Friedrich-Alexander-Universität Erlangen-Nürnberg
Naturwissenschaftliche Fakultät
Department Mathematik
Lehrstuhl für Angewandte Mathematik 1

Skript der Veranstaltung

Mathematik für Ingenieure C3 (IngMatC3)

gehalten im Wintersemester 2017/18
von Dr. Ilja Kröker

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

NATURWISSENSCHAFTLICHE
FAKULTÄT

Vorbemerkung

Dieses Skript enthält den Inhalt der Vorlesung „Mathematik für Ingenieure C3“ des Wintersemester 2017/18 bei Dr. Ilja Kröker. Es wurde anhand von Mitschriften in LaTeX gesetzt und erhebt daher weder Anspruch auf Korrektheit noch Vollständigkeit und ist ***offensichtlich inoffiziell***. Bei Unstimmigkeiten und evtl. vorhandenen Fehlern bitte ich um eine Email an untenstehende Adresse. Dieses Skript stellt damit insbesondere **keine** offizielle Veröffentlichung des Lehrstuhl für Angewandte Mathematik 1 am Department Mathematik der Friedrich-Alexander-Universität Erlangen-Nürnberg dar.

Florian Frank — florian.ff.frank@fau.de
Version vom 14. Oktober 2018

Literaturvorschläge

[Bor13] Wolfgang Borchers. *Skript zu der Veranstaltung „Mathematik für Ingenieure D3“*. WS12. FAU Erlangen-Nürnberg, 2013.

[Erd08] Laszlo Erdos. *Skript zu der Veranstaltung „Numerik I“*. SS08. LMU München, 2008.

[Gra01] Hans Grabmüller. *Skripte zu den Veranstaltungen „Mathematik für Informatiker“*. SS00, WS00. FAU Erlangen-Nürnberg, 2001.

[HK06] Horst W. Hamacher und Kathrin Klamroth. *Lineare Optimierung und Netzwerkoptimierung*. 2., verb. Aufl., zweisprachige Ausg. dt. engl. 2006, 240 S. ISBN: 978-3-8348-0185-2.

[Lie15] Frauke Liers. *Skript zur Veranstaltung „Mathematik für Ingenieure C3“*. WS14. FAU Erlangen-Nürnberg, 2015.

Inhaltsverzeichnis

KAPITEL 1

ANALYSIS IM $\mathbb{R}^n$

In der Vorgängerveranstaltung C2 haben wir uns mit den Grundzügen der eindimensionalen Analysis beschäftigt. Wir haben dann gegen Ende des Semesters diese Grundlagen auf die Analysis mehrerer Veränderlicher abgebildet, und haben in diesem Zusammenhang Folgen und Funktionen, Stetigkeits-, Differentations- und Integrationsregeln, sowie die Formel von Taylor im „$\mathbb{R}^n$“ kennengelernt. Dieses Kapitel soll nun an diesem Punkt ansetzen und eine „angewandte“ Seite dieser – auf den ersten Blick eher theoretisch wirkenden – Methoden darstellen.

1.1 Extremstellen, Extremwertaufgaben

In diesem ersten Teil wollen wir einerseits wiederholen, was vorliegen muss, damit wir von Extremstellen reden, und gleichzeitig ebenso erklären, wie wir Extremstellen von Funktionen im mehrdimensionalen Raum finden können. Wir erinnern uns damit an letztes Semester und geben erneut einen Satz mit hinreichenden und notwendigen Kriterien für Extremata im eindimensionalen Fall an.

Satz 1.1 (*Extremata - Hinreichende und notwendige Kriterien*)

Sei $f : D \subseteq \mathbb{R} \to \mathbb{R} \in \mathcal{C}^2(D)$ mit $x_0 \in D$. Dann gelten folgende Implikationen:

(a) $f'(x_0) = 0 \wedge f''(x_0) < 0 \Rightarrow x_0$ ist **lokales Maximum**
(b) $f'(x_0) = 0 \wedge f''(x_0) > 0 \Rightarrow x_0$ ist **lokales Minimum**

} hinreichende Kriterien

(c) x_0 ist lokale Extremstelle $\Rightarrow f'(x_0) = 0$
(d) x_0 ist lokales Maximum $\Rightarrow f'(x_0) = 0 \wedge f''(x_0) \leqslant 0$
(e) x_0 ist lokales Minimum $\Rightarrow f'(x_0) = 0 \wedge f''(x_0) \geqslant 0$

} notwendige Kriterien

Dabei bezeichnen die Implikationen (c) – (e) notwendige Kriterien, sprich Kriterien, die für die Nichtexistenz von Extremstellen verwendet werden können, und (a) und (b) hinreichende Kriterien, sprich solche, welche sich für den Nachweis der Existenz solcher Stellen eignen.

Beweis: Wir verzichten an dieser Stelle auf einen Beweis, da wir ihn in C2 bereits geführt haben. Die Idee bestand darin mit der Taylorentwicklung einer Funktion zu arbeiten und verschiedene Aussagen zu einem Widerspruch zu führen. □

Wir wollen nun also den Begriff der Extremstellen auf n-dimensionale Vektorräume verallgemeinern. Wir definieren:

Definition 1.1 (*Extremstellen*)

Sei $f : D \subseteq \mathbb{R}^n \to \mathbb{R} \in \mathcal{C}^2(D)$. Dann heiße $x_0 \in D$ ...

... eine **lokale** Maximal- oder Minimalstelle von f gdw. $\exists \varepsilon > 0 : \forall x \in K_\varepsilon(x_0) \cap D : f(x) \left\{ \begin{matrix} \leqslant \\ \geqslant \end{matrix} \right\} f(x_0)$

... eine **isolierte** (lokale) Maximal- oder Minimalstelle gdw. sie lokal ist, aber die Gleichheit nicht gilt.

... eine **globale** Maximal- oder Minimalstelle von f gdw. $K_\varepsilon(x_0) = \mathbb{R}^n$, also $K_\varepsilon(x_0) \cap D = D$.

Wir werden mit Satz 1.3 Satz 1.1 auf mehrere Raumdimensionen verallgemeinern. Die Überlegung – mit der wir uns im Beweis beschäftigen – zeigt, dass die Unterscheidung zwischen Minimal- und Maximalstelle allein vom Vorzeichen des quadratischen Terms abhängt, also ob $\langle \mathcal{H}f(\vec{\xi})(\vec{x}-\vec{x_0}),(\vec{x}-\vec{x_0})\rangle \gtrless 0$. Wir betrachten nun also ebensolche Ausdrücke quadratischer Matrizen genauer und definieren damit den folgenden Begriff der Definitheit:

Definition 1.2 (*Definitheit*)

Bezeichne $\langle\cdot,\cdot\rangle$ das euklidische Skalarprodukt und $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix. A heiße dann ...

- ... **(symmetrisch) positiv definit** gdw. $\forall\, \vec{v} \in \mathbb{R}^n \setminus \{\vec{0}\} : \langle A\vec{v},\vec{v}\rangle > 0$
- ... **(symmetrisch) negativ definit** gdw. $\forall\, \vec{v} \in \mathbb{R}^n \setminus \{\vec{0}\} : \langle A\vec{v},\vec{v}\rangle < 0$
- ... **(symmetrisch) positiv semidefinit** gdw. $\forall\, \vec{v} \in \mathbb{R}^n : \langle A\vec{v},\vec{v}\rangle \geqslant 0$
- ... **(symmetrisch) negativ semidefinit** gdw. $\forall\, \vec{v} \in \mathbb{R}^n : \langle A\vec{v},\vec{v}\rangle \leqslant 0$
- ... **indefinit** gdw. A weder positiv noch negativ semidefinit ist.

Aus dem ersten Semester bekannt sollte dabei die Form $\langle A\vec{v},\vec{v}\rangle$ sein, sie wird ebenfalls durch $\sum_{i,j=1}^{n} a_{ij}v_iv_j$ beschrieben. Als kleine Anmerkung nebenbei sei darauf hingewiesen, dass ein $f : \mathbb{R}^n \to \mathbb{R}$ mit $\vec{x} \mapsto \langle A\vec{x},\vec{x}\rangle$ auch **quadratische Form** heißt.
Bevor wir nun zu einem leichteren Kriterium für die Definitheit kommen, sei hier noch etwas zu obigen Begriffen angemerkt. Da das symmetrisch in Definition 1.2 hier explizit dabei steht, liegt die Frage nahe, ob es auch **nicht**symmetrische definite Matrizen gibt. In der Tat ließe sich der Begriff der Definitheit „sinnvoll“ erweitern, in dem man einfach fordert, dass eine nichtsymmetrische Matrix B genau dann definit ist, wenn es die symmetrische Matrix $B + B^T$ ist. Wir wollen uns mit dieser Ansichtsweise allerdings nicht weiter beschäftigen. Allgemein sei noch anzumerken, dass die Matrix in jedem Fall quadratisch sein muss.
Wir wissen, dass Matrixvektorprodukte nicht leicht und vor allem nicht leicht für **alle** Vektoren aus dem $\mathbb{R}^n$ zu berechnen sind. Manchmal ergibt der Nachweis der Definitheit über die Definitionsanwendung Sinn, aber oft führt der folgende Satz zu schnellerem Erfolg:

Satz 1.2 (*Kriterium für Definitheit*)

Sei $A \in \mathbb{R}^{n\times n}$ wieder symmetrisch und beschreiben $\lambda_1,\ldots,\lambda_n \in \underline{\underline{\mathbb{R}}}$ die **reellen** Eigenwerte von A. Dann gelten folgende Äquivalenzen:

(a) A ist $\left\{\begin{matrix}\text{positiv}\\ \text{negativ}\end{matrix}\right\}$ definit $\iff \forall\, i \in \{1,\ldots,n\} : \lambda_i \left\{\begin{matrix}>\\ <\end{matrix}\right\} 0$

(b) A ist $\left\{\begin{matrix}\text{positiv}\\ \text{negativ}\end{matrix}\right\}$ semidefinit $\iff \forall\, i \in \{1,\ldots,n\} : \lambda_i \left\{\begin{matrix}\geqslant\\ \leqslant\end{matrix}\right\} 0$

(c) A ist indefinit $\iff \exists\, i,j \in \{1,...,n\} : (\lambda_i > 0) \wedge (\lambda_j < 0)$, also $\lambda_i \cdot \lambda_j < 0$

Beweis: (zu a.1)
Wir rufen uns eine Eigenschaft aller symmetrischen Matrizen aus dem ersten Semester in Erinnerung:

$$\forall A \in \mathbb{R}^{n\times n} \text{ sym.} : \exists Q \in \mathbb{R}^{n\times n} : Q^{-1} = Q^T \wedge \forall \lambda_i \text{ EW von } A : p_A(\lambda_i) = 0 \wedge A = QDQ^{-1}$$

Dabei stellt D die Diagonalmatrix der Eigenwerte dar. Setzen wir dies nun in unser bisheriges Kriterium für Definitheit ein, so ergibt sich:

$$\left\langle A\vec{v},\vec{v}\right\rangle = \left\langle QDQ^{-1}\vec{v},\vec{v}\right\rangle = \left\langle DQ^T\vec{v},Q^T\vec{v}\right\rangle$$

Sei nun $\vec{x} \mapsto \vec{y} := Q^T\vec{v} \in \mathrm{Abb}(\mathbb{R}^n,\mathbb{R}^n)$ linear und als Abbildung $\mathbb{R}^n\backslash\{\vec{0}\} \to \mathbb{R}^n\backslash\{\vec{0}\}$ bijektiv, da $\det(Q^T) = \pm 1 \neq 0$. Damit sei:

$$\begin{aligned} A \text{ positiv definit} &\iff \langle A\vec{v},\vec{v}\rangle > 0 \qquad \forall\, \vec{v} \in \mathbb{R}^n\backslash\{\vec{0}\} \\ &\iff \langle D\vec{y},\vec{y}\rangle > 0 \qquad \forall\, \vec{y} \in \mathbb{R}^n\backslash\{\vec{0}\} \\ &\iff \sum_{i=1}^{n} \lambda_i \cdot y_i^2 > 0 \quad \forall\, \vec{y} \in \mathbb{R}^n\backslash\{\vec{0}\} \\ &\overset{(*)}{\iff} \lambda_i > 0 \qquad\qquad \forall\, i \in \{1,\dots,n\} \end{aligned}$$

Die meisten Äquivalenzen sind trivial herleitbar und einleuchtend, die letzte $(*)$ soll nun näher beleuchtet werden.

„$\Leftarrow$“ trivial.

„$\Rightarrow$“ über *reductio ad absurdum*: Sei $\lambda_j \leqslant 0$, wähle dazu $y_j := 1$ und $y_k := 0$ für alle $k \neq j$. Dann ist aber $\sum_{i=1}^{n} \lambda_i y_i^2 = \lambda_j \leqslant 0$ ↯

Die Überlegung für (a.2), (b) und (c) ist dieselbe, der Beweis für den Rest verläuft dann analog. □
Wir übertragen nun die Überlegung zu Satz 1.1 auf das Mehrdimensionale:
Dazu entwickeln wir f mittels Taylor am Punkt $(x_0, f(x_0))$ und erhalten:

$$(I) f(x) = f(x_0) + \langle \nabla f(\xi), x - x_0\rangle \qquad \text{(Taylor nullter Ordnung)}$$

$$(II) f(x) = f(x_0) + \langle \nabla f(x_0), x - x_0\rangle + \frac{1}{2}\langle \mathcal{H} f(x - x_0), x - x_0\rangle \qquad \text{(Taylor erster Ordnung)}$$

Dabei beschreibt $\langle\cdot,\cdot\rangle$ wieder das euklidische Skalarprodukt, ∇f den Gradienten und $\mathcal{H} f$ die (symmetrische) Hessematrix der Funktion f. Für ξ gilt weiterhin $\xi = x_0 + \tau(x - x_0)$ mit $\tau \in (0,1)$.
Sei nun also $x_0 \in \overset{\circ}{D}$ Extremstelle von f. Angenommen $\nabla f(x_0) \neq 0$, so muss ein $i \in \{1,\dots,n\}$ existieren mit $\partial_i f(x_0) \neq 0$. Wir führen die Annahme mittels Fallunterscheidung zum Widerspruch:
Fall a: $\partial_i f(x_0) > 0$

Das heißt, es existiert ein $\varepsilon_0 > 0$, so dass für alle $\xi \in K_{\varepsilon_0}(x_0) : \partial_i f(\xi) > 0$. Sei nun $0 < \xi \leqslant \varepsilon_0$ beliebig, aber fest. Sei des Weiteren $x := x_0 + (0,\dots,0,\underbrace{\varepsilon}_{\text{i-te Position}},0,\dots,0)^T$ und $x' := x_0 - (0,\dots,0,\overbrace{\varepsilon}^{\text{i-te Position}},0,\dots,0)^T$. Wir betrachten nun $f(x)$ und $f(x')$:

- $f(x) - f(x_0) \overset{(I)}{=} \left\langle \nabla f(\xi), \overbrace{x - x_0}^{\varepsilon e_i} \right\rangle = \partial_i f(\xi) \cdot \varepsilon > 0$ da sowohl $\partial_i f(\xi)$ als auch ε echt größer null sind.
- $f(x') - f(x_0) \overset{(I)}{=} \left\langle \nabla f(\xi), \underbrace{x' - x_0}_{-\varepsilon e_i} \right\rangle = -\partial_i f(\xi) \cdot \varepsilon < 0$

Gelten diese beiden Aussagen, so ist allerdings $f(x_0)$ kein Extrempunkt, wie von der Annahme vorausgesetzt. ↯

Fall b: $\partial_i f(x_0) < 0$ – *verläuft analog*

Aus beiden Widersprüchen ziehen wir den Schluss, dass die Annahme falsch sei, demnach folgt also

$\nabla f(x_0) = 0$. An Extremstellen gilt damit:

$$f(x) \overset{(II)}{=} f(x_0) + \underbrace{\left\langle \overbrace{\nabla f(x_0)}^{=0}, x - x_0 \right\rangle}_{=0} + \frac{1}{2}\langle \mathcal{H}f(x - x_0), x - x_0\rangle = f(x_0) + \frac{1}{2}\underbrace{\langle \mathcal{H}f(x - x_0), x - x_0\rangle}_{\substack{\text{Vorzeichen entscheidet,}\\ \text{ob Min. oder Max.}\\ \rightarrow\text{Definitheit}}}$$

Mit dieser Überlegung kommt man auf folgenden Satz über notwendige und hinreichende Kriterien für Extremstellen:

Satz 1.3 (*Notwendige und Hinreichende Kriterien für Extremstellen*)

Sei $f : D \subseteq \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(D), x_0 \in \mathring{D}$, so gelten folgende Implikationen:

(a) $\nabla f(x_0) = 0 \wedge \mathcal{H}f(x_0)$ ist **negativ** definit $\Rightarrow x_0$ ist (isoliertes) **lok. Maximum**

(b) $\nabla f(x_0) = 0 \wedge \mathcal{H}f(x_0)$ ist **positiv** definit $\Rightarrow x_0$ ist (isoliertes) **lok. Minimum**

(c) $\nabla f(x_0) = 0 \wedge \mathcal{H}f(x_0)$ ist **in**definit $\Rightarrow x_0$ ist kein Extremum, sondern Sattelpunkt

} hinreichende Kriterien

(d) x_0 ist lokale Extremstelle $\Rightarrow \nabla f(x_0) = 0$

(e) x_0 ist lokales Maximum $\Rightarrow \nabla f(x_0) = 0 \wedge \mathcal{H}f(x_0)$ ist negativ **semi**definit

(f) x_0 ist lokales Minimum $\Rightarrow \nabla f(x_0) = 0 \wedge \mathcal{H}f(x_0)$ ist positiv **semi**definit

} notwendige Kriterien

Beweis: Der Satz ergibt sich aus unserer vorangehenden Überlegung, ist damit auch korrekt. □

! Die bisherigen Kriterien gelten selbstverständlich nur für das Innere des Definitionsbereiches ($\mathring{D}$). Randpunkte sind extra zu beachten!

Wir wollen uns mit diesem Problem näher in Kapitel 1.2 beschäftigen. Die Idee ist es am Rand die Funktion unter gewissen Nebenbedingungen zu optimieren, wir stellen dazu ein Lagrangefunktional auf, was uns dann eine ähnliche Methodik erlaubt anzuwenden, wie wir sie hier schon kennengelernt haben.

Wie kommt man jetzt von lokalen zu globalen Extremstellen? Eine der letzten Fragen, die sich in diesem Zusammenhang noch stellt, ist, ob es möglich ist, von den eben gefundenen lokalen Extremstellen auch auf Existenz von globalen Extremstellen zu schließen.
Es ist klar, dass wenn x_0 eine **globale** Extremstelle ist, x_0 ebenso eine **lokale** Extremstelle ist. Damit gilt also:

i Globale Extremstellen werden immer an lokalen Extremstellen angenommen oder sie existieren **nicht**!

Aus C2 wissen wir, dass wenn f stetig ist und D kompakt, so muss f auf D ein globales Maximum und ein globales Minimum annehmen. Dies führt uns zu folgender Vorgehensweise:

1. Berechne alle lokalen Extremstellen (kritische Punkte **und** Randpunkte).
2. Teile die Menge an lokalen Extremstellen E in die Menge an lokalen Maxima $\mathfrak{M}$ und lokalen Minima $\mathfrak{m}$ mit $\mathfrak{M} \cap \mathfrak{m} = \emptyset$ und $\mathfrak{M} \cup \mathfrak{m} = E$ auf.
3. Falls D kompakt ist, ist $\max \mathfrak{M}$ das globale Maximum und $\min \mathfrak{m}$ das globale Minimum, andernfalls versuche die Eigenschaft des globalen Maxi- und Minimums *per Hand* zu zeigen. Sonderfälle hierzu werden wir in Kapitel 1.2 und 1.5 kennenlernen.

1.2 Extremwertaufgaben mit Nebenbedingungen

Nach einfachen Extremwertproblemen wollen wir uns nun mit Problemen beschäftigen, die unter gewissen Nebenbedingungen zu optimieren sind. Sei im folgenden eine kurze Motivation für dieses Thema gegeben:

(1) Bezeichnen $G := \{\vec{x} \in \mathbb{R}^3 | \vec{x} = (x,y,z)^T, z = g(x,y)\}$ und $F := \{\vec{x} \in \mathbb{R}^3 | \vec{x} = (x,y,z)^T, z = f(x,y)\}$ Flächen im dreidimensionalen Raum. Suche nun ein $x_1 \in G$ und ein $x_2 \in F$ mit minimalem Abstand, also:

B Finde das Minimum der Funktion $h : \mathbb{R}^6 \to \mathbb{R}$ mit $h(x_1,x_2) := \|(x_1,y_1,z_1) - (x_2,y_2,z_2)\|$ unter den Nebenbedingungen $z_1 := g(x_1,y_1)$ und $z_2 := f(x_2,y_2)$.

(2) Seien mit F und G zwei Kurven wie folgt gegeben, wobei der minimale Abstand gesucht ist:
$F : \begin{cases} y = f_1(x) \\ z = f_2(x) \end{cases}$ und $G : \begin{cases} x = g_1(y) \\ z = g_2(y) \end{cases}$ Es ergibt sich also das Problem:

B Finde das Minimum der Funktion $h : \mathbb{R}^6 \to \mathbb{R}$ mit $h(x_1,y_1,z_1,x_2,y_2,z_2) := \|(x_1,y_1,z_1) - (x_2,y_2,z_2)\|$ unter den Nebenbedingungen $y_1 = f_1(x_1)$, $z_1 := f_2(x_1)$, $x_2 = g_1(y_2)$ und $z_2 := g_2(y_2)$.

(3) Gesucht sind die globalen Extrema einer Funktion $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ mit D kompakt. Dieses Problem wird in „zwei Einheiten“ oder Einzelprobleme unterteilt, so dass zuerst alle Extrema im inneren des Definitionsbereichs gesucht werden (siehe hierfür Kapitel 1.1) und danach das Optimierungsproblem mit Nebenbedingung $x_0 \in \partial D$ gelöst wird.

Wir wollen diese Probleme durch Anwendung des folgenden Satzes lösen:

Satz 1.4 (*Lagrange-Multiplikatoren bei einer Nebenbedingung*)

Seien $f, g : D \subseteq \mathbb{R}^n \to \mathbb{R}$ mit $f, g \in \mathcal{C}^1(D)$. Dann ist eine **notwendige** Bedingung für beliebige Extremstellen x_0 von f unter der Nebenbedingung $g(x_0) = 0$, dass ...

(I) $\mathcal{L}(x_0,\lambda) := \nabla f(x_0) - \lambda \nabla g(x_0) = 0$

(II) $g(x_0) = 0$

Man bezeichnet $\lambda \in \mathbb{R}$ dann auch als **Lagrange-Multiplikator**.

Beweis: Eine **notwendige** Bedingung bedeutet, dass wir im Beweis ausgehen, dass x_0 eine Extremstelle von f unter der Nebenbedingung $g(x_0) = 0$ ist, und damit zeigen, dass die Bedingung gilt. Da $\nabla g(x_0) \neq 0$ können wir nach eventueller Umnummerierung der Koordinaten annehmen, dass

$$\frac{\partial g}{\partial x_1}(x_0) \neq 0.$$

Wir setzen nun $x_0 = (a_1, x_0')$, wobei $x_0' = (a_2, \ldots, a_n)$. Nach Satz gibt es nun eine offene Umgebung $U' \subseteq \mathbb{R}^{n-1}$ von x_0', sowie eine stetig differenzierbare Abbildung $\varphi : U' \to \mathbb{R}$ mit $\varphi(x_0') = a_1$ und $g(\varphi(x'), x') = 0$ für alle $x' \in U'$, wobei $\varphi(U') \times U' \subset D$ ist. Sei $M := \{x \in D : g(x) = 0\}$, so liefert der Satz ebenso, dass sich M lokal auch durch $M \cap (\varphi(U') \times U') := \{x \in \varphi(U') \times U' : x_1 = \varphi(x_2, \ldots, x_n)\}$ beschreiben lässt. Wir wenden nun die Kettenregel auf obige Gleichung an und erhalten:

$$0 = \frac{\partial g}{\partial x_1}(x_0)\frac{\partial \varphi}{\partial x_i}(x_0') + \frac{\partial g}{\partial x_i}(x_0) \quad \text{für } i = 2, \ldots, n \tag{1.1}$$

Wir betrachten jetzt die Funktion $F : U' \to \mathbb{R}$ mit $F(x_2, \ldots, x_n) = f(\varphi(x_2, \ldots, x_n), x_2, \ldots, x_n)$. Da f auf M ein lokales Extremum besitzt, besitzt F auf $U' \subseteq \mathbb{R}^{n-1}$ im Punkte x_0' ein Extremum, also gilt nach

Satz , dass

$$\frac{\partial F}{\partial x_i}(x_0') = 0 \quad \text{für } i = 2, \ldots, n.$$

Mit der Kettenregel folgt hieraus:

$$0 = \frac{\partial F}{\partial x_i}(x_0') = \frac{\partial f}{\partial x_1}(x_0)\frac{\partial \varphi}{\partial x_i}(x_0') + \frac{\partial f}{\partial x_i}(x_0) \tag{1.2}$$

Setzt man nun

$$\lambda := \frac{\partial f}{\partial x_1}(x_0)\left(\frac{\partial g}{\partial x_1}(x_0)\right)^{-1},$$

so folgt aus (1.1) und (1.2) die Behauptung:

$$\frac{\partial f}{\partial x_i}(x_0) = \lambda \frac{\partial g}{\partial x_i}(x_0) \quad \text{für } i = 1, \ldots, n$$

□

Auch wenn die Methode der Lagrange-Multiplikatoren an sich wie eine Allzweckwaffe scheint, so ist es im Allgemeinen **sehr schwierig** ein nichtlineares Gleichungssystem mit $n+1$ Gleichungen zu lösen. Seien hier noch ein paar Tipps und Tricks für das Umgehen mit Lagrangefunktionalen angemerkt:

i

- Man sollte generell darauf achten, keinen Fall zu vergessen! Vor allem beim Kürzen von Variablen (also auch λ) muss auf den Fall Variable $= 0$ aufgepasst werden.
- Generell werden hierbei **kritische Stellen** von $\mathcal{L}$ gesucht, $\mathcal{L}$ heißt auch **Lagrange-Funktional**
- Ist $f = g \circ \widetilde{f}$ mit $g : \mathbb{R} \to \mathbb{R}$ monoton, so haben f und $\widetilde{f}$ die gleichen Extrem**stellen**.
- Multiplikation mit Größen, die zu 0 gleich sind, ist *nicht strengstens verboten*. Dies erhöht allerdings die Anzahl an möglichen Extremstellen.
- Der Lagrangemultiplikator λ kann auch additiv hinzugefügt werden, welche Variante genommen wird/werden soll, hängt vom Problem ab.

Beispiel 1.1:

B Wir suchen die Extremstellen der Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ mit $f(x,y) := y - x^2$ unter der Nebenbedingung $x^2 + y^2 = 1$.

Anhand der Nebenbedingung können wir die notwendige Funktion $g : \mathbb{R}^2 \to \mathbb{R}$ aufstellen mit $g(x,y) := x^2 + y^2 - 1 \stackrel{!}{=} 0$. Damit stellen wir dann das Lagrangefunktional, dessen kritische Stellen wir bestimmen wollen, auf und erhalten:

$$\mathcal{L}(x,y,\lambda) = (y - x^2) - \lambda(x^2 + y^2 - 1)$$

und damit für den Gradienten

$$\nabla\mathcal{L}(x,y,\lambda) = \begin{pmatrix} -2\cdot x - 2\cdot\lambda\cdot x \\ 1 - 2\cdot\lambda\cdot y \\ -x^2 - y^2 + 1 \end{pmatrix} = \begin{pmatrix} -2\cdot x\cdot(1+\lambda) \\ 1 - 2\cdot\lambda\cdot y \\ -x^2 - y^2 + 1 \end{pmatrix} \stackrel{!}{=} \vec{0}.$$

Wir nehmen Gleichung 1 und erhalten damit zwei Möglichkeiten $x = 0$ oder $\lambda = -1$.

Wir müssen nun **beide** Möglichkeiten näher untersuchen:

Fall 1: $x = 0$ – Wir erhalten damit:

$$-y^2 + 1 = 0 \rightsquigarrow y = \pm 1$$

in Gl. 2 eingesetzt $\quad 1 \mp 2 \cdot \lambda = 0 \rightsquigarrow \lambda = \pm\frac{1}{2}$

Wir erhalten somit zwei **mögliche Extremstellen** mit $(0,1)$ und $(0,-1)$.

Fall 2: $\lambda = -1$ – Wir erhalten damit:

$$1 + 2 \cdot y = 0 \rightsquigarrow y = -\frac{1}{2}$$

in Gl. 2 $\quad -x^2 - \frac{1}{4} + 1 = 0 \rightsquigarrow x = \pm\frac{\sqrt{3}}{2}$

Wir erhalten somit also wieder zwei **mögliche Extremstellen** mit $\left(+\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)$ und $\left(-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)$.

Durch Einsetzen bestimmen wir Maximum und Minimum, also

$$f\left(\pm\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) = -\frac{1}{2} - \left(\pm\frac{\sqrt{3}}{2}\right)^2 = -\frac{1}{2} - \frac{3}{4} = -\frac{5}{4} \qquad f(0, \pm 1) = \pm 1 - 0^2 = \pm 1$$

Wir erkennen somit, dass f unter der Nebenbedingung g das Maximum 1 bei $(0,1)$ und das Minimum $-\frac{5}{4}$ bei $\left(\pm\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)$ annimmt, womit die Extremstellen bestimmt wären. ※

Satz 1.4 lässt sich selbstverständlich auf auf mehrere Nebenbedingungen erweitern, er sei hier der Vollständigkeit halber noch mit angegeben:

Satz 1.5 (*Lagrange-Multiplikatoren bei mehreren Nebenbedingungen*)

Sei $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, $g : D \to \mathbb{R}^m$ mit $m \leqslant n$. Seien $f, g \in \mathcal{C}^1(D)$. Ferner besitze f in ξ unter der Nebenbedingung $g(\xi) = 0$ eine Extremstelle. Dann ist eine **notwendige** Bedingung für die Existenz dieser Extremstelle, dass ...

(I) $\nabla f(x_0) = \sum_{i=1}^{m} \lambda_i \cdot \nabla g_i(\xi)$

(II) $g(\xi) = 0$

$\lambda \in \mathbb{R}^m$ beschreibt wieder den Lagrangemultiplikator, das dazugehörende Funktional ist $\mathcal{L}(\xi, \lambda) := f(\xi) - \langle \lambda, g(\xi) \rangle$.

Beweis: Der Beweis verläuft analog zu dem von Satz 1.4. □

1.3 Satz über implizite Funktionen

Ein weiteres sehr interessantes Thema, was nun näher beleuchtet werden soll, sind implizite Funktionen. Mit dem hauptsächlichen Satz in dieser Sektion ist es möglich Aussagen zu treffen, ob eine implizit gegebene Funktion – zumindest lokal – eindeutig aufgelöst werden kann. Anwendungsgebiete sind beispielsweise das Finden von (lokalen) Umkehrfunktionen oder aber das Überprüfen der „Glattheit“ einer Kurve. Wir beschäftigen uns somit mit folgender Problemstellung:

i Sei $f : \mathbb{R}^2 \to \mathbb{R}$ – oder auch allgemein $\mathbb{R}^{n+m} \to \mathbb{R}^m$ – eine hinreichend glatte Funktion und sei $(x, y) \in \mathbb{R}^{n+m}$ ein Punkt mit $f(x, y) = 0$. Gibt es nun „lokal“ um x_* eine Auflösungsfunktion $x \mapsto y(x)$, die die durch $f(x, y) = 0$ charakterisierte Menge darstellt?

Das heißt, gibt es ein $y(x)$, so dass für ein $\varepsilon > 0$ und alle $x_* \in K_\varepsilon(x)$ $f(x, y(x)) = 0$ gilt?

Grundlage ist also der Folgende Satz:

Satz 1.6 (*Satz über implizite Funktionen*)

Sei $f : D \subseteq \mathbb{R}^2 \to \mathbb{R}$, wobei D offen, $f \in \mathcal{C}^1(D)$ und $(x_0, y_0) \in D$ mit $f(x_0, y_0) = 0$ gelte. Es gelte ferner, dass

$$\frac{\partial f}{\partial y}(x_0, y_0) \neq 0.$$

Dann existiert eine Umgebung $U_\varepsilon = (x_0 - \varepsilon, x_0 + \varepsilon)$ von x_0 und genau eine Abbildung $x \mapsto y(x), U \to \mathbb{R}$ – die sogenannte **lokale Auflösungsfunktion** – derart, dass $f(x, y(x)) = 0$ und $y_0 = y(x_0)$ ist.
Die lokale Auflösungsfunktion ist dann stetig differenzierbar und es gilt

$$y'(x) = -\frac{\partial_x f(x, y(x))}{\partial_y f(x, y(x))}$$

Beweis: Wir wollen Satz 1.6 jetzt an dieser Stelle nicht rigoros und umfassend beweisen, da seine Aussage aus Satz 1.7 folgt. Vielmehr soll hier eine andere Idee vorgestellt werden, aus der dann numerische Verfahren folgen können.
Wir definieren uns zuerst eine Funktionenfolge $y_{n+1} := y_n(x) - \frac{f(x, y_n(x))}{\partial_y f(x, y_n(x))}$ mit $y_0(x) := y_0$. Man zeigt nun die Konvergenz auf einer $K_\varepsilon(x_0)$ unter den gegebenen Voraussetzungen. Wir dies möglich ist, erfahren wir in Kapitel 1.7 mit Satz 1.22. Die Grenzfunktion bekommt man durch das Anwenden des Grenzwertes. Nach einem Satz aus C2 gilt somit:

$$y(x) = y(x) - \frac{f(x, y(x))}{\partial_y f(x, y(x))},$$

also $f(x, y(x))$.
Die verwendete Funktionenfolge eignet sich am besten als **Iterationsvorschrift** für das Auffinden lokaler Umkehrfunktionen. □
Wir verallgemeinern den Satz also, womit wir uns dann mit nichtlinearen Gleichungen,

$$f(x) = 0,$$

beschäftigen, wobei $f = (f_1, \ldots, f_m)$ stetig differenzierbar auf einem Gebiet $G \subset \mathbb{R}^k$ und $m < k$ ist. Man hat dann m skalare Gleichungen für $k = m + n$ Variablen, $n > 0$, also ein „unterbestimmtes System“. Die Lösungsmenge sollte sich unter geeigneten Voraussetzunge durch n Parameter beschreiben lassen. Wir sagen voraus:

Satz 1.7 (*Satz über implizite Funktionen im mehrdimensionalen Fall*)

Sei $f : D \subseteq \mathbb{R}^{n+m} \to \mathbb{R}^m$, wobei D offen, $f \in \mathcal{C}^1(D)$, $n, m \in \mathbb{N}$ und $(x_0, y_0) \in D$ mit $x_0 \in \mathbb{R}^n$, $y_0 \in \mathbb{R}^m$ und $f(x_0, y_0) = 0$ gelte. Es gelte ferner, dass die $m \times m$-Matrix $\frac{\partial f}{\partial y}(x_0, y_0)$ regulär – also invertierbar – ist. Dann existiert eine Umgebung $U_\varepsilon = K_\varepsilon(x_0) \subset \mathbb{R}^n$ von x_0 und genau eine Abbildung $x \mapsto y(x), U \to \mathbb{R}^m$ – die sogenannte **lokale Auflösungsfunktion** – derart, dass für alle $x \in U$ gilt, dass $f(x, y(x)) = 0$ und $y_0 = y(x_0)$.
Die lokale Auflösungsfunktion ist dann stetig differenzierbar und es gilt

$$\mathcal{J}y(x) = \frac{dy}{dx}(x) = -\left[\frac{\partial f}{\partial y}(x, y(x))\right]^{-1} \cdot \frac{\partial f}{\partial x}(x, y(x))$$

Beweis: Sei $k := n + m$. Wir betrachten also nun ein Gebiet $G \subset \mathbb{R}^k = \mathbb{R}^n \times \mathbb{R}^m$ und eine stetig differenzierbare Abbildung $f = (f_1, \ldots, f_m) : G \to \mathbb{R}^m$, also ein System von m nichtlinearen Gleichungen für $n + m$ Variablen. Die Gleichungen schaffen Abhängigkeiten zwischen den Variablen.

Es gibt dafür zwar viele Möglichkeiten, wir untersuchen der Einfachheit halber zunächst nur die Situation, dass die Variablen $x_{n+1},\dots,x_{n+m}$ differenzierbar von den Variablen $x_1,\dots,x_n$ abhängen. Der Satz der ersten n Variablen fassen wir zu einem Vektor x, den der folgenden m Variablen zu einem Vektor y zusammen. Dann definieren wir weiter:

$$\frac{\partial f}{\partial x} := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \quad \text{und} \quad \frac{\partial f}{\partial y} := \begin{pmatrix} \frac{\partial f_1}{\partial x_{n+1}} & \cdots & \frac{\partial f_1}{\partial x_{n+m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_{n+1}} & \cdots & \frac{\partial f_m}{\partial x_{n+m}} \end{pmatrix}$$

Damit ist

$$\mathcal{J}_f(x,y) = \left(\frac{\partial f}{\partial x}(x,y) \;\middle|\; \frac{\partial f}{\partial y}(x,y) \right).$$

Damit die gemeinsame Nullstellenmenge $N = \{(x,y) \in G : f(x,y) = 0\}$ in der Nähe eines Punktes $(x_0,y_0) \in N$ lokal wie der Graph einer Abbildung $y = y(x)$ aussieht, darf sie in (x_0,y_0) keine vertikale Tangente besitzen. Das wiederum bedeutet – da N Niveaumenge von f ist –, dass kein vertikaler Vektor $(0,b)$ mit $b \neq 0$ simultan auf allen Gradienten $\nabla f_i(x_0,y_0)$, $i = 1,\dots,m$, senkrecht stehen darf. Und das bedeutet, dass

$$\frac{\partial f}{\partial y}(x_0,y_0) \cdot b^T \neq 0^T$$

für alle $b \neq 0$ gelten muss.
Ein Gleichungssystem der Gestalt $A \cdot z^T = 0^T$ mit einer Matrix $A \in \mathbb{R}^{m\times m}$ hat genau dann nur die triviale Lösung, wenn A regulär ist (siehe C1 für diesen Satz). Angewandt auf das obige Problem bedeutet das, dass $\det \frac{\partial f}{\partial y}(x_0,y_0) \neq 0$ sein sollte.
Aus der Gleichung $f(x,g(x)) \equiv 0$ folgt dann mit der Kettenregel:

$$0 = \frac{\partial f}{\partial x}(x,g(x)) \cdot E_k + \frac{\partial f}{\partial y}(x,g(x)) \cdot \mathcal{J}_g(x),$$

also

$$\mathcal{J}y(x) = -\left[\frac{\partial f}{\partial y}(x,y(x))\right]^{-1} \cdot \frac{\partial f}{\partial x}(x,y(x)).$$

Der Trick des restlichen Beweises besteht nun darin, den Raum um (x_0,y_0) herum so differenzierbar zu verbiegen, dass aus den Niveaumengen

$$N_c(f) := \{(x_1,\dots,x_k) : \forall i = 1,\dots,m : f_i(x_1,\dots,x_k) = c_i\}$$

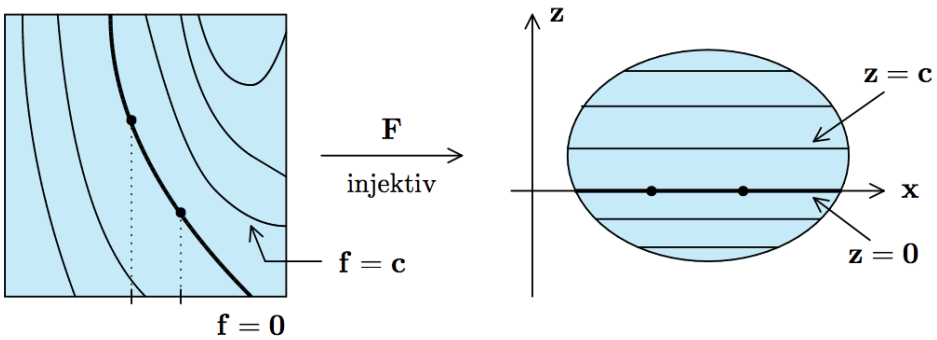
Ebenenstücke der Gestalt

$$E = \{(x_1,\dots,x_n,z_{n+1},\dots,z_k) : \forall i \in \{1,\dots,m\} : z_{n+i} = c_i\}$$

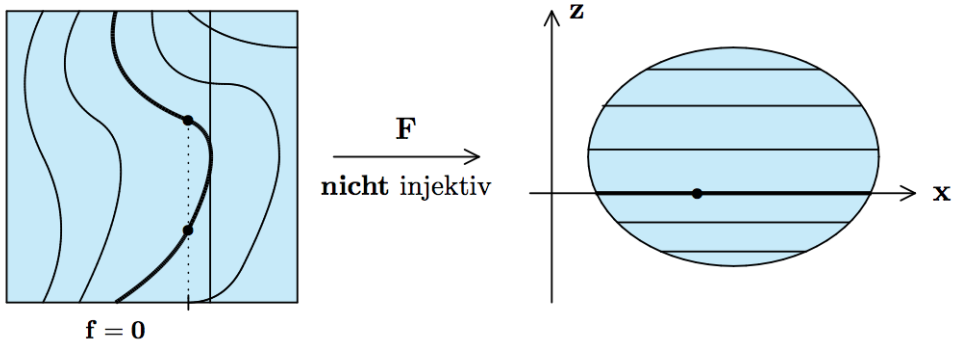
werden. Zu diesem Zweck definieren wir $F : B \to \mathbb{R}^k$ durch

$$F(x,y) := (x, f(x,y)).$$

Die Transformation kann man sich dann wie folgt vorstellen:



Die Punkte behalten bei dieser Transformation ihre x-Komponente, während ihre y-Komponente durch den Wert von f ersetzt wird. Das funktioniert aber nur, wenn verschiedene Punkte einer betroffenen Niveaufläche auch verschiedene x-Komponenten besitzen, wenn also die Niveaufläche keine vertikale Tangente besitzt. Dafür benötigen wir die vorausgesetzte Regularität von $\frac{\partial f}{\partial y}(x_0, y_0)$. Ein Fall, bei dem dies schiefgeht, könnte folgendermaßen aussehen:



Wir zeigen jetzt, dass F unter den Voraussetzungen des Satzes in der Nähe von (x_0, y_0) ein Diffeomorphismus[1] ist. Tatsächlich ist

$$\mathcal{J}_F(x_0, y_0) = \left(\begin{array}{c|c} E_k & 0 \\ \hline \frac{\partial f}{\partial x}(x_0, y_0) & \frac{\partial f}{\partial y}(x_0, y_0) \end{array}\right)$$

und daher

$$\det \mathcal{J}_F(x_0, y_0) = \det \frac{\partial f}{\partial y}(x_0, y_0) \neq 0.$$

Das bedeutet, dass F in (x_0, y_0) lokal umkehrbar ist. Wir setzten $H := F^{-1}$ (in der Nähe von $F(x_0, y_0)$). Weil F die ersten k Komponenten unverändert lässt, gilt das Gleiche für H. Also hat H die Gestalt

$$H(u, v) = (u, h(u, v)),$$

mit einer differenzierbaren Abbildung h.
Ist $f(x, y) = 0$, so ist $F(x, y) = (x, 0)$, also $(x, y) = F^{-1}(x, 0) = H(x, 0) = (x, h(x, 0))$. Deshalb setzen wir

$$y(x) := h(x, 0).$$

Offensichtlich ist g stetig differenzierbar. Ist $f(x, y) = 0$, so ist nach Konstruktion $y = y(x)$. Und umgekehrt ist

$$f(x, y(x)) = f(x, h(x, 0)) = f(F^{-1}(x, 0)) = 0.$$

Die Gleichung $f(x_0, y_0) = 0$ ergibt die Beziehung $y_0 = y(x_0)$. Wählt man nun die Umgebung U_ε klein genug, so ist alles gezeigt. □

Satz 1.8 (*Inverse Abbildung*)

Sei $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ mit D offen, $f \in \mathcal{C}^1(D)$ und $x_0 \in D$. Es sei die Jacobi-Matrix $\mathcal{J}_f(x_0) = \frac{\partial f}{\partial x}(x_0)$ invertierbar. Dann existiert eine Umgebung $U \subseteq D$ von x_0, so dass $f|_U : U \to f(U)$ umkehrbar (sprich bijektiv) ist.

Beweis: Wir können hier Satz 1.7 anwenden, wobei zu beachten ist, dass hier die Rollen von x und y vertauscht sind. Wir schreiben $F(x, y) = f(x) - y$. Ist nun $f(x_0) = \eta$ und die Matrix $\frac{\partial f}{\partial x}(\eta)$ regulär, so ist $F(x_0, \eta) = 0$ und $F_x(x_0, \eta)$ regulär. Nach Satz **??** gibt es daher eine Umgebung U von x_0 und eine Umgebung V von η sowie eine Funktion $g : V \to U$ stetig differenzierbar mit $F(g(y), y) = 0$ oder

[1] Ein **Diffeomorphismus** ist eine bijektive, stetig differenzierbare Abbildung, deren Umkehrabbildung auch stetig differenzierbar ist.

