g^{d-1}) \subset \mathrm{Ker}(g^d) = V.$$

iii) **Algorithmus:**

a) Wähle eine Hilfsbasis von $\mathrm{Ker}(g^{d-1})$ und ergänze diese zu einer Basis von V, also wähle eine Basis $(\overline{v}_d^{(1)}, \ldots, \overline{v}_d^{(s_d)})$ von ${}^{\mathrm{Ker}(g^d)}/_{\mathrm{Ker}(g^{d-1})}$ und wähle $v_d^{(1)}, \ldots, v_d^{(s_d)} \in \mathrm{Ker}(g^d)$ mit $v_d^{(j)} + \mathrm{Ker}(g^{d-1}) = \overline{v}_d^{(j)}$. Definiere $W_d := K\, v_d^{(1)} \oplus \ldots \oplus K\, v_d^{(s_d)}$.

b) Setze $v_i^{(j)} := g^{d-i}(v_d^{(j)})$ für alle $1 \leq i < d$ und $1 \leq j \leq s_d$. Definiere weiter $\mathcal{B}_d := (v_d^1, \ldots, v_1^{(1)}, \ldots, v_d^{(s_d)}, \ldots, v_1^{(s_d)})$.
Damit gilt $v_i^j \in \mathrm{Ker}(g^i) \setminus \mathrm{Ker}(g^{i-1})$ für alle $1 \leq i \leq d$, $1 \leq j \leq s_d$, denn

$$g^{i-1}(v_i^{(j)}) = g^{i-1}(g^{d-i}(v_d^(j))) = g^{d-1}(v_d^(j)) \neq 0$$
$$g^i(v_i^{(j)}) = g^d(v_d^{(j)}) = 0.$$

Es gilt also

$$\begin{array}{ccccccc}
\{0\} \subset \mathrm{Ker}(g) & \subset \ldots \subset & \mathrm{Ker}(g^\ell) & \subset \ldots \subset & \mathrm{Ker}(g^{d-1}) & \subset \mathrm{Ker}(g^d) = V. \\
\cup & & \cup & & \cup & \cup \\
v_1^{(1)} & & v_\ell^{(1)} & & v_{d-1}^{(1)} & v_d^{(1)} \\
\vdots & & \vdots & & \vdots & \vdots \\
v_1^{(s_d)} & & v_\ell^{(s_d)} & & v_{d-1}^{(s_d)} & v_d^{(s_d)}
\end{array}$$

c) Führe für $\ell = d-1, \ldots, 1$ die folgenden Schritte aus:

3i) Wähle eine Hilfsbasis von $\mathrm{Ker}(g^{\ell-1})$ und ergänze diese durch die schon bekannten $v_\ell^{(1)}, \ldots v_\ell^{(s_{\ell+1})}$ sowie neuen $v_\ell^{(1+s_{\ell+1})}, \ldots, v_\ell^{s_{\ell+1}}$

zu einer Basis von $\mathrm{Ker}(g^\ell)$. Wähle also eine Basis $(\overline{v}_\ell^{(1+s_{\ell+1})}, \ldots, \overline{v}_\ell^{(s_\ell)})$ von ${}^{\mathrm{Ker}(g^\ell)}\!/\!_{(\mathrm{Ker}(g^{\ell-1})+Kv_\ell^{(1)}+\ldots+Kv_\ell^{(s_{\ell+1})})}$ und wähle $v_\ell^{(j)} \in \mathrm{Ker}(g^\ell)$ mit

$$v_\ell^{(j)} + \mathrm{Ker}(g^{\ell-1}) + Kv_\ell^{(1)} + \ldots + Kv_\ell^{(s_{\ell+1})} = \overline{v}_\ell^{(j)}$$

für $j = 1 + s_{\ell+1}, \ldots, s_\ell$ (hier gilt $s_\ell > s_{\ell+1}$). Beachte dabei, dass die benötigte lineare Unabhängigkeit von $v_\ell^{(1)}, \ldots v_\ell^{(s_{\ell+1})}$ später gezeigt wird. Definiere nun $W_\ell := Kv_\ell^{(1+s_{\ell+1})} \oplus \ldots \oplus Kv_\ell^{(s_\ell)}$.