$f(g(y)) = y$. f ist damit bijektiv von $U' = g(V) \subset U$ nach V. Für die Funktionialmatrix von g erhalten wir

$$\frac{\partial g}{\partial y} = -\left(\frac{\partial F}{\partial x}\right)^{-1} \frac{\partial F}{\partial y} = \left(\frac{\partial f}{\partial x}\right)^{-1}.$$

Dies stellt eine Verallgemeinerung der wohlbekannten Formel für die eindimensionale Umkehrfunktion dar ($\rightarrow$ C2). f ist auf U definiert und stetig, U′ ist die Urbildmenge der offenen Menge V und damit selber offen. □

Korrolar 1.8 (*Iterationsverfahren*)

Anwendung des Iterationsverfahrens auf $\tilde{f}(x,y) := y - f(x)$ liefert:

$$\begin{aligned} x_{n+1}(y) &:= x_n(y) - \left[\frac{\partial f}{\partial x}(x_0, y_0)\right]^{-1} \tilde{f}(x_n(y), y) \\ &= x_n(y) - [\mathcal{J}f(x_0, y_0)]^{-1} \tilde{f}(x_n(y), y) \qquad \text{mit } x_0(y) := x_0. \end{aligned}$$

! Wenn also für alle $x \in D$ das Inverse der Jacobi-Matrix $[\mathcal{J}\vec{f}(x)]^{-1}$ existiert, so existiert auch für alle $x \in D$ eine lokale Umkehrfunktion. Dies muss aber noch lange **nicht** die Existenz einer globalen Umkehrfunktion bedeuten!

1.4 Parameterdarstellung von Kurven, Kurvenintegrale

Wir haben in Kapitel 1.3 gesehen, dass die Darstellung von Kurven meist nur iterativ und lokal geschehen kann. In diesem Kapitel wollen wir uns deswegen kurz anschauen, wie man diese Probleme lösen kann mittels der Parametrisierung von Kurven.

1.4.1 Grundlegendes

Definition 1.3 (*Parameterdarstellung einer Kurve*)

Eine **stetige** Abbildung $\gamma : \mathbb{R} \supseteq I \rightarrow \mathbb{R}^n$ mit I **abgeschlossen**, heiße **Parameterdarstellung** der Kurve $\Gamma : \gamma(I) \subset \mathbb{R}^n$.
Γ heiße $\mathcal{C}^1$-Kurve genau dann, wenn γ stetig differenzierbar sei. Γ heiße **geschlossen** genau dann, wenn $I = [a, b]$ mit $\gamma(a) = \gamma(b)$.

Satz 1.9 (*Tangente an eine Kurve in Parameterdarstellung*)

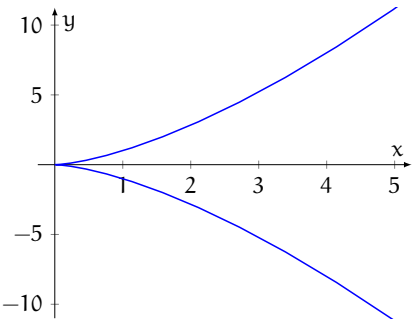
Eine durch γ parametrisierte $\mathcal{C}^1$-Kurve hat im Punkt $\gamma(t)$ die Tangentenrichtung $(\gamma(t))'$.

Beweis: Betrachte zunächst eine Sekante zwischen t und t_0, deren Richtung $\frac{\gamma(t)-\gamma(t_0)}{t-t_0}$ entspricht. Wir bilden nun den Grenzwert $t \rightarrow t_0$:

$$\lim_{t \to t_0} \frac{\gamma(t) - \gamma(t_0)}{t - t_0} = \gamma'(t_0)$$

Der Grenzwert muss dabei existieren, da Γ eine $\mathcal{C}^1$-Kurve ist. □
Wir wollen uns nun kurz anschauen wann eine Kurve eigentlich „glatt“ ist. Eine erste Idee wäre es, zu fordern, dass die Kurve eine $\mathcal{C}^1$-Kurve sein muss oder es ausreicht, wenn sie differenzierbar ist. Dass das aber nicht ausreicht zeigt folgendes Beispiel:

Beispiel 1.2:
Wir betrachten eine *Nellsche* Parabel (*siehe rechts*) mit der Parametrisierung $\gamma(t) = \begin{pmatrix} t^2 \\ t^3 \end{pmatrix}$. Man kann leicht nachrechnen, dass diese Kurve sogar die $\mathcal{C}^\infty$-Eigenschaft besitzt, jedoch an der Stelle $\gamma(0) = 0$ einen „Knick“ hat. Dies ist nur möglich, weil $\gamma'(0) = 0$.
Anschaulicher: Um „abrupte Bewegungen“ auf einer Kurve zu machen, muss man nicht zwingend auch „abrupte Lenkbewegungen“ machen. Es reicht aus, wenn man die Geschwindigkeit auf 0 abbremst und dann „sanfte“, „stetige“ Lenkbewegungen macht.



※

Wir fordern also für „glatte“ Kurven:

Definition 1.4 (*Regularität einer Kurve*)

Eine Parameterdarstellung $\gamma : I \to \mathbb{R}^n$ heiße „**glatt**“ oder „**regulär**“ gdw. $\gamma \in \mathcal{C}^1(I)$ und für alle $t \in I$ gilt, dass $\gamma'(t) \neq 0$.

Eine noch wichtigere Erkenntnis ist, dass Parametrisierungen **nicht** eindeutig sind. Als Beispiel lässt sich hier eine Einheitskreisparametrisierung nennen, es gibt aber unendlich viele davon. Wir fassen diese Erkenntnis nun in folgendem Satz zusammen:

Satz 1.10 (*Umparametrisierungen*)

Ist $\gamma : I \to \mathbb{R}^n$ die Parametrisierung einer Kurve und $u : \widetilde{I} \to I$ stetig und **sur**jektiv, sowie $\widetilde{I} \subset \mathbb{R}$ ein Intervall, so ist $\widetilde{\gamma} = \gamma \circ u : \widetilde{I} \to R^n$ dann ebenfalls eine Parametrisierung dieser Kurve. Umgekehrt hat **jede** Parametrisierung der Kurve die Form $\gamma \circ u$, wobei $u : \widetilde{I} \to I$ stetig und **sur**jektiv, sowie $\widetilde{I} \subset \mathbb{R}$ ein Intervall ist.
Eine Umparametrisierung einer **regulären** Kurve erfordert, dass $u \in \mathcal{C}^1$ sowie für alle $t \in \widetilde{I}$ gilt, dass $u'(t) \neq 0$. Somit ist u dann **global** streng monoton steigend oder fallend, wobei sich im letzteren Fall bei der Umparametrisierung der Durchlaufsinn der Kurve ändert.

Wir beschäftigen uns nun mit verschiedenen Maßen auf/von Kurven, zuerst mit der Bogenlänge. Dazu definieren wir wie folgt:

Definition 1.5 (*Bogenlänge einer Kurve*)

Sei $\gamma : [a, b] \to \mathbb{R}^n$ die Parametrisierung einer Kurve Γ. Gilt für jede Folge von Zerlegungen Z_k mit $|Z_k| := \max_{i \geqslant 1} t_i - t_{i-1} \to 0$, dass die Folge der Längen der Polygonzüge L_{Z_k} konvergent ist, und zwar **immer gegen denselben Wert**, dann heißt dieser Grenzwert

$$|\Gamma| := \lim_{|Z_k| \to 0} L_{Z_k} \quad \textbf{Bogenlänge der Kurve } \Gamma.$$

Γ heißt dann auch **rektifizierbar**.

Da diese Berechnung sehr umständlich ist, finden wir einen anderen Weg, die Bogenlänge zu berechnen über folgenden Satz:

Satz 1.11 (*Bogenlänge einer Kurve*)

Sei $\gamma : I \to \mathbb{R}^n$ eine Parametrisierung einer regulären Kurve Γ. Dann ist die Bogenlänge $|\Gamma|$ gegeben als

$$|\Gamma| = \int_I \left\| \gamma'(t) \right\| dt.$$

Beweis: Sei $I = [a, b]$ und beschreibe $\gamma : I \to \mathbb{R}^n$ eine reguläre Kurve Γ. Nach Satz 1.9 gilt dann

$$\gamma'(t) = \lim_{h \to 0} \frac{\gamma(t+h) - \gamma(t)}{h}.$$

Beschreibe $s : I \to \mathbb{R}^+$ die Weglänge, so ist s differenzierbar und es gilt:

$$s'(t) = \lim_{h\to 0} \frac{s(t+h) - s(t)}{h} = \left\|\gamma'(t)\right\|,$$

woraus unmittelbar folgt:

$$s(t) = |\Gamma| = \int_I s'(s)ds = \int_I \left\|\gamma'(s)\right\| ds$$

□

Beispiel 1.3: Betrachten wir als Beispiel die Parametrisierung der Kurve Γ $\gamma : [0, 2\pi] \to \mathbb{R}^2$ mit

$$\gamma(t) := r \cdot \begin{pmatrix} t - \sin t \\ 1 - \cos t \end{pmatrix}$$

Gesucht ist die Bogenlänge $|\Gamma|$. Wir berechnen also:

$$\gamma'(t) = r \cdot \begin{pmatrix} 1 - \cos t \\ \sin t \end{pmatrix}$$

Hiervon berechnen wir die Norm. Welche Norm wir dafür verwenden ist egal, da nach Satz aus Mathe C1, alle Normen auf dem $\mathbb{R}^n$ äquivalent sind.

$$\begin{aligned}\left\|\gamma'(t)\right\| &= \|r\| \cdot \sqrt{(1-\cos(t))^2 + \sin^2(t)} \\ &= \|r\| \cdot \sqrt{1 - 2\cdot\cos(t) + \cos^2(t) + \sin^2(t)} = \sqrt{2} \cdot \|r\| \cdot \sqrt{1-\cos t} \\ &= 2 \cdot \|r\| \cdot \left|\sin\left(\frac{t}{2}\right)\right|\end{aligned}$$

Damit berechnen wir dann mit Satz 1.11 die gesuchte Bogenlänge

$$\begin{aligned}|\Gamma| = \int_0^{2\pi} 2 \cdot \|r\| \cdot \left|\sin\left(\frac{t}{2}\right)\right| dt \quad &= 2 \cdot \|r\| \cdot \int_0^{2\pi} \left|\sin\left(\frac{t}{2}\right)\right| \\ = 2 \cdot \|r\| \cdot \left|-\cdot 2\cos\left(\frac{t}{2}\right)\right|\Big|_0^{2\pi} &= 4 \cdot \|r\| \cdot (-\cos(\pi) + \cos(0)) = 8 \cdot \|r\|\end{aligned}$$

※

1.4.2 Parametrisierungen nach der Bogenlänge

Wir wollen nun eine spezielle Parametrisierung näher betrachten, welche die Berechnung von Bogenlängen vereinfacht. Wir nennen sie deswegen auch „**Parametrisierung nach der Bogenlänge**“. Eine Motivation derselbigen lässt sich über die Krümmung definieren. Dabei bezeichnet die Krümmung κ das Folgende:

Definition 1.6 (*Krümmung* κ)

Die **Krümmung** κ ist definiert als der reziproke Wert des Radius R, des sich an die Kurve Γ anschmiegenden Kreises. Sei γ_0 eine Parametrisierung von Γ nach der Bogenlänge, so gilt:

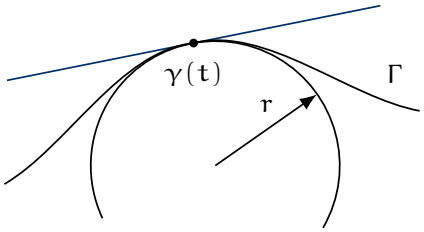
$$\kappa = \frac{1}{R(t)} = \left\|\gamma_0''(t)\right\|$$

Uns ist in diesem Moment streng genommen klar, dass wir an dieser Stelle Definition und Satz ein wenig vermischen, wir geben für die zweite Aussage $\kappa = \|\gamma_0''(t)\|$ deswegen nun einen Beweis an.
Beweis: Wir sehen ein, dass

$$1 = \|\gamma(s)\|^2 = \langle T(s), T(s)\rangle$$

Damit gilt aber auch

$$(\langle T(s), T(s)\rangle)' = \langle T(s), T'(s)\rangle + \langle T'(s), T(s)\rangle = 0,$$



womit dann aber aus $\langle T(s), T'(s)\rangle = 0$ folgt, dass

$$T'(s) = \kappa(s) \cdot \mathfrak{n}(s),$$

was wiederum bedeutet, dass

$$\kappa(s) = \|T'(s)\| = \|\gamma''(s)\|$$

□

Wir haben in Definition 1.6 bereits eine Parametrisierung nach Bogenlänge vorausgesetzt, wollen sie nun genauer spezifizieren und definieren.

Definition 1.7 (*Parametrisierung nach Bogenlänge*)

Wir suchen eine Parametrisierung $\gamma_0 : \widetilde{I} \to \mathbb{R}^n$ einer Kurve Γ, beschrieben durch $\gamma : I \to \mathbb{R}^n$ mit der Eigenschaft ...

$$\forall t \in \widetilde{I} : \|\gamma_0'(t)\| = 1 \tag{1.3}$$

Eine solche Parametrisierung γ_0 heiße dann **Parametrisierung nach der Bogenlänge**.

*Warum macht das Sinn?*Für beliebige Puntke $\gamma(t_1), \gamma(t_2)$ gilt:

$$\int_{t_1}^{t_2} \|\gamma_0'(t)\|\,dt = t_2 - t_1,$$

das heißt, $t_2 - t_1$ gibt genau den Abstand von $\gamma_0(t_1)$ zu $\gamma_0(t_2)$ „entlang der Kurve“ an. □

Wie findet man eine Solche?

Wir wissen aus Satz 1.10, dass es eine Darstellung $\gamma_0 = \gamma \circ u$ mit $u : \widetilde{I} \to I$ gibt. Aus Definition 1.7, Gleichung (1.3) folgt dann:

$$1 \overset{!}{=} \|\gamma_0'(t)\| = \|\gamma'(u(t)) \cdot u'(t)\| = |u'(t)| \cdot \|\gamma'(u(t))\|$$

Mit der zusätzlichen Forderung des sich nicht ändernden Durchlaufsinn ergibt sich folgende DGl.:

$$u'(t) = \frac{1}{\|\gamma'(u(t))\|}$$

Diese ist allerdings nicht trivial lösbar, genaueres dazu – unter anderem auch eine kurze Einführung in numerische Verfahren zum Lösen solcher Differentialgleichungen – aber erst in Kapitel 2.

1.4.3 Kurvenintegrale

Wir wollen nun die Bogenlänge verallgemeinern, jetzt soll eine Art „gewichtete Länge“ ausgerechnet werden. Ein Anwendungsbeispiel wäre ein Draht oder Faden mit variabler Stärke, wobei wir den Draht/Faden als parametrisierte Kurve γ betrachten, die Dichte – also $\frac{\text{Masse}}{\text{Stärke}}$ am Punkt $\gamma(t)$ sei dann beschrieben durch den drahtspezifischen Funktionswert $f(\gamma(t))$. Wie berechnet man nun die Gesamtmasse?

Wir definieren deswegen:

Definition 1.8 (*Kurvenintegral erster Art*)

Sei $\Gamma \subset \mathbb{R}^n$ eine reguläre Kurve mit Parametrisierung $\gamma : I \to \mathbb{R}^n$ und sei $f : \Gamma \to \mathbb{R}$ – eigentlich $f : \mathbb{R}^\kappa \supset D \supset \Gamma \to \mathbb{R}$ – eine **skalarwertige** Funktion. Dann bezeichnet – bei Existenz –

$$\int_\Gamma f \, ds := \int_I f(\gamma(t)) \left\|\gamma'(t)\right\| dt$$

das **Kurvenintegral erster Art** von f über Γ.

> **i** Für $f \equiv 1$ entspricht das Integral genau der Bogenlänge der Kurve, womit auch der ursprüngliche Kommentar, dass das Kurvenintegral erster Art nichts weiter als eine Verallgemeinerung der Bogenlängenberechnung sei, Sinn ergibt.

Ein anderes Anwendungsbeispiel sei die Bewegung einer Masse m durch ein Kraftfeld $x \mapsto F(x)$, $\mathbb{R}^3 \to \mathbb{R}^3$ – beispielsweise das Gravitationsfeld – entlang einer durch γ parametrisierten Kurve Γ. Welche Arbeit wird hierbei geleistet? Um diese Frage zu beantworten, verallgemeinern wir unser Kurvenintegral auf vektorwertige Funktionen. Wir definieren:

Definition 1.9 (*Kurvenintegral zweiter Art*)

Sei Γ eine reguläre Kurve mit Parametrisierung $\gamma : I \to \mathbb{R}^n$ und $F : \Gamma \to \mathbb{R}^n$ – eigentlich $F : \mathbb{R}^n \supset D \supset \Gamma \to \mathbb{R}^n$ – eine **vektorwertige** Funktion. Dann bezeichnet – bei Existenz –

$$\int_\Gamma F \bullet ds := \int_I \left\langle F(\gamma(t)), \gamma'(t) \right\rangle dt$$

das **Kurvenintegral zweiter Art** von F über Γ.

> **i** Das Kurvenintegral zweiter Art kann auch als Kurvenintegral erster Art der in Kurvenrichtung gerichteten Komponente der Funktion aufgefasst werden.

Eine sich jetzt noch stellende Frage ist, ob sich das Ergebnis des Kurvenintegrals im allgemeinen ändert, wenn man unterschiedliche Parametrisierungen verwendet. Wir stellen dazu folgenden Satz auf:

Satz 1.12 (*Unabhängigkeit des Kurvenintegrals von der Parametrisierung*)

Der Wert von $\int_\Gamma f \, ds$ ist unabhängig von der gewählten Parametrisierung $\gamma : I \to \mathbb{R}^n$.

Beweis: Betrachten wir zwei unabhängige Parametrisierungen $\gamma_1 : I_1 \to \mathbb{R}^n$, $\gamma_2 : I_2 \to \mathbb{R}^n$ der Kurve Γ. Da nach Definition 1.8 Γ regulär sein muss, existiert nach Satz 1.10 ein $u : I_1 \to I_2$ mit $u' \not\equiv 0$, so dass

$$\gamma_1 = \gamma_2 \circ u \tag{1.4}$$

Wir zeigen nun die Gleichheit:

$$\begin{aligned}\int_\Gamma f \, ds &= \int_{I_1} f(\gamma_1(t)) \cdot \left\|\gamma_1'(t)\right\| dt \overset{(1.4)}{=} \int_{I_1} f(\gamma_2 \circ u(t)) \cdot \left\|\gamma_2 \circ u'(t)\right\| dt \\ &= \int_{I_1} f(\gamma_2 \circ u(t)) \cdot \left\|\gamma_2'(u(t))\right\| \cdot \overbrace{\left|u'(t)\right|}^{u' \text{ ist skalar}} \, dt\end{aligned}$$

$$= \operatorname{sgn}(u') \cdot \int_{I_1} f(\gamma_2 \circ u(t)) \cdot \left\|\gamma_2'(u(t))\right\| \cdot u'(t)\,dt \qquad \text{Substituiere } s := u(t)$$

$$= \int_{I_2} f(\gamma_2(s)) \cdot \left\|\gamma_2'(s)\right\| ds$$

□

Ohne Beweis sei an dieser Stelle angemerkt, dass Satz 1.12 auch für das Kurvenintegral zweiter Art gilt.

1.4.4 Parametrisierungen von Flächen und Oberflächenintegrale

Neben eindimensionalen Kurven wollen wir uns jetzt mit Parametrisierungen von Flächen näher beschäftigen. Wir definieren dazu:

Definition 1.10 (*Parametrisierung einer Fläche*)

Eine stetige surjektive Abbildung $\gamma : \mathbb{R}^2 \supset M \to F \subset \mathbb{R}^n$ mit $(s,t) \mapsto \gamma(s,t)$ heiße **Parametrisierung der Fläche** $F \subset \mathbb{R}^n$.

Auch auf Flächen definieren wir – mit ähnlichen Motivationen – **Oberflächen**integrale erster und zweiter Art.

Definition 1.11 (*Oberflächenintegral erster Art*)

Sei F eine Fläche, die durch $\gamma : F \to \mathbb{R}^3$, $F \subset \mathbb{R}^2$ parametrisiert ist. Weiter sei $f : \mathbb{R}^3 \to \mathbb{R}$ **skalarwertig** und **stetig**. Dann ist – bei Existenz – das **(Ober-)Flächenintegral erster Art** von f über F definiert als

$$\int_F f \, do := \int_M f(\gamma(s,t)) \cdot \|\partial_s\gamma(s,t) \times \partial_t\gamma(s,t)\| \, ds \, dt$$

Dabei bezeichne $\times$ das Kreuzprodukt der Vektoren.

Definition 1.12 (*Oberflächenintegral zweiter Art*)

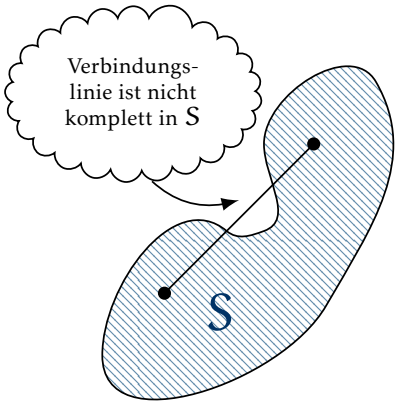
Sei F eine Fläche, die durch $\gamma : F \to \mathbb{R}^3$, $F \subset \mathbb{R}^2$ parametrisiert ist. Weiter sei $f : \mathbb{R}^3 \to \mathbb{R}$ ein **stetiges Vektorfeld** auf F. Dann ist – bei Existenz – das **(Ober-)Flächenintegral zweiter Art** von f über F definiert als

$$\int_F f \bullet do := \int_M \langle f(\gamma(s,t)), \partial_s\gamma(s,t) \times \partial_t\gamma(s,t)\rangle \, ds \, dt$$

Dabei bezeichne $\times$ wieder das Kreuzprodukt der Vektoren.

1.5 Konvexe, quadratische, linear-quadratische Optimierungsprobleme

Wir haben in Kapitel 1.1 notwendige und hinreichende Kriterien für *lokale* Extremstellen kennengelernt für Funktionen $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$. Bislang ist es uns nur im Falle von *kompakten Definitionsbereichen* gelungen von lokalen auf *globale* Extremata zu schließen. Wir wollen nun, indem wir *andere* Einschränkungen oder Annahmen an unseren Definitionsbereich und unsere Funktion stellen, stärkere Aussagen für mögliche Extremstellen erhalten. Dies werden beispielsweise *Eindeutigkeit* oder auch *Globalität* sein. Dazu definieren wir als erstes den Begriff der konvexen Menge.



Ein Beispiel für eine **nichtkonvexe** Menge.

Definition 1.13 (*Konvexe Menge*)

Sei $M \subseteq \mathbb{R}^n$ eine Menge. M heiße **konvex** genau dann, wenn mit beliebigen $x, y \in M$ auch die Verbindungsstrecke in M liegt, also

$$\forall x, y \in M : \forall \alpha \in [0,1] : \alpha x + (1-\alpha)y \in M$$

i

Interessanterweise gibt es in der Mathematik für Mengen nicht den Begriff der Konkavheit.

Jeder Vektorraum, der $\mathbb{R}$ enthält, ist konvex, ebenso alle Halbebenen und Halbräume.

Wir definieren dann weiter:

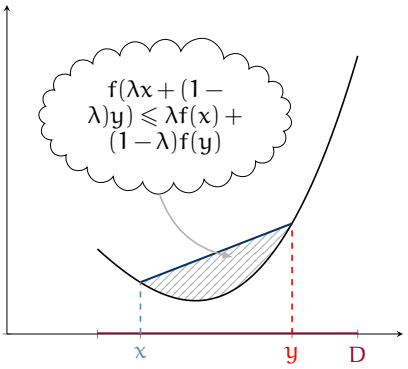
Definition 1.14 (*Konvexe Funktion*)

Sei $f : D \to \mathbb{R}$ mit $\emptyset \neq D \subseteq \mathbb{R}^n$ **konvex**. Wir sagen f ist **konvex** genau dann, wenn

$$\forall x, y \in D : \forall \lambda \in [0,1] : f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y)$$

f heiße **strikt konvex** genau dann, wenn

$$\forall x, y \in D : \forall \lambda \in [0,1] : f(\lambda x + (1-\lambda)y) \lneqq \lambda f(x) + (1-\lambda) f(y)$$



B

Beispiele für (strikt) konvexe Funktionen sind:

- Affin-Lineare Funktionen — konvex
- Normalparabel — strikt konvex
- Betragsfunktion — konvex
- e-Funktion — strikt konvex

Wir wollen nun untersuchen, wann eine Funktion konvex ist. Aussage darüber trifft der Folgende Satz:

Satz 1.13 (*Kriterium für Konvexität von Funktionen*)

Sei $D \subseteq \mathbb{R}^n$ eine konvexe Menge und sei $f \in \mathcal{C}^2(D)$. Dann gilt:

(1) f ist **konvex** $\iff$ Für alle $x \in D$ gilt, dass $\mathcal{H}f(x)$ positiv **semi**definit ist.

(2) f ist **strikt konvex** $\impliedby$ Für alle $x \in D$ gilt, dass $\mathcal{H}f(x)$ positiv **definit** ist.

(3) Die Umkehrung von (2) gilt **nicht**.

Beweis:

zu (1) Wir definieren uns zuerst folgenden Hilfssatz:

Hilfssatz 1 (*Hilfssatz zu Satz 1.13*)

Sei $f \in \mathcal{C}^2((a,b))$ mit $f : (a,b) \to \mathbb{R}$. f ist konvex genau dann, wenn $f'' \geqslant 0$ für alle $x \in (a,b)$.

Beweis:

„⇐“ Angenommen $f'' \geqslant 0$ auf (a,b). Dann ist f' monoton steigend auf (a,b). Für $a < x < y < b$, $0 < \lambda < 1$ und $z = (1-\lambda)x + \lambda y$ gilt dann:

- $f(z) - f(x) = \int_x^z f'(t)\,dt \leqslant f'(z)(z-x)$

  und
- $f(y) - f(z) = \int_z^y f'(t)\,dt \geqslant f'(z)(y-z)$

Da $z - x = \lambda(y-x)^2$ und $y - z = (1-\lambda)(y-x)$ folgt unmittelbar

$$\begin{array}{l|l} f(z) \leqslant f(x) + \lambda f'(z)(y-x) & \cdot\lambda \\ f(x) \leqslant f(y) - (1-\lambda)f'(z)(y-x) & \cdot(1-\lambda) \end{array}$$

Durch Addieren der beiden Ungleichungen ergibt sich auf der linken Seite quasi $f(z) = f(\lambda x + (1-\lambda)y)$, damit folgt die Konvexität nach Definition 1.14

„⇒“ Angenommen es existiere ein $x \in (a,b)$, für das $f''(x) < 0$ gelte. Dann wäre f'' negativ auf einem Subintervall (a',b') durch Kontinuität. Mit demselben Argument wie eben folgt auf (a',b'):

$$f(z) - f(x) > f'(z)(z-x) \text{ und } f(y) - f(z) < f'(z)(y-z)$$

und damit $f((1-\lambda)x+\lambda y) > (1-\lambda)f(x)+\lambda f(y)$, was aber der Konvexität von f widerspräche. ↯

□

Nun zum eigentlichen Beweis:
Die Konvexität von f auf D ist äquivalent zu setzen mit der Konvexität der Restriktion von f auf jedes Liniensegment in D, was dann der Konvexität der Funktion $g(\lambda) = f(y + \lambda z)$ auf dem konvexen Intervall $\left\{\lambda \middle| y + \lambda z \in D\right\}$ mit $y \in D$ und $y \in \mathbb{R}^n$. Es ist leicht zu sehen, dass

$$g''(\lambda) = \langle z, \mathcal{H}f(x) \cdot z \rangle,$$

mit $x = y + \lambda z$. Mit Hilfssatz 1 ergibt sich, dass g konvex ist gdw. $g''(\lambda) \geqslant 0$, also für jedes $y \in D$ und $z \in \mathbb{R}^n$. Damit ergibt sich $\langle z, \mathcal{H}f(x) \cdot z \rangle$ für jedes $z \in \mathbb{R}^n$ und $x \in D$, was genau der Definition der Semidefinitheit (Definition 1.2) entspricht.

zu (2) jetzt trivial, folgt aus (1).

zu (3) Betrachte $f(x) = x^4$, so ist die Hessematrix $\mathcal{H}f(0) = f''(0) = 0 \not> 0$, dennoch ist f strikt konvex – was leicht nachzuprüfen ist.

□

Wir kommen jetzt zu den Sätzen, die Aussagen über die Anzahl und Art von Extremstellen treffen. Wir gehen dabei immer von **konvexen Funktionen** aus.

Satz 1.14 (*Lokale Minimalstellen konvexer Funktionen*)

Sei f konvex, ...

(a) ...so ist **jede** lokale Minimalstelle immer auch **globale** Minimalstelle.

(b) ...so bilden **alle** lokalen Minimalstellen von f eine zusammenhängende, sogar eine **konvexe** Menge.

(c) ...so haben **alle** lokalen Minima den gleichen Wert.

Beweis:

(a) $\not\Leftarrow$: Sei x_1 lokale Minimalstelle, so ist x_1 sogar globale Minimalstelle. Beweis über *Reduktion auf das Absurde*:
Angenommen es existierte ein $x_2 \in D$, mit $f(x_2) < f(x_1)$. Betrachte dann die Verbindungslinie zwischen x_1 und x_2:

$$\forall \lambda \in [0,1] : f(\lambda x_1 + (1-\lambda)x_2) \overset{\text{konv.}}{\leqslant} \lambda f(x_1) + (1-\lambda) \underbrace{f(x_2)}_{<f(x_1)} < (\lambda + 1 - \lambda) f(x_1) = f(x_1)$$

Betrachte nun den Wert für $\lambda \to 1$:

$$\lim_{\lambda \to 1} \lambda x_1 + (1-\lambda)x_2 = x_1,$$

das heißt, in jeder ε-Umgebung um x_1 findet sich ein Punkt $\lambda x_1 + (1-\lambda)x_2$, so dass $f(\lambda x_1 + (1-\lambda)x_2) < f(x_1)$ $\lightning$ zu x_1 ist lokale Minimalstelle.

(b) $\not\Leftarrow$: Alle lokalen Minimalstellen bilden eine konvexe Menge
Seien $x_1 \neq x_2$ lokale Minimalstellen, so ist

$$f(\lambda x_1 + (1-\lambda)x_2) \overset{\text{konv.}}{\leqslant} \lambda f(x_1) + (1-\lambda) \underbrace{f(x_2)}_{=f(x_1), \text{ nach (c)}} = f(x_1) = f(x_2)$$

Da ferner nach (a) $f(x_1) = f(x_2)$ das globale Minimum ist, muss $f(\lambda x_1 + (1-\lambda)x_2) \geqslant f(x_1) = f(x_2)$, das heißt $f(\lambda x_1 + (1-\lambda)x_2) = f(x_1) = f(x_2)$, was bedeutet, dass $\lambda x_1 + (1-\lambda)x_2$ auch Minimalstelle ist.

(c) $\not\Leftarrow$: Alle lokale Minimalstellen haben den gleichen Funktionswert
Seien $x_1 \neq x_2$ lokale Minimalstellen, nach (a) sind x_1, x_2 sogar globale Minimalstellen, das heißt

$$\forall x \in D : f(x_1) \leqslant f(x_2) \text{ und } \forall x \in D : f(x_2) \leqslant f(x_1),$$

woraus die Gleichheit abzulesen ist. □

Satz 1.15 (*Eindeutigkeit des Minimums*)

Sei f strikt konvex, so hat f **höchstens** eine Minimalstelle

Beweis: Angenommen $x_1 \neq x_2$ seinen Minimalstellen von f. Nach Satz 1.14(c) gilt $f(x_1) = f(x_2)$, heißt:

$$f(\lambda x_1 + (1-\lambda)x_2) \underset{\text{konv.}}{\overset{\text{strikt}}{<}} \lambda f(x_1) + (1 - \lambda \underbrace{f(x_2)}_{=f(x_1)}) = f(x_1)$$

Andererseits ist $f(x_1)$ aber nach Satz 1.14(a) globale Minimalstelle. $\lightning$ □

Wir wollen nun noch einen Satz erklären, mit dem es uns möglich ist auf die Existenz von Minimalstellen zu schließen. Normale oder sogar strikte Konvexität reichen hierfür nicht aus, wie man sich ganz leicht am Beispiel der (strikt) konvexen Funktion $x \mapsto e^x$ klar machen kann.

Satz 1.16 (*Existenz des Minimums*)

Sei $f : \emptyset \neq D \subseteq \mathbb{R}^n \to \mathbb{R}$ stetig. Ferner habe f eine **nichtleere kompakte Levelmenge**

$$\emptyset \neq L_c := \left\{ x \in D \,\middle|\, f(x) \leqslant c \right\}.$$

Dann hat f **mindestens** eine globale Minimalstelle – und somit auch mindestens eine lokale Minimalstelle. Ist f sogar strikt konvex, ist diese eindeutig bestimmt.

Beweis: Sei $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ stetig, $c \in \mathbb{R}$, so dass $\emptyset \neq L_c := \{x \in D \mid f(x) \leqslant c\} \subseteq D$ kompakt ist. Betrachte dann die Restriktion von f auf L_c

$$f\Big|_{L_c} : L_c \to \mathbb{R}.$$

$f\Big|_{L_c}$ ist stetig und hat kompakten Definitionsbereich L_c. Nach dem Satz von Maximum und Minimum ($\to$ C2) hat $f\Big|_{L_c}$ auf jeden Fall ein Minimum, dieses ist $\leqslant c$. Da für alle $x \in D \backslash L_c$ gilt, dass $f(x) > c$, ist die Minimalstelle von $f\Big|_{L_c}$ gleichzeitig auch Minimalstelle von f. Ist f strikt konvex, muss die Minimalstelle dann nach Satz 1.15 auch eindeutig bestimmt sein. □

> **i** Für stetiges f und abgeschlossenes D sind die Levelgruppen **immer** abgeschlossen. Damit bleibt nur noch die Beschränktheit zu überprüfen.

Wir wollen uns jetzt noch mit ein paar Spezialfällen der Minimierungsprobleme näher beschäftigen.

1.5.1 Spezialfall: Quadratisches Optimierungsproblem

Sei $D := \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ symmetrisch positiv definit, $b \in \mathbb{R}^n$, $\zeta \in \mathbb{R}$ und bezeichne $\langle \cdot, \cdot \rangle$ wieder das euklidische Skalarprodukt. Sei

$$f(x) := \frac{1}{2} \langle Ax, x \rangle + \langle b, x \rangle + \zeta \tag{1.5}$$

Wir suchen das Minimum von f.
Lösung
Durch das Rechnen mit Komponenten ergibt sich für alle $x \in \mathbb{R}^n$:

$$\nabla f(x) = Ax + b$$
$$\mathcal{H} f(x) = A$$

Nach Voraussetzung an A ist f damit strikt konvex, dies folgt aus Satz 1.13. Nach Satz 1.14/1.15 gibt es höchstens eine Minimalstelle und diese ist global.
Mit $\nabla f(x_*) \stackrel{!}{=} 0$ ergibt sich:

$$x_* = -A^{-1} b$$

als kritische Stelle. Diese existiert, weil A, da symmetrisch positiv definit, nie den Eigenwert 0 besitzt, und deswegen immer invertierbar ist.

1.5.2 Linear-Quadratisches Minimierungsproblem

Sei $f : \mathbb{R}^n \to \mathbb{R}$ quadratisch und zu minimieren, also $f(x) = \frac{1}{2} \langle Ax, x \rangle + \langle b, x \rangle$ mit $A \in \mathbb{R}^{n \times n}$ symmetrisch positiv definit. f ist zu minimieren unter der Nebenbedingung $Bx = c$ mit $B \in \mathbb{R}^{m \times n}$ und $c \in \mathbb{R}^m$.
Wir bezeichnen mit $M := \{x \in \mathbb{R}^n \mid Bx = c\}$ die **zulässige Menge**, f heiße in diesem Zusammenhang auch **Ziel-** respektive **Kostenfunktion**. Es ist leicht einzusehen, dass M ein affin-linearer Unterraum des $\mathbb{R}^m$ ist, zudem ist es hierbei sinnvoll zu fordern, dass $m < n$.
Wir wissen, dass M und f (strikt) konvex sind, somit ist dann die Restriktion von f auf M $f\Big|_M$ ebenfalls strikt konvex. Damit hat $f\Big|_M$ nach Satz 1.15 höchstens eine Minimalstelle.

Die Lösung des Problems erfolgt durch die – uns bereits seit Kapitel 1.2 bekannte – Methode der Lagrangemultiplikatoren (Satz 1.5). Wir erhalten damit:

$$\forall i \in \{1,\dots,m\} : g_i(x) := B_i x - c_i \overset{!}{=} 0$$
$$\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) \overset{!}{=} 0$$

zusammengefasst dann

$$Ax + b + \sum_{i=1}^{m} \lambda_i B_i^T = 0 \text{ und } Bx - c = 0 \tag{1.6}$$

Wir erhalten damit das Folgende – zu lösende – Gleichungssystem:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} -b \\ c \end{pmatrix} =: \tilde{b}$$

1.5.3 Das Gradientenverfahren – Eine numerische Lösung des Optimierungsproblems

Wir wollen uns nun mit einer effizienten Vorgehensweise für das Lösen des LGS (1.6) beschäftigen. Effizienz ist beim alltäglichen Lösen von Linearen Gleichungssystemen sehr relevant, da die direkten[2] Löser meist laufzeittechnisch schlechter abschneiden. Wir überlegen uns jetzt eine Methode, die von der Idee bereits im zweiten Semester aufkam.
Sei $f : \mathbb{R}^n \to \mathbb{R}$ also nun differenzierbar. Um eine Minimalstelle von f zu finden ist das – im Allgemeinen **nichtlineare** – Gleichungssystem $\nabla f(x) = 0$ zu lösen, was nicht trivial machbar ist.
Denke aber ans zweite Semester zurück. Aus C2 wissen wir, dass der negative Gradient **immer** in die Richtung des **steilsten Abstiegs** zeigt. Sei also beispielsweise eine hinreichend gute Näherung x_0 an die Minimalstelle von f gegeben. Machen wir nun einen Schritt in Richtung von $-\nabla f(x_0)$ und wir erhalten – bei geeeigneter Schrittlänge – einen besseren Näherungswert. Dies führt zu folgender ersten Iterationsvorschrift:

$$x_{n+1} := x_n - \alpha_n \nabla f(x_n)$$

Wie ist α_n zu wählen?

Betrachten wir hierzu die Werte von f entlang der Geraden $\alpha \mapsto x_n - \alpha \nabla f(x_n)$, also die Abbildung $h : \mathbb{R} \to \mathbb{R}$ mit $h(\alpha) := f(x_n - \alpha \nabla f(x_n))$. Offensichtlich ist es sinnvoll $\alpha = \alpha_n$ als Minimalstelle dieser Funktion zu wählen. Dementsprechend ist folgende Gleichung zu lösen:

$$0 \overset{!}{=} h'(\alpha) = -\langle \nabla f(x_n - \alpha_n \nabla f(x_n)), \nabla f(x_n) \rangle \to \alpha_n \tag{1.7}$$

Diese Gleichung lässt sich aber nur nach α auflösen, wenn man Annahmen an f trifft. Sei f jetzt also quadratisch, damit ist $\nabla f(x) = Ax + b$. Eingesetzt in (1.7) ergibt:

$$0 = \langle A(x_n - \alpha \nabla f(x_n)) + b, Ax_n + b \rangle = \langle Ax_n + b, Ax_n + b \rangle - \alpha \langle A \nabla f(x_n), Ax_n + b \rangle$$

Damit ist α_n zu wählen als:

$$\alpha_n = \frac{\langle \nabla f(x_n), \nabla f(x_n) \rangle}{\langle A \nabla f(x_n), \nabla f(x_n) \rangle}, \text{ wobei } \nabla f(x_n) = Ax_n + b$$

[2] Ein Verfahren heiße dabei **direkt** genau dann, wenn es nach einer **endlichen** Anzahl an Rechenschritten zu einem **exakten** Ergebnis kommt.

Zusammengefasst ergibt sich folgendes Verfahren:

Verfahren 1.1 (*Gradienten-Verfahren*)

Sei f „glatt“ und x_n eine hinreichende Näherung an die Minimalstelle von f. Dann ist

$$x_{n+1} := x_n - \alpha_n \nabla f(x_n) \text{ mit bspw. } \alpha_n := \frac{\langle \nabla f(x_n), \nabla f(x_n) \rangle}{\langle A \nabla f(x_n), \nabla f(x_n) \rangle}$$

im Allgemeinen eine bessere Näherung.

Eine weitere interessante Fragestellung hier wäre die Konvergenz(geschwindigkeit) des Verfahrens. Bei quadratischen Modellproblemen hängt die Geschwindigkeit vom Verhältnis

$$\kappa := \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

von größtem zu kleinstem Eigenwert ab. Je größer dabei κ, desto schlechter ist das Konvergenzverhalten. Eine Verbesserung des Verfahrens stellt die „Methode der **konjungierten Gradienten**“ (*cg-Verfahren*) dar, dessen Konvergenzverhalten schwächer von κ abhängt.
Mehr dazu, wie man LGS numerisch lösen kann, kommt in Kapitel 1.7.4. An dieser Stelle beschäftigen wir uns dann auch nochmal kurz mit den Konvergenzverhalten.

1.6 Lineare Optimierung

Im letzten Optimierungskapitel, wollen wir uns nun näher mit affin-linearen Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ beschäftigen, die unter Einhaltung von ebenfalls affin-linearen Gleichungs- **sowie Ungleichungs**nebenbedingungen optimiert werden sollen.
Damit beschäftigen wir uns in diesem Kapitel mit der „**Linearen Optimierung**“ oder auch der „**Linearen Programmierung**“.
Wir werden uns nun verschiedene Begriffe an Folgendem Beispiel klar machen:
Beispiel 1.4: (*Produktplanung*)

B

Unternehmen U stellt 2 Produkte her, P_1 eine Glastür mit Alurahmen und P_2 ein Fenster mit Holzrahmen. Pro verkaufter „Türeinheit“ stellt sich ein Gewinn von 3GE, pro „Festereinheit“ ein Gewinn von 5GE ein. Die Produktion erfolgt in den 3 Fabriken ...

- ... F_1, stellt nur Alurahmen her, besitzt ein Produktionsvolumen von $4^{\text{Einheiten}/\text{Tag}}$,
- ... F_2, stellt nur Holzrahmen her, besitzt ein Produktionsvolumen von $6^{\text{Einheiten}/\text{Tag}}$, sowie
- ... F_3, stellt nur Glasscheiben her, besitzt ein Produktionsvolumen von $18^{\text{Einheiten}/\text{Tag}}$.

Es ist ferner bekannt, dass P_1 3 Glaseinheiten, P_2 nur 2 Glaseinheiten benötigt. Die Aufgabe ist es den Gewinn unter der Annahme, dass alle produzierten Einheiten auch verkauft werden, zu maximieren.

Wir stellen dazu ein Lineares Programm auf. Dazu identifizieren wir als erstes die Entscheidungsvariablen. Wir wählen x_1 für die Türeinheiten und x_2 für die Fenstereinheiten. Als nächstes stellen wir die Zielfunktion auf, die zu maximieren ist, in diesem Fall also die Gewinnfunktion:

$$f(x_1, x_2) = 3x_1 + 5x_2$$

Als letztes stellen wir die aus dem Text herauslesbaren Nebenbedingungen auf:

(i) $x_1 \leqslant 4$ — *Produktionsvolumen von* F_1

(ii) $x_2 \leqslant 6$ — *Produktionsvolumen von* F_2

(iii) $3x_1 + 2x_2 \leqslant 18$ — *Produktionsvolumen von* P_3 *mit eingerechnetem Glaseinheitenverbrauch von* P_1 *und* P_2

(iv) $x_1, x_2 \geqslant 0$ — *Es sollen Produkte verkauft werden*

In Kurzform ergibt sich somit:

$$\begin{aligned} & f(x) := \langle c, x\rangle \longrightarrow \max \\ \text{unter der NB} \quad & Ax \leqslant b \\ \text{mit} \quad & c := \begin{pmatrix} 3 \\ 5 \end{pmatrix}, \quad A := \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 3 & 2 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}, b := \begin{pmatrix} 4 \\ 6 \\ 18 \\ 0 \\ 0 \end{pmatrix} \end{aligned}$$

Das $\leqslant$ ist dabei *komponentenweise* zu verstehen.

Definition 1.15 (*Zulässige Menge*)

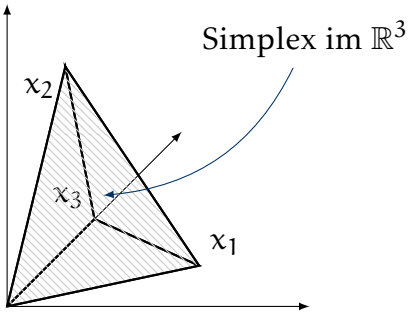
Die Menge aller $x \in \mathbb{R}^n$, die alle Nebenbedingungen erfüllt, heißt **zulässige Menge**. Die zulässige Menge von Linearen Programmen ist ein Schnitt von **Halbräumen**, sprich von Mengen der Form $\left\{x \in \mathbb{R}^n \middle| u(x) \leqslant c\right\}$, wobei $u : \mathbb{R}^n \to \mathbb{R}$ linear ist und $c \in \mathbb{R}$.

Definition 1.16 (*Polyeder, Simplex*)

Ein nichtleerer Schnitt von endlich vielen Halbräumen heißt **Polyeder**. Seien $n+1$ viele Punkte $p_i \in \mathbb{R}^n$ gegeben. Die von den Punkten „aufgespannte“ Menge

$$\left\{ \sum_{i=1}^{n+1} \alpha_i x_i \middle| \alpha_i \geqslant 0, \sum_{i=1}^{n+1} \alpha_i = 1 \right\}$$

heißt **Simplex**.



Korrolar D1.16

Jeder Polyeder ist konvex. Für $n = 2$ ist ein Simplex ein Dreieck im $\mathbb{R}^2$, für $n = 3$ ein Tetraeder im $\mathbb{R}^3$, allgemein für $n \in \mathbb{N}$ ist ein Simplex ein Polyeder.
Die zulässige Menge (nach Definition 1.15) ist konvex.

Beispiel 1.4 (*Fortsetzung*)**:** Bei unserem Beispiel haben wir nur zwei Entscheidungsvariablen, eine graphische Lösung ist also möglich.

Betrachten wir dazu eine Skizze des linearen Programms:

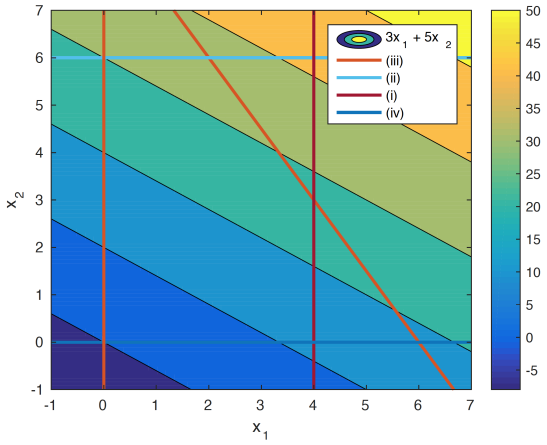


Abbildung 1.1: Skizze des linearen Programms aus Beispiel 1.4

Wir stellen fest:

(1) Die Lösung(en) liegen niemals nur im Inneren der zulässigen Menge, sondern immer auch am Rand. Generell kommen Lösungen im Inneren nur bei trivialen Funktionen ($f \equiv \text{const.}$) vor.

(2) Es kann passieren, dass es keine Lösung gibt, wenn …

- … die zulässige Menge leer ist *oder*
- … die zulässige Menge unbeschränkt ist, **kann** es passieren, dass es keine Lösung gibt ($\rightarrow$ die Lösung läge dann im „Unendlichen“)

(3) Wenn es eine/mehrere Lösung(en) gibt, so liegt mindestens eine Lösung auf einer Ecke der zulässigen Menge

(4) Wenn es mehrere Lösung(en) gibt, dann gibt es mehrere Ecken, die Lösungen sind, die Menge der Lösungen ist zusammenhängend.

Es reicht also die Ecken der zulässigen Menge nach Optimalstellen zu durchsuchen.

! Auch hier mag die Lösung auf den ersten Blick perfekt und simpel ausschauen, aber die Anzahl der Ecken wächst mit steigender Anzahl an Entscheidungsvariablen und steigender Anzahl an Nebenbedingungen enorm. Wir werden deswegen uns bemühen, die Anzahl an möglichen – auszuprobierenden – Ecken irgendwie einzugrenzen.

Eine Lösungsmethode dieses Problems ist der sogenannte Simplex-Algorithmus, den wir in Kapitel 1.6.3 genauer kennen lernen werden.
Wir wandeln dazu unser Lineares Programm in „Standardform“ um:

Definition 1.17 (*Allgemeine Form*)

Suche $x \in \mathbb{R}^n$, so dass $f(x) := \langle c, x \rangle \rightarrow \min$ oder $\rightarrow \max$,

$$\text{so dass} \quad \sum_{j=1}^{n} a_{ij} x_j = b_i \qquad \text{für } i = 1, \ldots, p$$

$$\sum_{j=1}^{n} a_{ij} x_j \leqslant b_i \quad \text{für } i = p+1, \ldots, q$$

$$\sum_{j=1}^{n} a_{ij} x_j \geqslant b_i \quad \text{für } i = q+1, \ldots, m$$

Definition 1.18 (*Standardform*)

Suche $x \in \mathbb{R}^n$, so dass $f(x) := \langle c, x\rangle \to \min$,

$$\begin{aligned} \text{so dass } Ax &= b \text{ und} \\ x &\geqslant 0, \\ \text{wobei } A \in \mathbb{R}^{m\times n}, m \in \mathbb{N} &\text{ und } \operatorname{rang}(A) = m \end{aligned}$$

Wir wollen nun ein Verfahren kennen lernen, mit dem eine Umwandlung von allgemeiner Form in Standardform möglich ist:

Verfahren 1.2 (*Allgemeine Form $\mapsto$ Standardform*)

(1) Ist das Maximum gesucht, so ersetze $f(x)$ durch

$$\widetilde{f}(x) := \langle -c, x\rangle \to \min$$

(2) Auch „$\geqslant$"-Nebenbedingungen werden mit -1 „durchmultipliziert", so dass nur noch „$\leqslant$"-Nebenbedingungen vorhanden sind.

(3) Die Ungleichungsnebenbedingungen $\sum_{j=1}^{n} a_{ij}x_j \leqslant b_i$ werden in Gleichungsnebenbedingungen umgeformt durch die Einführung von **Schlupfvariablen**, dann gilt:

$$\sum_{j=1}^{n} a_{ij}x_j + \widetilde{x}_i = b_i \wedge \widetilde{x}_i \geqslant 0$$

(4) **Jede** Variable x_i braucht eine Bedingung $x_i \geqslant 0$, dazu substituiere für jede Variable x_j **ohne** eine solche Bedingung:

$$x_j := x_j^+ - x_j^- \text{ mit } x_j^+, x_j^- \geqslant 0$$

(5) Zur Sicherstellung der linearen Unabhängigkeit der Zeilen von A. Wir wissen, dass sich lineare Abhängigkeit durch die Unlösbarkeit des LGS $Ax = b$ oder des Vorkommens redundanter Zeilen ausdrückt. Mache deswegen Folgendes:

- Bringe $Ax = b$ auf Stufenform $\left(\begin{array}{c|c} \widetilde{A} & \widetilde{b} \end{array}\right)$
- Ersetze im linearen Programm die Gleichung $Ax = b$ durch $\widetilde{A}x = \widetilde{b}$.
- Unterscheide nach Stufenform des LGS $\widetilde{A}x = \widetilde{b}$:
  - Hat das LGS Stufenform II, streiche alle Nullzeilen, die verbleibenden Zeilen sind dann linear unabhängig.
  - Hat das LGS Stufenform III, so ist das LGS unlösbar, womit die zulässige Menge leer ist, sprich das lineare Programm ist unlösbar.

Am Ende dieser 5 Schritte ist das lineare Programm in Standardform gem. Definition 1.18.

Beispiel 1.4 (*Fortsetzung*)**:**

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 3 & 2 \end{pmatrix}, b := \begin{pmatrix} 4 \\ 6 \\ 18 \end{pmatrix}, c := \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

Wir suchen das Maximum $\Rightarrow \tilde{c} = -c = \begin{pmatrix} -3 \\ -5 \end{pmatrix}$. Löse damit das LGS

$$\left(\tilde{A} \mid \tilde{b} \right) \rightsquigarrow \left(\begin{array}{ccccc|c} 1 & 0 & 1 & 0 & 0 & 4 \\ 0 & 1 & 0 & 1 & 0 & 6 \\ 3 & 2 & 0 & 0 & 1 & 18 \end{array} \right) \rightsquigarrow \left(\begin{array}{ccccc|c} 1 & 0 & 1 & 0 & 0 & 4 \\ 0 & 1 & 0 & 1 & 0 & 6 \\ 0 & -2 & -3 & 0 & 1 & 6 \end{array} \right) \rightsquigarrow \left(\begin{array}{ccccc|c} 1 & 0 & 1 & 0 & 0 & 4 \\ 0 & 1 & 0 & 1 & 0 & 6 \\ 0 & 0 & -3 & -2 & 1 & -2 \end{array} \right)$$

Damit ist das Optimum gefunden, der Gewinn wird maximiert bei 4 produzierten Türen und 6 produzierten Fenstern. ※

Wir setzten ab sofort – wenn nicht anders angegeben – ein Lineares Programm in Standardform voraus.

1.6.1 Wie findet man Ecken?

Wir wollen als erstes die Frage, wie man allgemeine Ecken der zulässigen Menge findet, beantworten. Betrachte dafür ersteinmal allgemein für $Ax = b$, $x \geqslant 0$ mit $A \in \mathbb{R}^{m \times n}$, $\operatorname{rang}(A) = m$, $x \in \mathbb{R}^n$ und $n \geqslant m$:

Da man m linear unabhängige Gleichungsbedingungen zu erfüllen hat, kann man $n-m$ Variablen zu null setzen – denn $\dim \ker(A) = n - m$ – und die übrigen m Variablen dann über die m Gleichungen berechnen. Mathematisch ausgedrückt:

$A = \left(A_B \mid A_N \right)$, wobei $A_B \in \mathbb{R}^{m \times m}$ und $A_N \in \mathbb{R}^{m \times (n-m)}$. Analog $x = \begin{pmatrix} x_B \\ x_N \end{pmatrix}$ mit $x_B \in \mathbb{R}^m$ und $x_N \in \mathbb{R}^{n-m}$.

Damit lässt sich dann das LGS $Ax = b$ schreiben als:

$$A_B x_B + A_N x_N = b \tag{1.8}$$

Ist nun A_B invertierbar – falls also die ersten m Spalten von A linear unabhängig waren – so lässt sich (1.8) nach x_B auflösen:

$$x_B = A_B^{-1}(b - A_N x_N) \tag{1.9}$$

Wir setzen also $x_N := 0$ und die übrigen Variablen zu $x_B = A_B^{-1} b$.

> **!** Da das Voraussetzen der linearen Unabhängigkeit der ersten m Spalten von A nur aus Vereinfachung geschah, brauchen wir noch einen allgemeinen Weg um **beliebige** Ecken zu finden.

Zuerst definieren wir erstmal verschiedene Begriffe:

Definition 1.19 (*Basis, Basisvariablen und die Basislösung eines in SF geg. LP*)

Sei $A = [a_1, \ldots, a_n] \in \mathbb{R}^{m \times n}$ mit $\operatorname{rang}(A) = m$ und $m \leqslant n$, so heiße eine Auswahl von m Spalten $a_{i_1}, \ldots, a_{i_m}$, welche linear unabhängig sind, aus den n Spalten $a_1, \ldots, a_n$ **Basis** des linearen Programms.

Zur Vereinfachung der Sprechweise wird manchmal auch die zugehörige Indexmenge $\mathcal{B}$ als Basis bezeichnet.

Die zugehörigen Komponenten von x, also $x_{i_1}, \ldots, x_{i_m}$ heißen **Basisvariablen**, der Rest **Nichtbasisvariablen**.

Die zugehörige **Basislösung** erhält man durch das „Ausnullen“ der Nichtbasisvariablen und das Bestimmen der Basisvariablen über die m Gleichungen.

> **i** Generell gilt: Jede Ecke der zulässigen Menge eines linearen Programms ist automatisch auch Basislösung, umgekehrt gilt dies aber **nicht** unbedingt!

Definition 1.20 (*Zulässigkeit einer Basislösung*)

Eine Basis/Basislösung heiße **zulässig** genau dann, wenn alle Komponenten der Basislösung

größer oder gleich 0 sind.

Betrachten wir nun ein einfaches **Beispiel 1.5:** Wähle $m = 1$, $n = 2$, $A = \begin{pmatrix} 1 & -1 \end{pmatrix}$ und $b = (0)$.

- Wähle die 1. Spalte als Basis $\rightsquigarrow$ Nichtbasisvariable $x_2 := 0$, so ergibt sich x_1 als Lösung von $x_1 - 1 \cdot 0 = 0$, das heißt $x_1 = 0$.
- Wähle nun die 2. Spalte als Basis $\rightsquigarrow$ Nichtbasisvariable $x_1 := 0$, so ergibt sich x_2 als Lösung von $0 - x_2 = 0$, sprich $x_2 = 0$.

Zwei verschiedene Basen liefern also dieselbe Ecke $(0,0)$. Wir schließen damit auf eine fehlende Injektivität der Abbildung. ※

! Die Zuordnung „Basis mit zulässiger Basislösung $\mapsto$ Ecke der zulässigen Menge“ ist im Allgemeinen **nicht** injektiv!

Das passiert genau dann, wenn neben den Nichtbasiskomponenten auch noch ein oder mehrere Basiskomponenten, dies sich als Lösung $x_b = A_B^{-1} b$ ergeben, „zufällig“ null sind. In diesem Fall gibt es also **mehrere** zulässige Basen, die ein und dieselbe Ecke beschreiben.

Betrachten wir nun noch zwei Existenzsätze für Basislösungen.

Satz 1.17 (*Existenz einer zulässigen Basislösung*)

Sei $M := \{x \in \mathbb{R}^n \mid Ax = b, x \geqslant 0\}$ die zulässige Menge eines linearen Programms in Standardform. Falls $M \neq \emptyset$, so gibt es (mindestens) **eine** zulässige Basislösung, andernfalls **nicht**.

Beweis: über Fallunterscheidung:

(1) $b = 0 \Rightarrow x = 0$ ist zulässige Basislösung.

(2) $b \neq 0$: Sei $x \in M$ mit kpositiven Komponenten, Œ sei die Nummerierung der Variablen so, dass

$$x_j \begin{cases} > 0 \\ = 0 \end{cases} \text{ für } \begin{matrix} j = 1, \ldots, k \\ j = k+1, \ldots, n \end{matrix}$$

(2a) $\{a_1, \ldots, a_k\}$ sind linear unabhängig.
Dann gilt – weil $\operatorname{rang}(A) = m - k \leqslant m$ und $\{a_1, \ldots, a_k\}$ kann durch $m - k$ Spalten $(A_{j_1}, \ldots, A_{j_{m-k}})$ zu einer Basis des Spaltenraumes ergänzt werden. x ist die Basislösung bzgl. $\mathcal{B} = \{1, \ldots, k, j_1, \ldots, j_{m-k}\}$. x ist außerdem zulässig, da $x_1, \ldots, x_k > 0$ und $x_{j_1}, \ldots, x_{j_{m-k}} = 0$ ist.

(2b) $\{a_1, \ldots, a_k\}$ sind linear abhängig.
Ist $A_k := (a_1, \ldots, a_k)$ die aus den Spalten $a_1, \ldots, a_k$ bestehende Matrix und $x_k = (x_1, \ldots, x_k)^T$, so gilt wegen $Ax = b$ und $x_{k+1} = \cdots = x_n = 0$

$$A_k \cdot x_k = b \tag{1.10}$$

Aufgrund der linearen Abhängigkeit von $\{a_1, \ldots, a_k\}$, existiert $\alpha = (\alpha_1, \ldots, \alpha_n)^T \neq 0$, so dass $A_k \alpha = 0$. Dann gilt:

$$\forall \delta \in \mathbb{R} : A_k(\delta \cdot \alpha) = 0 \tag{1.11}$$

Addition von (1.10) und (1.11) ergibt:

$$A_k(x_k + \delta\alpha) = b$$

Bezeichnen wir mit $x(\delta)$ den Vektor mit

$$x(\delta)_i = \begin{cases} x_i + \delta\alpha_i & \text{, wenn } i \in \{1, \ldots, k\} \\ 0 & \text{, sonst} \end{cases},$$

so ist $x(\delta) \in M$, das heißt für alle $i = 1,\dots,n$ gilt $x(\delta)_i \geqslant 0$ falls

$$\delta \geqslant -\frac{x_i}{\alpha_i} \quad \forall i = 1,\dots,k \text{ mit } \alpha_i > 0$$
$$\text{und } \delta \leqslant -\frac{x_i}{\alpha_i} \quad \forall i = 1,\dots,k \text{ mit } \alpha_i < 0$$

Für $\delta = \max\left\{-\frac{x_i}{\alpha_i} : \alpha_i > 0\right\} < 0$ oder $\delta = \min\left\{-\frac{x_i}{\alpha_i} : \alpha_i < 0\right\} > 0$ gilt $\delta \in \mathbb{R}$. Die entsprechende Lösung $x(\delta)$ hat höchstens $k-1$ positive Komponenten. Durch iterative Anwendung dieses Verfahrens erhalten wir – weil $b \neq 0$ nach spätestens $k-1$ Iterationen Fall (2a).

□

Satz 1.18 (*Fundametalsatz/Hauptsatz der linearen Optimierung*)

Sei (A,b,c) ein lineares Programm in Standardform mit $M := \left\{x \in \mathbb{R}^n \middle| Ax = b, x \geqslant 0\right\} \neq \emptyset$ nichtleerer zulässiger Menge und das lineare Programm $\min\left\{cx : Ax = b, x \geqslant 0\right\}$ sei beschränkt, so gibt es **eine optimale** zulässige Basislösung.

Beweis: Sei x eine Optimallösung und, wie im Beweis zu Satz 1.17,

$$x_j \begin{cases} > 0 \\ = 0 \end{cases} \text{ für } \begin{matrix} j = 1,\dots,k \\ j = k+1,\dots,n \end{matrix}$$

Sind $\{a_1,\dots,a_k\}$ linear unabhängig, so ist die Behauptung gezeigt. Andernfalls existiert ein $\alpha = (\alpha_1,\dots,\alpha_n)^T$ mit $\alpha \neq 0$, so dass

$$A_k \cdot \alpha = (a_1,\dots,a_k) \cdot \alpha = A_k(\delta\alpha) = 0$$

für alle $\delta \in \mathbb{R}$. Betrachte $x(\delta) \geqslant 0$ wie im Beweis zu Satz 1.17. Es gilt:

$$Ax(\delta) = b \text{ und } cx(\delta = cx + \delta\left(\sum_{i=1}^{k} \alpha_i c_i\right)$$

Wäre $\sum_{i=1}^{k} \alpha_i c_i > 0$, so wählen wir $\delta < 0$, wäre $\sum_{i=1}^{k} \alpha_i c_i < 0$, so $\delta > 0$. In **beiden** Fällen wäre $cx(\delta) < cx$ im Widerspruch zur Optimalität von x. Damit gilt $\sum_{i=1}^{k} \alpha_i c_i = 0$. Wie im Beweis zu Satz 1.17 wählen wir δ so, dass $x(\delta)$ nur noch $k-1$ positive Komponenten hat. Da $cx(\delta) = cx$, erhalten wir somit nach maximal $k-1$ Iterationen dieser Prozedur eine optimale zulässige Basislösung. □

Korrolar 1.18

Es reicht „nur“ alle Basislösungen durchzuprobieren, auf Zulässigkeit zu überprüfen und f an der Stelle auszuwerten um das Minimum von f_M zu finden.

! Einziges Problem an der Sache ist: Es kann bis zu $\binom{n}{m}$ viele Basisvektoren geben!

Eine Verbesserung wäre es, wenn wir von einer gegebenen **zulässigen** Basislösung/Ecke x ausgehen, wir zu einer „benachbarten“ zulässigen Basislösung x_* „wandern“, die $f(x_*) < f(x)$ erfüllt.
Was heiße hierbei benachbart? Man kommt von der alten zur neuen Basis, indem man **nur einen der Basisvektoren ersetzt**. Man spricht dabei von einem sogenannten **Basiswechsel** oder auch **Basisaustauschschritt**.
Ist nun kein solcher Basiswechsel mehr möglich, so hat man eine Lösung gefunden. Wir wollen uns mit diesem Basiswechsel in folgendem Kapitel näher befassen.

1.6.2 Basiswechsel

Sei die Basis $\mathcal{B}$ mit Basislösung x gegeben. Wie testet man effizient, ob bei einem Basiswechsel $\mathcal{B} \to \widetilde{\mathcal{B}}$, $x \mapsto \widetilde{x}$ $f(\widetilde{x}) < f(x)$ gilt und $\widetilde{\mathcal{B}}$ zulässig ist?
Sei $\mathcal{B} := \{b_1, \dots, n_m\} \subseteq \{1, \dots, n\}$ also eine **Basis**. Damit ist die aus den zugehörigen Spalten von A bestehende Matrix $A_{\mathcal{B}} \in \mathbb{R}^{m \times m}$ invertierbar. Seien $\mathcal{N} := \{n_1, \dots, n_{n-m}\} \subseteq \{1, \dots, n\}$ die übrigen, sprich die Nichtbasisindizes. Dies heißt $\mathcal{B}$ und $\mathcal{N}$ bilden eine Partiton von $\{1, \dots, n\}$. Sei $A_{\mathcal{N}} \in \mathbb{R}^{m \times (n-m)}$ der „Rest" der Matrix A.
Die Komponenten der Vektoren x, c werden dementsprechend ebenso aufgeteilt in $x_{\mathcal{B}}, c_{\mathcal{B}} \in \mathbb{R}^m$ und $x_{\mathcal{N}}, c_{\mathcal{N}} \in \mathbb{R}^{n-m}$. Wir schreiben – mit der Elimination von $x_{\mathcal{B}}$ mittels $A_{\mathcal{B}} x_{\mathcal{B}} + A_{\mathcal{N}} x_{\mathcal{N}} = b$ an Stelle $(*)$ – die Evaluation von f als:

$$\begin{aligned} f(x) = & & \langle c, x \rangle &= \langle c_{\mathcal{B}}, x_{\mathcal{B}} \rangle + \langle c_{\mathcal{N}}, x_{\mathcal{N}} \rangle \\ \overset{(*)}{=} & & \left\langle c_{\mathcal{B}}, A_{\mathcal{B}}^{-1} b - A_{\mathcal{B}}^{-1} A_{\mathcal{N}} x_{\mathcal{N}} \right\rangle &+ \langle c_{\mathcal{N}}, x_{\mathcal{N}} \rangle \\ = & & \left\langle c_{\mathcal{B}}, A_{\mathcal{B}}^{-1} b \right\rangle + \left\langle c_{\mathcal{B}}, -A_{\mathcal{B}}^{-1} A_{\mathcal{N}} x_{\mathcal{N}} \right\rangle &+ \langle c_{\mathcal{N}}, x_{\mathcal{N}} \rangle \\ = & & \left\langle c_{\mathcal{B}}, A_{\mathcal{B}}^{-1} b \right\rangle + \left\langle c_{\mathcal{N}} - (A_{\mathcal{B}}^{-1} A_{\mathcal{N}})^T c_{\mathcal{B}}, x_{\mathcal{N}} \right\rangle & \end{aligned} \tag{1.12}$$

Wir betrachten nun den Basiswechsel $\mathcal{B} \to \widetilde{\mathcal{B}}$, $\mathcal{N} \to \widetilde{\mathcal{N}}$, $x \mapsto \widetilde{x}$. Wir benutzen dazu den Zusammenhang aus (1.12) um sowohl $f(x)$ vor dem Basiswechsel als auch $f(\widetilde{x})$ nach dem Basiswechsel – mit festem $\mathcal{B}$ – zu berechnen.

Sei j_* derjenige Index, der zur Basis hinzugefügt wird, k_* der Index, der entfernt wird, also

$$\{j_*\} = \widetilde{\mathcal{B}} \setminus \mathcal{B} = \mathcal{N} \setminus \widetilde{\mathcal{N}} \qquad \{k_*\} = \widetilde{\mathcal{N}} \setminus \mathcal{N} = \mathcal{B} \setminus \widetilde{\mathcal{B}}$$

Damit wird die j-te Komponente vor dem Basiswechsel null und beim Basiswechsel freigegeben ($x_{j_*} = 0$, x_{k_*} beliebig), die k_*-te Komponente kann vor dem Basiswechsel $\neq 0$ sein, sie wird beim Basiswechsel allerdings auf null gesetzt.

Geometrische Interpretation

Die Wahl der Indizes j_* und k_* entspricht geometrisch gesehen der Wahl der Kante entlang derer wir uns von x aus bewegen.

Für x **vor** dem Basiswechsel gilt $x_{\mathcal{N}} = 0$ und damit nach (1.12):

$$f(x) = \left\langle c_{\mathcal{B}}, A_{\mathcal{B}}^{-1} b \right\rangle$$

Für x **nach** dem Basiswechsel gilt $\widetilde{x_{\mathcal{N}}} = (0, \dots, 0, x_{j_*}, 0, \dots, 0)^T$, damit nach (1.12):

$$f(x) = \left\langle c_{\mathcal{B}}, A_{\mathcal{B}}^{-1} b \right\rangle + (c_{\mathcal{N}} - (A_{\mathcal{B}}^{-1} A_{\mathcal{N}})^T c_{\mathcal{B}})_{j_*} \widetilde{x}_{j_*}$$

Das Anwachsen des Funktionswertes berechnet sich somit durch

$$\Delta f_{j_*} := f(\widetilde{x}) - f(x) = \underbrace{(c_{\mathcal{N}} - (A_{\mathcal{B}}^{-1} A_{\mathcal{N}})^T c_{\mathcal{B}})_{j_*}}_{\Delta_a^1} \overbrace{\widetilde{x}_{j_*}}^{\Delta_b^1} \overset{!}{<} 0 \qquad (\Delta^1)$$

Da $\widetilde{x}$ **zulässig** sein soll, muss $\widetilde{x}_{j_*} \geqslant 0$ gelten. Wir fassen die Überlegungen in einem Satz auf:

Satz 1.19 (*Kriterium der reduzierten Kosten*)

Wir erhalten als Kriterium, dass der Wert der Zielfunktion bei Basiswechseln fällt:

$$(c_{\mathcal{N}} - (A_{\mathcal{B}}^{-1} A_{\mathcal{N}})^T c_{\mathcal{B}})_{i_*} \overset{!}{<} 0$$

Beim Basiswechsel ist also ein Indes $j_* \in \mathbb{N}$ zur Basis hinzuzufügen, so dass obige Bedingung erfüllt ist. Gibt es kein solches j_*, so ist x bereits minimal.

Beweis: Siehe Herleitung von (Δ^1) oben. □

Was wir bei unserer bisherigen Überlegung noch außer Acht gelassen haben sind die Werte von $\widetilde{x}_{j_*}$ und k_*. Wir überlegen deshalb:
Die j_*-te Komponente von x war vor dem Basiswechsel null, alls Basiskomponenten waren größer oder gleich null – weil x zulässig. Nach dem Basiswechsel soll die j_*-te Komponente größer oder gleich null sein, und zwar so, dass

(1) **alle** Basiskomponenten größer oder gleich null sind (*Zulässigkeit*) und

(2) (*mindestens*) eine genau null wird, der Index dieser Komponente wird dann das k_* sein.

Bedingung (2) stellt ...

- ...aus **geometrischer** Sicht sicher, dass wir die Kante **genau** bis zu einer Ecke laufen und **nicht** weiter
- ...aus **analytischer** Sicht sicher, dass wir einen Indes k_* finden, den wir aus der Basis streichen können.

Da $\widetilde{x}_{\mathcal{B}} = A_{\mathcal{B}}^{-1}b - A_{\mathcal{B}}^{-1}A_{\mathcal{N}}\widetilde{x_{\mathcal{N}}} = A_{\mathcal{B}}^{-1}b - A_{\mathcal{B}}^{-1}(A_{\mathcal{N}})_{j_*}x_{j_*} \overset{!}{\geqslant} 0$, wobei $(A_{\mathcal{N}})_{j_*}$ die j_*-te Spalte von $A_{\mathcal{N}}$ sein soll, lautet (1) und (2) also ...

$$\forall k \in \mathcal{B} : (\underbrace{A_{\mathcal{B}}^{-1}b}_{=x_{\mathcal{B}}})_k \geqslant \widetilde{x}_{j_*}(A_{\mathcal{B}}^{-1}(A_{\mathcal{N}})_{j_*})_k,$$

wobei mindestens eine der Ungleichungen mit Gleichheit erfüllt sein soll. Links steht dabei $(x_{\mathcal{B}})_k \geqslant 0$, da x zulässig ist. Solche k, für die $(A_{\mathcal{B}}^{-1}(A_{\mathcal{N}})_{j_*})_k \leqslant 0$, liefern keine Beschränkung für $\widetilde{x}_{j_*}$ und sind zu **ignorieren**.
Wir erfüllen (1) und (2) durch die **Quotientenregel der linearen Programmierung**.

Definition 1.21 (*Quotientenregel der linearen Programmierung*)

Wir legen $\widetilde{x}_{j_*}$ und k_* wie folgt fest:

$$\begin{aligned} \widetilde{x}_{j_*} &:= \min_{k\in\mathcal{B}} \left\{ \frac{(x_{\mathcal{B}})_k}{(A_{\mathcal{B}}^{-1}(A_{\mathcal{N}})_{j_*})_k} \middle| (A_{\mathcal{B}}^{-1}(A_{\mathcal{N}})_{j_*})_k > 0 \right\} \\ k_* &:= \text{derjenige Index } k, \text{ für die das obige Minimum angenommen wird, das heißt,} \\ &\quad \text{so dass } \widetilde{x}_{j_*} = \frac{(x_{\mathcal{B}})_{k_*}}{(A_{\mathcal{B}}^{-1}(A_{\mathcal{N}})_{j_*})_{k_*}}, (A_{\mathcal{B}}^{-1}(A_{\mathcal{N}})_{j_*})_{k_*} > 0 \text{ und } k_* \in \mathcal{B}. \end{aligned}$$

Falls das Minimum über die leere Menge gebildet wird – das heißt für alle $k \in \mathcal{B} : (A_{\mathcal{B}}^{-1}(A_{\mathcal{N}})_{j_*})_k \leqslant 0$, bedeutet dies, dass es **keinerlei** Einschränkung an $\widetilde{x}_{j_*} \geqslant 0$ gibt, also für beliebig großes $\widetilde{x}_{j_*} \geqslant 0$ die zulässige Menge M nicht verlassen wird.
Aufgrund der Wahl j_*, so dass die Kostenfunktion f in diese Richtung fällt , gilt dann $\inf f_M = -\infty$.

1.6.3 Der Simplex-Algorithmus

Wir wollen nun die Ergebnisse der letzten beiden Kapitel zu einem Verfahren fusionieren, mit dem lineare Programme in Standardform dann leichter und standardisierter zu lösen sind. Gegeben sei also folgendes Verfahren:

Verfahren 1.3 (*Simplex-Algorithmus*)

(1) Suche eine zulässige Basis $\mathcal{B}$ und die zugehörige Basislösung x, oder entscheide, dass es keine solche gibt.
($\rightarrow$ *Ist etwa* $M = \emptyset$, *so gibt es* ***keine*** *Lösung*)

(2) (2a) Bilde aus A, x, c unter Verwendung der Indexmengen $\mathcal{B}$ und $\mathcal{N} = \{1,\dots,n\}\backslash\mathcal{B}$ die Größen $A_\mathcal{B}, A_\mathcal{N}, c_\mathcal{B}, c_\mathcal{N}, x_\mathcal{B}, x_\mathcal{N}$ gemäß Gleichung (1.9) mit Satz danach.

(2b) Falls für alle $j \in \mathcal{N}$ gilt, dass $(c_\mathcal{N})_j \geqslant ((A_\mathcal{B}^{-1}A_\mathcal{N})^T c_\mathcal{B})_j$, so ist das Ziel erreicht. Die Lösung ist dann $f(x) = \left\langle c_\mathcal{B}, A_\mathcal{B}^{-1}b \right\rangle$. (**ENDE**)

(2c) Wähle ein $j_* \in \mathcal{N}$ mit $(c_\mathcal{N})_{j_*} \geqslant ((A_\mathcal{B}^{-1}A_\mathcal{N})^T c_\mathcal{B})_{j_*}$

(2d) Falls für alle $k \in \mathcal{B}$ gilt, dass $(A_\mathcal{B}^{-1}(A_\mathcal{N})_{j_*})_k \leqslant 0$, so ist f nach unten unbeschränkt, damit ist $\inf_{x\in M} f = -\infty$. (**ENDE**)

(2e) Setze $\widetilde{x}_{j_*}$ und k_* wie in Definition 1.21 beschrieben.

(2f) Steiche j_* aus $\mathcal{N}$ und füge es zu $\mathcal{B}$ hinzu, streiche k_* aus $\mathcal{B}$ und füge es zu $\mathcal{N}$ hinzu.

(2g) Gehe zu Schritt (2a).

i Ein kleiner Tipp zum Durchführen dieses Verfahrens: Gibt es in Schritt (2c) ein j_*, das die Bedingung in (2d) erfüllt, so ist dieses j_* – der Effizienz wegen – zu präferieren.

1.6.3.1 Schritt 2 des Verfahrens

In der Praxis gehen wir jedoch wie folgt vor: Wir erstellen ein „Tableau“, und bilden damit Schritt (2) komplett auf „Gauß“ähnliche Operationen ab. Wir definieren das Tableau wie folgt:

Definition 1.22 (*Simplex-Tableau*)

$$\begin{aligned} T &:= \left(\begin{array}{c|c} c^T - c_\mathcal{B}^T A_\mathcal{B}^{-1}A & -c_\mathcal{B}^T A_\mathcal{B}^{-1}b \\ \hline A_\mathcal{B}^{-1} & A_\mathcal{B}^{-1}b \end{array}\right) = \left(\begin{array}{c|c} c^T - c_\mathcal{B}^T A_\mathcal{B}^{-1}A & -f(x) \\ \hline A_\mathcal{B}^{-1}A & x_\mathcal{B} \end{array}\right) \\ &=: (t_{ij})_{i=0..m, j=1..(n+1)} \in \mathbb{R}^{(m+1)\times(n+1)} \end{aligned}$$

Interpretation des Simplex-Tableaus

(1) Der Bereich $\left(\begin{array}{c|c} A_\mathcal{B}^{-1}A & A_\mathcal{B}^{-1}b \end{array}\right)$ entsteht, indem man das LGS $Ax = b$ mit **elementaren Gaußoperationen** zu $A_\mathcal{B}^{-1}Ax = A_\mathcal{B}^{-1}b$ umformt. ($\rightarrow$ er „speichert“ also die Gleichungsnebenbedingungen)
Da für Basisspalten $j \in \mathcal{B}$ $A_\mathcal{B}^{-1}A_{b_j} = \vec{e}_j$ gilt, enthält der Block $A_\mathcal{B}^{-1}A \in \mathbb{R}^{m\times n}$ in den Basisspalten gerade die Einheitsvektoren $\vec{e}_1, \dots, \vec{e}_m \in \mathbb{R}^m$.

! Ist der untere Bereich berechnet, kann man auch die oberste Zeile des Schemas leicht ausrechnen. Jede Zeile – *bis auf die oberste* – repräsentiert *eine* Nebenbedingung, damit sind **elementare Zeilenoperationen** in diesem Bereich also erlaubt, weil diese das LGS nur in äquivalente umwandeln.

(2) Der Zeilenvektor $c^T - c_{\mathcal{B}}^T A_{\mathcal{B}}^{-1} A$ enthält in Basisspalten (Spaltenindex $j = b_i \in \mathcal{B}$) genau die Einträge

$$t_{0,j} = (c^T - c_{\mathcal{B}}^T A_{\mathcal{B}}^{-1} A)_j = c_j - c_{\mathcal{B}}^T \underbrace{A_{\mathcal{B}}^{-1} A_{b_i}}_{=\vec{e}_i} = c_j - c_{b_i} = 0,$$

und in Nichtbasisspalten $j = n_j \in \mathcal{N}$ enthält er – nach Satz 1.19 – die reduzierten Kosten, die bei der Wahl $j_* = j$ anfielen. Welche Nichtbasisspalte j im anstehenden Basiswechsel zu einer Basisspalte werden kann, erkennt man also daran, dass die obere Zeile von T in der betreffenden Spalte einen negativen Eintrag hat, das heißt $t_{0,j} < 0$. Damit gilt auch: Gilt für alle j, dass $t_{0,j} \geqslant 0$, so hat man eine Lösung gefunden.

(3) Rechts oben steht, bis auf das Vorzeichen, der aktuelle Wert von f, am Ende des Algorithmus also das gesuchte Minimum.

(4) Der Wahl des Index $k_* \in \mathcal{B}$ erfolgt nach der Quotientenregel mittels Tableau. In der zuvor ausgewählten Spalte j_* – die also zuoberst mit einem negativen Eintrag t_{0,j_*} beginnt – betrachte die positiven Einträge $t_{j,j_*} > 0$. Falls keiner positiv ist, ist die Lösung $\inf f|_M = -\infty \longrightarrow$ **Abbruch**
Andernfalls dividiere für Zeilen j, in denen $t_{j,j_*} > 0$ gilt, den entsprechenden Eintrag von $x_{\mathcal{B}}$ in der rechten Spalte, also $t_{j,n+1}$, durch eben diesen positiven Wert. Bestimme dann die Zeile $j_* = j$, für die der Quotient $t_{j,n+1}/t_{j,j_*}$ **minimal** wird.

(5) Nachdem man j_* und k_*, für die der Basiswechsel durchgeführt werden soll, ermittelt hat, muss der eigentliche Basiswechsel im Tableau durchgeführt werden. Gehe dazu wie folgt vor:

→ In der j_*-ten Spalte muss aus dem Vektor des linken unteren Blocks der Vektor $\vec{e}_{k_*}$ entstehen. **Dies geschieht dabei durch elementare Zeilenoperationen.** Diese sind *per se erlaubt*, da sie lediglich die Nebenbedingung $Ax = b$ äquivalent umformen. Diejenige Spalte, in der bisher der Vektor $\vec{e}_{k_*}$ stand, verändert sich im Allgemeinen dabei **wie gewünscht zu einem *Nicht*basisvektor**, die übrigen $\vec{e}_i$ verändern sich aber **nicht**.

→ *Was passiert beim Basiswechsel mit der obersten Zeile?*
Basisspalten starten mit einem Nulleintrag, das heißt da die j_*-te Spalte zur Basispalte wird, muss t_{0,j_*} mindestens eine Null enthalten. Man kann sich nun überlegen, dass man die oberste Zeile **genauso wie die anderen Zeilen behandeln *muss***, das heißt mittels **elementaren Zeilenumformungen** bei t_{0,j_*} eine Null erzeugt. Die dabei auftretenden Veränderungen der Einträge in der obersten Zeile $t_{0,j}$ sind genau die richtigen.

! Nach dem Basiswechsel muss das Tableau wieder in **allen** Basisspalten Standardbasisvektoren, oberhalb derer eine Null steht, enthalten. Diese Struktur darf **unter *keinen* Umständen** verloren gehen!

Deshalb sind Umformungen, welche eine Zerstörung der – wie oben beschriebenen – Struktur mit sich führen **NICHT ERLAUBT**!

Beispiel 1.6: Gegeben sei ein lineares Programm in Standardform mit $f(x) = \langle c, x\rangle, Ax = b, x \geqslant 0$, wobei

$$c^T = (-3,5,0,0,0), \quad A = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 4 & 4 & 1 & 1 & 1 \\ 3 & 2 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ 34 \\ 18 \end{pmatrix}$$

Man kann leicht erkennen, dass die Spalten 3, 4 und 5 linear unabhängig sind. Damit bilden sie auch eine Basis, ob diese aber zulässig ist, ist noch unklar. Wir betrachten deswegen $(\ A \mid b\ )$:

$$(\ A \mid b\ ) = \left(\begin{array}{ccccc|c} 1 & 0 & 1 & 0 & 0 & 4 \\ 4 & 4 & 1 & 1 & 1 & 34 \\ 3 & 2 & 0 & 0 & 1 & 18 \end{array}\right) \xrightarrow{\text{II}-(\text{I}+\text{III})} \left(\begin{array}{ccccc|c} 1 & 0 & 1 & 0 & 0 & 4 \\ 0 & 2 & 0 & 1 & 0 & 12 \\ 3 & 2 & 0 & 0 & 1 & 18 \end{array}\right)$$

Wir haben damit die „letzten“ drei Spalten auf die Einheitsmatrix gebracht, erkennen damit $x_{\mathcal{B}}$ an der – tatsächlich – letzten Spalte $\begin{pmatrix} 4 \\ 12 \\ 18 \end{pmatrix} = \underbrace{\begin{pmatrix} x_3 & x_4 & x_5 \end{pmatrix}^T}_{\geqslant 0} = x_{\mathcal{B}}$ und $x_{\mathcal{N}} = \begin{pmatrix} x_1 & x_2 \end{pmatrix}^T = 0$, womit die Basis sogar ist. Wir rechnen nun die restlichen des Tableaus aus.

$$f(x) = \left\langle c_{\mathcal{N}}, \underbrace{x_{\mathcal{N}}}_{=0} \right\rangle + \left\langle \underbrace{c_{\mathcal{B}}}_{=0}, x_{\mathcal{B}} \right\rangle = 0,$$

womit auch $-f(x) = 0$ folgt. Die reduzierten Kosten werden ebenfalls über die bekannte Formel ausgerechnet, es ist:

$$c^T - \overbrace{c_{\mathcal{B}}^T}^{=0} A_{\mathcal{B}}^{-1} A = \begin{pmatrix} -3 & -5 & 0 & 0 & 0 \end{pmatrix}^T$$

Wir stellen damit das Simplex-Tableau für unsere Startbasis auf:

1. Basisaustauschschritt:

Wähle eine Spalte j_* mit **negativen** reduzierten Kosten – diese sind in der **obersten** Zeile ablesbar – zur Aufnahme in die Basis. Hier sind Spalten $j_* = 1$ oder $j_* = 2$ möglich. Wir wählen, aus einem uns noch zu bestimmenden Grund (siehe Degeneriertheit und Gegenmaßnahmen), hier $j_* = 1$.
Die Quotientenregel liefert dann, dass Spalte $k_* = 3$ die Basis verlässt.

$$\left(\begin{array}{ccccc|c} \boxed{-3} & -5 & 0 & 0 & 0 & 0 \\ \hline \boxed{1} & 0 & \boxed{1} & 0 & 0 & 4 \\ 0 & 2 & 0 & 1 & 0 & 12 \\ 3 & 2 & 0 & 0 & 1 & 18 \end{array}\right) \begin{array}{lll} \text{QR:} & & \\ \leftarrow & 4 \div 1 = & \textcircled{4} \\ \leftarrow & 18 \div 3 = & 6 \end{array}$$

$A_{\mathcal{B}}^{-1} A_{\mathcal{N}}$ $\quad$ $x_{\mathcal{B}}$

2. Basisaustauschschritt:

Die j_*-te Spalte hat jetzt durch Zeilenoperationen die Form der k_*-ten angenommen, wodurch sich $x_{\mathcal{B}}$ zu $\begin{pmatrix} x_1 & x_4 & x_5 \end{pmatrix}^T$ verändert hat.
Wähle jetzt $j_* = 2$ zur Aufnahme in die Basis. Die Quotientenregel liefert dann, dass Spalte $k_* = 5$ die Basis verlässt. In der zweiten Spalte ist somit $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ zu erzeugen.

$$\left(\begin{array}{ccccc|c} 0 & \boxed{-5} & 3 & 0 & 0 & 12 \\ \hline 1 & 0 & 1 & 0 & 0 & 4 \\ 0 & 2 & 0 & 1 & 0 & 12 \\ 0 & \boxed{2} & -3 & 0 & \boxed{1} & 6 \end{array}\right) \begin{array}{lll} \text{QR:} & & \\ \leftarrow & 12 \div 2 = & 6 \\ \leftarrow & 6 \div 2 = & \textcircled{3} \end{array}$$

3. Basisaustauschschritt:

Die j_*-te Spalte hat jetzt durch Zeilenoperationen die Form der k_*-ten angenommen, wodurch sich $x_{\mathcal{B}}$ zu $\begin{pmatrix} x_1 & x_4 & x_2 \end{pmatrix}^T$ verändert hat.
Wähle jetzt $j_* = 3$ zur Aufnahme in die Basis. Die Quotientenregel liefert dann, dass Spalte $k_* = 4$ die Basis verlässt. In der dritten Spalte ist somit $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ zu erzeugen.

$$\left(\begin{array}{ccccc|c} 0 & 0 & \boxed{-9/2} & 0 & 5/2 & 27 \\ \hline 1 & 0 & 1 & 0 & 0 & 4 \\ 0 & 0 & \boxed{3} & \boxed{1} & -1 & 6 \\ 0 & 1 & -3/2 & 0 & 1/2 & 3 \end{array}\right) \begin{array}{lll} \text{QR:} & & \\ \leftarrow & 4 \div 1 = & 4 \\ \leftarrow & 6 \div 3 = & \textcircled{2} \\ \leftarrow & \text{negativ} & \text{✘} \end{array}$$

$$\left(\begin{array}{ccccc|c} 0 & 0 & 0 & 3/2 & 1 & \boxed{36} \\ \hline 1 & 0 & 0 & -1/3 & 1/3 & 2 \\ 0 & 0 & 1 & 1/3 & -1/3 & 2 \\ 0 & 1 & 0 & 1/2 & 0 & 6 \end{array}\right)$$

Die j_*-te Spalte hat jetzt durch Zeilenoperationen die Form der k_*-ten angenommen, wodurch sich $x_{\mathcal{B}}$ zu $\begin{pmatrix} x_1 & x_3 & x_2 \end{pmatrix}^T$ verändert hat.
Da die oberste Zeile nur Einträge > 0 (sprich *echt größer null* besitzt), ist der Algorithmus am Ende angelangt. Bilde nun keine reduzierten Kosten mehr.