3ii) Setze $v_i^{(j)} := g^{\ell-i}(v_\ell^{(j)})$ für alle $1 \leq i \leq \ell$, $1 + s_{\ell+1} \leq j \leq s_\ell$ und $\mathcal{B}_\ell := \mathcal{B}_{\ell+1} \cup \{v_\ell^{(1+s_{\ell+1})}, \ldots, v_1^{(1+s_{\ell+1})}, \ldots, v_\ell^{(s_\ell)}, \ldots, v_1^{(s_\ell)}\}$.

Wir zeigen, dass $\mathcal{B} := \mathcal{B}_1$ die gesuchte Basis ist.

iv) Wir behaupten, dass für alle $1 \leq \ell \leq d$

$$\mathrm{Ker}(g^\ell) = \mathrm{Ker}(g^{\ell-1}) \oplus g^{d-\ell}(W_d) \oplus \ldots \oplus g(W_{\ell+1}) \oplus W_\ell$$

gilt. Dies zeigen wir mit Hilfe der absteigenden Induktion nach ℓ.

$\ell = d$, $\mathrm{Ker}(g^\ell) = V = \mathrm{Ker}(g^{\ell-1}) \otimes W_d$ nach iii)a).
Induktionshypothese: $\mathrm{Ker}(g^{\ell+1}) = \mathrm{Ker}(g^\ell) \oplus g^{d-\ell-1}(W_d) \oplus \ldots \oplus W_{\ell+1}$.
Induktionsschritt:

- Aus $W_i \subset \mathrm{Ker}(g^i)$ folgt $g^{i-\ell}(W_i) \subset \mathrm{Ker}(g^\ell)$.
- Die Abbildung $g : (g^{d-\ell-1}(W_d) \oplus \ldots \oplus W_{\ell+1}) \to V$, $v \mapsto g(v)$ ist injektiv, denn sei $v \in \mathrm{Ker}(g) \cap (g^{d-\ell-1}(W_d) \oplus \ldots \oplus W_{\ell+1})$, so folgt mit $\mathrm{Ker}(g) \subset \mathrm{Ker}(g^\ell)$ und der Induktionshypothese $v \in \mathrm{Ker}(g^\ell) \cap (g^{d-\ell-1}(W_d) \oplus \ldots \oplus W_{\ell+1}) = \{0\}$. Somit zerfällt das Bild $g^{d-\ell}(W_d) \oplus \ldots \oplus g(W_{\ell+1})$ in die entsprechende direkte Summe.
- Es gilt $\mathrm{Ker}(g^{\ell-1}) \cap (g^{d-\ell}(W_d) \oplus \ldots \oplus g(W_{\ell+1}) = \{0\}$, denn sei $w \in (g^{d-\ell}(W_d) \oplus \ldots \oplus g(W_{\ell+1})$ mit $v = g(w) \in \mathrm{Ker}(g^{\ell-1})$, so folgt $w \in \mathrm{Ker}(g^\ell)$ und mit der Induktionshypothese $w = 0$ sowie $v = 0$. Somit ist die Summe $\mathrm{Ker}(g^{\ell-1}) \oplus g^{d-\ell}(W_d) \oplus \ldots \oplus g(W_\ell)$ direkt.
- Nach Konstruktion ist $\mathrm{Ker}(g^\ell) = \big(\mathrm{Ker}(g^{\ell-1}) \oplus g^{d-\ell}(W_d) \oplus \ldots \oplus g(W_{\ell+1})\big) \oplus W_\ell$.

v) W_i hat die Basis $v_i^{k(i+1)+\ell}, \ldots, v_i^{k(i)}$, denn für alle $\ell \geq 1$ ist

$$g^{i-\ell} : W_i \xrightarrow{g} g(W_i) \xrightarrow{g} \ldots \xrightarrow{g} g^{i-\ell}(W_i)$$

ein Isomorphismus. Somit hat $g^{i-\ell}(W_i)$ die Basis $v_i^{k(i+1)+\ell}, \ldots, v_i^{k(i)}$.

vi) Der Vektorraum

$$\begin{aligned} V &= \mathrm{Ker}(g^d) \\ &= \mathrm{Ker}(g^{d-1}) \oplus W_d \\ &= \mathrm{Ker}(g^{d-2}) \oplus g(W_d) \oplus W_d \oplus W_{d-1} \\ &= \dots \\ &= (W_d \oplus g(W_d) \oplus \dots \oplus g^{d-1}(W_d)) \oplus \dots \oplus (W_2 \oplus g(W_2)) \oplus W_1 \end{aligned}$$

hat Basis $\mathcal{B}$.