Es ergibt sich eine Optimalstelle $x = \begin{pmatrix} 2 & 6 & 2 & 0 & 0 \end{pmatrix}^T$ und ein Minimum von $f(x) = -36$. ※

Definition 1.23 (*Degeneriertheit*)

Eine zulässige Basislösung x, bei der mehr als $(n-m)$ Elemente null sind, heiße **degeneriert**. Ein lineares Programm heiße **nicht degeneriert**, wenn **keine** degenerierte **zulässige** Basislösung existiert.

Im Allgemeinen gilt, dass für **nicht degenerierte lineare Programme** der Wert der Zielfunktion während der Simplexiteration streng monoton fallend, womit jede der $\binom{n}{m}$-vielen Basislösungen **höchstens** einmal angelaufen werden kann. Dies ist auch im *schlimmsten Fall* das Terminationsmaß, im Durchschnitt ist das Verfahren aber deutlich schneller.
Bei **degenerierten linearen Programmen** *kann* es passieren, dass der Algorithmus das „Kreisen“ anfängt, das heißt er zyklisch immer wieder die gleichen Basen durchläuft – die allesamt dieselbe Polyederecke repräsentieren. Man mag meinen, dass degenerierte lineare Programme eher die Seltenheit sind, aber vor allem in Anwendungsbereichen finden sich solche Programme häufiger. Eine Strategie zum „nicht kreiseln lassen“ ist es **von allen möglichen Indizes j_* und k_* *immer* den *kleinsten* auszuwählen**. Damit gilt auch obiges Terminationsmaß für degenerierte lineare Programme.

1.6.3.2 Schritt 1 des Verfahrens

Was bislang noch gefehlt hat, ist das Finden eines geeigneten Startwerts, also einer beliebigen *zulässigen* Basislösung. Dies ist im Allgemeinen auch nicht trivial machbar, es sei denn man möchte alle möglichen Basislösungen durchprobieren. Vorneweg sei ein extrem einfacher Fall noch erwähnt. Sollte die allgemeine Form des linearen Programms **nur** Ungleichungsnebenbedingungen enthalten, so bilden **immer** die Schlupfvariaben aus Verfahren 1.2, Schritt (3) eine zulässige Basis.

i Man mag sich fragen, warum die Schlupfvariablen immer eine zulässige Basis bilden. Die Begründung liegt auf der Hand, denn für Schlupfvariablen ist $A_{\mathcal{B}} = E_n$ und $x_{\mathcal{B}} = A_{\mathcal{B}}^{-1} b = E_n b = b \geqslant 0$.

OBdA. sei nun $b \geqslant 0$, wie findet man nun einen geeigneten Startwert im allgemeinen Fall *möglichst effizient*?
Wir betrachten dazu zum gegebenen linearen Programm in Standardform das Folgende Hilfsproblem:

$$\begin{array}{rcll} & \text{Suche } x \in \mathbb{R}^n, \widetilde{x} \in \mathbb{R}^m, \text{ so dass} & & \\ & \widetilde{f}(x, \widetilde{x}) := \sum_{i=1}^{m} \widetilde{x}_i & \to \min, & \\ \text{so dass} & Ax + E_m \widetilde{x} & = b & , \\ \text{und} & x & \geqslant 0 & , \\ \text{und} & \widetilde{x} & \geqslant 0 & . \end{array} \qquad (\widetilde{\text{LP}})$$

Für das Hilfsproblem $(\widetilde{\text{LP}})$ kann man jetzt problemlos eine zulässige Basislösung angeben,

Wähle dazu einfach $x_{\mathcal{B}} := \widetilde{x}$ und $x_{\mathcal{N}} := x$. Mit $x_{\mathcal{N}} := 0$ folgt dann $x_{\mathcal{B}} = b \geqslant 0$ und damit ist die Basislösung auch zulässig. Ganz nebenbei hat $(\widetilde{LP})$ übrigens immer eine Lösung: Obige zulässige Basislösung zeigt, dass die zulässige Menge definitiv nicht leer ist und aufgrund der Form von $\widetilde{f}$ ist $\inf \widetilde{f}\big|_M = -\infty$ offensichtlich unmöglich.

sowie folgt aus der Lösung von $(\widetilde{LP})$ eine zulässige Basislösung für das eigentliche Problem.

Wie ermittelt man aus $(x,\widetilde{x})$ von $(\widetilde{LP})$ einen Startwert von unserem eigentlichen linearen Programm? Wir betrachten zwei Fälle:

Fall 1: $\widetilde{f}(x,\widetilde{x}) > 0$, $\sum_i \widetilde{x}_i > 0$ *und* $Ax + \widetilde{x} = b$, *so ist die zulässige Menge leer und es gibt keine Lösung zum linearen Programm. Denn angenommen sie wäre nicht leer, so gäbe es ein x_* mit $Ax_* = b$ und $x_* \geqslant 0$. Mit $\widetilde{x}_* = 0$ gilt dann $(x_*,\widetilde{x}_*) \in M_{(\widetilde{LP})}$ und $\widetilde{f}(x_*,\widetilde{x}_*) = \sum_i \widetilde{x}_i = 0 < \widetilde{f}(x,\widetilde{x})$, was einen Widerspruch zur Optimalität darstellt.*

Fall 2: $\widetilde{f}(x,\widetilde{x}) = 0$ *und* $Ax+\widetilde{x} = b$ *mit* $x \geqslant 0$. *Folgender Schluss ist dann möglich. Da $\sum_i \widetilde{x}_i = 0$ folgt unmittelbar, dass $\forall i : \widetilde{x}_i = 0$, also $Ax = b$ mit $x \geqslant 0$, was bedeutet, dass x ein zulässiger Punkt ist und die zulässige Menge von unserem ursprünglichem linearen Programm nicht leer ist.*

Eine der letzten nun noch offenen Fragen um den Simplexalgorithmus ist, um das ursprüngliche lineare Programm starten zu können, man neben der Angabe des Startwertes x auch noch die Indexmengen $\mathcal{B},\mathcal{N} \subset \{1,\dots,n\}$ herausfinden müsste. Auch das ist jedoch kein Problem. Der Endzustand $(x,\widetilde{x})$ unseres angepassten linearen Programmes $(\widetilde{LP})$ enthält m Basisspalten. Die Basisspalten sind im Allgemeinen über sowohl x als auch $\widetilde{x}$ verteilt. Es kann also sein, dass im x-Block weniger als m Basisspalten enthalten sind. Übernehme für den Startwert von unserem linearen Programm nun **alle** Basisspalten des x-Blocks. Sind das weniger als m, so füge sukzessive solche Spalten von den x hinzu, so dass die m sich ergebenden Basisspalten **linear unabhängig** sind[3]. Die dabei übriggebliebenen $n-m$ Spalten sind die Nichtbasisspalten. Da diese Spalten auch in unserem modifizierten linearen Programm $(\widetilde{LP})$ Nichtbasisspalten waren, ist dafür gesorgt, dass $x_{\mathcal{N}} = 0$.

i Der hier beschriebene Simplexalgorithmus (Verfahren 1.3) ist ein **Zwei-Phasen-Algorithmus**. In Phase I wird – mittels einem Tableau für $(\widetilde{LP})$ – ein Startwert für Phase II ermittelt, in welcher dann das lineare Programm mittels Tableau gelöst wird.

Wir schließen damit unsere Einführung in die lineare Optimierung ab, das nächste Kapitel beschäftigt sich mit dem Finden von Lösungen von Gleichungen der Form $f(x) = x$ und schließt mit zwei numerischen Verfahren zum Lösen von linearen Gleichungssystemen ab.

[3] Die funktioniert nach dem Basisergänzungssatz, den wir in Mathe C1 bereits kennengelernt haben, immer. Der Basisergänzungssatz sagt effektiv aus, dass wenn man ein Erzeugendensystem eines Vektorraumes hat und darin ein Teilsystem von linear unabhängigen Vektoren, man dieses Teilsystem durch Hinzunahme weiterer Vektoren aus dem gegebenen Erzeugendensystem immer zu einer Basis machen kann.

1.7 Fixpunktiterationen

Nach den Optimierungsproblemen der letzten Kapitel soll es jetzt um Gleichungen der Form

$$f(x) = x$$

mit einer beliebigen Abbildung $f \in \mathrm{Abb}(M, M)$, wobei $M \neq \emptyset$ gehen. Wir wollen erkennen, wann solche Gleichungen eindeutig bestimmbare Lösungen besitzen und unsere Ergebnisse auf das Nullstellenproblem übertragen. Wir werden dazu das aus dem zweiten Semester bekannte Newton-Verfahren anders herleiten und interpretieren, sowie zum Abschluss des Kapitels der Analysis zwei numerische Verfahren kennenlernen, mit denen wir lineare Gleichungssysteme schneller und effizienter lösen können. Wir wollen uns in diesem Zusammenhang ebenso mit den Konvergenzeigenschaften dieser Verfahren näher befassen. Fangen wir zuerst aber mit den grundlegenden Definitionen an.

1.7.1 Grundlegendes und der Fixpunktsatz von Banach

Definition 1.24 (*Norm — Wiederholung aus C1*)

Sei V ein $\mathbb{K}$-Vektorraum. Eine Abbildung $\|\cdot\| \in \mathrm{Abb}(V, \mathbb{K})$ heiße **Norm** genau dann, wenn für alle $v, u \in V$ und $\lambda \in \mathbb{K}$ die folgenden Eigenschaften erfüllt sind:

1. $\|v\| \geqslant 0$
2. $\|v\| = 0 \Leftrightarrow v = 0$
3. $\|\lambda v\| = |\lambda| \|v\|$
4. $\|v + u\| \leqslant \|v\| + \|u\|$

Definition 1.25 (*Fixpunkt*)

Sei $M \neq \emptyset$ und $\Phi \in \mathrm{Abb}(M, M)$ eine Abbildung. Ein $x \in M$, das den Zusammenhang

$$\Phi(x) = x \tag{1.13}$$

erfülle heiße **Fixpunkt** von Φ.

Wir betrachten also eine Abbildung $\Phi : M \to M$. Sei $M \subseteq V$, wobei V ein normierter Vektorraum ist. Wir betrachten für einen Startwert $x_0 \in M$ eine rekursiv definierte Folge

$$x_{n+1} := \Phi(x_n), \tag{$*$}$$

und stellen uns die Frage, was der Grenzwert der Folge sei. Aussage darüber verschafft der folgende Satz:

Satz 1.20 (*Fixpunkt*)

Falls die oben definierte Folge $(x_n)_{n \in \mathbb{N}}$ konvergiert mit $\lim\limits_n x_n := x_*$ und Φ stetig sei, so ist der Grenzwert x_* **immer ein Fixpunkt** von Φ.

$$\Phi(x_*) = x_*$$

Beweis: Zu zeigen ist hier die Fixpunkteigenschaft des Grenzwertes. Aufgrund der Stetigkeit, existiert nach dem Epsilon-Delta-Kriterium ein ε und ein δ, so dass, für ein $x \in V$ mit $\|x - x_*\| < \delta$ gilt

$$\|\Phi(x_*) - \Phi(x)\| < \varepsilon.$$

Wegen der Konvergenz von der Folge gilt dann, dass ebenso ein $n_0 \in \mathbb{N}$ existieren muss, so dass für alle n, die echt größer als n_0 sind gilt, dass $\|\Phi(x_*) - \Phi(x_n)\| < \frac{\varepsilon}{2}$ und $\|x_{n-1} - x_*\| \leqslant \frac{\varepsilon}{2}$. Damit gilt dann:

$$\|\Phi(x_*) - x_*\| = \|\Phi(x_*) - x_{n+1} + x_{x+1} - x_*\| \leqslant \|\Phi(x_*) - x_{n+1}\| + \|x_{x+1} - x_*\| \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

Ist die Folge nun konvergent, so folgt $\|\Phi(x_*) - x_*\| = 0$, was nach dem vierten Normaxiom nur bedeuten kann, dass $\Phi(x_*) - x_* = 0$, woraus die Behauptung direkt abzulesen ist. □

Definition 1.26 (*Fixpunktiteration*)

Sei $\Phi \in \mathrm{Abb}(M, M)$ mit $M \subseteq V$, wobei V ein normierter Vektorraum ist und $x_0 \in M$. Die Berechnung

$$x_{n+1} := \Phi(x_n) \qquad (*)$$

heiße **Fixpunktiteration**.

Die Fixpunktiteration hat ihren Namen daher, dass wenn Φ stetig ist und die Fixpunktiteration konvergiert, so der Grenzwert immer ein Fixpunkt von Φ ist. Wir wollen uns jetzt mit der Frage beschäftigen, wann Fixpunktiterationen überhaupt konvergieren. Ein wichtiges Kriterium dafür ist der Bachach'sche Fixpunktsatz, auf den dieses Kapitel hinzielen möchte. Eine Voraussetzung dafür sind sogenannte Banachräume. Diese sollten aus C1 bekannt sein, wir wiederholen sie an dieser Stelle aber nocheinmal.

Definition 1.27 (*Vollständigkeit, Banachraum*)

Ein normierter $\mathbb{K}$-VR $(V, \|\cdot\|)$ heiße **vollständig** genau dann, wenn jede Cauchyfolge in V gegen ein Element in V konvergiert.
Ein Vektorraum heiße **Banachraum** genau dann, wenn er normiert und vollständig ist.
Ein Banachraum, dessen Norm durch ein Skalarprodukt erzeugt wird, heiße **Hilbertraum**.

B

Es ist leicht zu sehen, dass $\mathbb{R}^n$ bezüglich einer jeden Norm ein Banachraum ist. Mit der $\|\cdot\|_2$-Norm ist er sogar ein Hilbertraum.

Ein weiteres interessanteres Beispiel sind Funktionenräume wie beispielsweise $V - \{f \in \mathrm{Abb}([a,b], \mathbb{R}), f \text{ ist stetig}\}$. Betrachten wir V zusammen mit der Maximumsnorm $\|\cdot\|_\infty$, so ist der Raum **vollständig**. Betrachteten wir ihn mit der ersten Norm

$$\|f\|_1 := \int_a^b |f(x)|\,dx$$

, so ist der Raum **kein** Banachraum, aufgrund von fehlender Vollständigkeit. Man betrachtete dazu einfach die Funktionenfolge

$$f_n(x) = \begin{cases} 1 & , x \leqslant 0 \\ x/n & , 0 < x \leqslant \frac{1}{n} \\ 0 & , \text{sonst} \end{cases}$$

Wir wissen aus dem zweiten Semester, dass jedes f_n hier zwar stetig ist, aber die Grenzfunktion nicht. Dennoch ist f_n eine Cauchyfolge bezüglich der $\|\cdot\|_1$-Norm.

Ein anderes Beispiel sei mit $L^1([a,b]) := \left\{f \in \mathrm{Abb}([a,b], \mathbb{R}) \middle| \int_a^b |f(x)|\,dx < \infty\right\}$. Mit dem normalen Riemannintegral, welches wir in C2 näher kennengelernt haben, existieren auch hier Grenzfunktionen, die nicht im Raum liegen, wie beispielsweise die Dirichletfunktion. Nimmt man anstatt der Riemannintegrale allerdings die Lebesgue'schen Integrale, so ergibt sich – interessanterweise – ein vollständiger Raum, sogar ein Banachraum. Es ist sogar ein Hilbertraum unter Verwendung eines bestimmten Skalarproduktes möglich.

Nun zurück zur Frage, wann und unter welchen Bedingungen Fixpunktiterationen konvergieren. Man kann sich *graphisch* überlegen, dass die „Steilheit" von Φ dabei eine Rolle spielt. Wir definieren:

Definition 1.28 (*Kontraktion*)

Sei $(V, \|\cdot\|)$ ein $\mathbb{R}$-Vektorraum, $M \subseteq V$ und $\Phi \in \mathrm{Abb}(M, V)$. Wir sagen Φ heiße **Kontraktion** genau dann, wenn eine Konstante $0 \leqslant k < 1$ existiert, so dass für alle $x, y \in M$ gelte, dass

$$\|\Phi(x) - \Phi(y)\| \leqslant k\|x - y\|. \tag{K}$$

Eine Konstante k, welche diesen Zusammenhang erfüllt, heißt auch **Kontraktionskonstante** von Φ.

i Anschaulich bedeutet es, wenn Φ eine Kontraktion ist, dass die Bilder von Φ näher beieinander liegen als die Urbilder von Φ.

Korrolar D1.28 (*Stetigkeit*)

Eine jede Kontraktion ist stetig.

Das in Definition 1.28 definierte Kriterium für Kontraktionen (K) ist nur sehr umständlich zu überprüfen. Folgender Satz vereinfacht dieses Kriterium, schränkt aber gleichzeitig auch unsere Auswahlmenge an Funktionen ein:

Satz 1.21 (*Kontraktionskriterium*)

Im Fall $V = \mathbb{R}$ ist für die Kontraktionseigenschaft hinreichend, dass f differenzierbar ist mit $k_* := \sup_{x \in M} |f'| < 1$. k_* ist dann **Kontraktionskonstante** von Φ.

Beweis: Der Mittelwertsatz für $\Phi \in \mathrm{Abb}(M, \mathbb{R})$ mit $M \subseteq \mathbb{R}$ ergibt, dass für alle $x, y \in M$ ein $\xi \in M$ existiert, so dass $\Phi(x) - \Phi(y) = \Phi'(\xi) \cdot (x - y)$. Daraus folgt, dass

$$|\Phi(x) - \Phi(y)| = \left|\Phi(\xi)'\right| \cdot |x - y| \leqslant \sup_{z \in M} \left|\Phi(z)'\right| \cdot |x - y|.$$

Man erkennt nun leicht, dass Φ genau dann eine Kontraktion ist, well $\sup_{z \in M} |\Phi(z)'| < 1$ ist, womit die Behauptung gezeigt ist. □

i Es ist sicherlich auch möglich Satz 1.21 auch auf $V = \mathbb{R}^n$ anzupassen, dann hieße die Bedingung $\sup_{x \in M} \|\mathcal{J}f(x)\| < 1$. In diesem Fall ist es aber notwendig zu spezifizieren, welche Matrixnorm man verwendet und überhaupt Normen auf Matrizen ersteinmal zu definieren.

Wir kommen nun zu einer Art „Hauptsatz“ in unserem Kapitel über Fixpunktiterationen, dem Fixpunktsatz von Banach. Mit ihm wollen alles, was wir bislang über Fixpunkte und ihre Iterationen gelernt haben zusammenfassen um eine möglichst allgemeine Aussage über die Konvergenz von Fixpunktiterationen zu treffen.

Satz 1.22 (*Fixpunktsatz von* Banach)

Sei $(V, \|\cdot\|)$ ein **Banachraum** und sie $\emptyset \neq M \subseteq V$ eine **abgeschlossene Teilmenge** von V. Sei $\Phi \in \mathrm{Abb}(M, V)$ **selbstabbildend**, sprich $\Phi(M) \subseteq M$, sowie eine **Kontraktion**. Dann hat Φ genau einen Fixpunkt $x_* \in M$ und x_* ist Grenzwert der Folge $(*)$, wobei $x_0 \in M$ **beliebig** ist.
Sei k zudem noch eine Kontraktionskonstante von Φ, so fällt der Approximationsfehler in jedem Iterationsschritt um mindestens den Faktor k, also

$$\|x_{n+1} - x_*\| \leqslant k\|x_n - x_*\| \text{ damit auch } \|x_n - x_*\| \leqslant k^n \|x_0 - x_*\|$$

Beweis: Wir sehen schnell, dass ausgrund der Selbstabbildungseigenschaft von Φ die Folge (x_n), $x_{n+1} := \Phi(x_n)$ mit einem $x_0 \in M$, wohldefiniert ist. Sei also nun $x_0 \in M$ wohldefiniert ist. Sei also

nun $x_0 \in M$. Nachdem Φ kontrahierend ist, muss ein $\lambda \in [0,1)$, so dass für alle $x, y \in M$ gilt, dass

$$\|\Phi(x) - \Phi(y)\| \leqslant \lambda \cdot \|x - y\|.$$

Mit Kombination dieser Eigenschaften folgt:

$$\begin{aligned}
\|x_{n+1} - x_n\| &= \|\Phi(x_n) - \Phi(x_{n-1})\| \\
&\leqslant \lambda \cdot \|x_n - x_{n-1}\| && = \lambda \cdot \|\Phi(x_{n-1}) - \Phi(x_{n-2})\| \\
&\leqslant \lambda^2 \cdot \|x_n - x_{n-1}\| && = \lambda^2 \cdot \|\Phi(x_{n-1}) - \Phi(x_{n-2})\| \\
&\vdots \\
&\leqslant \lambda^n \cdot \|x_1 - x_0\|
\end{aligned}$$

Diese Ungleichung ist auch der Grund für die obige Fehlerabschätzung. Unter Verwendung der Dreiecksungleichung und dem vorherigen Resultat gilt für $0 \leqslant n < m$, dass

$$\begin{aligned}
\|x_m - x_n\| &\leqslant \|x_m - x_{m-1}\| + \cdots + \|x_{n+1} - x_n\| \\
&\leqslant \lambda^{m-1} \|x_1 - x_0\| + \cdots + \lambda^n \|x_1 - x_0\| \\
&= \lambda^n \cdot \sum_{i=0}^{m-1-n} \lambda^i \cdot \|x_1 - x_0\| \\
&\overset{(*)}{=} \lambda^n \cdot \frac{1 - \lambda^{m-n}}{1 - \lambda} \cdot \|x_1 - x_0\| \\
&= \frac{\lambda^n - \lambda^m}{1 - \lambda} \|x_1 - x_0\| \leqslant \frac{\lambda^n}{1 - \lambda} \|x_1 - x_0\|,
\end{aligned}$$

wobei an der Stelle $(*)$ der Grenzwert der geometrischen Reihe verwendet wurde, da $0 \leqslant \lambda < 1$ ist. Unter der Voraussetzung, dass man $n_0(\varepsilon) \in \mathbb{N} \setminus \{0\}$ nun so wählt, dass für alle $\varepsilon > 0$

$$\frac{\lambda^{n_0(\varepsilon)}}{1 - \lambda} \|x_1 - x_0\| < \varepsilon$$

gilt, so ist (x_n) eine Cauchyfolge. Da M eine Teilmenge des – insbesondere vollständigen – Banachraums $(V, \|\cdot\|)$, besitzt (x_n) **genau** einen Grenzwert $x_* \in M$, *da alle* $x_i \in M$ *und* M *abgeschlossen.* Da Φ als Kontraktion stetig ist (siehe Korrolar D1.28), gilt, dass der Grenzwert der Fixpunktiteration der Fixpunkt sein muss: $\Phi(x_i) = x_*$. Dieser ist dann sogar **eindeutig**. Denn angenommen er wäre es nicht, das heißt es gäbe ein $\widetilde{x} \in M$ mit $\widetilde{x} \neq x_*$, aber $\Phi(\widetilde{x}) = \widetilde{x}$, dann gilt:

$$\|\widetilde{x} - x_*\| = \|\Phi(\widetilde{x}) - \Phi(x_*)\| \leqslant \lambda \cdot \|\widetilde{x} - x_*\| < \|\widetilde{x} - x_*\|,$$

da Φ eine Kontraktion ist und somit $\lambda < 1$. Dies führt aber genau zu einem Widerspruch, aus welchem dann folgt, dass x_* einzelner, eindeutiger Fixpunkt der Kontraktion Φ ist. □

i

Satz 1.22 ist „**konstruktiv**“, damit erklärt er insbesondere auch den Weg zum Fixpunkt und nicht nur Existenz oder Eindeutigkeit.

Falls $M = V$, so ist sowohl die Eigenschaft der Abgeschlossenheit, als auch die Selbstabbildungseigenschaft **trivialerweise** erfüllt.

1.7.2 Zusammenhang zwischen dem Nullstellenproblem und dem Newton-Verfahren

Eine bislang noch unbeantwortete, aber durchaus berechtigte, Frage ist die Motivation für Fixpunkte. Hier lassen sich die verschiedensten Punkte anführen, wie zum Beispiel das Anwenden auf Nullstellenprobleme, die wir in diesem Kapitel behandeln werden, andererseits kann man Fixpunkte und Fixpunktiterationen als Grundlage von Lösungsverfahren für lineare Gleichungssysteme verwenden – was wir in Kapitel 1.7.4 sehen werden – und ebenso können sie die Grundlage für den ein oder anderen Beweis darstellen, was bereits in dem Kurzbeweis zu Satz 1.6 geschehen ist und auch noch an anderer Stelle – beim Beweis des Satzes 2.12 – deutlich gemacht wird. Beschränken wir uns an dieser Stelle aber erstmal mit normalen Nullstellenproblemen. Wir definieren das Nullstellenproblem wie folgt:

Definition 1.29 (*Nullstellenproblem*)

Sei $f \in \mathrm{Abb}(\mathbb{R},\mathbb{R})$ gegeben und stetig – *gegebenenfalls sogar stetig differenzierbar*. Gesucht ist ein/das $x_* \in \mathbb{R}$ mit $f(x_*) = 0$.

Wir wollen nun dieses Problem in ein Fixpunktproblem umwandeln. Es ergibt sich:

Definition 1.30 (*Das Fixpunkt basierte Nullstellenproblem*)

Für vorgegebenes $f \in \mathrm{Abb}(\mathbb{R},\mathbb{R})$, suche ein $\Phi \in \mathrm{Abb}(\mathbb{R},\mathbb{R})$, so dass

$$f(x) = 0 \iff \Phi(x) = x,$$

das heißt x_* ist Nullstelle von f genau dann, wenn x_* ein Fixpunkt von Φ ist.

Die in Definition 1.30 gestellte Aufgabe mag auf den ersten Blick nahezu trivial erscheinen, so es doch viele solcher Φ-Funktionen gibt, die die obige Bedingung auch tatsächlich erfüllen. Unser eigentliches Ziel ist es jedoch ein Φ zu finden, das obiger Bedingung genügt und **gleichzeitig** noch eine **Kontraktion** ist.

Man mag sich wundern, warum wir eine Kontraktion suchen, dabei muss man nur einen Blick in Satz 1.22 werfen, in dessen eine Kontraktion zu den vielen Voraussetzungen zählt. Wir wollen hier auch Satz 1.22 anwenden, da wir letztenendes ja das Problem lösen wollen und dies –für uns zumindest – nur mit einer Kontraktion auf die wir den Fixpunktsatz von Banach anwenden können gelingt.

Veranschaulichung

Um Definition 1.30 zu erfüllen, addiere x auf beiden Seiten. $f(x_*) + x_* = x_*$ ist das Fixpunktproblem von $\Phi(x) := f(x) + x$. Wir rechnen – gemäß Satz 1.21 – $\Phi'(x) = f'(x) + 1$. Damit Φ eine Kontraktion ist, muss Φ' in einer gewählten Umgebung M, in der f selbstabbildend ist, betragsmäßig echt kleiner als 1 sein. Wir erkennen, dass dies nur auf wenige Funktionen f zutrifft.

Ebenso ist $-f(x_*) + x_*$ das Fixpunktproblem zu $\Phi(x) := x - f(x)$. Auch hier bestimmen wir $|\Phi'(x)| = |1 - f'(x)|$. Es muss dann $|\Phi'| \leqslant k < 1$ für ein $k \in [0,1)$ in einer selbstabbildenden Umgebung $K_\varepsilon(x_*)$ gelten.

1.7.2.1 Allgemeiner Ansatz I: Multiplikation mit einer Konstanten λ

Der erste Ansatz an eine Kontraktion Φ zu kommen ist ähnlich dem Beispiel oben:

Definition 1.31 (*„Erweiterung" des Fixpunktbasierten Nullstellenproblems*)

Multipliziere, für ein $f \in \mathrm{Abb}(\mathbb{R},\mathbb{R})$ mit $f(x) = 0$ zunächst mit einem $\lambda \neq 0$ und addiere dann x. Wir erhalten damit:

$$f(x) = 0 \Longleftrightarrow \underbrace{x + \lambda \cdot f(x)}_{=:\Phi(x)} = x$$

Damit Φ eine Kontraktion ist, muss gelten, dass in $K_\varepsilon(x_*)$ $|\Phi'| = |1 + \lambda f'(x)| \leqslant k < 1$ mit selbstabbildenden Umgebung $K_\varepsilon(x_*)$. Damit liegt es nahe das λ wie folgt zu wählen:

$$\lambda := -\frac{1}{f'(x_*)},$$

denn dann ist zumindest $\Phi'(x_*) - 0$ und wenn f' und damit auch Φ' stetig ist, ist $|\Phi'(x)|$ auch in der Umgebung um x_* kleiner als 1. Das einzige Problem hierbei ist, dass x_* bei der Bestimmung von λ a priori **nicht bekannt** ist. Wir wählen deswegen eine Näherung anstatt.
Beispiel 1.7: Es soll eine Nullstelle x_* von $f(x) := x^2 - 8$ berechnet werden. Das Nullstellenproblem für f ist äquivalent zum Fixpunktproblem für $\Phi(x) := x + \lambda f(x)$ mit $\lambda \neq 0$. Wähle für λ also $-\frac{1}{f'(x_*)}$, sprich eine Näherung an die Nullstelle. Wir schätzen $f(3) \approx 0$. Wähle also $\lambda = -\frac{1}{6}$. Damit sieht unser Fixpunktproblem wie folgt aus:

$$\Phi(x) = x - \frac{x^2 - 8}{6}$$

Wir iterieren also:

$$\begin{aligned}
x_1 = \Phi(x_0) &= 3 - \frac{9-8}{6} &&= 2,8\overline{3} \\
x_2 = \Phi(x_1) &= 2,8\overline{3} - \frac{(2,8\overline{3})^2 - 8}{6} &&\approx 2,828707 \\
x_3 = \Phi(x_2) & &&\approx 2,828442929
\end{aligned}$$

Im Vergleich dazu der exakte Wert $x_* = \sqrt{8} \approx 2,828\mathbf{427125}\ldots$.
Wir vermuten aufgrund der sinkenden Differenz zwischen x_n und x_*, dass die Folge (x_n) gegen die gesuchte Nullstelle von f konvergiert. Wir wollen dies nun mathematisch korrekt unter Verwendung des Banach'schen Fixpunktsatzes (Satz 1.22) zeigen.
Wir überprüfen dazu Schritt für Schritt die Vorasussetzungen des Satzes:

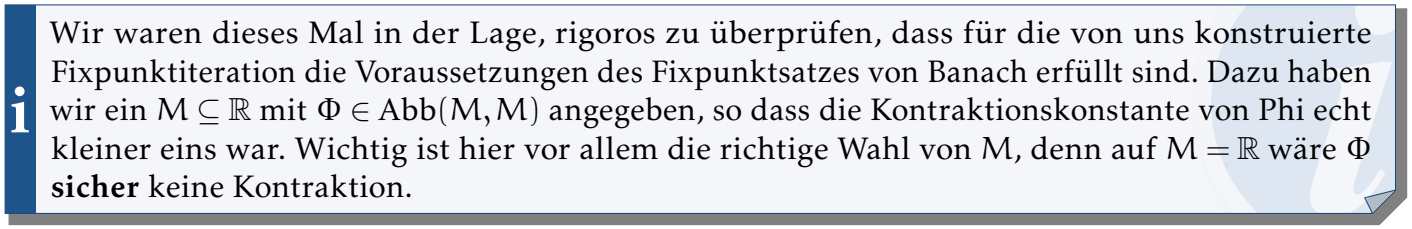
- $V = \mathbb{R}$, $(V, \|\cdot\|)$ ist Banachraum, da $\mathbb{R}$ vollständig und normiert ✓
- Sei $M = [2,3]$. $M \neq \emptyset$ und $M \subset V = \mathbb{R}$, und M ist abgeschlossen ✓
- Für $M = [2,3]$ ist $\Phi'(x) = 1 - \frac{x}{3}$, also $\Phi'(3) = 0 \leqslant \Phi' \leqslant \frac{1}{3} = \Phi'(2)$. Damit ist Φ monoton fallend auf M, weswegen gilt: $M \ni \Phi(2) = 2\frac{2}{3} \leqslant \Phi(x)\Big|_M \leqslant \Phi(3) = 2\frac{5}{6} \in M$, woraus die Selbstabbildungseigenschaft folgt. ✓
- Da $\Phi''(x) = -\frac{1}{3}$, ist Φ' ist damit monoton steigend auf M, somit ist $\sup_{x \in M} |\Phi'(x)| = \frac{1}{3} =: k < 1$, Φ ist Kontraktion mit Kontraktionskonstante $k = \frac{1}{3}$. ✓

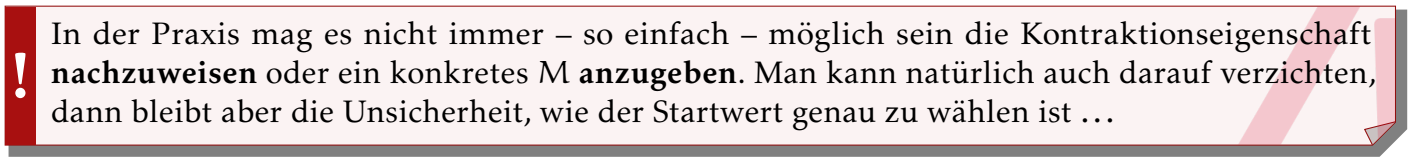
Damit muss Φ gegen die gesuchte Nullstelle konvergieren.

Wir ermitteln zum Schluss noch den Approximationsfehler:

n	x_n	$\lvert x_n - x_* \rvert$
0	3	0,17157
1	2,833333333	0,00490622
2	2,828703703	0,000276578
3	2,828442929	0,000015804
4	2,828428028	0,000000900
$\vdots$	$\vdots$	$\vdots$
	$\to \sqrt{8} =: x_*$	$\to 0$

Wir erkennen, dass der Appoximationsfehler durch k beschränkt ist.





Konvergezgeschwindigkeit Laut Theorie fällt der Fehler pro Schritt um den Faktor k. Wie eben aber schon angemerkt, kann es unter Umständen relativ schwer sein dieses k zu berechnen – auch wegen der Voraussetzung an ein bekanntes M. Wir können jedoch – a posteriori – leicht eine „**asymptotische Kontraktionskonstante**“ $k_* := \text{„}\Phi'(x_*)\text{“}$ angeben, die im Grenzwert $x_n \to x_*$ die Fehlerreduktion beschreiben sollte, sofern $\Phi \in \mathcal{C}^1$ versteht sich. Wir definieren deswegen wie folgt:

Definition 1.32 (*Kontraktionskonstante und asymptotische Kontraktionskonstante*)

Für ein stetig differenzierbares $\Phi \in \mathrm{Abb}(M, \mathbb{R})$ mit $M \subseteq \mathbb{R}$ und einem Fixpunkt $x_* \in M$ heiße

$$k := \sup_{x \in M}$$

die **Kontraktionskonstante**.
Des Weiteren heiße bei gleichem Φ, M und x_*

$$k_* = \left|\Phi'(x_*)\right|$$

die **asymptotische Kontraktionskonstante**. Dabei ist $k_* \leqslant k$.

Erläuterung 1.32

Es gilt $k_* \leqslant k$.
k gibt an, um wieviel sich der Fehler pro Iterationsschritt $x_{n+1} := \Phi(x_n)$ **mindestens** verringert:

$$|x_{n+1} - x_*| \leqslant k \cdot |x_n - x_*|$$

k_* hingegen gibt – bei hinreichend großem n mit $x_n \approx x_*$ – im Allgemeinen einen besseren **Schätzwert** für die **zu erwartende Fehlerreduktion** pro Iterationsschritt ab:

$$|x_{n+1} - x_*| \approx k_* \cdot |x_n - x_*|$$

Bei unserem Beispiel wäre $k_* = \left|\Phi'(\sqrt{8})\right| = 1 - \frac{1}{3}\sqrt{8} \approx \frac{1}{17{,}5}$. In der Tat fällt der Fehler von x_1 auf x_2, x_2 auf x_3 und x_3 auf x_4 um ziemlich genau den Faktor k_*. Lediglich der Startwert ist zu weit von x_* entfernt, hier ist k_* **nicht gültig**.

1.7.2.2 Allgemeiner Ansatz II: Multiplikation mit einer Funktion

Ein weiterer Ansatz das fixpunktbasierte Nullstellenproblem zu lösen ist die Multiplikation mit einer vorgegebenen Funktion g. Wir definieren also auch hier:

Definition 1.33 (*„Zweite Erweiterung" des fixpunktbasierten Nullstellenproblems*)

Anstatt mit einem $\lambda \neq 0$ multipliziere $f(x) = 0$ mit einer Funktion $h \in \mathrm{Abb}(\mathbb{R}, \mathbb{R}$, wobei $h \not\equiv 0$, und addiere anschließend den Linearterm x.

$$f(x) = 0 \Longleftrightarrow \underbrace{x + h(x) \cdot f(x)}_{=:\Phi(x)} = x$$

Damit ist das Nullstellenproblem äquivalent zum Fixpunktproblem mit $\Phi(x) := x + h(x) \cdot f(x) \overset{!}{=} x$.

Um eine möglichst gute Konvergenz zu erhalten, minimieren wir k_*:

$$k_* = \left|\Phi'(x_*)\right| = \left|1 + h'(x_*) \cdot \underbrace{f(x_*)}_{=0} + h(x_*) \cdot f'(x_*)\right| = \left|1 + h(x_*) \cdot f'(x_*)\right|$$

Wir sehen k_* wird minimal – sprich nimmt den Wert null an – genau dann, wenn $h(x_*) = -\frac{1}{f'(x_*)}$. Wir wählen deshalb:

$$h(x) := -\frac{1}{f'(x)},$$

somit ist

$$\Phi(x) = x - \frac{f(x)}{f'(x)}.$$

Unsere Iteration lautet damit:

$$\boxed{x_{n+1} = \Phi(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}}$$

Diese kennen wir so bereits aus dem zweiten Semester, als wir das Newton-Verfahren kennengelernt haben. Wir rekapitulieren und erweitern unsere Verfahrensdefinition aus C2:

Verfahren 1.4 (*Das Newton-Verfahren als Fixpunktiteration*)

Das Newton-Verfahren

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

zur Bestimmung einer Nullstelle x_* von f ist eine Fixpunktiteration, und zwar für die Funktion

$$\Phi(x) := x - \frac{f(x)}{f'(x)}.$$

Die asymptotische Kontraktionskonstante ist hierbei 0.

Beispiel 1.7 (*Fortsetzung*)**:** Sei f wieder wie oben, unser Φ bestimmt sich dann durch

$$\Phi(x) = \frac{1}{2}x + \frac{4}{x}.$$

Die Fixpunktiteration durch Newton liefert:

n	x_n	$\lvert x_n - x_* \rvert$
0	3	$1,7157 \cdot 10^{-1}$
1	2,8333333333333333	$4,90621 \cdot 10^{-3}$
2	2,8284313725490196078	$4,24780 \cdot 10^{-6}$
3	2,8284271247293798213	$3,18972 \cdot 10^{-12}$
4	2,8284271247261900976	$1,79859 \cdot 10^{-24}$
$\vdots$	$\vdots$	$\vdots$

※

Wir erkennen hier:

> **i** Das Newton-Verfahren hat eine Fehlerreduktion der Form $\|x_{n+1} - x_*\| \approx c \cdot \|x_n - x_*\|^2$, sollte man $f \in \mathcal{C}^2$ und $f'(x_*) \neq 0$ voraussetzen.

Auch diese Erkenntnis haben wir im zweiten Semester aus dem Newton-Verfahren gezogen. Wir definieren deswegen ebenso:

Definition 1.34 (*Lineare und quadratische Konvergenz*)

Sei $(x_n) \subset (V, \|\cdot\|)$ eine Folge mit Granzwert $x_* \in V$ und V sei normiert. Das Verfahren heiße …

… **linear konvergent** genau dann, wenn für ein $k < 1$ der Zusammenhang $\lvert x_n - x_* \rvert \leqslant k \cdot \lvert x_{n-1} - x_* \rvert$ erfüllt ist.

… **quadratisch konvergent** genau dann, wenn für ein $c > 0$ der Zusammenhang $\|x_n - x_*\| \approx c \cdot \|x_{n-1} - x_*\|^2$ erfüllt ist.

An dieser Stelle folgt nun ein abschließender Satz dieses Kapitels, hier an dieser Stelle allerdings ohne Beweis, eine Beweisidee sei mit dem Satz von Taylor genannt:

Satz 1.23 (*Hinreichendes Kriterium für quadratisches Konvergenzverhalten*)

Eine Fixpunktiteration sei mindestens **quadratisch konvergent** genau dann, wenn die asymptotische Kontraktionskonstante k_* gleich null it.

> **i** Fixpunktiterationen sind also im Allgemeinen – sofern die Voraussetzungen des Satzes 1.22 erfüllt sind – **linear konvergent**, das heißt der Fehler reduziert sich pro Iterationsschritt mindestens um die Kontraktionskonstante k. Als gute Näherung für die Fehlerabschätzung kann – bei hinreichender Nähe zu x_* – auch die asymptotische Kontraktionskonstante k_* verwendet werden.

1.7.3 Verallgemeinerung des Newton-Verfahrens auf den $\mathbb{R}^m$

Wir sehen in der Praxis häufig nichtlineare Gleichungssysteme, die zu lösen sind. Beispiele hierfür haben wir bereits in Kapitel 1.1 mit $\nabla f \stackrel{!}{=} 0$ oder 1.2 mit dem Lagrangeformalismus gesehen. Wir wollen nun allgemein solche Systeme beschreiben und uns dann ein numerisches Verfahren zur Lösung ebendieser Systeme überlegen.

Definition 1.35 (*Nichtlineares System*)

Sei $f \in \mathrm{Abb}(\mathbb{R}^n, \mathbb{R}^n)$. Wir sagen zu der Gleichung

$$f(x) \stackrel{!}{=} 0,$$

sie sei ein **nichtlineares System aus m Gleichungen**, wenn jede Komponente aus f eine dieser Gleichungen beschreibt.

Definition 1.35 beschreibt damit ein Nullstellenproblem im $\mathbb{R}^m$. Wir wollen auch in diesem Fall ein Newtonverfahren herleiten.

Möglichkeit 1 — Linearisierung Wir argumentieren ganz analog wie im zweiten Semester. Verwenden wir die Taylorentwicklung um $x = x_n$. Die aktuelle Iterierte sei damit

$$\begin{aligned} f(x) = & \underbrace{T_n(x)}_{\to \text{Linearisierung}} + \underbrace{f_R(\xi)}_{\to \text{Restterm}} \\ & = f(x_n) + (\mathcal{J}f(x_n)) \cdot (x - x_n) + f_R(\xi) \end{aligned}$$

Statt $f(x)$ setzen wir nun die Linearisierung gleich null und erhalten damit einen Näherungswert, der die neue Iterierte wird, also $T(x_{n+1}) \stackrel{!}{=} 0$. Ist nun $\mathcal{J}f(x_n)$ invertierbar, so folgt:

$$x_{n+1} := x_n - [\mathcal{J}f(x_n)]^{-1} f(x_n)$$

Möglichkeit 2 — Modellierung als Fixpunktproblem Das Nullstellenproblem für f ist äquivalent zum Fixpunktproblem für

$$\Phi(x) := x - [\mathcal{J}f(x)]^{-1} f(x),$$

vorausgesetzt das Inverse von $\mathcal{J}f(x)$ existiert. Das heißt: $\Phi(x) = x \Leftrightarrow f(x) = 0$
Damit ergibt sich die zugehörige Fixpunktiteration:

$$x_{n+1} := \Phi(x_n) = x_n - [\mathcal{J}f(x_n)]^{-1} f(x_n)$$

Wir definieren damit das Newtonverfahren im $\mathbb{R}^n$ wie folgt:

Verfahren 1.5 (*Newtonverfahren im* $\mathbb{R}^n$)

Sei $f \in \text{Abb}(\mathbb{R}^n, \mathbb{R}^n) \in \mathcal{C}^2(D)$ und es sei die Jacobimatrix $\mathcal{J}f(x)$ invertierbar für alle $x \in \mathbb{R}^n$. So lautet das **Newtonverfahren** zur Bestimmung einer Nullstelle von f:

$$x_{n+1} := x_n - [\mathcal{J}f(x_n)]^{-1} \cdot f(x_n)$$

Das Verfahren ist – wie im skalaren Fall – **lokal quadratisch konvergent**.

Definition 1.36 (*Lokale quadratische Konvergenz*)

Es existiere eine Umgebung $U_\varepsilon(x_*)$ um x_* derart, dass wenn der Startwert $x_0 \in U_\varepsilon(x_*)$ aus ebendieser Umgebung gewählt sei, das Verfahren dann **quadratisch konvergent** heiße.
Lokal ist in diesem Fall dann tatsächlich als Umgebung zu verstehen.

! Man kennt im Allgemeinen U, respektive die Größe von U, **nicht** oder hat nur sehr schwere, sehr rechenaufwändige Formeln für sie.

Man kann die in Verfahren 1.5 getroffene Voraussetzung an die Invertierbarkeit von $\mathcal{J}f$ wie folgt abschwächen:

- Ein erster Gedanke ist, dass es per se ersteinmal reicht, wenn $\mathcal{J}f(x)$ nur für alle x aus der Umgebung $U_\varepsilon(x_*)$ der Nullstelle invertierbar ist. Die lokale quadratische Konvergenzeigenschaft bleibt so erhalten.
- Es reicht sogar, da $\mathcal{J}f \in \mathcal{C}^1$, dass $\mathcal{J}f(x_*)$ invertierbar ist, um auf lokale quadratische Konvergenz in einer Umgebung $U_\varepsilon(x_*)$ zu schließen.

Wie invertiert man die Jacobimatrix effizient? Eine sich jetzt noch der Effektivitätssteigerung bemühte Frage wäre die effiziente Invertierung der Jacobimatrix $\mathcal{J}f$. Die Antwort auf die Frage ist das Auslassen eines jeglichen Invertierschrittes. Dies ist möglich, da es reicht ein lineares Gleichungssystem zu lösen, um den Newtonschritt durchzuführen. Wir schreiben dazu die Iterationsvorschrift um zu

$$(\mathcal{J}f)(x_n) \cdot (x_{n+1} - x_n) = -f(x_n)$$

Wir führen dann die Hilfsgröße Δx mit der Eigenschaft $\Delta x = x_{n+1} - x_n$ ein und erhalten somit:

Verfahren 1.6 (*„Effizienteres" Newtonverfarhren im* $\mathbb{R}^n$)

1. Löse das lineare Gleichungssystem $(\mathcal{J}f)(x_n) \cdot \Delta x = -f(x_n)$
2. Setze $x_{n+1} := x_n + \Delta x$

1.7.4 Fixpunktverfahren für Gleichungssysteme

Wir kennen bereits ein – **allgemeingültiges** – **terminierendes** Verfahren zum Lösen von linearen Gleichungssystemen. Man stellt sich also jetzt die Frage, warum es numerische Verfahren für ein solches – in unseren Augen – *trivial erscheinendes* Problem gesucht werden. Wir geben an dieser Stelle zwei Motivationspunkte:

1. Das Gaußverfahren hat eine Laufzeitkomplexität von $\mathcal{O}(n^3)$, wie wir im ersten Semester festgestellt haben, und hat damit unter Umständen fÜr manche – häufig in der Praxis vorkommende – lineare Gleichungssysteme einen **vergleichsweise hohen Rechenaufwand**. Numerische Verfahren können hier unter Umständen schneller sein.

2. Der sogenannte **Auslöschungseffekt** kann auftreten.
   Durch die Verwendung von Gleitkommazahlen, welche dem Informatiker aus verschiedenen Veranstaltungen bereits bekannt sein sollten, werden Zahlen aufgrund der **festen** Mantisse mit einem gewissen relativen **Rundungsfehler** gespeichert. Der Fehler ist ursprünglich die Anzahl an Mantissenziffern, kann aber durch gewisse Rechenoperationen **drastisch erhöht werden**. Man nennt diesen Effekt dann **Auslöschung**. Wir betrachten dazu folgendes **Beispiel 1.8:** 16-ziffrige Mantisse

$$0,\underbrace{7948236243749276}_{\text{Fehler von } 10^{-16}} - 0,\underbrace{7948236243749241}_{\text{Fehler von } 10^{-16}} = \underbrace{0,35}_{\text{Fehler von } 10^{-2}} \cdot 10^{-14}$$

※

Der Auslöschungseffekt führt auch zu großen Problemen beim Lösen von linearen Gleichungssystemen, denn bei großem n sind derart viele Rechenoperationen notwendig, dass die Wahrscheinlichkeit eines Auslöschungseffekts **sehr hoch** ist. Man mag zwar Techniken zur Reduzierung dieser Wahrscheinlichkeit finden, jedoch gibt es auch lineare Gleichungssysteme, bei denen durch die **nicht entfernbare** Auslöschung, die numerisch ermittelte Lösung **nicht mal im entferntesten** mit der tatsächlichen übereinstimmt.
Fixpunktiterationen sind, durch die wegen der Kontraktionseigenschaft auftretende Dämpfung eines eventuell auftretenden Fehlers mit einem Faktor $k < 1$, weniger anfällig für große Auslöschungseffekte.

Unser Konzept zur Herleitung iterativer Verfahren für lineare Gleichungssysteme Das lineare Gleichungssystem $Ax = b$ mit $A \in \mathbb{R}^{n\times n}$ soll in ein Fixpunktproblem umgewandelt werden. Wir verwenden einen ähnlichen Ansatz wie schon in Kapitel 1.7.2.1. Dazu suchen wir uns eine Matrix $M \in \mathbb{R}^{n\times n}$, multiplizieren diese mit $b - Ax = 0$ und addieren anschließend x. Wir erhalten dadurch

$$x - MAx + Mb = x \Leftrightarrow \underbrace{(E_n - MA)x + Mb}_{=:\Phi(x)} = x,$$

womit unser Φ festgesetzt ist. Damit nun die Kontraktionseigenschaft erfüllt ist, und damit auch die Voraussetzungen des Satzes 1.22, muss

$$\|\Phi(x) - \Phi(y)\| = \|(E_n - MA)(x-y)\| \overset{!}{\leqslant} k\|x-y\|$$

gelten. Ideal wäre dafür selbstverständlich die Wahl von M als das Inverse der Matrix A. Gangbar ist dieser Ansatz aber nur dann, wenn man eine Näherung M an die Matrix A^{-1} kennt. Stellen wir A als Summe von Diagonalmatrix D und Restmatrix R dar, so können wir – *unter der Bedingung, dass R „klein" sei* – M als das Inverse von D setzen, da $D^{-1} \approx A^{-1}$, wobei sich D^{-1} trivial berechnen lässt. Damit gilt dann

$$\begin{aligned}\Phi(x) = (E_n - MA)x + Mb \qquad &= (E_n - \overbrace{D^{-1}}^{M}\underbrace{(D+R)}_{A})x + \overbrace{D^{-1}}^{M}b \\ = (E_n - E_n - D^{-1}R)x + D^{-1}b &= -D^{-1}Rx + D^{-1}b. \qquad (\Phi_{LGS})\end{aligned}$$

Wir beschäftigen uns nun mit zwei Verfahren, welche ebendiese Idee aufgreifen, das Jacobi- und Gauß-Seidel-Verfahren.

1.7.4.1 Jacobi-Verfahren

Verfahren 1.7 (*Jacobi-Verfahren / Gesamtschrittverfahren*)

Gegeben sei ein lineares Glechungssystem $Ax = b$ mit $A = (a_ij) \in \mathbb{R}^{\ltimes} \times \ltimes$ mit $n \in \mathbb{N}$. Seien alle Diagonaleinträge a_{ii} ungleich 0. Wir konstruieren uns gemäß dem Beispiel oben (Φ_{LGS}) ein $\Phi(x) := -D^{-1}Rx + D^{-1}b$. Die Fixpunktiteration $x_{m+1} := \Phi(x_m)$ mit dem so kontstruierten Φ heiße dann **Jacobi**- oder auch **Gesamtschrittverfahren**. Ein Iterationsschritt sehe dabei wie folgt für alle $i = 1, \ldots, n$ aus:

$$(x_{m+1})_i := \frac{1}{a_{ii}}\left(b_i - \sum_{j\in\{1,\ldots,n\}\setminus\{i\}} a_{ij}(x_m)_j\right)$$

Beispiel 1.9: Sei

$$A = \begin{pmatrix} 4 & 1 \\ -1 & 2 \end{pmatrix}, b = \begin{pmatrix} 6 \\ 4 \end{pmatrix},$$

und die Lösung des Gleichungssystems $Ax = b$

$$x_* = \begin{pmatrix} \frac{8}{9} \\ \frac{22}{9} \end{pmatrix}.$$

Wir wählen den Startwert $x_0 = 0$ und definieren den Fehler der Stufe n als $e_n = \|x_n - x_*\|_\infty = \max_i \{|x_{n,i} - x_{*,i}|\}$.
Wir wollen eine näherungsweise Lösung mittels Jacobi-Verfahren (Verfahren 1.7) bestimmen. Dazu

führen wir die ersten vier Schritte durch:

$$x_1 = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix}\left(\begin{pmatrix} 6 \\ 4 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 6/4 \\ 4/2 \end{pmatrix} = \begin{pmatrix} 3/2 \\ 2 \end{pmatrix} \quad \text{(Schritt 1)}$$

$$x_2 = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix}\left(\begin{pmatrix} 6 \\ 4 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 3/2 \\ 2 \end{pmatrix}\right) = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix}\left(\begin{pmatrix} 6 \\ 4 \end{pmatrix} - \begin{pmatrix} 2 \\ -3/2 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 11/9 \end{pmatrix} \quad \text{(Schritt 2)}$$

$$x_3 = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix}\left(\begin{pmatrix} 6 \\ 4 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 11/9 \end{pmatrix}\right) = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix}\left(\begin{pmatrix} 6 \\ 4 \end{pmatrix} - \begin{pmatrix} 11/9 \\ -1 \end{pmatrix}\right) = \begin{pmatrix} 13/16 \\ 5/2 \end{pmatrix} \quad \text{(Schritt 3)}$$

$$x_4 = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix}\left(\begin{pmatrix} 6 \\ 4 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 13/16 \\ 5/2 \end{pmatrix}\right) = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix}\left(\begin{pmatrix} 6 \\ 4 \end{pmatrix} - \begin{pmatrix} 5/2 \\ -13/16 \end{pmatrix}\right) = \begin{pmatrix} 7/8 \\ 77/32 \end{pmatrix} \quad \text{(Schritt 4)}$$

Wir stellen damit die Fehlerwerte auf …

n	e_n
0	$22/9$
1	$11/18$
2	$11/36$
3	$11/144$
4	$11/288$

… und erkennen, dass nach dem vierten Schritt nur noch eine Abweichung von weniger als vier Prozent vorhanden ist. ※

Konvergenz des Jacobiverfahrens Neben Satz 1.22 wollen wir uns nun mit Satz 1.24 eine einfachere Methode überlegen, um die Konvergenz des Jacobiverfahrens (1.7) nachzuweisen. Zuerst definieren wir aber Normen auf Matrizen.

Definition 1.37 (*Matrixnorm*)

Seien V, W normierte Vektorräume und die Abbildung $F \in \mathrm{Abb}(V, W)$ linear (also $F \in L(V, W)$. Dann heiße

$$\|F\| = \sup_{v \in V \setminus \{0\}} \frac{\|Fv\|_W}{\|v\|_V} = \sup_{\|v\|_V = 1} \|Fv\|_W$$

die **Operator-** oder auch **Matrixnorm** von F.

Definition 1.38 (*Spektralradius*)

Sei $A \in \mathbb{R}^{m \times m}$, so heiße

$$\rho(A) := \max_{i=1,\dots,m} |\lambda_i(A)|,$$

wobei $\lambda_i(A)$ der i-te Eigenwert von A sei, der **Spektralradius** von A.

B

$\|\cdot\|_1$ — Spaltensummennorm

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 = \max_{j=1,\dots,m} \sum_{i=1}^{m} |a_{ij}|$$

$\|\cdot\|_2$ — Spektralnorm

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\rho(A^H A)}$$

Dabei ergibt sich für ein invertierbares A der Zusammenhang $\|A\|_2 = \rho(A)$

$\|\cdot\|_\infty$ — Zeilensummennorm

$$\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{m} |a_{ij}|$$

Wir definieren noch eine weitere Eigenschaft von Matrizen, deren Diagonaldominanz, mittels dessen uns der Ausdruck der Bedingung aus Satz 1.24 leichter fällt.

Definition 1.39 (*Diagonaldominanz*)

Sei $A \in \mathbb{R}^{n\times n}$, so heiße A …

… **(zeilenweise) diagonaldominant** genau dann, wenn für alle $i \in \{1,\dots,n\}$

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

gilt.

… **(zeilenweise) *schwach* diagonaldominant** genau dann, wenn für alle $i \in \{1,\dots,n\}$

$$|a_{ii}| \geqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

gilt.

… **spaltenweise diagonaldominant** genau dann, wenn für alle $j \in \{1,\dots,n\}$

$$|a_{jj}| > \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}|$$

gilt.

… **spaltenweise *schwach* diagonaldominant** genau dann, wenn für alle $j \in \{1,\dots,n\}$

$$|a_{jj}| \geqslant \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}|$$

gilt.

Damit ist es uns nun möglich Satz 1.24 zu formulieren.

Satz 1.24 (*Hinreichendes Konvergenzkriterium*)

Sei Φ wie in (Φ_{LGS}), dann ist Φ eine Kontraktion bezüglich der $\|\cdot\|_\infty$-Norm auf ganz $\mathbb{R}^n$, falls die Systemmatrix A **diagonaldominant** ist. Damit konvergiert Verfahren 1.7.

Beweis: Wir betrachten wieder die Iterationsformel aus Verfahren 1.7

$$x_{m+1,i} := \frac{1}{a_{ii}} \left(b_i - \sum_{j \in \{1,\dots,n\} \setminus \{i\}} a_{ij} x_{m,i} \right),$$

sowie die Differenz zweier – so erzeugten – Vektoren x und y

$$x_{m+1,i} - y_{m+1,i} = \frac{1}{a_{ii}} \left(- \sum_{j \in \{1,\dots,n\} \setminus \{i\}} a_{ij} (x_{m,i} - y_{m,i}) \right).$$

Damit folgt dann:

$$|x_{m+1,i} - y_{m+1,i}| \leqslant \overbrace{\underbrace{\frac{1}{a_{ii}} \cdot \sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{ij}|}_{<1} \cdot |x_{m,j} - y_{m,j}|}^{=:k} \leqslant k \cdot |x_{m,j} - y_{m,j}|,$$

wobei $k < 1$ wegen der diagonaldominanz hält und mit diesem Zusammenhang folgt dann unmittelbar, dass

$$\|x_{m+1} - y_{m+1}\|_\infty \leqslant k \cdot \|x_m - y_m\|_\infty,$$

was der Definition einer Kontraktion (Definition 1.28) entspricht. □

Korrolar 1.24

Falls die Systemmatrix A diagonaldominant sei, so ist Φ eine Kontraktion bezüglich **jeder** Norm, da im $\mathbb{R}^n$ alle Normen äquivalent sind.

Effizienzsteigerung beim Jacobiverfahren Vor allem bei den mittels Diskretisierung von partiellen Differentialgleichungen entstehenden linearen Gleichungssystemen ist zwar das Jacobiverfahren im Allgemeinen konvergent, jedoch ist die Kontraktionskonstante k nahe 1. Ein Lösungsvorschlag, der hier näher diskutiert werden soll, ist die **Relaxation**.
Wir wissen, dass $x_{n+1} = D^{-1}b - D^{-1}Rx_n$. Damit folgt unmittelbar, dass

$$x_{n+1} = D^{-1}(D - A)x_n + D^{-1}b = (E_n - D^{-1}A)x_n + D^{-1}b = x_n + \underbrace{D^{-1}(b - Ax_n)}_{\Delta x}.$$

Wir gewichten Δx dann mit einem Parameter $\omega \in (0,2)$ und erhalten so

$$\begin{aligned} x_{m+1} &= x_m + \omega \Delta x_m = x_m + \omega D^{-1}(b - Ax_m) = (\underbrace{E_n - \omega D^{-1}A}_{=:M(\omega)})x_m + \omega D^{-1}b \\ &= M(\omega)x_m + \omega D^{-1}b, \end{aligned}$$

womit dann die neue Fixpunktiteration durch

$$\Phi(x) = M(\omega)x + \omega D^{-1}b \tag{1.14}$$

bestimmt ist. An dieser Stelle sei noch erwähnt:

Definition 1.40 (*Über- und Unterrelaxation*)

Sei Φ wie in (1.14), so heiße das Verfahren ein …

… **Überrelaxationsverfahren** genau dann, wenn ω echt **größer** als 1 ist und

… **Unterrelaxationsverfahren** genau dann, wenn ω echt **kleiner** als 1 ist.

Meistens wählt man $\|M\|_2 < 1$, mit $\|M\|_2 = \rho(M(\omega))$ – da M invertierbar ist, da dann neue Aussagen über $\rho(M(\omega))$ möglich sind. Generell gilt, dass der optimale Wert von ω – an dieser Stelle ohne Beweis, da dies nicht Hauptthema des Semesters ist – sich durch

$$\omega_{opt} = \frac{2}{2 - \lambda_{min} - \lambda_{max}},$$

wobei λ_{min} den minimalen und λ_{max} den maximalen Eigenwert von $M(\omega)$ beschreiben, beschreiben lässt.

1.7.4.2 Gauß-Seidel-Verfahren – Einzelschrittverfahren

Eine weitere Überlegung zur Verbesserung des Jacobiverfahrens ist es bei der Berechnung von $x_{m+1,i}$ die bereits berechneten Komponenten wieder zu verwenden. Denn bei der Berechnung der i-ten Komponente sind die Komponenten $j < i$ des **neuen** Vektors x_{m+1} schon bekannt, es ergibt sich damit folgendes Verfahren.

Verfahren 1.8 (*Gauß-Seidel-Verfahren*)

Gegeben sei ein lineares Glechungssystem $Ax = b$ mit $A = (a_ij) \in \mathbb{R}^{\kappa} \times \kappa$ mit $n \in \mathbb{N}$. Seien des Weiteren alle Diagonaleinträge a_{ii} ungleich 0. Ein Verfahren heiße **Gauß-Seidel**- oder auch **Einzelschrittverfahren** genau dann, wenn es zur Berechnung der i-ten Komponente des neuen Vektors die Iterationsvorschrift

$$x_{m+1,i} := \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} \mathbf{x_{m+1,j}} - \sum_{j=i+1}^{n} a_{ij} x_{m,j} \right),$$

bei der der Unterschied zum Jacobiverfahren (1.7) **blau** hervorgehoben wurde, verwendet.

Beispiel 1.10: Sei das lineare Gleichungssystem

$$\begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix} x = \begin{pmatrix} 6 \\ 8 \end{pmatrix}$$

gegeben. Lösen Sie das lineare Gleichungssystem mit dem Gauß-Seidel-Verfahren auf zwei Ziffern ($\to 10^{-2}$) genau und verwenden Sie als Startwert $x_0 = 0$. Wir betrachten zur Lösung also einerseits

$$x_{n,1} = \frac{1}{4}(6 - 1 \cdot x_{n-1,2})$$

und andererseits

$$x_{n,2} = \frac{1}{3}(8 - 2 \cdot x_{n,1}).$$

Wir erkennen durch Gauß schnell, dass $x_* = (\ 1 \quad 1\ )^T$ ist und rechnen damit aus:

n	$x_{n,1}$	$x_{n,2}$	$\Delta = \|x_{n+1} - x_n\|_\infty$	$\|x_n - x_*\|_\infty$
1	1,5	1,66	1,66	0,5
2	1,08	1,94	0,42	0,08
3	1,01	1,99	0,07	0,01
4	1,0025	1,998	0,008	0,0025

Damit sind wir nach dem vierten Schritt fertig, da der Unterschied echt kleiner als ein Hundertstel ist. ※

Anmerkungen

Auf der rechten Seite der Iteration stehen wirklich nur bekannte Größen

Das Gauß-Seidel-Verfahren lässt sich platzsparender implementieren, bei der Berechnung von $x_{m+1,i}$ ist $x_{m,i}$ überschreibbar, der Wert wird nicht mehr gebraucht.

Das Gauß-Seidel-Verfahren konvergiert im Allgemeinen (bei tridiagonalen Matrizen) um den Faktor **zwei schneller** als das Jacobiverfahren. Einen Beweis dieses Satzes findet sich in Plato, Numerische Mathematik kompakt mit Korollar 10.42.

i Auch das Gauß-Seidel-Verfahren kann man relaxiert auffassen, man nennt dies dann sukzessiv überrelaxiert *„successive overrelaxion“* (SOR). Die Idee hierbei ist den Zielwert entweder **bewusst** zu „überschießen“ oder **bewusst** das „Überschießen“ abzuschwächen. Auch hier gilt dann $\omega \in (0,2)$, sonst divergiert das Verfahren. Man erhält dann:

$$x_n = H_\omega x_{n-1} + \omega(D + \omega L)^{-1} b,$$

wobei H_ω definiert ist als

$$H_\omega := (D + \omega L)^{-1}\left[(1-\omega)D - \omega R\right] = E_n - \omega(D + \omega L)^{-1} A.$$

Damit sind wir am Ende der reinen Analysis angekommen, das folgende Kapitel beschäftigt sich nun mehr mit einem Gleichungstyp ähnlich wie in Kapitel 1.3.

KAPITEL 2

Gewöhnliche Differenzialgleichungen

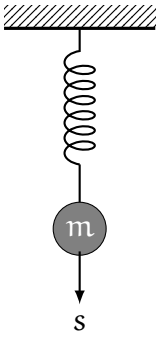
In Kapitel 1.3 haben wir Gleichungen der Form $F(x,y) = 0$ studiert, die unter geeigneten Auflösbarkeitsbedingungen implizit eine Funktion $y = f(x)$ definieren. Treten in der Funktion F aber außer der gesuchten Funktion $y(x)$ auch noch deren (**partielle**) **Ableitungen** nach den Koordinaten $x_1, \ldots, x_n$ auf, so heiße die Gleichung $F = 0$ eine **Differentialgleichung** (DGl.). Ist $x \in \mathbb{R}$, so heiße die Differnetialgleichung **gewöhnlich**, andernfalls **partiell**. Wir wollen uns in diesem Kapitel aber ausschließlich – bis auf einen kleinen Ausblick in Kapitel 2.6.2 – mit **gewöhnlichen** Differentialgleichungen beschäftigen.

2.1 Einführung, Beispiele, grobe Klassifizierung

2.1.1 Motivation und Einführung an anwendungsorientierten Beispielen

2.1.1.1 Das Hook'sche Gesetz – Ein Federschwinger

Sei ein Körper κ_m der Masse m an einer elastischen Feder befestigt. Gesucht ist der Ort $s(t)$ des Körpers zum Zeitpunkt t.
Nach dem Hook'schen Gesetz ist die rücktreibende Kraft F_{Feder} proportional zur Auslenkung $s(t)$, es ergibt sich also als Formel



$$F_{\text{Feder}} = -k \cdot s(t).$$

Nach Newton ist die Beschleunigung $\ddot{s}(t)$ proportional zur Kraft $F_a(t) = m \cdot \ddot{s}(t)$. Damit ergibt sich eine Differentialgleichung der Form

$$\ddot{s}(t) + \omega^2 s(t) = 0$$

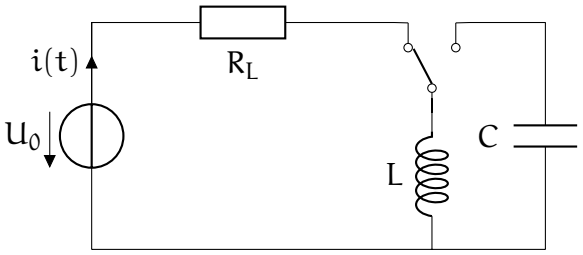
2.1.1.2 Der RL-Stromkreis

Sei $i(t)$ die Stromstärke und $u(t)$ die anliegende Spannung zum Zeitpunkt t, R ein ohm'scher Widerstand und L die Induktivität der Spule, so ergibt sich die Differentialgleichung:

$$u(t) = L \cdot \dot{i}(t) + R \cdot i(t)$$

2.1.1.3 Der elektromagnetische Schwingkreis

Sei $i(t)$ die Stromstärke zum Zeitpunkt t, R ein ohm'scher Widerstand, C die Kapazität des Kondensators und L die Induktivität der Spule, so wird die Stromstärke implizit durch Folgende Differentialgleichung beschrieben:



$$LC \cdot i'' + RC \cdot i' + i = 0$$

2.1.1.4 Populationsdynamik

Sei $s_p(t)$ die Größe einer Population p zum Zeitpunkt t. Man kann verschiedene Arten des Wachstums beschreiben, wovon zwei interessante hier vorgestellt werden:

2.1.1.4.1 Exponentielles Wachstum

Sei die Wachstumsrate proportional zur Populationsgröße, so ergibt sich

$$s_p'(t) = k \cdot s_p(t)$$

Man erkennt mit dem Wissen aus dem letzten Semester schnell, dass $s_p(t) = s_0 \cdot \exp(k \cdot t)$ die Lösungsschar der Differentialgleichung ist.

2.1.1.4.2 Logistisches Wachstum

Sei die Wachstumsrate nun proportional zu $\overline{s} - s_p(t)$, wobei $\overline{s}$ eine Wachstumsschranke darstellt, so ergibt sich:

$$s_p'(t) = k \cdot \overline{s} \cdot s_p(t) - k \cdot (s_p(t))^2$$

2.1.1.5 Die Potentialgleichung

$$\Delta f(x,y,z) = \frac{\partial^2}{\partial x^2} f(x,y,z) + \frac{\partial^2}{\partial y^2} f(x,y,z) + \frac{\partial^2}{\partial z^2} f(x,y,z)$$

2.1.1.6 Jäger-Beute-Modelle (*nach Volterra-Lotka*)

Sei α die Reproduktionsrate der Beute ohne Störung, β die Sterberate der Beute pro Räuber/Jäger, γ die Sterberate der Räuber ohne Störung und δ die Reproduktionsrate der Räuber pro Beutelebewesen. Sei $x(t)$ der Bestand an Beutelebewesen und $y(t)$ der Bestand an Jägern zum Zeitpunkt t, so ergibt sich Folgendes System von Differentialgleichungen:

$$\begin{cases} x'(t) &= \alpha x - \beta xy \\ y'(t) &= \delta xy - \gamma y \end{cases}$$

2.1.1.7 Der „freie“ Fall *ohne* Widerstände

Werde ein Körper κ_1 von einer Höhe h_0 mit konstanter Beschleunigung g von einem anderen Körper κ_B – mit $m(\kappa_B) >>> m(\kappa_1)$ – angezogen (wir sagen dann, der Körper κ_1 fällt zum Körper κ_B) und passiere dies in einem luftleeren Raum, so dass wir eventuelle Reibungen aufgrund deren Minimalität außer Acht lassen können, so ergibt sich folgende Differentialgleichung für die aktuelle Höhe:

$$\ddot{h}(t) = g$$

Dies ist vor allem durch Newtons zweites Gesetz der Mechanik begründet. Damit ergibt sich als Lösung der Differentialgleichung:

$$h(t) = \int \left(\int s(t) dt \right) dt = \int \left(\int g dt \right) dt = \int (g \cdot t + c_1)\, dt = \frac{1}{2} \cdot g \cdot t^2 + c_1 \cdot t + c_2$$

An diesem Beispiel wird vor allem die Wichtigkeit des Anfangswertproblemes (AWP) bewusst, welches wir in Definition 2.6 näher definieren werden. Nur durch Anfangswerte ist eine den angegebenen Sachverhalt korrekt modellierende Funktion anzugeben. Mit Starthöhe h_0 und Startgeschwindigkeit v_0 ergibt sich dann:

$$h(t) = \frac{1}{2} \cdot g \cdot t^2 + v_0 \cdot t + s_0$$

2.1.2 Grobe Klassifizierung

Definition 2.1 (*Gewöhnliche Differentialgleichung*)

Sei $n \in \mathbb{N}$. Ist $F \in \mathrm{Abb}(\mathbb{R}^{n+2}, \mathbb{R})$ eine gegebene **skalare** Funktion, so heiße die Gleichung

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

eine **implizite gewöhnliche Differentialgleichung n-ter Ordnung** für eine gesuchte Funktion $y = y(x)$. Sie heiße **gewöhnlich**, da $x \in \mathbb{R}$ skalar ist.

Beispiel 2.1: Ein Beispiel hierfür ist die Euler'sche Dgl. mit

$$F(x, y, y', \ldots, y^{(n)}) := \left(\sum_{k=0}^{n} a_k \cdot (c \cdot x + d)^k \cdot y^{(k)}(x) \right) - b(x),$$

wobbei $cx + d$ echt größer als 0 sei. ※

Definition 2.2 (*Explizite gewöhnliche Differentialgleichung*)

Sei $n \in \mathbb{N}$. Ist $F \in \mathrm{Abb}(\mathbb{R}^{n+1}, \mathbb{R})$ eine gegebene **skalare** Funktion, so heiße die Gleichung

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

eine **explizite gewöhnliche Differentialgleichung n-ter Ordnung** für eine gesuchte Funktion $y = y(x)$. Sie heiße **gewöhnlich**, da $x \in \mathbb{R}$ skalar ist.

Hierfür sei ein Beispiel in 2.1.1.4.1 gegeben.

Definition 2.3 (*Autonome Differentialgleichung*)

Sei $n \in \mathbb{N}$. Ist $f \in \mathrm{Abb}(\mathbb{R}^n, \mathbb{R})$ eine gegebene **skalare** Funktion, so heiße die Gleichung

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

eine **autonome explizite gewöhnliche Differentialgleichung n-ter Ordnung** für eine gesuchte Funktion $y = y(x)$.

Auch hier sei ein Beispiel in 2.1.1.4.1 gegeben.

Definition 2.4 (*Differentialgleichungssystem*)

Sei $n \in \mathbb{N}$. Ist $f \in \mathrm{Abb}(\mathbb{R} \times \mathbb{R}^{m \times (n+1)}, \mathbb{R}^{<})$ eine gegebene **vektorwertige** FUnktion, so heiße die Gleichung

$$f(x, y, \ldots, y^{(n)}) = 0$$

ein **implizites Differentialgleichungssystem n-ter Ordnung** für eine gesuchte **vektorwertige** Funktion $y = y(x) \in \mathrm{Abb}(\mathbb{R}, \mathbb{R}^m)$.
Analog spricht man bei Gleichungen der Form

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

von einem **expliziten Differentialgleichungssystem n-ter Ordnung** für eine gesuchte **vektorwertige** Funktion $y = y(x) \in \mathrm{Abb}(\mathbb{R}, \mathbb{R}^m)$.

Ein Beispiel für ein Differentialgleichungssystem wurde in 2.1.1.6 mit dem Jäger-Betue-Modell gegeben.

Definition 2.5 (*Allgemeine und partikuläre Lösungen*)

Eine Lösung $y = y(x, c_1, \ldots, c_n)$ heiße **allgemeine Lösung** der Differentialgleichung n-ter

Ordnung, wenn jede spezielle Lösung durch geeeignete Wahl der Konstanten $c_1,\dots,c_n$ aus der Lösung y konstruiert werden kann.
Eine jede auf solche Weise für spezielle Werte der freien Konstanten $c_1,\dots,c_n$ aus der allgemeinen Lösung konstruierte Lösung heiße **partikuläre Lösung** der Differentialgleichung.

Dass die explizite Differentialgleichung unter gewissen allgemeinen Voraussetzungen an f stets eine allgemeine Lösung haben wird, sehen wir in Kapitel 2.3. Der allgemeine Fall soll hier aber nicht näher thematisiert werden. Darüber hinaus werden wir uns auf einige spezielle Typklassen einschränken, für die eine **vollständige** Lösungstheorie existiert.
Das Anwendungsfeld für Differentialgleichungen ist – wie bereits in Kapitel 2.1.1 gesehen – sehr weiträumig und umfasst unter anderem die Geometrie, die verschiedensten Teildisziplinen der Physik, die Biomathematik oder die technischen und ingenieurswissenschaftlichen Disziplinen. In diesen Anwendungsbereichen ist man meist aber nur an partikulären Lösungen interessiert, die zum Beispiel durch Anfangsbedingungen eines physikalischen Systems oder Bedingungen an den Rändern des Betrachtungsintervalls aus der allgemeinen Lösung selektiert werden. Wir unterscheiden also und wollen definieren:

Definition 2.6 (*Anfangswertproblem und Randwertproblem*)

Eine Differentialgleichung/ Ein Differentialgleichungssystem $y^{(n)} = f(x, y, \dots, y^{(n-1)})$ zusammen mit den Forderungen $y(x_0) = y_0.y'(x_0) = y_1,\dots,y^{(n-1)}(t_0) = y_{n-1}$, wobei $t_0, y_0, \dots, y_{n-1}$ fest gegeben ist, heiße **Anfangswertproblem n-ter Ordnung** (AWP).
Ist auf einem Intervall $I = [a, b]$ jedoch eine Lösung $y \in \mathcal{C}^n(T)$ gesucht, deren Funktionswerte in den Randpunkten $a, b \in I$ vorgegeben sind (Linearkombinationen oder nichtlineare Relationen dieser Funktionswerte sind ebenfalls möglich), so verstehen wir das Problem als **Randwertproblem n-ter Ordnung** (RWP).

Wir wollen uns hier – im Rahmen dieser Grundlagenveranstaltung – nur mit Anfangswertproblemen in einem hinreichend allgemeinen Rahmen beschäftigen, für Randwertprobleme sei an dieser Stelle auf spätere vertiefendere Veranstaltung hingewiesen.
Wir wollen nun kurz den Begriff und die Definition des Anfangswertproblems näher motivieren. Sei dazu noch einmal das Beispiel 2.1.1.4.1 ranzuziehen. Wir sahen ein, dass für die Differentialgleichung $y'(x) = ky(x)$ die **Schar** an Funktionen $y_c(x) = c \cdot \exp(k \cdot x)$, wobei $c \in \mathbb{R}$ beliebig, die Lösung beschreibt. Damit ist $y_c(x)$ nach Definition 2.5 eine allgemeine Lösung der Differentialgleichung.
Wollen wir hier zu einer – tatsächlich – **eindeutigen** Lösung kommen, so geben wir einen **festen** Anfangswert $y(x_0) = y_0$ für gegebenes,. **festes** $x_0, y_0 \in \mathbb{R}$ vor und erhalten somit als partikuläre Lösung

$$y(x) = y_0 \cdot \exp(k \cdot (x - x_0)).$$

Wir wollen später in Kapitel 2.3 erkennen, dass „harmlose“ Anforderungen an f reichen sollen, um mit einem Anfangswert das Anfangswertproblem einer Differentialgleichung, oder auch eines Systems von Differentialgleichungen, **erster** Ordnung exakt und eindeutig zu lösen. Das motiviert den Begriff des Anfangswertproblems **erster** Ordnung.
Eine Frage, die sich nun stellt, ist, ob sich obige Motivation auch auf Differentialgleichungen(-systeme) höherer Ordnung übertragen lässt. Man betrachte – für eine erste Idee – das relativ triviale Beispiel 2.1.1.7. Wir haben aus der Differentialgleichung **zweiter** Ordnung $h''(t) = g$ durch zweifaches Integrieren die allgemeine Lösung

$$h(t) = \frac{g}{2}t^2 + c_1 t + c_0$$

gefunden. Durch das „sinnvolle“ Wählen der Parameter c_0 und c_1 als Anfangshöhe und -geschwindigkeit haben wir dann die partikuläre Lösung

$$h(t) = \frac{g}{2}t^2 + v_0 t + s_0$$

erhalten. Wir erkennen, die allgemeine Form enthielt **zwei** Parameter, also haben wir auch **zwei** Anfangsbedingungen gestellt.
Gilt das auch für andere – weniger triviale – Differentialgleichungen n-ter Ordnung, dass man n Anfangsbedingungen stellen kann?
Wir wollen dazu das folgende wichtige Konzept betrachten:

Lemma 2.1 (*Umwandlung Differentialgleichungen n-ter Ordnung*)

Sei $y^{(n)} = f(x, y, \ldots, y^{(n-1)})$ eine skalare Differentialgleichung n-ter Ordnung. Man setze

$$y_{neu} = \begin{pmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \end{pmatrix}$$

und leite damit ein System erster Ordnung $y'_{neu} = \widetilde{f}(x, y_{neu})$ für die gesuchte Funktion $y_{neu} = y_{neu}(x)$ her.

Beweis: Wir beweisen das Lemma durch Angabe der Konstruktion von $\widetilde{f}$. Sei also die Differentialgleichung wie oben und y_{neu} definiert als

$$y_{neu} := \begin{pmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \end{pmatrix},$$

so ergibt sich für y'_{neu}

$$y'_{neu} = \begin{pmatrix} y' \\ y'' \\ \vdots \\ y^{(n-1)} \\ y^{(n)} \end{pmatrix} = \begin{pmatrix} y_{neu,2} \\ y_{neu,3} \\ \vdots \\ y_{neu,n} \\ f(x, y_{neu,1}, \ldots, y_{neu,n}) \end{pmatrix},$$

womit dann für $\widetilde{f}$ gilt, dass

$$\widetilde{f}(x, y) = \begin{pmatrix} y_2 \\ y_3 \\ \vdots \\ y_n \\ f(x, y_1, \ldots, y_n) \end{pmatrix}$$

und sich die Differentialgleichung schreiben lässt als

$$y'_{neu} = \widetilde{f}(x, y_{neu}(x))$$

□

Ein Beispiel für diese Technik soll nun mit der Umwandlung der Differentialgleichung des Federschwingers unter Berücksichtigung der Reibung gegeben werden, wir rechnen also **Beispiel 2.2:**

$$x''(t) = -\frac{c}{m}x'(t) - \frac{k}{m}x(t)$$

Wir definieren also nach Lemma 2.1 x_{neu} als

$$x_{neu}(t) := \begin{pmatrix} x(t) \\ x'(t) \end{pmatrix}$$

und wandeln damit die Differentialgleichung in das folgende System um

$$x'_{neu} = \begin{pmatrix} x'_{neu,1} \\ x'_{neu,2} \end{pmatrix} = \begin{pmatrix} x_{neu,2} \\ -\frac{c}{m}x_{neu,2} - \frac{k}{m}x_{neu,1} \end{pmatrix}$$

※

Damit sei der Begriff und die Definition ausreichend motiviert, die Definition sollte sich jetzt auch leicht erschließen. Bei Systemen n-ter Ordnung zu m Gleichungen sei dann auch klar, dass ein System erster Ordnung mit $n \cdot m$ Gleichungen entsteht.
Die zentralen Fragen, mit denen wir uns nun im Folgenden näher beschäftigen wollen lauten:

(1) Existiert überhaupt eine Lösung?
Eine Antwort auf diese Frage findet sich in Kapitel 2.3

(2) Ist diese Lösung eindeutig?
Eine Antwort auf diese Frage findet sich in Kapitel 2.3

(3) Hängt diese Lösung stetig von den Parametern der Aufgabe (Anfangswerte, Randwerte, ...) ab?

(4) Wie bestimmt man die Lösung?
Eine Antwort auf diese Frage findet sich in den Kapiteln 2.2, 2.4, 2.5 und 2.6

Wir wollen Probleme, bei denen Fragen (1) – (3) bejaht werden können, als **korrekt gestellt** bezeichnen.
Oft gelingt es aber auch Differentialgleichungen explizit zu bestimmen, man kann dann die Fragen (1) – (4) simultan beantworten. Dieser Fall soll nun näher in Kapitel 2.2 betrachtet werden.

2.2 Elementare Lösungsverfahren für skalare Differenzialgleichungen erster Ordnung

Wir wollen uns für den Moment mit expliziten Differentialgleichungen erster Ordnung beschäftigen, sprich mit Gleichungen der Form

$$\frac{dy}{dx} =: y' = f(x,y). \qquad (*)$$

In diesem Kapitel werden wir uns auf **spezielle** rechte Seiten $f(x,y)$ beschränken.

2.2.1 Typus A: Differentialgleichungen mit getrennten Variablen

Seien die stetigen Funktionen $h, g \in \mathrm{Abb}(\mathbb{R},\mathbb{R})$ gegeben, wobei für alle y aus $\mathbb{D}_g$ $g(y)$ ungleich 0 gelte. Sei des Weiteren $f(x,y)$ darstellbar als $h(x) \cdot g(y)$. Zur Lösung der Differentialgleichung

$$y' = h(x) \cdot g(y)$$

wollen wir das Verfahren der *„Trennung der Veränderlichen (TdV)"* kennenlernen und verwenden.

Verfahren 2.1 (*Trennung der Veränderlichen*)

Seien $h, g \in \mathrm{Abb}(\mathbb{R}, \mathbb{R})$ und $f(x,y) = h(x) \cdot g(y) = y'$ die Differentialgleichung für eine gesuchte Funktion $y = y(x)$. Wir betrachten nun folgende Fälle:

(a) $g(y_*) = 0$, so ist $y(x) :\equiv y_*$

(b) Für alle y aus $\mathbb{D}_g$ gelte $g(y) \neq 0$, so gilt unter Anwendung der Substitutionsregel

$$\frac{1}{g(y)} \cdot y' = h(x)$$

oder

$$\int \frac{y'}{g(y)} dy = \int h(x) dx + c$$

ist allgemeine Lösung der Differentialgleichung.

Betrachte nun das Anfangswertproblem mit der Anfangsbedingung $y(x_0) = y_0$ und $g(y_0) \neq 0$. Auf Intervallen mit $y(x) \neq 0$ gilt, dass für jeden Punkt x_0 aus dem Definitionsbereich $\mathbb{D}_h$ hat die Anfangswertaufgabe **genau eine Lösung** $y(x)$, welche implizit durch

$$\int_{y_0}^{y(x)} \frac{1}{g(t)} dt = \int_{x_0}^{x} h(s) ds$$

beschrieben wird.

Anmerkungen

Lösungen eines solchen Anfangswertproblems sind im Allgemeinen **lokale Lösungen**, sie existieren lediglich auf einem Intervall $I(x_0)$ in der Umgebung $K^{\varepsilon}_{x_0}$ um x_0, in der $g(y(x)) \neq 0$ für alle x aus dem Intervall gilt. Selten gilt $I(x_0) = \mathbb{R}$.

Was passiert nun genau bei $g(y_0) = 0$?
Wir haben diesen Fall bislang nur stiefmütterlich behandelt, was sich jetzt auch nicht besonders ändert. Wir geben hier nur eine kleine Startidee, näheres kann in Vertiefungsveranstaltungen mitgenommen werden.
Offensichtlich ist nun eine Lösung durch $y^*(x) := y_0$ gegeben. Existieren aber nun **Berührungspunkte** (x, y_0) der Geraden $y^*(x)$ mit den Lösungskurven, so kann man in (x, y_0) von dieser Geraden **stetig differenzierbar** in eine **andere** Lösungskurve überwechseln, das Anfangswertproblem ist also **mehrdeutig** lösbar. Wir wollen uns den genauen *mathematischen Grund* hier an dieser Stelle, wie eben erwähnt, nicht näher überlegen, aus diesem folgt allerdings dann unmittelbar ein **hinreichendses Eindeutigkeitskriterium**.