Nun zeigen wir die Eindeutigkeit der $s_d, \dots, s_1$: Es gilt

$$\begin{aligned} s_d &= r_d = \dim(W_d) = \dim(V) - \dim(\mathrm{Ker}(g^{d-1})), \\ s_{d-1} &= r_{d-1} + r_d = \dim(W_{d-1} \oplus g(W_d)) = \dim(\mathrm{Ker}(g^{d-1})) - \dim(\mathrm{Ker}(g^{d-2})), \\ &\vdots \\ s_1 &= r_1 + \dots + r_d = \dim(W_1 \oplus \dots \oplus g^{d-1}(W_d)) = \dim(\mathrm{Ker}(g^1)). \end{aligned}$$

Somit sind rekursiv $r_d, \dots, r_1$ eindeutig durch g bestimmt:

$$r_\ell = s_\ell - s_{\ell+1} = 2\dim(\mathrm{Ker}(g^\ell)) - \dim(\mathrm{Ker}(g^{\ell-1})) - \dim(\mathrm{Ker}(g^{\ell+1}))$$

für alle $1 \le \ell \le d$. ∎

Wir haben also folgenden Algorithmus zur Bestimmung der Jordan-Normalform: Sei $\mathbf{A} \in K^{n\times n}$ nilpotent.

0. Berechne $\mathbf{A}^2$, $\mathbf{A}^3$, … , $\mathbf{A}^d$ mit $d := \min\{i \in \mathbb{N} \mid \mathbf{A}^i \equiv 0\}$. Bestimme die Dimension und Basis von $\mathrm{Ker}(\mathbf{A}^\ell)$, $1 \le \ell \le d-1$. Setze

$$s_\ell := \dim(\mathrm{Ker}(\mathbf{A}^\ell)) - \dim(\mathrm{Ker}(A^{\ell-1})) \qquad 1 \le \ell \le d$$

(beachte dabei $A^0 := Id_n$). Setze weiter $r_\ell := s_\ell - s_{\ell+1}$ für $1 \le \ell \le d$.

1. Ergänze die Basis von $\mathrm{Ker}(A^{d-1})$ durch $v_d^{(1)}, \dots, v_d^{(s_d)}$ zu einer Basis von $\mathrm{Ker}(\mathbf{A}^d) = K^n$.

2. Setze $v_i^{(j)} := \mathbf{A}^{d-i} v_d^{(j)}$ für alle $1 \le i < d$ und alle $1 \le j \le s_d$. Setze $\mathcal{B}_d := \{v_d^{(1)}, \dots, v_1^{(1)}, \dots, v_d^{(s_1)}, \dots, v_1^{(s_1)}\}$.

3. Für $\ell = d-1, \dots, 1$ tue folgendes:

   – Falls $r_\ell = 0$ gilt, mache nichts und gehe zum nächsten ℓ

– Falls $r_\ell > 0$ gilt:

3 a) Ergänze die Basis von $\mathrm{Ker}(g^{\ell-1})$ durch die schon bekannten $v_\ell^{(1)}, \ldots, v_\ell^{(s_{\ell+1})}$ sowie die neuen $v_\ell^{(1+s_{\ell+1})}, \ldots, v_\ell^{(s_\ell)}$ zu einer Basis von $\mathrm{Ker}(g^\ell)$.

3 b) Setze $\mathcal{B}_\ell := \mathcal{B}_{\ell+1} \cup \{v_\ell^{(1+s_{\ell+1})}, \ldots, v_\ell^{(s_\ell)}, \ldots, v_\ell^{(s_\ell)}, \ldots, v_1^{(s_\ell)}\}$

– Erhalte $\mathcal{B} = \mathcal{B}_1$ Basis von K^n mit $M_{\mathcal{B}}^{\mathcal{B}}(f_{\mathbf{A}}) = \mathbf{B}^{-1}\mathbf{A}\mathbf{B}$ wie in Satz 1.3.