Satz 2.2 (*Hinreichendes Eindeutigkeitskriterium*)

Seien stetige Funktionen $f, g \in \mathrm{Abb}(\mathbb{R}, \mathbb{R})$ sowie ein Punkt $y_0 \in \mathbb{D}_g$ mit $g(y_0) = 0$ gegeben. Sei das Anfangswertproblem definiert als $y' = f(x) \cdot g(y)$ mit $y(x_0) = y_0$. Hinreichend für die **eindeutige** Lösbarkeit des Anfangswertproblems ist, dass das folgende uneigentliche Integral **nicht existiert**:

$$\lim_{\varepsilon \to 0^{\pm}} \int_{y_0 + \varepsilon}^{y(x)} \frac{1}{g(t)} dt$$

Beweis: Aufgrund der *für uns* mangelnden Relevanz und des Nichtvorhandenseins des Satzes im aktuellen Curriculum soll an dieser Stelle auf einen Beweis verzichtet werden. □

Beispiel 2.3: Sei

$$y'(x) = x \cdot y^2(x), y(x_0) = y_0 \neq 0.$$

Wir wollen die Differentialgleichung durch das Verfahren „Trennung der Veränderlichen" versuchen zu lösen. (Verfahren 2.1) Wir stellen leicht fest, dass

$$g(y(x)) = (y(x))^2$$

und

$$h(x) = x.$$

Wir lösen also das Anfangswertproblem mit:

$$\begin{aligned}
\text{AWP} \Leftrightarrow \quad \frac{y'}{y^2} &= x & \Leftrightarrow \quad \int_{y_0}^{y(x)} \frac{1}{\eta^2} d\eta &= \int_{x_0}^{x} \chi d\chi \\
\Leftrightarrow \quad -\frac{1}{\eta}\Big|_{y_0}^{y(x)} &= \frac{1}{2}\chi^2\Big|_{x_0}^{x} & \Leftrightarrow \quad \frac{1}{y_0} - \frac{1}{y(x)} &= \frac{1}{2}(x^2 - x_0^2) \\
\Leftrightarrow \quad y(x) &= \left(\frac{1}{y_0} - \frac{1}{2}(x^2 - x_0^2)\right)^{-1}
\end{aligned}$$

Man kann sich nun ebenfalls noch die Frage stellen, was $I(x_0)$ ist, dazu betrachten wir:

$$\frac{1}{y_0} - \frac{1}{2}x^2 + \frac{1}{2}x_0^2 \stackrel{!}{=} 0 \Leftrightarrow x^2 = x_0^2 + \frac{2}{y_0}.$$

Gilt nun:

- $x_0^2 + \frac{2}{y_0} < 0$, also $-\frac{2}{x_0^2} < y_0 < 0$, so ist die Gleichung nie erfüllt, also gilt $I(x_0) = \mathbb{R}$
- $x_0^2 + \frac{2}{y_0} > 0$, so muss zur Existenz $t \neq \pm\sqrt{x_0^2 + \frac{2}{y_0}}$ gelten, damit ist

$$I(x_0) = \begin{cases} \left(\sqrt{x_0^2 + \frac{2}{y_0}}, \infty\right) & \text{falls } y_0 < 0 \\ \left(-\sqrt{x_0^2 + \frac{2}{y_0}}, \sqrt{x_0^2 + \frac{2}{y_0}}\right) & \text{falls } y_0 > 0 \end{cases}$$

※

Beispiel 2.4: Die Differentialgleichung

$$y'(x) = 2x \cdot \sqrt{y(x)}$$

ist ebenfalls vom obigen Typ mit

$$f(x) := 2x, g(y) := \sqrt{y},$$

wobei für alle x gelte, dass $y(x) > 0$. Es folgt durch TdV

$$2\sqrt{y} = \int \frac{dy}{\sqrt{y}} = \int 2x \, dx = x^2 + c,$$

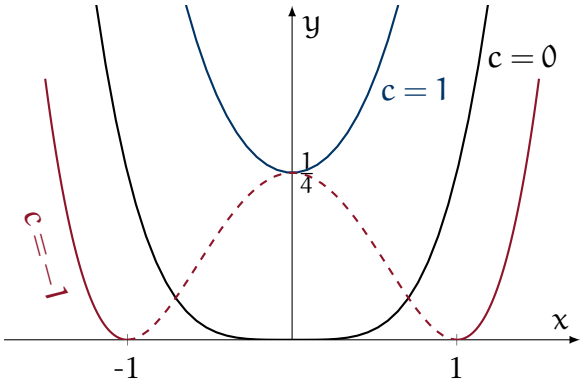
und somit

$$y(x) = \frac{1}{4}(x^2 + c)^2, (x^2 + c) \geqslant 0.$$

Sei nun $y(0) = \frac{1}{4}$ das zugehörige Anfangswertproblem, so gilt:

$$y(0) = \frac{1}{4}(0^2 + c)^2 = \frac{c^2}{4} \stackrel{!}{=} \frac{1}{4},$$

also $c = 1$, weil $(x^2 + c) \geqslant 0$ gelten muss.
Ein Ausschnitt aus der Lösungsschar sei beispielhaft rechts skizziert. Dabei sei zu beachten, dass $c = -1$ **nicht** Teil der Lösungsschar ist. Die Lösung des AWP ist **farbig** markiert.



※

2.2.2 Typus B: Homogene Differentialgleichungen

Eine Differentialgleichung

$$y' = f\left(\frac{y}{x}\right) \tag{Δ}$$

mit $x \neq 0$ heiße, vorausgesetzt $f \in \mathrm{Abb}(\mathbb{R}, \mathbb{R})$ sei stetig, **homogen**. Wir führen diese Differentialgleichung nun mittels Substitution auf eine Differentialgleichung des Typs A, welchen wir in Kapitel 2.2.1 behandelt haben, zurück. Dazu wählen wir folgenden Ansatz:
Sei

$$u(x) := \frac{y(x)}{x},$$

so ergibt sich

$$y(x) = x \cdot u(x), y'(x) = x \cdot u'(x) + u(x)$$

und somit die Differentialgleichung

$$u'(x) = \frac{f(u(x)) - u(x)}{x}$$

mit dem Lösungsansatz

$$\int \frac{du}{g(u) - u} = \int \frac{dx}{x} = \ln|x| + c.$$

Mit $g(u) - u \neq 0$ erhält man eine Lösung $u(x)$, durch Rücksubstitution dann $y(x) = x \cdot u(x)$. ISt $g(u) - u$ nicht nicht 0, so gilt an Stelle $g(c) = c$ $u' \cdot x = 0$, also ist $u(x) = c$ und somit $y(x) = c \cdot x$ Lösung der Differentialgleichung.
Zur näheren Erläuterung der Bezeichnung einer „homogenen Differentialgleichung“ definieren wir:

Definition 2.7 (*Homogene Funktion*)

Eine skalare Funktion $f \in \mathrm{Abb}(\mathbb{R}^n, \mathbb{R})$ heiße **homogen vom Grade** $p \in \mathbb{N}$ genau dann, wenn

$$\forall 0 \neq \lambda \in \mathbb{R}. \forall x \in \mathbb{D}(f). f(\lambda \cdot x) = \lambda^p \cdot f(x)$$

Anmerkung

Ist $f \in \mathrm{Abb}(\mathbb{R}^2, \mathbb{R})$ homogen vom Grade 0, so gilt also $f(x,y) = f(\lambda x, \lambda y)$ für alle $\lambda \in \mathbb{R}\backslash\{0\}$. Damit wäre also $(*)$ eine homogene Differentialgleichung.

Mit der Spezifikation $\lambda = \frac{1}{x}$ und $x \neq 0$ resultiert

$$f(x,y) = f\left(1, \frac{y}{x}\right) =: g\left(\frac{y}{x}\right)$$

Man nennt (Δ) deswegen auch die **Normalform einer homogenen Differentialgleichung**.

Man kann nun einfach erkennen, dass eine Variablentransformation $z := \lambda \cdot x$ auf

$$y'(z) = \frac{d}{dx}\left(\frac{1}{\lambda} y(\lambda x)\right) = g\left(\frac{\frac{1}{\lambda} y}{z}\right)$$

führt. Dies führt dann weiter zu folgendem interessanten Lemma.

Lemma 2.3 (*Lösungen der homogenen Differentialgleichung*)

Ist $y_1(x)$ eine Lösung der homogenen Differentialgleichung (Δ), so trifft das auch auf die Funktion

$$y_2(x) := \frac{1}{\lambda} y_1(\lambda x)$$

für jedes $\lambda \in \mathbb{R}\backslash\{0\}$ zu. Man erhält somit die allgemeine Lösung

$$y_a(x,c) := \frac{1}{c} y_p(c \cdot x)$$

aus eine jeden beliebigen **partikulären** Lösung $y_p(x)$.

Beispiel 2.5: Sei

$$y' = \frac{y}{x} - \frac{x^2}{y^2} = g\left(\frac{y}{x}\right),$$

mit $g(z) = z - z^{-2}$ und $y(1) = 1$ das gegebene Anfangswertproblem, so ergibt sich mit Substitution durch $u(x) = \frac{y(x)}{x}$ die Differentialgleichung

$$u'(x) = \frac{1}{x}(g(u(x)) - u(x)) = \frac{1}{x} \cdot (u(x) - u^{-2}(x) - u(x)) = -\frac{1}{x \cdot u^2(x)},$$

welche dann durch das Anwendung der TdV durch

$$\frac{1}{3} u^3 = \int \frac{du}{u^{-2}} = -\int \frac{dx}{x} = -\ln|x| + c,$$

also mit

$$u(x) = \sqrt[3]{c - \ln|x|},$$

also

$$y(x) = x \cdot \sqrt[3]{c - \ln|x|}$$

gelöst wird. Die Lösung des Anfangswertproblems ergibt sich dann mit

$$y(1) = 1 \cdot \sqrt[3]{c - \ln|1|} = \sqrt[3]{c} \stackrel{!}{=} 1,$$

also $c = 1$. Damit gilt die Lösung des Anfangswertproblems ist beschrieben durch

$$y_p(x) = x \cdot \sqrt[3]{1 - \ln|x|} \quad \text{für } 0 < x < \sqrt[3]{e}$$

※

Beispiel 2.6: Sei

$$x^2 y' = a^2 x^2 + y^2 + xy$$

mit $x, a \neq 0$, a konstant, gegeben. Wir erhalten mit Division durch x^2 die Differentialgleichung

$$y' = a^2 + \left(\frac{y}{x}\right)^2 + \frac{y}{x},$$

die wir durch Substitution durch $u(x) := \frac{y}{x}$ und dem Zusammenhang $y' = xu' + u$ auf die Form

$$u'(x) = \frac{a^2 + u^2(x)}{x}$$

bringen. Wir lösen nun mit TdV:

$$\int \frac{du}{a^2 + u^2} = \int \frac{dx}{x} = \ln|x| + c^* \overset{(**)}{=} \ln|c \cdot x|,$$

wobei an Stelle $(**)$ der **gültige (!)** Zusammenhang $c^* = \ln|c|$ verwendet wurde. Daraus ergibt sich unmittelbar

$$\begin{array}{rcll} \frac{1}{a} \cdot \arctan\left(\frac{u}{a}\right) & = & \ln|cx| & \Big| \cdot a, \tan(\dots), \cdot a \\ \Leftrightarrow u(x) & = & a \cdot \tan\left(\ln|cx| \cdot a\right), & \end{array}$$

was dann schlussendlich zur Lösung

$$y(x) = a \cdot x \cdot \tan\left(\ln|cx| \cdot a\right)$$

mit $x, a \neq 0$ und $c \neq 0$ führt. ※

Beispiel 2.7: Zu bestimmen ist die Lösung des Anfangswertproblems

$$y' = \frac{y}{x} - \left(\frac{y}{x}\right)^2,$$

wobei $x \neq 0$ und $y(1) = y_0$. Wir substituieren $u(x) := \frac{y(x)}{x}$ und erhalten somit die Differentialgleichung

$$u' = -\frac{u^2}{x} = -\frac{1}{x} \cdot u^2 =: f(x) \cdot \underbrace{g(u)}_{=:u^2}.$$

Die Anfangsbedingung wird somit zu $u(1) = y_0$. Damit ist $g(y_0) = 0$ genau für $y_0 = 0$. Man sieht dann durch fortgeschrittene Betrachtung, dass $y(x) := 0$ eine eindeutige Lösung ist.
Für $y_0 \neq 0$ führe eine TdV durch:

$$\frac{1}{y_0} - \frac{1}{u(x)} = \int_{y_0}^{u} \frac{dt}{t^2} = -\int_1^x \frac{ds}{s} = -\ln|x|$$

Damit gilt dann, dass

$$u(x) = \frac{y_0}{\ln|x| \cdot y_0 + 1}$$

und somit

$$y(x) = \frac{y_0 \cdot x}{\ln|x| \cdot y_0 + 1}$$

eine partikuläre Lösung des Anfangswertproblems ist. Zusammengefasst gilt:

$$y(x) = \begin{cases} \frac{y_0 \cdot x}{\ln|x| \cdot y_0 + 1} & \text{falls } y_0 \neq 0 \\ 0 & \text{falls } y_0 = 0 \end{cases}$$

※

2.2.3 Typus C: Differentialgleichungen aus Linearkombinationen

Für $a, b, c \in \mathbb{R}$ fest, aber beliebig, $b \neq 0$ und $f \in \mathrm{Abb}(\mathbb{R}, \mathbb{R})$ stetig betrachten wir die Differentialgleichung

$$y' = f(ax + by + c),$$

welche mittels des Ansatzes

$$u(x) := ax + by(x) + c, \;\; u' = a + b \cdot y' = a + b \cdot f(u)$$

in eine Differentialgleichung mit getrennten Variablen überführt wird. Mit einer Lösung $u(x)$ ergibt sich dann eine Lösung

$$y(x) = \frac{u(x) - ax - c}{b}.$$

Beispiel 2.8: Sei

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

also $a = b = 1$, $c = 0$ und $f(z) = z^2$. Setze u also auf

$$u(x) := x + y(x)$$

und somit

$$u' = u^2 + 1.$$

Wir lösen die substituierte Differentialgleichung mit TdV:

$$\arctan(u(x)) = \int \frac{du}{u^2 + 1} = \int 1 dx = x + c$$

Damit ergibt sich dann eine Lösung u als $u(x) = \tan(x + c)$. Durch Rücksubstitution erhalten wir somit die allgemeine Lösung y als

$$y(x) = \tan(x - c) - x.$$

※

2.2.4 Typus D: Lineare skalare Differentialgleichungen erster Ordnung

Definition 2.8 (*Lineare skalare Differentialgleichungen erster Ordnung*)

Eine Differentialgleichung der Form

$$y' = a(x) \cdot y(x) + b(x) \tag{$**$}$$

heiße auch **lineare skalare Differentialgleichung erster Ordnung**. Sie heiße **homogen** genau dann, wenn $b(x) :\equiv 0$, andernfalls **inhomogen**.

Definition 2.9 (*Lineare Differentialgleichungen*)

Sei $I \subset \mathbb{R}$ ein Intervall und seien $a_0, \ldots, a_n, b \in \mathrm{Abb}(I, \mathbb{R})$ stetig. Dann heiße eine Differentialgleichung der Form

$$\sum_{i=0}^{n} a_i(x) y^{(i)} = b(x) \tag{2.1}$$

eine **lineare Differentialgleichung n-ter Ordnung**. Ist das sogenannte **Störglied** $b(x)$ null – sprich für alle $x \in I$ gelte, dass $b(x) = 0$ – so nennt man die Differentialgleichung **homogen**, andernfalls **inhomogen**.
Man nennt die Differentialgleichung *normiert* genau dann, wenn $a_n(x) :\equiv 1$.

Man bezeichnet mit

$$L := \sum_{i=1}^{n} a_i(x)\frac{d^i}{dx^i} + a_0(x)$$

den **Differentialoperator** und schreibt damit für eine solche Differentialgleichung auch kurz

$$Lx = b.$$

Satz 2.4 (*Überlagerungssatz*)

Sind $y_1(x), \ldots, y_k(x)$ Lösungen der Differentialgleichungen $Ly = b_i$, $i = 1, \ldots, k$, so ist für beliebige Konstanten $c_1, \ldots, c_k$ die Funktion

$$\sum_{i=1}^{k} c_i y_i(x)$$

Lösung der Differentialgleichung

$$Ly = \sum_{i=1}^{k} c_i b_i.$$

Beweis: folgt direkt aus der Linearität des Ableitungsoperators. □

Satz 2.5 (*Struktur des Lösungsraums*)

Die allgemeine Lösung einer inhomogenen linearen Differentialgleichung $Ly = b$ erhält man als Summe einer partikulären Lösung dieser Gleichung und der allgemeinen Lösung y_H der zugehörigen homogenen Differentialgleichung $Ly = 0$.

Beweis:

(1) $\mathbb{Z}$: $y_P + y_H \in \mathbb{L}(Ly = b)$
Dazu betrachten wir $L(y_P + y_H) = Ly_P + Ly_H = b + 0 = b$ ✓

(2) $\mathbb{Z}$: Ist y_A eine beliebige Lösung von $Ly = b$, so ist $y_A - y_P \in y_H$.
Wir betrachten $L(y_A - y_P) = Ly_A - Ly_P = b - b = 0$, woraus direkt folgt, dass $y_A - y_P \in y_H$ ✓

□

Wir erhalten damit folgendes interessantes Korollar für lineare Differentialgleichungen erster Ordnung:

Korrolar 2.5 (*Struktur des Lösungsraums*)

(a) Die Lösungsmenge L_{hom} einer linearen homogenen Differentialgleichung erster Ordnung ist ein **Vektorraum**.

(b) Die Lösungsmenge L_{inhom} einer linearen **in**homogenen Differentialgleichung erster Ordnung ist ein **affiner Raum**, es gilt: Sei y_P eine beliebige partikuläre Lösung der inhomogenen Differentialgleichung und L_{hom} der Lösungsraum der zugehörigen homogenen Differentialgleichung so ist

$$L_{inhom} = \{y_P\} + L_{hom}.$$

Satz/Korollar 2.5 gibt uns die Information, dass die Lösungskonstruktion für eine inhomogenen lineare Differentialgleichung erster Ordnung in die zwei Teilaufgaben (H) und (P) zerfällt:

(H) Bestimme den Unterraum L_{hom}, heißt die Lösungsgesamtheit der homogenen Differentialgleichung $y' + a(x)y = 0$.

(P) Bestimme **eine** Partikulärlösung y_p der inhomogenen Differentialgleichung $y' + a(x)y = b(x)$.

Teilaufgabe (H) – Lösen der homogenen Differentialgleichung

Satz 2.6 (L_{hom})

Für ein gegebenes $a \in \mathcal{C}(I)$ hat die homogene Differentialgleichung $y' + a(x)y = 0$ genau die Lösungen

$$y_h(x) := c \cdot \exp\left(\int a(x)dx\right) \qquad (***)$$

resp. bei einem Anfangswertproblem

$$y_h(x) := c \cdot \exp\left(\int_{x_0}^{x} a(t)dt\right).$$

Damit folgt, dass für den Lösungsvektorraum gilt:

$$L_{hom} = \operatorname{span}\left\{\exp\left(\int a(x)dx\right)\right\}$$

Beweis: (*Variante 1 — Rechnung mit der TdV*)
Es ist die Lösung der Differentialgleichung

$$y_h' = a(x)y_h \tag{2.2}$$

gesucht, was eine Differentialgleichung mit getrennten Veränderlichen ist und somit mit dem Verfahren der Trennung der Veränderlichen gelöst werden kann. Wir schreiben also Gleichung (2.9) um zu

$$a(x) = \frac{y_h'}{y_h}, \tag{2.3}$$

was dann unter Anwendung des Verfahrens äquivalent ist zu der Darstellung

$$\int \frac{dy_h}{y_h} = \int a(x)dx \tag{2.4}$$

Wir können dies umstellen nach y_h, es ergibt sich somit:

$$(2.4) \Rightarrow \quad \ln|y_h| = \int a(x)dx + \widetilde{c} \tag{2.5}$$

$$\Leftrightarrow \quad |y_h| = \exp\left(\int a(x)dx + \widetilde{c}\right) \quad \Leftrightarrow \quad |y_h| = \exp(\widetilde{c}) \cdot \exp\left(\int a(x)dx\right) \tag{2.6}$$

$$\Rightarrow \quad y_h = \underbrace{\pm e^{\widetilde{c}}}_{=:c} \cdot \exp\left(\int a(x)dx\right) \tag{2.7}$$

$$\Leftrightarrow \quad y_h = c \cdot \exp\left(\int a(x)dx\right) \tag{2.8}$$

□

Beweis: (*Variante 2 — „cleveres Hinsehen“*)
Es ist immernoch die Lösung der Differentialgleichung

$$y_h' = a(x)y_h \tag{2.9}$$

gesucht. Man kann leicht erkennen, dass

$$u(x) := \exp\left(\int a(x)dx\right) \tag{2.10}$$

aufgrund der Eigenschaften der exp-Funktion *nullstellenfrei* ist, damit gilt äquivalent zur Gleichung $y' + a(x)y = 0$:

$$\forall x \in I. \quad 0 = \exp\left(\int a(x)dx\right)(y' + a(x)y) = \frac{d}{dx}\left(\exp\left(\int a(x)dx\right) \cdot y\right) \tag{2.11}$$

Aus (2.11) folgt dann mit einem Satz aus vergangenem Semester, dass

$$\exp\left(\int a(x)dx\right) \cdot y = c = \text{konst.},$$

womit der Satz bewiesen ist. □

Anmerkung

In Satz 2.6 bezeichnet

$$\int a(x)dx$$

eine konkrete, spezifische und *feste* Stammfunktion von a.

Aus Satz 2.6 folgt dann auch unmittelbar:

$$\dim(L_{hom}) = 1$$

Teilaufgabe (P) – Herausfinden einer partikulären Lösung Teilaufgabe (P) ist bei Kenntnis von y_h stets konstruktiv lösbar. Man bedient sich dazu des auf Lagrange zurückgehenden Verfahrens der „**Variation der Konstanten** *(VdK)*“.
Wir fassen dazu das c in Gleichung $(***)$ als **differenzierbare Funktion, abhängig von der Veränderlichen x** auf und erhalten somit für y_p

$$y_p(x) = c(x) \cdot \exp\left(\int a(x)dx\right) \tag{2.12}$$

mit der Ableitung y_p'

$$y_p'(x) = c'(x) \cdot \exp\left(\int a(x)dx\right) + c(x) \cdot a(x) \cdot \exp\left(\int a(x)dx\right). \tag{2.13}$$

Diese Gleichungen werden nun in die Differentialgleichung $(**)$ eingesetzt und wir erhalten die finale Gleichung

$$c'(x) \cdot \exp\left(\int a(x)dx\right) = b(x), \tag{2.14}$$

aus der dann schlussendlich folgt, dass

$$c(x) = \int_{x_0}^{x} \frac{b(t)}{y_h(t)}dt. \tag{2.15}$$

Wir setzen nun (2.15) in den allgemeinen Ansatz (2.12) ein und erhalten somit für die partikuläre Lösung

$$y_p(x) = \int_{x_0}^{x} b(t) \cdot \frac{y_h(x)}{y_h(t)} dt. \tag{2.16}$$

Aus Satz 2.6 kennen wir y_h es gilt also:

$$\frac{y_h(x)}{y_h(t)} = \frac{\exp\left(\int_{x_0}^{x} a(t)dt\right)}{\exp\left(\int_{x_0}^{t} a(s)ds\right)} = \exp\left(\int_{t}^{x} a(s)ds\right). \tag{2.17}$$

Mit (2.17) ergibt sich somit insgesamt für y_p die Lösung

$$y_p(x) = \int_{x_0}^{x} b(t) \cdot \left(\exp\left(\int_{t}^{x} a(s)ds\right)\right) dt. \tag{2.18}$$

Wir fassen unsere Ergebnisse also in einem abschließendem Satz zusammen:

Satz 2.7 (*Lösung der allg. lin. inhom. gew. DGl. erster Ordnung*)

Gegeben seien ein Intervall $I \subset \mathbb{R}$ und Funktionen $a, b \in \mathrm{Abb}(\mathbb{R}, \mathbb{K})$, mit $a, b \in \mathcal{C}(I)$. Sei

$$A(x) := \int a(x)dx$$

eine **feste** Stammfunktion von a. Dann hat die Differentialgleichung

$$L_1 y := y' + a(x)y = b(x) \qquad , x \in I$$

die allgemeine Lösung

$$y_p(x) \in \{\lambda y_h(x) + y_p(x) : \lambda \in \mathbb{R}\}$$
$$= \left\{ \lambda \cdot \exp\left(\int_{x_0}^{x} a(t)dt\right) + \int_{x_0}^{x} b(t) \cdot \left(\exp\left(\int_{t}^{x} a(s)ds\right)\right) dt \;\middle|\; x \in I, x_0 \in I \text{ bel., aber fest}, \lambda \in \mathbb{R} \right\}.$$

Das Anfangswertproblem

$$L_1 y = b, x \in I, y(x_0) = y_0, x_0 \in I$$

ist stets eindeutig lösbar mit der Lösung

$$y(x) = y_0 \cdot \exp\left(\int_{x_0}^{x} a(t)dt\right) + \int_{x_0}^{x} b(t) \cdot \left(\exp\left(\int_{t}^{x} a(s)ds\right)\right) dt$$

Beweis: siehe Herleitung oben. □

Beispiel 2.9: Zu bestimmen sei die Lösung der inhomogenen linearen Differentialgleichung

$$y' = 2xy(x) + x \cdot \exp\left(x^2\right).$$

Wir teilen die Aufgabe wieder in seine Teilaufgaben (H) und (P) auf:

(H) Zu lösen sei nun

$$y' = 2xy,$$

wir verwenden dazu wieder TdV:

$$\begin{aligned} & y' && = && 2xy \\ \Rightarrow \quad & \frac{y'}{y} && = && 2x \\ \Rightarrow \quad & y_h(x) && = && c \cdot \exp\left(x^2\right), c \in \mathbb{R}. \end{aligned}$$

(P) Wir bestimmen eine partikuläre Lösung mittels der Variation der Konstanten (VdK), also dem Ansatz

$$y_p(x) := c(x) \cdot y_h(x).$$

$$\begin{aligned} \Rightarrow \quad c'(x) \cdot \exp\left(x^2\right) &= x \cdot \exp\left(x^2\right) \\ \Rightarrow \quad c'(x) &= \frac{x \cdot \cancel{\exp\left(x^2\right)}}{\cancel{\exp\left(x^2\right)}} \\ \Rightarrow \quad c(x) &= \int x dx = \frac{1}{2}x^2 \\ \Rightarrow \quad y_p(x) &= \frac{1}{2}x^2 \cdot \exp\left(x^2\right) \end{aligned}$$

Damit kann die allgemeine Lösung der inhomogenen Differentialgleichung mit

$$y(x) = \frac{1}{2}x^2 \cdot \exp\left(x^2\right) + c \cdot \exp\left(x^2\right),$$

wobei $c \in \mathbb{R}$, aufgestellt werden. ※

2.2.5 Typus E: Die Bernoulli-Differentialgleichung

Seien $p, q \in \mathrm{Abb}(\mathbb{R}, \mathbb{R})$ stetige Funktionen, $\lambda \in \mathbb{R} \backslash \{0, 1\}$, so heiße die Differentialgleichung

$$y' + p(x)y = q(x)y^\lambda$$

auch **Bernoulli'sche Differentialgleichung**. Für ein ganzzahliges λ **können** Lösungen $y < 0$ sinnvoll sein.
Wir verwenden hier stets den Ansatz

$$z(x) := y^{1-\lambda}(x) \quad \text{mit} \quad z'(x) = (1-\lambda)y^{-\lambda}(x) \cdot y'(x),$$

um die Differentialgleichung mit

$$z' + (1-\lambda)p(x)z = (1-\lambda)q(x)$$

in eine lineare Differentialgleichung umzuwandeln, welche dann mit den Möglichkeiten aus Kapitel 2.2.4 zu lösen ist.

2.2.6 Typus F: Die Euler'sche Differentialgleichung

Seien $a_k(x) := a_k x^k \in \mathrm{Abb}(\mathbb{R},\mathbb{R})$ $n+1$ viele stetige Funktionen, so heiße die Differentialgleichung

$$\sum_{k=0}^{n} a_k(x) y^{(k)}(x) = 0$$

auch **Euler'sche Differentialgleichung**.
Wir substituieren hier mit $u(x) = y(e^x)$ und bringen die Differentialgleichung damit auf die Form

$$\sum_{k=0}^{n} \widetilde{a_k} u^{(k)} = 0.$$

Dies ist dann eine lineare Differentialgleichung n-ter Ordnung mit *konstanten* Koeffizienten, ihre Lösungsweise werden wir in Kapitel 2.5 genauer behandeln.

2.3 Existenz und Eindeutigkeit von Lösungen von Anfangswertproblemen

Anfangswertprobleme entstehen im Allgemeinen aus Anwendungen heraus mittels mathematischer Modellierung. Häufig erwartet man, dass das Modell dann **genau** eine Lösung hat. Dies zu überprüfen mag in einfachen Fälle – *wie in Kapitel 2.2* – noch relativ einfach möglich sein, bei recht trivial erscheinenden Differentialgleichungen – wie beispielsweise $y' = x^2 + y^2$ – ist aber die Lösungsgesamtheit nicht mehr in geschlossener Form analytisch bestimmbar. Deshalb ist es von Interesse allgemeine Existenz- und Eindeutigkeitsaussagen bereitzustellen, die wenigstens eine theoretische Garantie der Existenz/Eindeutigkeit von Lösungen geben.

2.3.1 Existenz von Lösungen

Wir betrachten in diesem Abschnitt das Anfangswertproblem

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

mit einer lediglich *stetigen* Funktion f. Unser Ziel besteht nun darin, zumindest die Existenz einer Lösung dieses Anfangswertproblems nachzuweisen. Wir betrachten also folgende Existenzaussage:

Satz 2.8 (*Existenzsatz von Peano*)

Sei $x_0 \in \mathbb{R}$, $y_0 \in \mathbb{R}^n$, $X > x_0$, $R > 0$, $f \in \mathrm{Abb}(G,\mathbb{R}^n)$, wobei

$$G \supseteq [x_0, X] \times K_R(y_0)$$

und f stetig ist. So besitzt das Anfangswertproblem

$$y' = f(x,y), \quad y(x_0) = y_0 \tag{2.19}$$

mindestens eine Lösung auf dem Intervall $[x_0, x_0+\varepsilon]$, wobei wir

$$\varepsilon := \min\left\{X - x_0, \frac{R}{M}\right\} \quad \text{mit} \quad M = \max\left\{\|f(x,y)\| \;\middle|\; (x,y) \in G\right\}$$

gesetzt haben und für den trivialen Fall $M = 0$ die Konvention „$R/M := \infty$" verwenden, und diese Lösung ist stetig differenzierbar.

Wir wollen nun diesen Existenzsatz beweisen, dazu braucht es allerdings unter anderem das „Euler-Cauchy'sches Polygonzugverfahren“ (Verfahren 2.3), welches wir in Kapitel 2.6.1 erst kennenlernen werden. Wir verweisen an dieser Stelle auf den späteren Zeitpunkt, verwenden allerdings die Grundidee des Verfahrens um den Satz 2.8 jetzt an dieser Stelle zu beweisen.
Beweis: Für $M = 0$ ist $f(x,y) \equiv 0$ und damit $y(x) \equiv y_0$ eine Lösung des Anfangswertproblems (2.19). Im Folgenden sei damit also Œ $M > 0$. Wir gliedern den Beweis nun in fünf Schritte.

Schritt 1: Wir konstruieren auf dem Intervall $[x_0, x_0+\varepsilon]$ zunächst eine Folge von Eulerpolygonen $p^{(k)}$. Dazu sei bemerkt, dass die stetige Funktion f auf dem kompakten Zylinder G natürlich gleichmäßig stetig ist. Daher gibt es auch zu jedem $k \in \mathbb{N}$ ein $\delta = \delta_k > 0$, so dass für alle $(x,y), (\widetilde{x},\widetilde{y}) \in G$ die folgende Implikation gilt:

$$|x-\widetilde{x}| \leqslant \delta_k \wedge \|y-\widetilde{y}\| \leqslant \delta_k \Longrightarrow \|f(x,y)-f(\widetilde{x},\widetilde{y})\| \leqslant \frac{1}{k}. \tag{2.20}$$

Da $\left\{\frac{\varepsilon}{m}\right\}$ eine Nullfolge ist, gibt es zu jedem $k \in \mathbb{N}$ ein $m = m_k \in \mathbb{N}$, so dass für $h := h_k := \frac{\varepsilon}{m_k}$ die Ungleichung

$$0 < h_k < \min\left\{\delta_k, \frac{\delta_k}{M}\right\} \tag{2.21}$$

gilt. Für jedes $k \in \mathbb{N}$ und die gerade definierte zugehörige Schrittweite $h = h_k$ definieren wir dann induktiv die $m+1$ „Eckpunkte“ (x_i, y_i) mit $0 \leqslant i \in \mathbb{N} \leqslant +m$ des zu konstruierenden Eulerpolygons durch $x_i := x_0 + i \cdot h$ mit $0 \leqslant i \in \mathbb{N} \leqslant +m$ und

$$y_i := y_{i-1} + hf(x_{i-1}, y_{i-1}) \qquad \text{für } i = 1, \ldots, m.$$

Auf diese Weise erhalten wir für jedes $k \in \mathbb{N}$ ein Eulerpolygon $p^{(k)}(x)$, welches sich stückweise auf den Teilintervallen $[x_i, x_{i+1}]$ mit $0 \leqslant i \in \mathbb{N} < +m$ wie folgt definiert:

$$p^{(k)}(x) := y_i + (x-x_i)f(x_i, y_i) \qquad \text{für } i = 0, 1, \ldots, m-1. \tag{2.22}$$

Dann gilt insbesondere auch

$$p^{(k)}(x_i) = y_i \quad \text{für } i = 0, \ldots, m. \tag{2.23}$$

Dabei ist die hier beschriebene Konstruktion möglich, weil die Punkte (x_i, y_i) für alle $i = 0, \ldots, m$ in G und somit im Definitionsbereich von f liegen; für $i = 1, \ldots, m$ gilt nämlich

$$\begin{aligned}
\|y_i - y_0\| &= \left\| \sum_{j=0}^{i-1} y_{j+1} - y_j \right\| \\
&\overset{(2.23)}{=} \left\| \sum_{j=0}^{i-1} p^{(k)}(x_{j+1}) - y_j \right\| \\
&\overset{(2.22)}{=} \left\| \sum_{j=0}^{i-1} (x_{j+1} - x_j) f(x_j, y_j) \right\| \\
&\leqslant M \cdot \sum_{j=0}^{i-1} (x_{j+1} - x_j) \\
&\leqslant M \cdot \varepsilon \\
&\leqslant R.
\end{aligned}$$

Schritt 2: Wir wollen nun zeigen, dass eine gleichmäßig konvergente Teilfolge der betrachteten Eulerpolygone $p^{(k)}$ existiert. Dazu wenden wir folgenden Satz an:

Satz 2.9 (*Satz von Arzelà-Ascoli*)

Gegeben seien ein kompaktes Intervall $I := [a,b] \subseteq \mathbb{R}^n$ sowie eine Folge $\{f_k\}$ von stetigen Funktionen $f_k \in \mathrm{Abb}(I,\mathbb{R}^n)$ mit den beiden folgenden Eigenschaften:

(a) Die Folge $\{f_k\}$ ist (punktweise) **gleichmäßig beschränkt**, das heißt zu jedem $x \in I$ existiert ein $S = S(x) > 0$ mit der Eigenschaft

$$\forall k \in \mathbb{N}. \|f_k(x)\| \leqslant S.$$

(b) Die Folge $\{f_k\}$ ist **gleichgradig stetig**, das heißt zu jedem $\varepsilon > 0$ existiert ein $\delta = \delta(\varepsilon) > 0$ derart, dass die folgende Implikation für alle $x_1, x_2 \in I$ gilt:

$$\forall k \in \mathbb{N}. |x_1 - x_2| < \delta \Longrightarrow \|f_k(x_1) - f_k(x_2)\| < \varepsilon$$

Dann besitzt die Folge $\{f_k\}$ eine auf $I = [a,b]$ *gleichmäßig konvergente* Teilfolge.

i Der Begriff der **gleichgradigen Stetigkeit** ist im Prinzip eine Verallgemeinerung des Begriffs der gleichmäßigen Stetigkeit einer Funktion. Denn läge statt einer Funktionenfolge f_k nur eine *einzelne* Funktion f vor, welche der entsprechenden Bedingung

$$|x_1 - x_2| < \delta \Longrightarrow \|f(x_1) - f(x_2)\| < \varepsilon$$

für alle $x_1, x_2 \in I$ genügen soll, so läge eben genau die Definition der gleichmäßigen Stetigkeit vor. Gleichgradige Stetigkeit besagt daher, dass **alle** Elemente f_k einer Funktionenfolge $\{f_k\}$ gleichmäßig stetig sind, und dass sie dies *gleichgradig* sind, in dem Sinne, als dass die Wahl von δ eben nicht vom speziellen Element f_k abhängt.

Beweis: Wir merken zuerst an, dass die Menge $\mathbb{Q} \cap [a,b]$ der in dem Intervall $[a,b]$ liegenden rationalen Zahlen **abzählbar** ist, etwa

$$\mathbb{Q} \cap [a,b] = \left\{ r_i \;\middle|\; i = 1,2,\ldots \right\}.$$

Wir zerlegen den Beweis wieder in zwei Schritte:

Schritt 1: Im ersten Beweisschritt benutzen wir ein Diagonalfolgenargument, um eine Teilfolge $\{g_k\}$ von $\{f_k\}$ zu finden, die in allen Punkten r_k *punktweise konvergiert*. Nach Voraussetzung (a) ist zunächst die Folge $\{f_k(r_1)\}$ beschränkt im $\mathbb{R}^n$. Nach dem Satz von Bolzano-Weierstraß besitzt sie daher eine konvergente Teilfolge. Es gibt also eine Teilfolge $\{f_k^{(1)}\}$ von $\{f_k\}$ und einen Grenzwert $f_1^* \in \mathbb{R}^n$ mit

$$f_k^{(1)}(r_1) \longrightarrow f_1^* \qquad \text{für } k \to \infty.$$

Wiederum nach Voraussetzung (a) ist aber auch die Folge $\{f_k^{(1)}(r_2)\}$ beschränkt im $\mathbb{R}^n$. Nach Bolzano-Weierstraß existiert daher eine weitere Teilfolge $f_k^{(2)}$ von $f_k^{(1)}$ sowie ein Grenzwert f_2^* mit

$$f_k^{(2)}(r_2) \longrightarrow f_2^* \qquad \text{für } k \to \infty.$$

So fortfahrend, erhält man induktiv für jedes $\ell \in \mathbb{N}$ Teilfolgen $\left\{f_k^{(\ell)}\right\}$ von $\left\{f_k\right\}$ und Grenzwerte $f_\ell^* \in \mathbb{R}^n$ mit den folgenden Eigenschaften:

$$\begin{array}{rcl} f_1^{(1)}(r_1), f_2^{(1)}(r_1), f_3^{(1)}(r_1), \ldots & \to & f_1^*, \\ f_1^{(2)}(r_2), f_2^{(2)}(r_2), f_3^{(2)}(r_2), \ldots & \to & f_2^*, \\ f_1^{(3)}(r_3), f_2^{(3)}(r_3), f_3^{(3)}(r_3), \ldots & \to & f_3^*, \\ \vdots & & \vdots \end{array} \tag{2.24}$$

Dabei ist $\left\{f_k^{(\ell+1)}\right\}$ eine Teilfolge von $\left\{f_k^{(\ell)}\right\}$ für jedes $\ell \in \mathbb{N}$. Für beliebiges $\ell \in \mathbb{N}$ gilt daher

$$\left\{f_k^{(\ell)}(r_j)\right\} \to f_j^* \quad \text{für alle } j = 1, \ldots, \ell.$$

Betrachten wir nun die Diagonalfolge

$$g_k := f_k^{(k)} \quad \text{für } k = 1, 2, \ldots,$$

so ist

$$g_k(r_j) \to f_j^* \quad \text{für alle } j = 1, 2, \ldots, \tag{2.25}$$

denn die Folge $\{g_1, g_2, \ldots\}$ ist, zumindest von ihrem j-ten Glied an, eine Teilfolge der j-ten Zeile von (2.24) und somit sicherlich in r_j konvergent. Damit ist der erste Beweisschritt abgeschlossen.

Schritt 2: Wir zeigen nun, dass die im ersten Schritt konstruierte Folge $\{g_k\}$ auf dem Intervall $[a, b]$ gleichmäßig konvergiert.
Sei dazu $\varepsilon > 0$ beliebig gegeben. Wähle ein zugehöriges $\delta > 0$ gemäß Voraussetzung (b). Offenbar gibt es dann endlich viele Punkte $x_1, \ldots, x_s \in \mathbb{Q} \cap [a, b]$, so dass zu jedem $x \in [a, b]$ ein x_j mit $j \in \{1, ..., s\}$ existiert mit

$$|x - x_j| < \delta \tag{2.26}$$

Wegen (2.25) sind die Folgen $\{g_k(x_j)\}$ dann für jedes $j = 1, \ldots, s$ konvergent, insbesondere handelt es sich daher um Cauchyfolgen. Da es sich hier um endlich viele Punkte x_j handelt, existiert daher eine zwar von ε, nicht jedoch von dem speziellen j abhängige Zahl $N \in \mathbb{N}$ mit

$$\left\|g_k(x_j) - g_m(x_j)\right\| < \varepsilon \quad \text{für alle } k, m \geqslant N \text{ und alle } j = 1, \ldots, s.$$

Wegen der Voraussetzung (b) ergibt sich unter Berücksichtigung von (2.26) dann für alle $x \in [a, b]$ (wobei x_j einer der am dichtesten an x liegenden Punkte sei):

$$\begin{array}{rcl} \|g_k(x) - g_m(x)\| & \leqslant & \left\|g_k(x) - g_k(x_j)\right\| + \left\|g_k(x_j) - g_m(x_j)\right\| + \left\|g_m(x_j) - g_m(x)\right\| \\ & < & \varepsilon + \varepsilon + \varepsilon \\ & = & 3\varepsilon \end{array}$$

für alle $k, m \geqslant N$. Dies impliziert offenbar die gleichmäßige Konvergenz der Folge $\{g_k\}$ auf $[a, b]$.

□

Wir betrachten jetzt also die spezielle Funktionenfolge der Eulerpolygone $p^{(k)}$. Wir stellen zuerst fest, dass diese Folge **gleichmäßig beschränkt** ist, da alle Polygone im kompakten Zylinder G verlaufen. Zum Nachweis der gleichgradigen Stetigkeit zeigen wir, dass die Ungleichung

$$\left\|p^{(k)}(x_1) - p^{(k)}(x_2)\right\| \leqslant M|x_1 - x_2| \quad \text{für alle } k \in \mathbb{N} \tag{2.27}$$

und alle $x_1, x_2 \in [x_0, x_0 + \varepsilon]$ erfüllt ist. Daraus kann man dann die gleichgradige Stetigkeit schließen. Seien also $x_1, x_2 \in [x_0, x_0 + \varepsilon]$ beliebig, aber fest, gegeben. Liegen beide Punkte x_1, x_2 im gleichen Intervall $[x_i, x_{i+1}]$, so gilt

$$\left\| p^{(k)}(x_1) - p^{(k)}(x_2) \right\| = \|(x_1 - x_i) f(x_i, y_i) - (x_2 - x_i) f(x_i, y_i)\| \leqslant M |x_1 - x_2|.$$

Andernfalls ist $t_1 \in [t_i, t_{i+1}]$ und $t_2 \in [t_j, t_{j+1}]$ für gewisse Indizes i, j, wobei Œ angenommen werden darf, dass $i < j$. Unter Verwendung des bereits erledigten Falles folgt dann

$$\begin{aligned}
& \left\| p^{(k)}(x_1) - p^{(k)}(x_2) \right\| \\
\leqslant\ & \left\| p^{(k)}(x_1) - p^{(k)}(x_{i+1}) \right\| + \sum_{\ell=i+1}^{j-1} \left\| p^{(k)}(x_\ell) - p^{(k)}(x_{\ell+1}) \right\| + \left\| p^{(k)}(x_j) - p^{(k)}(x_2) \right\| \\
\leqslant\ & M \cdot \left((x_{i+1} - x_1) + \sum_{\ell=i+1}^{j-1} (x_{\ell+1} - x_\ell) + (x_2 - x_j) \right) \\
=\ & M(x_2 - x_1) \\
=\ & M |x_2 - x_1|.
\end{aligned}$$

Wir zeigen mit diesem Ergebnis nun noch, dass die Funktionenfolge auch gleichgradig stetig ist. Sei dazu $\alpha > 0$ beliebig, aber fest, gegeben. Setze dann

$$\delta := \delta(\alpha) := \frac{\alpha}{M}.$$

Dann gelte für alle $p^{(k)}$ und für alle $x_1, x_2 \in [a, b]$ mit $|x_1 - x_2| < \delta$ die Abschätzung

$$\left\| p^{(k)}(x_1) - p^{(k)}(x_2) \right\| \leqslant M |x_2 - x_1| \leqslant M\delta = \alpha,$$

womit die gleichgradige Stetigkeit der Funktionenfolge $\left\{p^{(k)}\right\}$ bereits bewiesen ist. Damit wenden wir Satz 2.9 an, womit folgt, dass eine gleichmäßig konvergente Teilfolge der betrachteten Eulerpolygone $p^{(k)}$ existiert. Wir wollen nun im weiteren Verlauf nur noch diese Teilfolge betrachten, welche wir der Einfachheit halber wieder mit $\left\{p^{(k)}\right\}$ bezeichnen werden.

Schritt 3: Wir zeigen hier, dass das Eulerpolygon $p^{(k)}$ der Abschätzung

$$\left\| p^{(k)}(x) - f(x, p^{(k)}(x)) \right\| \leqslant \frac{1}{k} \quad \text{für alle } x \in [x_0, x_0 + \varepsilon] \setminus \{x_i \mid i = 0, \ldots, m\}$$

genügt und somit zumindest näherungsweise als eine Lösung der Differentialgleichung angesehen werden kann. In der Tat gilt auf jedem der offenen Teilintervalle (x_i, x_{i+1}) nämlich

$$\left\| p^{(k)}(x) - y_i \right\| \overset{(2.22)}{=} |x - x_i| \cdot \|f(x_i, y_i)\| \leqslant h_k M \overset{(2.21)}{\leqslant} \delta_k$$

sowie

$$|x - x_i| < h_k \overset{(2.21)}{\leqslant} \delta_k.$$

Hieraus folgt dann

$$\left\| p^{(k)}(x) - f(x, p^{(k)}(x)) \right\| \overset{(2.22)}{=} \left\| f(x_i, y_i) - f(x, p^{(k)}(x)) \right\| \overset{(2.20)}{\leqslant} \frac{1}{k}, \tag{2.28}$$

womit diese Aussage bewiesen wäre.

Schritt 4: Wir verifizieren in diesem Teil des Beweises die Gültigkeit von

$$\left\| p^{(k)}(x) - y_0 - \int_{x_0}^{x} f\left(t, p^{(k)}(t)\right) dt \right\| \leqslant \frac{1}{k} |x - x_0| \tag{2.29}$$

für alle $x \in [x_0, x_0 + \varepsilon]$ und alle $k \in \mathbb{N}$. Sei dazu $x \in [x_i, x_{i+1}]$ für ein $i \in \{0, 1, \ldots, m-1\}$ beliebig, aber fest, gegeben. Dann folgt

$$\begin{aligned}
& \left\| p^{(k)}(x) - y_0 - \int_{x_0}^{x} f\left(t, p^{(k)}(t)\right) dt \right\| \\
= & \left\| p^{(k)}(x) - y_i - \int_{x_i}^{x} f\left(t, p^{(k)}(t)\right) dt + \sum_{j=0}^{i-1} \left(y_{j+1} - y_j - \int_{x_j}^{x_{j+1}} f\left(t, p^{(k)}(t)\right) dt \right) \right\| \\
\overset{(2.23)}{=} & \left\| \int_{x_i}^{x} \left(p^{(k)}(t) - f\left(t, p^{(k)}(t)\right) \right) dt + \sum_{j=0}^{i-1} \int_{x_j}^{x_{j+1}} \left(p^{(k)}(t) - f\left(t, p^{(k)}(t)\right) \right) dt \right\| \\
\overset{(2.28)}{\leqslant} & \int_{x_i}^{x} \frac{1}{k} dt + \sum_{j=0}^{i-1} \int_{x_j}^{x_{j+1}} \frac{1}{k} dt \\
= & \frac{1}{k} |x - x_0|,
\end{aligned}$$

so dass (2.29) tatsächlich erfüllt ist.

Schritt 5: Bevor wir jetzt tatsächlich die Behauptung zeigen benötigen wir noch einen weiteren Hilfssatz.

Satz 2.10 (*Äquivalenz von Anfangswertproblem und Integralgleichung*)

Sei $I \subseteq \mathbb{R}$ ein beliebiges Intervall und $x_0 \in I$. Betrachte dann das Anfangswertproblem

$$y'(x) = f(x, y(x)) \forall x \in I, y(x_0) = y_0 \tag{2.30}$$

für eine gegebene stetige Funktion $f \in \mathrm{Abb}(I \times \mathbb{R}^n, \mathbb{R}^n)$ und einen Anfangswert $y_0 \in \mathbb{R}^n$. Dann sind die beiden folgenden Aussagen äquivalent:

(a) $y \in \mathrm{Abb}(I, \mathbb{R}^n)$ ist stetig differenzierbar und löst das Anfangswertproblem (2.30)

(b) $y \in \mathrm{Abb}(I, \mathbb{R}^n)$ ist stetig und löst die **Integralgleichung**

$$y(x) = y_0 + \int_{x_0}^{x} f(s, y(s)) ds \tag{2.31}$$

Beweis: Sei $y \in \mathrm{Abb}(I, \mathbb{R}^n)$ zunächst eine Lösung des Anfangswertproblems (2.30). Mit dem Hauptsatz der Differential- und Integralrechnung erhalten wir durch Integration der Differentialgleichung daher

$$y(x) = c + \int_{x_0}^{x} y'(s) ds = c + \int_{x_0}^{x} f(s, y(s)) ds,$$

und aus $y(x_0) = y_0$ folgt für die noch freie Integrationskonstante $c \in \mathbb{R}^n$ unmittelbar $c = y_0$. Also ist y eine (stetige) Lösung der Integralgleichung (2.31).

Sei umgekehrt $y \in \mathrm{Abb}(I, \mathbb{R}^n)$ jetzt eine mindestens stetige Lösung der Integralgleichung (2.31). Dann ist der Integrand insbesondere eine stetige Funktion, und wir erhalten aus dem Hauptsatz der Differential- und Integralrechnung durch Ableiten aus der Identität (2.31) sofort

$$y'(x) = f(x, y(x)) \forall x \in I.$$

Ebenso ergibt sich aus (2.31) unmittelbar $y(x_0) = y_0$, so dass y in der Tat eine stetig differenzierbare Lösung des Anfangswertproblems (2.30) ist. □

Damit zurück zum Beweis der eigentlichen Behauptung. Sei dazu $p^*(x)$ die Grenzfunktion der auf $[x_0, x_0 + \varepsilon]$ gleichmäßig konvergenten (Teil-)Folge der Eulerpolygone $p^{(k)}(x)$. Wegen der gleichmäßigen Konvergenz dieser Folge und (2.20) konvergiert die Folge $\left\{f(x, p^{(k)}(x))\right\}$ für alle $x \in [x_0, x_0 + \varepsilon]$ gleichmäßig gegen $f(x, p^*(x))$. Der Grenzübergang $k \to \infty$ in der Beziehung (2.29) liefert dann die Identität

$$\left\| p^*(x) - y_0 - \int_{x_0}^{x} f(t, p^*(t))\, dt \right\| = 0 \quad \text{für alle } x \in [x_0, x_0 + \varepsilon].$$

Aus Satz 2.10 folgt daher die Behauptung, denn p^* ist als gleichmäßiger Grenzwert einer konvergenten (Teil-)Folge stetiger Funktionen selbst wieder stetig.

□

! Die Stetigkeitsforderung an f kann nicht ersatzlos gestrichen werden! Allerdings ist das Fehlen von Stetigkeit **kein** Ausschlusskriterium für eine Lösung!

2.3.2 Eindeutigkeit von Lösungen

Eine Frage, die sich jetzt noch stellt, ist, ob die Anforderungen an f, die im Satz 2.8 getroffen wurden, nicht auch ausreichen, um Eindeutigkeit zu fordern, also ob es Anfangswertprobleme gibt, die mehrere Lösungen haben, obwohl die Existenz nach Satz 2.8 gesichert ist. Wir betrachten dazu **Beispiel 2.10:**

$$y' = \sqrt{|y|} \quad y(0) = 0$$

Es ist leicht zu sehen, dass die rechte Seite $f(y) := |y|^{1/2}$ offensichtlich auf $\mathbb{R} \times \mathbb{R}$ stetig ist, damit ist Satz 2.8 anwendbar. Suchen wir nun nach Lösungen, so erhalten wir mit einer Trennung der Veränderlichen:

$$\pm 2(\pm y)^{1/2} = \int \frac{dy}{\sqrt{|y|}} = \int 1\, dx = x + c \quad \text{für } y \gtrless 0 \text{ und } c \in \mathbb{R},$$

was wir weiter vereinfachen können zu

$$y(x) = \pm \frac{1}{4}(x + c)^2 \quad \text{für } x + c \gtrless 0.$$

Damit stellen wir fest, dass es neben der offensichtlichen Lösung $y_t(x) :\equiv 0$ auch beispielsweise das obige y oder ein y_a mit

$$y_a(x) = \left\{ \begin{array}{ll} 0 & \text{für } x \leqslant c \\ \frac{1}{4}(x + c)^2 & \text{für } x > c \end{array} \right\}, \text{ wobei } c \geqslant 0 \text{ beliebig, aber fest, sei}$$

Lösungen des Anfangswertproblems. Es gibt also keine **eindeutige** Lösung, es kommt stattdessen zu einer **Bifurkation**. ※

Wir schließen aus diesem Beispiel, dass für eine **Eindeutigkeit** von Lösungen die zugehörige Funktion eine „stärkere“ Bedingung als *Stetigkeit* erfüllen muss. Wir wollen nun eine solche Bedingung angeben, welche dann auch zentrale Voraussetzung im Eindeutigkeitssatz (Satz 2.12) wird, und definieren deswegen den im zweiten Semester eingeführten Begriff der Stetigkeit weiter.

Definition 2.10 (*Lipschitz-Stetigkeit*)

Sei $(V, \|\cdot\|)$ ein normierter Vektorraum, $\emptyset \neq M \subseteq V$, $f \in \mathrm{Abb}(M, V)$ und sei $I \subseteq \mathbb{D}(f)$. f heiße **lipschitzstetig** auf I genau dann, wenn eine **Lipschitzkonstante** $L > 0$ existiert, so dass

$$\forall x, x_0 \in I. \|f(x) - f(x_0)\| \leqslant L \cdot \|x - x_0\|$$

Gilt $I = \mathbb{D}(f)$ so heiße f auch **gleichmäßig** lipschitzstetig. L heiße in jedem Fall **Lipschitzkonstante** von f auf I.

Korrolar D2.10

Wir stellen leicht fest, dass eine **jede** Kontraktion lipschitzstetig ist. Ebenfalls gelte, dass eine Funktion eine Kontraktion ist genau dann, wenn sie lipschitzstetig mit einer Lipschitzkonstanten $L_k < 1$ ist.

Anmerkung

Aus Schritt 2 des Beweises zu Satz 2.8 wissen wir, dass aus Lipschitzstetigkeit einer Funktionenfolge automatisch deren gleichgradige Stetigkeit folgt. Wir haben in diesem Schritt ebenfalls gezeigt, dass die Funktionenfolge der Eulerpolygone lipschitzstetig mit der Lipschitzkonstanten $M = \max\left\{ \|f(x,y)\| \;\middle|\; (x,y) \in G \right\}$ ist.

Wir wollen jetzt noch einen Zusammenhang zwischen Stetigkeit/Differenzierbarkeit und Lipschitzstetigkeit herstellen.

Satz 2.11 (*Zusammenhang der Lipschitzstetigkeit zur Stetigkeit und Differenzierbarkeit*)

(a) Sei f wie oben lipschitzstetig mit Lipschitzkonstante L, so ist f auch stetig.

(b) Sei $f \in \mathrm{Abb}(M, \mathbb{R})$ differenzierbar mit beschränkter Ableitung oder sei f stetig differenzierbar auf kompakter Menge, so ist f auch lipschitzstetig.

Beweis:

(a) Sei f wie oben lipschitzstetig mit Lipschitzkonstante L. Wähle dann $\delta := \frac{\varepsilon}{L}$, so gelte für alle $x, x_0 \in M$ mit $\|x - x_0\| \leqslant \delta = \frac{\varepsilon}{L}$, dass

$$\|f(x) - f(x_0)\| \leqslant L \cdot \|x - x_0\| \leqslant L \cdot \delta = L \cdot \frac{\varepsilon}{L} = \varepsilon,$$

womit die Stetigkeit durch das ε-δ-Kriterium nachgewiesen wurde.

(b) Wir betrachten hier nur den Fall, dass f eine beschränkte Ableitung hat, denn ist f stetig differenzierbar auf kompakter Menge, so ist die Ableitung von f auf dieser kompakten Menge – nach Satz von vorherigem Semester – auch beschränkt. Seien $x, x_0 \in M$ also beliebig, aber fest. Nach dem Mittelwertsatz gilt somit für ein $\xi \in (x, x_0)$, dass $f(x) - f(x_0) = f'(\xi) \cdot (x - x_0)$. Dementsprechend gilt somit für $|f(x) - f(x_0)|$, dass

$$|f(x) - f(x_0)| = \left| f'(\xi) \cdot (x - x_0) \right| \leqslant \left(\max_{\xi \in (x, x_0)} \left| f'(\xi) \right| \right) \cdot |x - x_0|.$$

Wir wissen, dass $\max\limits_{\xi \in (x, x_0)} \left| f'(\xi) \right|$ existieren muss, da f' beschränkt ist und somit laut Satz aus vergangenem Semester ein Maximum und Minimum existiert. Damit folgt dann automatisch die Lipschitzstetigkeit mit der Lipschitzkonstanten $\max\limits_{\xi \in (x, x_0)} \left| f'(\xi) \right|$.

□

Die Umkehrungen der Aussagen des Satzes gelten im Allgemeinen **nicht**! Man kann leicht triviale Gegenbeispiele angeben.
Für die erste Aussage betrachte man $f : x \mapsto \sqrt{|x|}$. Man erkennt leicht, dass f zwar stetig ist, mit scharfem Blick auch, dass f nicht lipschitzstetig ist. Wir führen wieder eine Reduktion aufs Absurde und nehmen an, dass f lipschitzstetig wäre. Man setze nun $x_{0_n} = 0$ und $x_n = 1/n^2$ und stelle damit fest, dass

$$|f(x_n) - f(x_{0_n})| \leqslant L \cdot |x_n - x_{0_n}| \quad \Longleftrightarrow \quad \left|\frac{1}{\sqrt{n^2}}\right| \leqslant L \cdot \left|\frac{1}{n^2}\right|,$$

was wiederum äquivalent zur Aussage

$$1 \leqslant L \cdot \frac{1}{n}$$

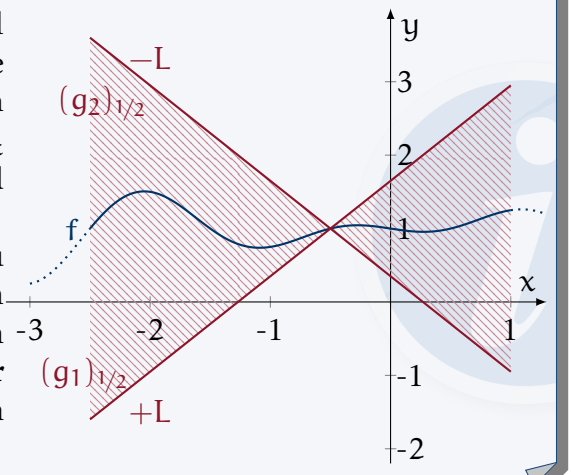
ist und, da L nicht durch n verändert wird, zu einem Widerspruch führt.
Für die zweite Aussage betrachte mann $g : x \mapsto |x|$ und stelle leicht fest, dass g lipschitzstetig mit Konstante 1 ist. Aus vergangenem Semester wissen wir allerdings auch, dass g an der Stelle $x_* = 0$ *nicht* differenzierbar ist.

Anmerkung

Auch wenn die Umkehrung der zweiten Aussage so einfach nicht funktioniert, ist bei einer lipschitzstetigen Funktion, voraussgesetzt ihre Ableitung existiert, diese auch garantiert beschränkt. Dies ist eine einfache Folgerung aus der Definition der Lipschitzstetigkeit.

Veranschaulichung

Wir wollen uns nun die Lipschitzstetigkeit einmal genauer veranschaulichen. Wir zeichnen dazu die Funktion f mit zwei Geraden $(g_1)_x$ und $(g_2)_x$ in ein gemeinsames Koordinatensystem ein. Die Geraden $(g_1)_x$ und $(g_2)_x$ haben die respektive Steigung von $\pm L$ und haben den Punkt $P_L(x, f(x))$ gemein.
f sei nun lipschitzstetig in diesem Punkt P_L genau dann, wenn der Graph von f **immer** zwischen den beiden Geraden liegt. f ist dann lipschitzstetig auf einem Intervall I genau dann, wenn der Graph von f **immer** zwischen den Geraden $(g_1)_x$ und $(g_2)_x$, für alle x aus dem Intervall I, liegt.



Wir wollen einen Eindeutigkeitssatz angeben, der zusätzlich neben der Eindeutigkeit der Lösung ebenfalls deren Existenz fordert.

Satz 2.12 (*Satz von Picard-Lindelöf*)

Sei $x_0 \in \mathbb{R}$, $y_0 \in \mathbb{R}^n$, $X > x_0$ und $f \in \mathrm{Abb}([x_0, X] \times \mathbb{R}^n, \mathbb{R}^n)$ stetig und sei f *lipschitzstetig* bezüglich seinem *zweiten* Argument und *gleichmäßig lipschitzstetig* bezüglich seinem *ersten*, es gelte also

$$\exists L > 0. \forall x_1, x_2 \in \mathbb{R}^n, x \in [x_0, X]. \|f(x, x_1) - f(x, x_2)\| \leqslant L \cdot \|x_1 - x_2\|.$$

Dann habe das Anfangswertproblem

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

eine **eindeutig** bestimmte, *stetig differenzierbare* Lösung $y = y(x)$.

Beweis: Wir wollen an dieser Stelle einen großen Bogen hin zum Banach'schen Fixpunktsatz (Satz 1.22) spannen und mit diesem dann die Eindeutigkeit und Existenz einer Lösung nachweisen. Dazu wandeln wir mit Satz 2.10 das gegebene Anfangswertproblem in eine Integralgleichung (ein *Integralgleichungsproblem*) um.
Man kann nun erkennen, dass durch die Vorschrift

$$(\Phi(y))(x) := y_0 + \int_{x_0}^{x} f(t, y(t)) dt, \quad y \in M, x \in [x_0, x_0 + a]$$

eine Abbildung $\Phi \in \mathrm{Abb}(M, M)$ erklärt ist, die genau einen Fixpunkt $y = \Phi(y) \in M$ besitzen soll, damit das Anfangswertproblem genau eine Lösung hat. Um den Banach'schen Fixpunktsatz (Satz 1.22) anwenden zu können, benötigen wir noch einen Banachraum M.
Wir wählen dazu den Funktionenraum $M := \mathcal{C}^0([x_0, X], \mathbb{R}^n) := \{y \in \mathrm{Abb}([x_0, X], \mathbb{R}^n) \mid y \text{ ist stetig}\}$ und versehen ihn mit der gewichteten Metrik

$$d_\alpha(y_1, y_2) := \max_{x \in [x_0, X]} \exp(-\alpha(x - x_0)) \cdot \|y_1(x) - y_2(x)\|_\infty$$

respektive

$$\|y\|_{\infty,\alpha} := \max_{x \in [x_0, X]} \exp(-\alpha(x - x_0)) \cdot \|y(x)\|_\infty,$$

wobei α echt größer als die Lipschitzkonstante L der Funktion f gewählt wird. $(M, \|\cdot\|_{\infty,\alpha})$ ist vollständig und normiert, damit ein Banachraum, M ist eine abgeschlossene Teilmenge von M und Φ ist selbstabbildend. Damit Satz 1.22 angewandt werden kann, ist noch die Kontraktionseigenschaft von Φ zu zeigen. Wähle also $x \in [x_0, X]$ beliebig, aber fest, so gelte:

$$\begin{aligned}
\|(\Phi(y_1))(x) - (\Phi(y_2))(x)\|_\infty &= \left\| \left(\cancel{y_0} + \int_{x_0}^{x} f(t, y_1(t)) dt \right) - \left(\cancel{y_0} + \int_{x_0}^{x} f(t, y_2(t)) dt \right) \right\|_\infty \\
&\leqslant \int_{x_0}^{x} \|f(t, y_1(t)) - f(t, y_2(t))\|_\infty dt \\
&\overset{(*)}{\leqslant} \int_{x_0}^{x} L \cdot \|y_1(t) - y_2(t)\|_\infty dt \\
&= L \cdot \int_{x_0}^{x} \overbrace{\|y_1(t) - y_2(t)\|_\infty}^{= d_\alpha(y_1, y_2)} \cdot \underbrace{\exp(-\alpha(t - x_0)) \cdot \exp(+\alpha(t - x_0))}_{=1} dt,
\end{aligned}$$

das heißt dann ebenso

$$\begin{aligned}\|(\Phi(y_1))(x)-(\Phi(y_2))(x)\|_\infty &\leqslant L\cdot\|y_1-y_2\|_{\infty,\alpha}\int_{x_0}^{x}\exp(\alpha(t-x_0))\,dt\\ &= L\cdot\|y_1-y_2\|_{\infty,\alpha}\cdot\left(\frac{1}{\alpha}\cdot(\exp(\alpha(x-x_0))-1)\right)\\ &\leqslant \frac{L}{\alpha}\cdot\|y_1-y_2\|_{\infty,\alpha}\cdot\exp(\alpha(x-x_0)),\end{aligned}$$

wobei an Stelle $(*)$ die Eigenschaft der Lipschitzstetigkeit von f herangezogen wurde. Daraus folgt dann mit der Division von $\exp(\dots)$, dass

$$\|\Phi(y_1)-\Phi(y_2)\|_{\infty,\alpha}\leqslant\frac{L}{\alpha}\cdot\|y_1-y_2\|_{\infty,\alpha}.$$

Das bedeutet, dass Φ eine Kontraktion ist, da $\alpha > L$.
Damit sind die Voraussetungen des Satzes 1.22 erfüllt, somit gilt nach Satz 1.22, dass das Integralgleichungsproblem eine eindeutig bestimmte Lösung hat, nach Satz 2.10 folgt damit ebenso, dass das Anfangswertproblem genau eine eindeutig bestimme Lösung hat, die stetig differenzierbar ist. □

An dieser Stelle sei auch nochmal der *„konstruktive“* Aspekt des Satzes angemerkt. Wie bereits schon zu Satz 1.22 angemerkt, besitzt der Banach'sche Fixpunktsatz eine Iterationsvorschrift zum Finden des Fixpunktes. Durch *sukzessive Approximation* kann der Fixpunkt der Abbildung Φ durch die Iterationsvorschrift

$$y_{n+1}(x) := (\Phi(y_n))(x) = y_0+\int_{x_0}^{x} f(t,y_n(t))dt \quad \text{für } n\in\mathbb{N} \text{ und } y_0(x) := y_0 \tag{2.32}$$

näherungsweise berechnet werden. Ein solches Verfahren hat aber gewisse negative Eigenschaften, welche wir in Kapitel 2.6.1 neben weiteren numerischen Verfahren betrachten werden.
Wir werden nun noch zwei Varianten des Satzes 2.12 kennen lernen, die in der Praxis häufiger Anwendung finden, da mehr Funktionen die Anforderungen erfüllen.

Satz 2.12 (*Satz von Picard-Lindelöf, Variante 2*)

Der obige Satz sei nahezu zu übernehmen. Man betrachte jetzt allerdings an des *„Streifens“* statt einen Zylinder Z_b mit

$$Z_b = \left\{(x,y)\;\middle|\; x_0\leqslant x\leqslant x_0+\lambda \text{ und } \|y-y_0\|<b\right\}\subset\mathbb{R}^{n+1}.$$

Man beachte nun, dass Lösungen gegebenenfalls nicht mehr auf dem ganzen Intervall $[x_0,\mathcal{X}]$ existieren, sondern nur auf einem Intervall $[x_0,x_0+\lambda]$ mit $\lambda := \min\left\{\mathcal{X}-x_0,\frac{b}{m}\right\}$, wobei m definiert ist als

$$m := \max_{(x,y)\in Z_b}\|f(x,y)\|,$$

welches gegebenenfalls kleiner ist. Ferner liefert der Satz in seiner neuen Form, dass die Lösung immer bis zum Rand des Zylinders geht.

Beweis: Erfülle f nun die Lipschitzbedingung nur auf dem Zylinder Z_b mit $b<\infty$ und gelte

$$k := \max_{(x,y)\in Z_b}\|f(x,y)\|.$$

Damit unsere Iteration (2.32) nun im $(k+1)$-ten Schritt nicht über den Zylinder „hinausschießt“, muss die Beziehung

$$\|y_{k+1}-y_0\| = \left\|\int_{x_0}^{x} f(t,y(t))dt\right\| \leqslant k(x-x_0)\leqslant Kc\overset{!}{\leqslant} b, \quad \text{wobei } x_0\leqslant x\leqslant x_0+c,$$

gelten. Das heißt, es muss $c \leqslant b/k$ sichergestellt sein. Da x selbstverständlich auch nicht über das Stetigkeitsintervall $[x_0, x_0 + \lambda]$ der Funktion f hinausreichen darf, können Lösungen des Anfangswertproblems im Allgemeinen nur auf dem Intervall

$$x_0 \leqslant x \leqslant x_0 + \underbrace{\min\left\{\lambda, \frac{b}{k}\right\}}_{=:c}$$

erwartet werden.
An diesen Überlegungen ändert sich nichts, wenn man anstelle des Intervalls $[x_0, x_0 + c]$ das symmetrische Intervall $[x_0 - c, x_0 + c]$ zugrundelegt. □
Neben dieser stärkeren Variante existiert auch noch folgende Variante, die die ursprüngliche Bedingung sehr abschwächt:

Satz 2.12 (*Satz von Picard-Lindelöf, Variante 3*)

Sei nun $f \in \mathrm{Abb}(U, \mathbb{R}^n)$ mit $U \subseteq \mathbb{R} \times \mathbb{R}^n$, sowie U eine Umgebung des Anfangswertes von (x_0, y_0). Für jedes $(x, y) \in U$ gebe es zusätzlich eine Umgebung $V_{x,y} \subseteq U$ und eine Zahl $L_{x,y} > 0$, so dass f einer lokalen Lipschitzbedingung auf $V_{x,y}$ genügt, also

$$\forall (x_1, y_1), (x_2, y_2) \in V_{x,y}.\ \|f(x_1, y_1) - f(x_2, y_2)\| \leqslant L_{x,y} \cdot \|y_1 - y_2\|$$

gilt. Dann bleibt die Aussage des Satzes erhalten; die Lösung geht so sicher „bis zum Rand“ von U.

i Falls U unbeschränkt ist, kann die Aussage, dass die Lösung sicher **„bis zum Rand“** ginge, bedeuten, dass einerseits die Lösung auf ganz $(-\infty, \infty)$ existiert, oder auch, dass die Lösung ein sogenanntes „Blow-Up-Verhalten“ zeigt, sprich eine Polstelle hat. In einem solchen Fall ist das Modell noch einmal zu überdenken, denn explosionsartige Verhaltensmuster (Änderungsprozesse) treten so eigentlich im Allgemeinen nicht auf.

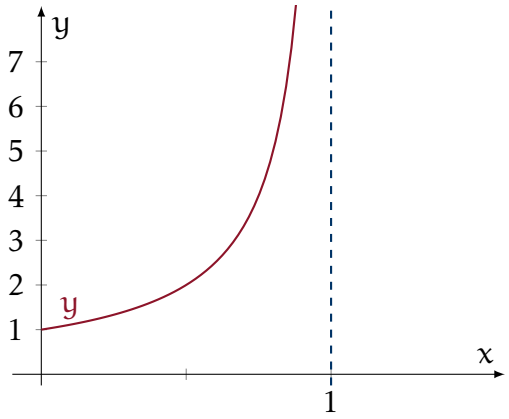
Beweisidee: Wir geben an dieser Stelle eine reine Beweisidee an, keinen fertig ausformulierten Beweis. Die grundlegende Idee hierbei ist es mehrere Lösungen „zusammenzustückeln“, das heißt man starte mit (x_0, y_0), finde eine Lösung auf der Umgebung V_{x_0,y_0}, welche dann Œ zylinderförmig gewählt werden kann. Diese Lösung geht dann unter Verwendung der vorherigen Variante bis zum Rand des Zylinders. Der Endpunkt kann aber dann wieder als neuer Startpunkt verwendet werden. Dies wiederholt man *„bis zum Rand“* der Umgebung U. □
Wir wollen nun die Anwendbarkeit des Satzes 2.12 an einem Beispiel betrachten. Man betrachte also
Beispiel 2.11: Gegeben sei das Anfangswertproblem

$$y' = y^2, \quad y(0) = 1.$$

$f(x, y) = y^2$ ist auf der unbeschränkten Menge $M := [x_0, X] \times \mathbb{R}$ stetig, damit existiert nach Satz 2.8 zumindest eine Lösung. f ist auf M allerdings nicht lipschitzstetig, da die Ableitung auf M unbeschränkt ist. Damit ist Variante 1 des Satzes 2.12 **nicht** anwendbar.
Auf der *kompakten* Menge $M_k := [0, X] \times [1 - R, 1 + R]$ ist $\frac{\partial}{\partial y} f$ stetig, somit auch beschränkt[1] und somit ist f auf M_k lipschitzstetig. Damit ist Variante 2 des Satzes 2.12 anwendbar, eine Lösung existiert damit zumindest auf $\left[0, \min\left\{X, \frac{R}{L}\right\}\right]$.
Auf $U = \mathbb{R} \times \mathbb{R}$ ist f **lokal** lipschitzstetig, womit Variante des Satzes 2.12 anwendbar ist, die Lösung, deren Existenz und Eindeutigkeit folgt, geht ganz sicher von „Rand bis zum Rand“ des $\mathbb{R} \times \mathbb{R}$.
Wir überprüfen unsere theoretischen Überlegungen, indem wir die tatsächliche Lösung mit einer Trennung der Veränderlichen ausrechnen:

[1] Dies folgt aus dem Satz aus C2 für stetige Funktionen auf kompakten Mengen

$$\begin{aligned}\int_0^{y(x)} \frac{d\eta}{\eta^2} &= \int_0^x 1\,dt \\ -\frac{1}{\eta}\Big|_1^{y(x)} &= x \\ -\frac{1}{y(x)} &= x-1 \\ y(x) &= \frac{1}{1-x}\end{aligned}$$



Wir stellen hier auch fest, dass die erste Variante des Satzes nicht klappen hätte können, da wir an der Stelle 1 eine Polstelle haben. Man spricht hierbei von einem explosiven („*blow-up*“) Verhalten. ※

2.4 Lineare Differenzialgleichungssysteme erster Ordnung

Nachdem wir uns jetzt in Kapitel 2.3 mit Existenz- und Eindeutigkeitsfragen für Anfangswertprobleme beschäftigt haben, wollen wir nun weitere, vertiefte Überlegungen zur Lösungsgesamtheit eines Differentialgleichungssystems anstellen. Zu diesem Punkt lassen sich – abgesehen von ebigen Sätzen 2.8 – 2.12 – nur relativ wenige allgemeine Aussagen treffen. Im Spezialfall der Systeme erster Ordnung allerdings sind sehr spezielle Aussagen möglich. Wir definieren deswegen:

Definition 2.11 (*Lineares System von Differentialgleichungen erster Ordnung*)

Sei die Matrixfunktion $A \in \mathrm{Abb}(\mathbb{R}, \mathbb{K}^{n\times n})$ und die Vektorfunktion $b \in \mathrm{Abb}(\mathbb{R}, \mathbb{K}^n)$ gegeben. Wir wollen ein System der Form

$$y'(x) = A(x)\cdot y(x) + b(x) \tag{2.33}$$

ein **lineares Differentialgleichungssystem erster Ordnung** nennen. Gilt $b(x) \equiv 0$, so heiße es **homogen**, andernfalls **inhomogen**. Das System

$$y'(x) = A(x)\cdot y(x) \tag{2.34}$$

heiße das zu (2.33) zugehörige **homogene System**.

i Sei an dieser Stelle noch einmal eine kurze Anmerkung zu Existenz und Eindeutigkeit von möglichen Lösungen erlaubt:
Es ist zeigbar, dass solange A und b stetig seien, die „rechte Seite“ (das f) ebenfalls stetig und sogar *lokal* lipschitzstetig bezüglich ihrem zweiten Argument y sei. Damit ist Variante 3 des Satz 2.12 zu zugehörigem Anfangswertproblem anwendbar, es *existiert* also eine Lösung und diese ist *eindeutig* bestimmt und verläuft von „Rand zu Rand des $\mathbb{R}\times\mathbb{R}^n$“.

2.4.1 Die Struktur der Lösungsmenge

Wir haben uns bereits in Kapitel 2.2.4 (insbesondere mit Satz 2.5) mit der Struktur von Lösungsräumen von Differentialgleichungen beschäftigt. Wir haben das Fazit gezogen, dass effektiv es ausreicht *eine* partikuläre Lösung mit allen homogenen Lösungen zu kennen, um *alle* Lösungen zu kennen. Es sei leicht und damit dem Leser überlassen sich die Analogie bei Systemen zu verdeutlichen. Wir erhalten damit folgenden Satz:

Satz 2.13 (*Struktur des Lösungsraums linearer Differentialgleichungssysteme erster Ordnung*)

Gegeben seien ein Intervall $I \subseteq \mathbb{R}$ und Funktionen $A \in \mathrm{Abb}(I, \mathbb{K}^{n\times n})$ und $v \in \mathrm{Abb}(I, \mathbb{K}^n)$ mit

$A, b \in \mathcal{C}^0(I)$, so gelten folgende Aussagen:

(a) Die Lösungen $y_h \in \mathcal{C}^1(I, \mathbb{K}^n)$ des homogenen Systems bilden einen Unterraum des $\mathrm{Abb}(I, \mathbb{K}^n)$. Damit ist L_{hom} ein Vektorraum.

(b) Ist $y_p \in \mathcal{C}^1(I, \mathbb{K}^n)$ eine beliebig, aber feste, partikuläre Lösung des inhomogenen Systems, so ist die allgemeine Lösung des Systems der affine Unterraum

$$L_{inhom} = \{y_p\} + L_{hom}.$$

Beweis: Der Beweis verläuft hier analog zum Beweis des Satzes 2.5, weswegen wir an dieser Stelle auf diesen verweisen. □

Auch hier zerfällt die Lösungskonstruktion für das System (2.33) in die zwei folgenden Teilaufgaben:

(H) Bestimme den Unterraum L_{hom}, heißt die Lösungsgesamtheit des homogenen Differentialgleichungssystems (2.34).

(P) Bestimme **eine** partikuläre Lösung y_p des inhomogenen Differentialgleichungssystems (2.33).

Wir wollen nun, bevor wir uns mit den einzelnen Aufgaben beschäftigen, noch klären, welche Dimension der Unterraum L_{hom} hat. Wir stellen dazu einen – auf den ersten Blick vielleicht unzusammenhängenden – Satz auf:

Satz 2.14 (*Lösung des Anfangswertproblems*)

Auf dem Intervall $I := [x_0, x_0 + a]$ seien **stetige** Koeffizientenfunktionen $a_{jk} \in \mathcal{C}^0(I)$ der Matrixfunktion $A(x) = (a_{jk}(x))$ gegeben. Sei des Weiteren $\mathcal{B} := \{b_1, b_2, \ldots, b_n\}$ eine Basis des Raumes der Anfangswerte mit $y_i(x)$ als zum Anfangswert b_i gehörende Lösung.
Dann existiere zu jeder Anfangsbedingung $y(x_0) = y_0$ genau eine Lösung $y \in \mathcal{C}^1(I, \mathbb{K}^n)$ des homogenen Systems. Zum Anfangswert

$$y_0 = \sum_{i=1}^{n} \alpha_i \cdot b_i$$

gehört dann die Lösung

$$y(x) = \sum_{i=1}^{n} \alpha_i \cdot y_i(x)$$

.

Beweis: Wir setzen

$$L := \max_{x \in I} \left(\sum_{j,k=1}^{n} \left| a_{jk}(x) \right|^2 \right)^{1/2} = \max_{x \in I} \|A(x)\|_F$$

sowie $f(x, y) := A(x)y$ für $y \in \mathbb{K}^n$. Dann genügt f den Voraussetzungen des Satzes 2.12, insbesondere auch der Lipschitzbedingung mit der gerade definierten Konstanten L. Daraus folgt Existenz und Eindeutigkeit einer Lösung.
Werden nun die n linear unabhängigen Vektoren b_j der Basis $\mathcal{B}$ als Anfangsvektoren $y_0 = b_j$ eingesetzt, so resultieren n Lösungen $y_j \in \mathcal{C}^1(I, \mathbb{K}^n)$, wobei $1 \leqslant j \leqslant n$, des homogenen Systems (2.34) mit den Anfangswerten $y_j(x_0) = b_j$.
Aus der linearen Unabhängigkeit im Punkt $x = x_0$ folgt dann auch die generelle lineare Unabhängigkeit dieser n Lösungen. Wir wissen, dass eine beliebige Lösung $y_h \in \mathcal{C}^1(I, \mathbb{K}^n)$ sicher einen Anfangswert $y_h(x_0) = (c_1, \ldots, c_n)^T \in \mathbb{K}^n$ stetig annimmt. Aus der eben gezeigten Eindeutigkeit

der Lösungen, muss damit dann

$$y_h = \sum_{j=1}^{n} c_j \cdot y_j(x)$$

gelten, woraus direkt folgt, dass das homogene System keine weiteren linear unabhängigen Lösungen haben kann. □
Aus dem eben geführten Beweis können wir dann folgenden Satz schließen:

Satz 2.15 (*Struktur des homogenen Lösungsvektorraums*)

Seien dieselben Voraussetzungen wie in Satz 2.14 gegeben. So bilden die zu den Anfangswerten gehörenden Lösungen y_j mit $1 \leqslant j \leqslant n$ eine Basis von L_{hom}.
L_{hom} wird damit von genau n linear unabhängigen Vektoren y_j aufgespannt.

Beweis: Der Satz folgt tatsächlich direkt aus den Aussagen im Beweis zu Satz 2.14. □

Korrolar 2.15

Insbesondere gilt damit

$$\dim(L_{hom}) = n.$$

Wir wollen nun weiter definieren:

Definition 2.12 (*Fundamentalsysteme und* Wronski-*Matrizen*)

(a) Ein System von n linear unabhängigen Lösungen y_j mit $1 \leqslant j \leqslant n$ des homogenen Differentialgleichungssystems (2.34) heiße ein **Fundamentalsystem**. Die einzelnen Lösungen eines solchen Fundamentalsystems heißen **Fundamentallösungen**.

(b) Bilden die Lösungen $y_1, \ldots, y_n$ des homogenen Differentialgleichungssystems (2.34) ein solches Fundamentalsystem, so heiße die Matrix(-funktion)

$$W(x) := (y_1(x), \ldots, y_n(x)) \in \mathbb{K}^{n \times n}$$

eine **Fundamentalmatrix** oder auch **Wronski-Matrix** des Differentialgleichungssytems. Die zu W zugehörige Determinante $\det(W(x))$ heiße **Wronski-Determinante**.

Man kann erkennen, dass sich die Lösung eines Anfangswertproblems durch $W(x_0) \cdot \vec{\alpha} = y_0$ ausdrücken lässt. Man verweise hier auf einen späteren Zeitpunkt, an dem wir diesen Zusammenhang noch einmal Revue passieren lassen. Mit ein bisschen Nachdenken ist es dann möglich, zu überprüfen, ob n Lösungen ein Fundamentalsystem bilden. Es ergeben sich folgende Äquivalenzen, mit denen eine Prüfung an einer beliebig, aber festen, Stelle x_* möglich wird.

Lemma 2.16 (*Äquivalenz Fundamentalsystem und* Wronski-*Determinante*)

Sei $x_* \in I$ beliebig, aber fest, so gelte für Lösungen der Differentialgleichungen y_j mit $1 \leqslant j \leqslant n$:

$$\begin{aligned} y_1, \ldots, y_n \in \mathrm{Abb}(I, \mathbb{K}^n) \text{ bilden ein FS} &\iff y_1(x_*), \ldots, y_n(x_*) \text{ bilden eine Basis des } \mathbb{K}^n \\ &\iff W(x_*) \text{ ist invertierbar} \\ &\iff \det(W(x_*)) \neq 0 \end{aligned}$$

Beweis: trivial. Man verweise an dieser Stelle auf das Skript zur Erstsemestermathematikveranstaltung, andererseits auch auf die Vorüberlegungen und vorangestellten Definitionen und auf „gesunden Menschenverstand“. □
Wir werden uns nun mit Lösungsmethoden für die Teilaufgaben (H) und (P) beschäftigen.

2.4.2 Teilaufgabe (H) — Bestimmung der Lösung des homogenen Differentialgleichungssystems

Aus Satz 2.15 folgt unmittelbar, dass um L_{hom} aufzustellen es reicht eine Basis, und damit ein Fundamentalsystem, von L_{hom} zu kennen. Wir wollen uns damit auf die Suche nach einem Verfahren machen, mit dem wir *leicht* ein Fundamentalsystem berechnen können. Bei dieser Aufgabe kommen wir schnell zu einem unumgänglichen Problem, denn bei einer von der Veränderlichen x abhängenden Systemmatrix A ist so kein Verfahren zur Berechnung eines Fundamentalsystems angebbar. Es existiert allerdings ein Verfahren – das D'Alembert*'sche Reduktionsverfahren* – es möglich macht mit einer (*geratenen*) Lösung das System von dem $\mathbb{K}^n$ auf den $\mathbb{K}^{n-1}$ zu reduzieren. Mit diesem wollen wir uns an dieser Stelle allerdings nicht näher beschäftigen. Hängt jedoch die Systemmatrix A eben *nicht* von der Veränderlichen x ab, so existieren durchaus Verfahren, mit welchen eine Bestimmung des Fundamentalsystems möglich ist. Mit eben diesen wollen wir uns näher befassen, deswegen sei im Folgenden Œ A *nicht* von x abhängig.
Wir suchen also nach einem Fundamentalsystem für das Differentialgleichungssystem

$$y' = A \cdot y \tag{2.35}$$

Dazu tasten wir uns von den „leichtesten" Klassen an Systemmatrizen – sprich den Klassen mit den *striktesten* Voraussetzungen – bis hin zu den „schwersten" respektive „allgemeinsten" Klassen an Systemmatrizen voran.

2.4.2.1 Systemmatrix A ist eine Diagonalmatrix

Sei also $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Das System (2.35) lässt sich damit dann umschreiben zu

$$y'(x) = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} \cdot y(x) \Leftrightarrow \begin{cases} y_1'(x) = \lambda_1 \cdot y_1(x) \\ \vdots \\ y_n'(x) = \lambda_n \cdot y_n(x) \end{cases},$$

womit das System in n skalare Probleme zerfällt, die alle *unabhängig* voneinander sind. Diese wiederum werden „leicht" nach den Varianten, welche in Kapitel 2.2 vorgestellt wurden, gelöst, wir erhalten damit dann als Lösung

$$y_i(x) = c_i \cdot \exp(\lambda_i \cdot x), c_i \in \mathbb{K} \text{ beliebig, für } 1 \leqslant i \leqslant n.$$

Demnach hat unser System dann die Lösungsmenge

$$\begin{aligned} L_{hom} &= \left\{ y \in \operatorname{Abb}(I, \mathbb{K}^n \;\middle|\; y(x) = \begin{pmatrix} c_1 \cdot \exp(\lambda_1 \cdot x) \\ \vdots \\ c_n \cdot \exp(\lambda_n \cdot x) \end{pmatrix}, c_1, \ldots, c_n \in \mathbb{K} \right\} \\ &= \left\{ y \in \operatorname{Abb}(I, \mathbb{K}^n \;\middle|\; y(x) = \sum_{i=1}^{n} c_i \cdot \nu_i, c_1, \ldots, c_n \in \mathbb{K}, \nu_i \text{ enthält } \exp(\lambda_i \cdot x) \text{ an Position } i \right\} \\ &= \operatorname{span}\left\{ \begin{pmatrix} c_1 \cdot \exp(\lambda_1 \cdot x) \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ c_n \cdot \exp(\lambda_n \cdot x) \end{pmatrix} \right\}, \end{aligned}$$

womit folgt, dass die gefundenen Lösungen eine Basis des $\mathbb{K}^n$, womit sie nach Lemma 2.16 dann ein Fundamentalsystem bilden. Wir fassen zusammen:

Sollte $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ gelten, so ist $\exp(\lambda_i \cdot x) \cdot e_i$ für $1 \leqslant i \leqslant n$ ein Fundamentalsystem.

Dabei sind die e_i genau die Eigenvektoren von A.

2.4.2.2 Systemmatrix A ist diagonalisierbar

Wir wenden nun eine ähnliche Strategie an, wie sie bereits in Kapitel 1.1 verwendet wurde, um die Definitheit von symmetrischen Matrizen zu untersuchen. Anders als in Kapitel 1.1 allerdings müssen wir jetzt die Diagonalisierbarkeit von A extra fordern, da diese nicht herzuleiten ist.
Sei A also nun diagonalisierbar, es existiert also ein Q, dessen Spalten den zu A zugehörigen Eigenvektoren entsprechen, so dass

$$A = QDQ^{-1} \quad \text{mit } D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

wobei λ_i mit $1 \leqslant i \leqslant n$ die Eigenwerte von A darstellen. Wir setzen nun diesen Zusammenhang in die Differentialgleichung (2.35) ein und erhalten damit:

$$\begin{aligned}
& y'(x) &&= QDQ^{-1} \cdot y && \mid \cdot Q^{-1} \text{ „von links“} \\
\iff\ & \frac{d}{dx} Q^{-1} y(x) &&= DQ^{-1} \cdot y(x) && \mid u(x) := Q^{-1} \cdot y(x) \text{ „Substitution“} \\
\iff\ & u(x) &&= D \cdot u(x) && (2.36)
\end{aligned}$$

Damit liegt (2.36) in diagonalisierter Form vor, und kann mit den Methoden aus Sektion 2.4.2.1 gelöst werden. Wir erhalten damit ein Fundamentalsystem für (2.36) mit

$$u_i(x) = \exp(\lambda_i \cdot x) \cdot e_i,$$

wobei $1 \leqslant i \leqslant n$ und e_i die Standardbasisvektoren des $\mathbb{K}^n$ und damit die Eigenvektoren von D sind. Durch eine Rücksubstitution $y_i := Q \cdot u(x)$ folgt dann für y_i

$$y_i(x) = Q \cdot u_i(x) = Q \cdot e_i \cdot \exp(\lambda_i \cdot x) = q_i \cdot \exp(\lambda_i \cdot x),$$

für $1 \leqslant i \leqslant n$ und q_i als Eigenvektor von A zum zugehörigen Eigenwert λ_i. Wir können also sowohl Sektion 2.4.2.1 als auch Sektion 2.4.2.2 in folgendem Satz zusammenfassen:

Satz 2.17 (*Fundamentalsystem bei einer diagonalisierbaren Systemmatrix*)

Sei eine Differentialgleichung wie in (2.35) gegeben und A diagonalisierbar . Seien $\lambda_1, \ldots, \lambda_n$ die – *nicht notwendigerweise verschiedenen* – Eigenwerte von A und sei $\{q_1, \ldots, q_n\}$ eine Basis des $\mathbb{K}^n$, so dass q_i Eigenvektor zum Eigenwert λ_i ist, so sei durch

$$\exp(\lambda_1 \cdot x) \cdot q_1, \ldots, \exp(\lambda_n \cdot x) \cdot q_n \tag{2.37}$$

ein Fundamentalsystem für das lineare Differentialgleichungssystem (2.35) gegeben.

Beweis: Der Satz gibt sich als Konsequenz der obigen Überlegungen. □

Anmerkungen

Es sei das hinreichende Kriterium für die Diagonalisierbarkeit, welches wir bereits seit dem ersten Semester kennen, sich in Erinnerung zu rufen: Eine Matrix ist diagonalisierbar genau dann, wenn für *jeden* Eigenwert λ_i die algebraische und geometrische Vielfachheit übereinstimmen. Dies wiederum ist genau dann der Fall, wenn es eine Basis des $\mathbb{K}^n$ aus Eigenvektoren gibt.

i Eben genannter Satz 2.17 liefert ebenfalls eine Erklärung, warum bei linearen *skalaren* Differentialgleichungen höherer Ordnung mit konstanten Koeffizienten ein sogenannter „Exponentialansatz“ – also ein $y(x) = c \cdot \exp(\lambda \cdot x)$, wobei $c \in \mathbb{K}$ und λ zu bestimmen ist – weiterhilft. Dies ist der Tatsache geschuldet, dass eben solche Differentialgleichungen in lineare Systeme erster Ordnung umgewandelt werden können – das wissen wir bereits durch Lemma 2.1 –, denn diese Systeme haben dann mit Satz 2.17 eben Lösungen der Form $\exp(\lambda \cdot x) \cdot q$, wobei q ein Vektor ist. Für das ursprüngliche *skalare* Problem braucht es dann von dem Lösungsvektor nur die erste Komponente, es ergibt sich somit eine Lösung der Form $y(x) = c \cdot \exp(\lambda \cdot x)$.

Im diagonalisierbaren Fall reicht es also, die Eigenwerte und Eigenvektoren von A zu kennen, um damit dann ein Fundamentalsystem aufzustellen. Wir betrachten dazu **Beispiel 2.12:** Zu berechnen ist ein Fundamentalsystem für ein Differentialgleichungssystem wie in (2.35) mit

$$A = \begin{pmatrix} 7 & -5 \\ 10 & -8 \end{pmatrix}$$

und zusätzlich die Lösung des Anfangswertproblems mit $y(0) = y_0 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$.

Wir bestimmen also das Fundamentalsystem, dazu berechnen wir zuerst Eigenwerte und -vektoren. Wir stellen also das charakteristische Polynom auf und berechnen die Nullstellen ebendieses:

$$p(\lambda) = (7-\lambda)\cdot(-8-\lambda) + 50 = \lambda^2 + \lambda - 6 \overset{!}{=} 0$$

Wir schlussfolgern mit der abc-Formel/Mitternachtsformel:

$$\lambda_{1/2} = -\frac{1}{2} \pm \underbrace{\sqrt{\frac{1}{4} + 6}}_{=\sqrt{\frac{25}{4}}} = -\frac{1}{2} \pm \frac{5}{2}$$

und somit

$$\lambda_1 = 2 \quad \text{und} \quad \lambda_2 = -3.$$

Wir berechnen nun die Eigenvektoren zu den Eigenwerten λ_1 und λ_2. Das heißt wir bestimmen die Vektoren v mit der Eigenschaft

$$(A - E_2 \cdot \lambda) \cdot v = \begin{pmatrix} 7-\lambda & -5 \\ 10 & -8-\lambda \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \vec{0}.$$

$\lambda_1 = 2$: Wir lösen also das Lineare Gleichungssystem

$$\begin{pmatrix} 7-2 & -5 \\ 10 & -8-2 \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

$$\left(\begin{array}{cc|c} 5 & -5 & 0 \\ 10 & -10 & 0 \end{array}\right) \begin{array}{l} \longleftarrow \\ \mid \cdot \frac{1}{10} \end{array} \rightsquigarrow \left(\begin{array}{cc|c} 1 & -1 & 0 \\ 5 & -5 & 0 \end{array}\right) \begin{array}{l} \\ \cdot(-5) \\ + \end{array} \rightsquigarrow \left(\begin{array}{cc|c} 1 & -1 & 0 \\ 0 & 0 & 0 \end{array}\right)$$

Wir erhalten somit den Zusammenhang

$$v_1 - v_2 = 0 \quad \Leftrightarrow \quad v_1 = v_2$$

Ein Eigenvektor ist damit beispielsweise

$$v_{\lambda_1} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$\lambda_2 = -3$: Wir lösen also das Lineare Gleichungssystem

$$\begin{pmatrix} 7-(-3) & -5 \\ 10 & -8-(-3) \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

$$\left(\begin{array}{cc|c} 10 & -5 & 0 \\ 10 & -5 & 0 \end{array}\right) \begin{array}{l} |\cdot\frac{1}{5} \\ |\cdot\frac{1}{5} \end{array} \quad \rightsquigarrow \quad \left(\begin{array}{cc|c} 2 & -1 & 0 \\ 2 & -1 & 0 \end{array}\right) \begin{array}{l} \\ \leftarrow + \end{array}{}^{\cdot(-1)} \quad \rightsquigarrow \quad \left(\begin{array}{cc|c} 2 & -1 & 0 \\ 0 & 0 & 0 \end{array}\right)$$

Wir erhalten somit den Zusammenhang

$$2 \cdot v_1 - v_2 = 0 \quad \Leftrightarrow \quad v_1 = \frac{v_2}{2}$$

Ein Eigenvektor ist damit beispielsweise

$$v_{\lambda_2} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Mit den Eigenvektoren bestimmen wir dann das Fundamentalsystem, welches aus y_1 und y_2 mit

$$y_1(x) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot \exp(2 \cdot x) \quad \text{und} \quad y_2(x) = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \exp(-3 \cdot x)$$

besteht, und somit dann auch die *allgemeine* Lösung als

$$y(x) = c_1 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot \exp(2 \cdot x) + c_2 \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \exp(-3 \cdot x),$$

mit $c_1, c_2 \in \mathbb{R}$. Wir können nun dann das Anfangswertproblem für $x = 0$ lösen, auch dies passiert wieder durch das Lösen eines linearen Gleichungssystems.

$$\begin{pmatrix} 5 \\ 2 \end{pmatrix} \stackrel{!}{=} y(0) = c_1 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2 \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

beschreibt die zu lösende Gleichung, welche das lineare Gleichungssystem

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$$

induziert.

$$\left(\begin{array}{cc|c} 1 & 1 & 5 \\ 1 & 2 & 2 \end{array}\right) \begin{array}{l} \\ \leftarrow + \end{array}{}^{\cdot(-1)} \quad \rightsquigarrow \quad \left(\begin{array}{cc|c} 1 & 1 & 5 \\ 0 & 1 & -3 \end{array}\right)$$

Damit schließen wir dann, dass $c_1 = 8$ und $c_2 = -3$ standhalten muss, damit das Anfangswertproblem erfüllt ist. Eine partikuläre Lösung des Anfangswertproblems ist also

$$y(x) = 8 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot \exp(2 \cdot x) - 3 \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \exp(-3 \cdot x)$$

※

Was passiert bei komplexen Eigenwerten einer rein reellen Systemmatrix? Das letzte Problem welchem wir uns, bei diagonalisierbaren Matrizen, stellen ist ein für die Modellierung realer Sachverhalte zentrales Problem. Betrachten wir dazu einmal **Beispiel 2.13:** Gesucht sei eine rein reelle Lösung für die Differentialgleichung

$$y' = A \cdot y,$$

wobei die Systemmatrix A definiert ist als

$$A = \begin{pmatrix} -3 & -2 \\ 5 & 3 \end{pmatrix}.$$

Eine solche muss nach Theorie existieren, wir wollen uns damit auf die Suche nach einem rein reellen Fundamentalsystem der Systemmatrix A machen. Wir bestimmen dazu als erstes die Eigenwerte der Systemmatrix A, indem wir die Nullstellen des charakteristischen Polynoms finden.

$$p(\lambda) = (-3-\lambda)\cdot(-3-\lambda) - (-2\cdot 5) = \lambda^2 + 1 \stackrel{!}{=} 0$$

Wir sehen sofort, dass $\pm i$ die *einzigen* Nullstellen des charakteristischen Polynoms sind, sowie ebenfalls, dass $\pm i \notin \mathbb{R}$, obwohl die Systemmatrix nur reelle Einträge hat. Wir wollen nun dennoch ein reelles Fundamentalsystem bestimmen, was mit komplexen Eigenwerten mit unserer bisherigen Methode ja „*ausgeschlossen*“ ist.
Wir erkennen durchaus, dass es selbst bei relativ simplen Matrizen zu komplexen Eigenwerten kommen kann, obwohl wir ein reelles Fundamentalsystem haben wollen. Wir befassen uns deswegenn mit folgendem Satz:

Satz 2.18 (*Umwandlung eines komplexen in ein reelles Fundamentalsystem*)

Seien $a, b \in \mathbb{R}, s, r \in \mathbb{R}^n$ beliebig. Ist dann eine Fundamentallösung $y_1(x) = \exp(\lambda \cdot x) \cdot q$ mit $q = r + s \cdot i$ und $\lambda = a + b \cdot i$ eines rein *reellen* Differentialgleichungssystems echt komplex, so ist auch $y_2(x) = \exp(\overline{\lambda} \cdot x) \cdot \overline{q}$ mit $\overline{\lambda} = a - b \cdot i$ und $\overline{q} = r - s \cdot i$ eine Fundamentallösung und y_1 ist das konjungiert komplexe zu y_2.
Die Funktionen

$$\begin{aligned} y_{1,reell}(x) := \Re(y_1) &= \frac{1}{2}(y_1(x) + y_2(x)) \\ &= (r \cdot \cos(b \cdot x) - s \cdot \sin(b \cdot x)) \cdot \exp(a \cdot x) \\ &= [\Re(q) \cdot \cos(\Im(\lambda) \cdot x) - \Im(q) \cdot \sin(\Im(\lambda) \cdot x)] \cdot \exp(\Re(\lambda) \cdot x) \end{aligned}$$

und

$$\begin{aligned} y_{2,reell}(x) := \Im(y_1) &= \frac{1}{2 \cdot i}(y_1(x) - y_2(x)) \\ &= (s \cdot \cos(b \cdot x) + r \cdot \sin(b \cdot x)) \cdot \exp(a \cdot x) \\ &= [\Im(q) \cdot \cos(\Im(\lambda) \cdot x) + \Re(q) \cdot \sin(\Im(\lambda) \cdot x)] \cdot \exp(\Re(\lambda) \cdot x) \end{aligned}$$

sind dann die reellen Fundamentallösungen, die die komplexen ersetzen können. Ersetzt man so alle komplexen Paare durch die somit gefundenen reellen Paare, erhält man ein rein reelles Fundamentalsystem.

Beweis: Wir wissen aus dem ersten Semester, dass komplexe Nullstellen eines reellen Polynoms immer „paarweise“ auftreten, sprich zu einer komplexen Nullstelle z muss auch die konjungiert komplexe Zahl $\overline{z}$ als Nullstelle desselben Polynoms existieren. Damit folgt dann die gleiche Aussage mit komplexen Eigenwerten bei reellen Matrizen. Wir haben damit dann geschlussfolgert, dass nicht nur Eigenwerte in konjungierten Paaren auftreten, sondern auch die zugehörigen Eigenvektoren konjungiert komplex sind. Man bekommt somit *nicht reelle* Fundamentallösungen *immer* paarweise als

$$y_1(x) = \exp(\lambda \cdot x) \cdot q \quad \text{und} \quad y_2(x) = \exp(\overline{\lambda} \cdot x) \cdot \overline{q}$$

Man zerlege nun $\lambda \in \mathbb{C}$ und $q \in \mathbb{C}^n$ in Real- und Imaginärteil mit

$$\lambda = a + b \cdot i \quad \text{und} \quad q = r + s \cdot i,$$

wobei $a, b \in \mathbb{R}$ und $r, s \in \mathbb{R}^n$, und erhalte somit die Darstellungen für y_1 mit

$$\begin{aligned} y_1(x) &= (r + s \cdot i) \cdot \exp((a + b \cdot i) \cdot x) \\ &= (r + s \cdot i) \cdot (\cos(b \cdot x) + i \cdot \sin(b \cdot x)) \exp(a \cdot x) \\ &= \underbrace{(r \cdot \cos(b \cdot x) - s \cdot \sin(b \cdot x)) \cdot \exp(a \cdot x)}_{=:\Re(y_1(x))} + i \cdot \underbrace{(s \cdot \cos(b \cdot x) + r \cdot \sin(b \cdot x)) \cdot \exp(a \cdot x)}_{=:\Im(y_1(x))} \end{aligned}$$

und für y_2 mit

$$\begin{aligned} y_2(x) &= (r - s \cdot i) \cdot \exp((a - b \cdot i) \cdot x) \\ &= (r - s \cdot i) \cdot (\cos(b \cdot x) - i \cdot \sin(b \cdot x)) \exp(a \cdot x) \\ &= \underbrace{(r \cdot \cos(b \cdot x) - s \cdot \sin(b \cdot x)) \cdot \exp(a \cdot x)}_{=:\Re(y_2(x))} - i \cdot \underbrace{(s \cdot \cos(b \cdot x) + r \cdot \sin(b \cdot x)) \cdot \exp(a \cdot x)}_{=:\Im(y_2(x))}. \end{aligned}$$

Beim Addieren/Subtrahieren der beiden Fundamentallösungen erhalten wir etwas rein reelles oder rein komplexes, sprich es gilt

$$y_{1,reell} := \frac{1}{2}(y_1 + y_2) = \Re(y_1) \in \mathbb{R} \quad \text{und} \quad y_{2,reell} := \frac{1}{2 \cdot i}(y_1 - y_2) = \Im(y_1) \in \mathbb{R}.$$

Wir sehen schnell ein, dass solche Linearkombinationen von Lösungen wieder Lösungen sind, und die Fundamentallösungen somit einfach ersetzen können. □

Wir wollen nun mit vorherigem Beispiel fortfahren, rechnen also **Beispiel 2.13** (*Fortsetzung*)**:** Wir berechnen nun also als erstes ein komplexes Fundamentalsystem und damit die Eigenvektoren zu den Eigenwerten $\pm i$:

$\lambda_1 = -i$: Wir lösen also das Lineare Gleichungssystem

$$\begin{pmatrix} -3+i & -2 \\ 5 & 3+i \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

$$\left(\begin{array}{cc|c} -3+i & -2 & 0 \\ 5 & 3+i & 0 \end{array}\right) \overset{\leftarrow\rfloor}{\leftarrow\rfloor} \rightsquigarrow \left(\begin{array}{cc|c} 5 & 3+i & 0 \\ -3+i & -2 & 0 \end{array}\right) \overset{\cdot(\frac{3-i}{5})\ |\cdot\frac{1}{5}}{\rfloor_{+}\leftarrow\rfloor} \rightsquigarrow \left(\begin{array}{cc|c} 1 & \frac{3+i}{5} & 0 \\ 0 & 0 & 0 \end{array}\right)$$

Wir erhalten somit den Zusammenhang

$$v_1 + \frac{3+i}{5} \cdot v_2 = 0 \quad \Leftrightarrow \quad v_1 = -\frac{3+i}{5} \cdot v_2$$

Ein Eigenvektor ist damit beispielsweise

$$v_{\lambda_1} = \begin{pmatrix} -\frac{3+i}{5} \\ 1 \end{pmatrix}$$

$\lambda_2 = i$: Wir lösen also das Lineare Gleichungssystem

$$\begin{pmatrix} -3-i & -2 \\ 5 & 3-i \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

$$\left(\begin{array}{cc|c} -3-i & -2 & 0 \\ 5 & 3-i & 0 \end{array}\right) \overset{\leftarrow\rfloor}{\leftarrow\rfloor} \rightsquigarrow \left(\begin{array}{cc|c} 2 & -1 & 0 \\ 2 & -1 & 0 \end{array}\right) \overset{\cdot(\frac{3+i}{5})\ |\cdot\frac{1}{5}}{\rfloor_{+}\leftarrow\rfloor} \rightsquigarrow \left(\begin{array}{cc|c} 1 & \frac{3-i}{5} & 0 \\ 0 & 0 & 0 \end{array}\right)$$

Wir erhalten somit den Zusammenhang

$$v_1 + \frac{3-i}{5} \cdot v_2 = 0 \quad \Leftrightarrow \quad v_1 = -\frac{3-i}{5} \cdot v_2$$

Ein Eigenvektor ist damit beispielsweise

$$v_{\lambda_2} = \begin{pmatrix} -\frac{3-i}{5} \\ 1 \end{pmatrix}$$

Mit den Eigenvektoren können wir dann das – wohlgemerkt *komplexe* –Fundamentalsystem bestimmen. Jenes enthält y_1 und y_2 mit

$$y_1 = \left[\begin{pmatrix} -\frac{3}{5} \\ 1 \end{pmatrix} - i \cdot \begin{pmatrix} \frac{1}{5} \\ 0 \end{pmatrix}\right] \cdot \exp(-i \cdot x) \quad \text{und} \quad y_2 = \left[\begin{pmatrix} -\frac{3}{5} \\ 1 \end{pmatrix} + i \cdot \begin{pmatrix} \frac{1}{5} \\ 0 \end{pmatrix}\right] \cdot \exp(i \cdot x).$$

Mit Satz 2.18 berechnen wir schlussendlich dann das rein reelle Fundamentalsystem, es gilt dann also

$$FS_{reell} = \left\{ \left[\begin{pmatrix} -3/5 \\ 1 \end{pmatrix} \cdot \cos(-x) - \begin{pmatrix} 1/5 \\ 0 \end{pmatrix} \cdot \sin(-x)\right], \left[\begin{pmatrix} 1/5 \\ 0 \end{pmatrix} \cdot \cos(-x) + \begin{pmatrix} -3/5 \\ 1 \end{pmatrix} \cdot \sin(-x)\right] \right\}.$$

※

2.4.2.3 Systemmatrix A ist nicht diagonalisierbar

In dieser letzten Sektion wollen wir uns mit den restlichen Matrizen beschäftigen, also diejenigen, welche nicht diagonalisierbar sind. In solchen Fällen müssen wir uns erstmal anderen Methodiken bedienen, um ein Fundamentalsystem zu berechnen. Auch hier erinnern wir uns an das erste Semester und die Äquivalenzen

$$\begin{aligned} A \in \mathbb{K}^{n\times n} \text{ ist } \textit{nicht} \text{ diagonalisierbar} \quad &\Leftrightarrow \quad \sum_{\lambda_i \text{ ist EW}} \dim(\mathrm{Eig}(\lambda_i)) < n \\ &\Leftrightarrow \quad \exists \lambda.\lambda \text{ ist Eigenwert und } \mathrm{geomVfht}(\lambda) < \mathrm{algVfht}(\lambda). \end{aligned} \tag{2.38}$$

Gleichzeitig stellen wir fest, dass Lösungen der Bauart wie in (2.37) – also $\exp(\lambda_i \cdot x)\cdot q_i$ mit q_i als dem zum Eigenwert λ_i zugehörigen Eigenvektor – auch weiterhin Lösungen sind. Wegen (2.38) erhalten wir aber echt weniger als n viele solcher Lösungen und damit kein vollständiges Fundamentalsystem. Wir wollen deswegen nun den Begriff des Eigenvektors „verallgemeinern“ und definierern dazu:

Definition 2.13 (*Hauptvektoren*)

Ein Vektor $0 \neq v \in \mathbb{K}^n$ heiße **Hauptvektor der Stufe** $k \in \mathbb{N}$ zum Eigenwert λ der Matrix $A \in \mathbb{K}^{n\times n}$ genau dann, wenn

$$(A - \lambda E_n)^k \cdot v = 0 \quad \text{und} \quad (A - \lambda E_n)^{k-1} \cdot v \neq 0$$

Aus dieser Definition resultieren nun einige Folgerungen:

Korrolar 2.13.1

Die Hauptvektoren erster Stufe einer Matrix zum Eigenwert λ sind genau dessen Eigenvektoren .

Korrolar 2.13.2

Ist $v \neq 0$ ein Hauptvektor der Stufe $k \geqslant 2$ zum Eigenwert λ, so ist $(A - \lambda E_n)v$ ein Hauptvektor der Stufe $k-1$ zum selben Eigenwert λ.

Beweis: Setze $w := (A - \lambda E_n)v$, so ist $(A - \lambda E_n)^{k-1} w = 0$, aber $(A - \lambda E_n)^{k-2} w \neq 0$, damit dann auch $w \neq 0$. □

Korrolar 2.13.3

Die Hauptvektoren der Stufe $k \geqslant 2$ zum Eigenwert λ sind von den Hauptvektoren der Stufen $k' \leqslant k-1$ zum selben Eigenwert λ linear unabhängig.

Beweis: Seien $v_1, \ldots, v_r$ mit $r \leqslant n$ die Hauptvektoren der Stufen $k' \leqslant k-1$, so folgte aus dem Ansatz der linearen Abhängigkeit

$$w = \sum_{j=1}^{r} \mu_j v_j \quad \text{mit} \sum_{j=1}^{r} |\mu_j| \neq 0$$

die widersprüchliche Bedingung

$$(A - \lambda E_n)^{k-1} w = \sum_{j=1}^{r} \mu_j \underbrace{(A - \lambda E_n)^{k-1} \cdot v_j}_{=0} = 0. \lightning$$

□

Korrolar 2.13.4

Sei $H_k := \left\{ v \in \mathbb{K}^n \;\middle|\; (A - \lambda E_n)^k v = 0 \right\} = \ker\left((A - \lambda E_n)^k\right)$ die Menge aller Hauptvektoren vom

Grad $\leqslant k$ zum Eigenwert λ mit der 0, so gelte für alle $k \in \mathbb{N}$:

$$\{0\} \subset \mathrm{Eig}(\lambda) = H_1(\lambda) \subseteq H_2(\lambda) \subseteq \cdots \subseteq H_k(\lambda) \subseteq \mathbb{K}^n$$

Korrolar 2.13.5

Die Gesamtheit aller Hauptvektoren der Stufe $\leqslant k$ zum Eigenwert λ spannt den Unterraum $\ker(A - \lambda E_n)^k \subset \mathbb{K}^n$ auf.

Korrolar 2.13.6

Ist $H_k(\lambda) = H_{k+1}(\lambda)$ für ein $k \in \mathbb{N}$ erfüllt, so folgt

$$\forall r \geqslant k. H_r(\lambda) = H_{r+1}(\lambda)$$

Beweis:

$\subseteq$ trivial, folgt aus Korollar 2.13.4.

$\supseteq$ Wir zeigen dies durch vollständige Induktion.

IA $(r = k)$: trivial, dies ist genau die Voraussetzung.

IS $(r \to r+1)$:
Gelte schon $H_{r+1}(\lambda) \subseteq H_r(\lambda)$ für ein $r > k$. Sei $v \in \ker(A - \lambda E_n)^{r+2} = H_{r+2}(\lambda)$, also $0 = (A - \lambda E_n)^{r+2} v = (A - \lambda E_n)^{r+1}(A - \lambda E_n)v$, womit $(A - \lambda E_n)v \in H_{r+1}(\lambda)$ gilt. Mit der IV folgt dann ebenfalls, dass $(A - \lambda E_n)v \in H_r(\lambda)$, was allerdings implziert, dass $(A - \lambda E_n)^{r+1} v = 0$, woraus folgt, dass $v \in H_{r+1}(v)$.

□

Korrolar 2.13.7

Aus den Korollaren 2.13.3 bis 2.13.6 folgt die Existenz eines kleinsten Index $f = f(\lambda)$, für den gelte, dass

$$\{0\} \subset \mathrm{Eig}(\lambda) = H_1(\lambda) \subset H_2(\lambda) \subset \cdots \subset H_f(\lambda) = H_{f+1}(\lambda), \tag{2.39}$$

wobei die ersten f Inklusionen echt sind.

Wir definieren weiter:

Definition 2.14 (*Fittingindex und Haupträume*)

Die durch die Eigenschaft (2.39) charakterisierte Zahl $f = f(\lambda) \in \mathbb{N}$ heiße der **Fittingindex** – oder kurz auch der **Index** – des Eigenwertes λ einer Matrix $A \in \mathbb{K}^{n \times n}$. Für diesen Index gilt stets $f(\lambda) \leqslant n$.
Der Unterraum $H_f(\lambda) = \ker(A - \lambda E_n)^f \subseteq \mathbb{K}^n$ heiße der **verallgemeinerte Eigenraum** oder auch **Hauptraum** von A zum Eigenwert λ.

Der folgende Satz soll nun die Frage beantworten, wie groß $\dim(H_f(\lambda))$ ist:

Satz 2.19 (*Zusammenhang Fittingindex und algebraische Vielfachheit*)

Sei $\varphi \in \mathrm{End}(V)$ auf einem $\mathbb{K}$-Vektorraum V mit $\dim(V) < \infty$. Gegeben seien dann die Darsstellungsmatrix $A \in \mathbb{K}^{n \times n}$ der Linearen Abbildung φ und ein Eigenwert λ_* von A mit $\mathrm{geomVfht}(\lambda_*) < \mathrm{algVfht}(\lambda_*) = k$. Dann gebe es zum Eigenwert λ_* genau k linear unabhängige Hauptvektoren von A und somit eine Zahl $f \leqslant n$ mit

$$\mathrm{geomVfht}(\lambda_*) = \dim(H_1(\lambda_*)) \lneqq \dim(H_2(\lambda_*)) < \cdots < \dim(H_f(\lambda_*)) = \dim(H_{f+1}(\lambda_*)) = k.$$

Beweis: Die Existenz einer solchen Zahl f ist durch Korollar 2.13.7 gesichert, da V endlichdimensional ist. Ebenfalls ist die linke Seite der Kette „geomVfht$(\lambda_*) = \dim(H_1(\lambda_*))$“ trivial, da $H_1(\lambda_*)$ genau dem Eigenraum des Eigenwertes λ_* entspricht und die Gleichung somit mit Satz aus dem ersten Semester gilt. Die Kette sei ebenfalls mit Korollar 2.13.7 begründet, genauso wie $\dim(H_i(\lambda_*)) = \dim(H_{i+1}(\lambda_*))$ für $i \geqslant f$. Wir wollen also nun zeigen, dass $\dim(H_f(\lambda_*)) = k$ gilt. Wir nutzen deswegen die Stelle hier um einige Begrifflichkeiten aus dem ersten Semester nocheinmal zu wiederholen respektive zu erweitern:

Definition 2.15 (*Invariante Unterräume, Direkte Summen und Polynomringe*)

(a) Sei $\mathbb{K}$ ein Körper, V ein $\mathbb{K}$-Vektorraum und $\varphi \in \mathrm{Lin}(V,V)$ eine lineare Abbildung. Dann heiße ein Untervektorraum $U \subseteq V$ **φ-invariant** genau dann, wenn gelte, dass

$$\varphi(U) \subseteq U.$$

(b) Es sei $\mathbb{K}$ ein Körper und V ein $\mathbb{K}$-Vektorraum. Es sei $U_1, \ldots, U_m$ eine Familie von Untervektorräumen von V. Man sagt, dass V die direkte Summe der U_i ist, wenn die beiden folgenden Bedingungen erfüllt sind.

- $\forall v \in V.\ \exists u_i \in U_i, 1 \leqslant i \leqslant m.\ v = \sum_{i=1}^{m} u_i$
- $\forall 1 \leqslant i \leqslant m.\ U_i \cap \left(\sum_{j \neq i} U_j \right) = \{0\}$

(c) Ein Polynomring $\mathbb{K}[X]$ über dem Körper $\mathbb{K}$ bestehe aus allen Polynomen der Form

$$P = \sum_{i=0}^{n} a_i \cdot X^i,$$

wobei $n \in \mathbb{N}$ und $a_i \in \mathbb{K}$ für alle i gelten muss. Es ist eine komponentenweise Addition und eine Multiplikation, mit einer distributiven Fortsetzung der Regel

$$X^n \cdot X^m = X^{n+m},$$

auf $\mathbb{K}[X]$ definiert.

(d) Sei $\mathbb{K}$ ein Körper und $\mathbb{K}[X]$ der auf dem Körper $\mathbb{K}$ definierte Polynomring. Ein Polynom $T \in \mathbb{K}[X]$ heiße Teiler eines Polynoms $P \in \mathbb{K}[X]$ genau dann, wenn ein Polynom $Q \in \mathbb{K}[X]$ existiert, so dass

$$P = T \cdot Q$$

gelte.

(e) Seien $P_1, \ldots, P_n \in \mathbb{K}[X]$ Polynome auf dem Polynomring über dem Körper $\mathbb{K}$. Ein weiteres Polynom $T \in \mathbb{K}[X]$ heiße *gemeinsamer Teiler* der Polynome $P_1, \ldots, P_n$ genau dann, wenn T Teiler eines jeden Polynoms P_i mit $1 \leqslant i \leqslant n$ ist.

(f) Seien $P_1, \ldots, P_n \in \mathbb{K}[X]$ Polynome auf dem Polynomring über dem Körper $\mathbb{K}$. $P_1, \ldots, P_n$ heißen *teilerfremd* genau dann, wenn sie außer Konstanten ($c \neq 0$) keine weiteren gemeinsamen Teiler besitzen.

Wir wollen ebenso den Begriff der **adjungierten Matrix** nochmal auffrischen und definieren deswegen:

Definition 2.16 (*Minoren und adjungierte Matrizen*)

Sei $A \in \mathbb{K}^{n\times n}$ mit einem $n \geqslant 2$. Dann heiße die Matrix $A(s,t) \in \mathbb{K}^{(n-1)\times(n-1)}$, welche durch das Streichen der s-ten Zeile und t-ten Spalte von A entstand, ein **Minor** von A.
Man bilde dann die Matrix $B = (b_{i,j})$ mit

$$b_{i,j} = (-1)^{i+j} \cdot \det(A(j,i))$$

für $1 \leqslant i,j \leqslant n$, so heiße B die zu A **adjungierte Matrix** und wird auch gern als $\mathrm{Adj}(A)$ bezeichnet.

Wir stellen in diesem Zusammenhang folgende Aussage auf:

Satz 2.20 (*Satz der Adjungierten*)

Für eine Matrix $A \in \mathbb{K}^{n\times n}$ mit $n \geqslant 2$, ihrer Determinante $\det(A)$, sowie ihrer Adjungierten $\mathrm{Adj}(A)$ gelte

$$\det(A) \cdot E_n = A \cdot \mathrm{Adj}(A) = \mathrm{Adj}(A) \cdot A.$$

Beweis: Sei $C = (c_{i,j}) = \mathrm{Adj}(A) \cdot A$, so gelte

$$\begin{aligned} c_{i,j} &= \sum_{k=1}^{n} b_{i,k} \cdot a_{k,j} \\ &= \sum_{k=1}^{n} (-1)^{i+k} \cdot \det(A(k,i)) \cdot a_{k,j}. \end{aligned}$$

Nun gilt aber ebenso, dass $\det(A(k,i)) = (-1)^{i+k} \cdot \det \widetilde{A}_{k,i}$, wobei $\widetilde{A}_{k,i}$ definiert ist, als

$$\widetilde{A}_{k,i} = \begin{pmatrix} a_{1,1} & \dots & a_{1,i-1} & 0 & a_{1,i+1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{k-1,1} & \dots & a_{k-1,i-1} & 0 & a_{k-1,i+1} & \dots & a_{k-1,n} \\ a_{k,1} & \dots & a_{k,i-1} & 1 & a_{k,i+1} & \dots & a_{k,n} \\ a_{k+1,1} & \dots & a_{k+1,i-1} & 0 & a_{k+1,i+1} & \dots & a_{k+1,n} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & \dots & a_{n,i-1} & 0 & a_{n,i+1} & \dots & a_{n,n} \end{pmatrix},$$

denn durch $(k-1)$ Zeilenvertauschungen und $(i-1)$ Spaltenvertauschungen kann man $\widetilde{A}_{k,i}$ auf die Form

$$\left(\begin{array}{c|c} 1 & * \\ \hline 0 & A(k,i) \end{array}\right)$$

brigen und dafür gilt bekanntermaßen, dass

$$\det(\widetilde{A}_{k,i}) = (-1)^{i-1+k-1} \cdot \det\left(\begin{array}{c|c} 1 & * \\ \hline 0 & A(k,i) \end{array}\right) = (-1)^{i+k} \cdot \det(A(k,i)).$$

Also folgt damit dann unmittelbar, dass

$$\begin{aligned}
c_{i,j} &= \sum_{k=1}^{n} (-1)^{i+k} \cdot \det(A(k,i)) \cdot a_{k,j} \\
&= \sum_{k=1}^{n} (-1)^{i+k} \cdot (-1)^{i+k} \det(\widetilde{A}_{k,i}) \cdot a_{k,j} \\
&= \sum_{k=1}^{n} \det \begin{pmatrix}
a_{1,1} & \dots & a_{1,i-1} & 0 & a_{1,i+1} & \dots & a_{1,n} \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
a_{k-1,1} & \dots & a_{k-1,i-1} & 0 & a_{k-1,i+1} & \dots & a_{k-1,n} \\
a_{k,1} & \dots & a_{k,i-1} & a_{k,j} & a_{k,i+1} & \dots & a_{k,n} \\
a_{k+1,1} & \dots & a_{k+1,i-1} & 0 & a_{k+1,i+1} & \dots & a_{k+1,n} \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
a_{n,1} & \dots & a_{n,i-1} & 0 & a_{n,i+1} & \dots & a_{n,n}
\end{pmatrix} \\
&= \det \left(\begin{array}{ccc|c|ccc}
a_{1,1} & \dots & a_{1,i-1} & a_{1,j} & a_{1,i+1} & \dots & a_{1,n} \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
a_{k-1,1} & \dots & a_{k-1,i-1} & a_{k-1,j} & a_{k-1,i+1} & \dots & a_{k-1,n} \\
a_{k,1} & \dots & a_{k,i-1} & a_{k,j} & a_{k,i+1} & \dots & a_{k,n} \\
a_{k+1,1} & \dots & a_{k+1,i-1} & a_{k+1,j} & a_{k+1,i+1} & \dots & a_{k+1,n} \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
a_{n,1} & \dots & a_{n,i-1} & a_{n,j} & a_{n,i+1} & \dots & a_{n,n}
\end{array}\right) \\
&= \begin{cases} 0 & , \text{ wenn } i \neq j \\ \det(A) & , \text{ wenn } i = j \end{cases}.
\end{aligned}$$

Daraus schließen wir dann schlussendlich, dass $C = \det(A) \cdot E_n$ gelte. Die andere Aussage zeigt man dann analog. □

Neben dem Satz der Adjungierten lernen wir nun einen Satz kennen, der für den folgenden Beweis relativ wichtig ist.

Satz 2.21 (*Satz von Cayley-Hamilton*)

Sei also $\mathbb{K}$ ein Körper und $A \in \mathbb{K}^{n\times n}$ eine Matrix auf $\mathbb{K}$ mit charakteristischem Polynom $p_A(\lambda)$. Dann erfülle A den Zusammenhang

$$p_A(A) = A^n + a_1 \cdot A^{n-1} + \ldots + a_{n-1} \cdot A + a_n \cdot E_n = 0.$$

Beweis: Für $n = 1$ ist der Satz trivial. Sei also Œ $n \geqslant 2$. Dann betrachte man die Matrix $\mathrm{Adj}(\lambda E_n - A)$. Alle Einträge sind Polynome vom Grad $\xi \leqslant n-1$, es sind *Minoren* von $(\lambda E_n - A)$. Wir können also schreiben:

$$\mathrm{Adj}(\lambda E_n - A) = \sum_{i=0}^{n-1} B_i \cdot \lambda^i \quad \text{für } B_i \in \mathbb{K}^{n\times n}.$$

Mit dem Satz der Adjungierten (Satz 2.20) folgt dann

$$(\lambda E_n - A) \cdot \mathrm{Adj}(\lambda E_n - A) = (\lambda E_n - A) \cdot \left(\sum_{i=0}^{n-1} B_i \cdot \lambda^i \right) = p_A(\lambda) \cdot E_n.$$

Daraus schließen wir, dass

$$\underbrace{\sum_{i=0}^{n-1} B_i \cdot \lambda^{i-1}}_{=\sum_{i=1}^{n} B_{i-1}\cdot\lambda^i} - \sum_{i=0}^{n-1} A B_i \lambda^i = p_A(\lambda) \cdot E_n.$$

Koeffizientenvergleich liefert dann:

$$\underbrace{\begin{array}{lrcrcll|l}
(\lambda^n): & B_{n-1} & & & = & E_n & \cdot A^n \\
(\lambda^{n-1}): & B_{n-2} & - & AB_{n-1} & = & a_1 E_n & \cdot A^{n-1} \\
\vdots & \vdots & & \vdots & & \vdots & \\
(\lambda^1): & B_0 & - & AB_1 & = & a_{n-1} E_n & \cdot A^1 \\
(\lambda^0): & & - & AB_0 & = & a_n E_n & \cdot A^0 = E_n
\end{array}}_{\Downarrow}$$

$$\overbrace{\begin{array}{lrcrcll}
(\lambda^n): & A^n B_{n-1} & & & = & & A^n \\
(\lambda^{n-1}): & A^{n-1} B_{n-2} & - & A^n B_{n-1} & = & a_1 & A^{n-1} \\
\vdots & & \vdots & & \vdots & & \vdots \\
(\lambda^1): & AB_0 & - & A^2 B_1 & = & a_{n-1} & A \\
(\lambda^0): & & - & AB_0 & = & a_n & E_n
\end{array}}$$

Eine Addition der Gleichungen liefert dann die zu zeigende Aussage mit

$$\underbrace{A^n + a_1 \cdot A^{n-1} + \ldots + a_{n-1} \cdot A + a_n \cdot E_n}_{= p_A(A)} = 0.$$

□

Korrolar 2.21

Sei $\varphi \in \mathrm{Lin}(V,V)$ mit V einem endlichdimensionalen $\mathbb{K}$-Vektorraum, so gelte für das charakteristische Polynom, dass

$$p_\varphi(\varphi) = 0.$$

Seien noch folgende Sätze bekannt:

Satz 2.22 (*Satz über die teilerfremde Zerlegung*)

Sei $\varphi \in \mathrm{Lin}(V,V)$, wobei V ein $\mathbb{K}$-Vektorraum sei und $\mathbb{K}$ ein Körper, sowie $p_A = P \cdot Q$ eine Faktorzerlegung des charakteristischen Polynoms in *teilerfremde* Polynome $P, Q \in \mathbb{K}[X]$. So gelte die direkte Summenzerlegung

$$V = \ker(P(A)) \oplus \ker(Q(A)),$$

wobei diese Räume φ-invariant sind. Die Einschränkung $P(\varphi)\big|_{\ker(Q(\varphi))}$ ist dann bijektiv.

Beweis: Nach dem Lemma von Bezout[2] (Lemma 3.11) gibt es Polynome $S, T \in \mathbb{K}[X]$ mit

$$S \cdot P + T \cdot Q = 1.$$

Sei $U = \ker P(\varphi)$ und $W = \ker Q(\varphi)$, sowie $v \in V$. Nach dem Satz von Cayley-Hamilton (Satz 2.21) und dessen Korollar gelte somit

$$0 = p_\varphi(\varphi) = (P(\varphi) \circ Q(\varphi))(v) = P(\varphi)(Q(\varphi)(v)),$$

[2]Das Lemma von Bezout wird in Kapitel 3 noch eine große Rolle spielen. Da die Aussage hier aber sowieso nur eine Hilfsaussage ist, verschieben wir den Beweis des Lemmas auf einen späteren Zeitpunkt.

und somit ist $\operatorname{im} Q(\varphi) \subset \ker P(\varphi)$ und umgekehrt. Aus

$$\begin{aligned} v &= \mathrm{Id}_V(v) \\ &= (SP + TQ)(\varphi)(v) \\ &= S(\varphi)(P(\varphi)(v)) + T(\varphi)(Q(\varphi)(v)) \\ &= P(\varphi)(S(\varphi)(v)) + Q(\varphi)(T(\varphi)(v)) \end{aligned}$$

ist abzulesen, dass der linke Summand zu $\operatorname{im} P(\varphi) \subseteq \ker(Q(\varphi))$, der rechte zu $\operatorname{im} Q(\varphi) \subseteq \ker P(\varphi)$ gehört. Es liegt also eine Summenzerlegung vor, welche auch direkt ist, da auch $P(\varphi)(v) = Q(\varphi)(v) = 0$ sofort $v = 0$ folgt. Dass die Räume φ-invariant sind ist trivial. Zu $v \in \ker Q(\varphi)$ ist

$$v = S(\varphi)(P(\varphi)(v)) + T(\varphi)(Q(\varphi)(v)) = S(\varphi)(P(\varphi)(v)) = P(\varphi)(S(\varphi)(v)),$$

das heißt es gelte $\operatorname{im} P(\varphi) = \ker Q(\varphi)$ und somit ist $P(\varphi)\big|_{\ker(Q(\varphi))}$ auch surjektiv, damit bijektiv. □
Wir stellen nun noch einen letzten Satz vor, welchen wir anschließend im Beweis des eigentlichen Satzes 2.19 verwenden wollen:

Satz 2.23 (*Satz der direkten Summenzerlegung*)

Sei V ein endlichdimensionaler $\mathbb{K}$-Vektorraum und $\varphi \in \mathrm{Lin}(V,V)$. Es sei $V = U \oplus W$ eine direkte Summenzerlegung von V in φ-invariante Unterräume. Dann gelte für das charakteristische Polynom

$$p_\varphi(\lambda) = p_{\varphi|_U}(\lambda) \cdot p_{\varphi|_W}(\lambda)$$

Beweis: Sei $\{u_1, \ldots, u_k\}$ eine Basis von U und $\{w_1, \ldots, w_m\}$ eine Basis von W, so dass $\mathcal{B} = \{u_1, \ldots, u_k, w_1, \ldots, w_m\}$ eine Basis von V ist. Bezüglich der Basis $\mathcal{B}$ wird φ durch die Blockmatrix $M = \begin{pmatrix} C & 0 \\ 0 & D \end{pmatrix}$ beschrieben, wobei C $\varphi|_U$ und D $\varphi|_W$ beschreibt. Damit gilt nach Satz aus C1, dass $p_\varphi(\lambda) = p_M(\lambda) = \det(\lambda E_n - M) \overset{*}{=} \det(\lambda E_n - C)\det(\lambda E_n - D) = p_{\varphi|_U}(\lambda) \cdot p_{\varphi|_W}(\lambda)$ □
Wir kehren nun zum Beweis des eigentlichen Satzes zurück. Wir schreiben also das charakteristische Polynom zu φ als $p_\varphi(\lambda) = (\lambda - \lambda_*)^k \cdot Q$, wobei $(\lambda - \lambda_*)^k$ in Q nicht als Linearfaktor auftaucht. Das heißt ebenso $k = \mathrm{algVfht}(\lambda_*)$. Dann sind $P = (\lambda - \lambda_*)^k$ und Q teilerfremd und mit Satz 2.22 dann

$$V = \ker P(\varphi) \oplus \ker Q(\varphi)$$

und

$$P(\varphi) = (\varphi - \lambda_* \cdot E_n)^k : \ker Q(\varphi) \to \ker Q(\varphi)$$

ist eine Bijektion. Es sei ferner $H_f(\xi) = \ker P(\varphi)$, wobei „$\supseteq$“ klar ist und sich „$\subseteq$“ daraus ergibt, dass höhere Potenzen von $(\lambda - \lambda_*)$ wegen der eben erwähnten Bijektivität auf $\ker Q(\varphi)$ keine weiteren Elemente annulieren. Mit Satz 2.23 folgt dann für p_λ, dass

$$p_\lambda = p_1(\lambda) \cdot p_2(\lambda),$$

wobei p_1 das charakteristische Polynom zu $\varphi|_{H_f}$ und p_2 zu $\varphi|_{\ker Q(\varphi)}$ ist. Da $(\varphi - \lambda_*)^k$ auf H_f eine Nullabbildung ist, ist das Minimalpolynom zu $\varphi|_{H_f}$ und damit auch das charakteristische Polynom p_1 eine Potenz von $(\lambda - \lambda_*)$, wir setzten

$$p_1 = (\lambda - \lambda_*)^d, \text{ wobei } d = \dim(H_f(\lambda_*)) \text{ sei.}$$

Damit gilt auch automatisch $d \leqslant k$, da p_1 ein Teiler von p_φ ist. Gelte nun $d < k$, so müsste λ_* eine Nullstelle von p_2 sein und λ_* wäre ein Eigenwert von $\varphi|_{\ker(Q(\varphi))}$. Dies ist aber ein Widerspruch zu der Aussage, dass $P(\varphi)$ auf diesem Raum eine Bijektion ist. □
Gelte zudem aber noch folgender Satz:

Satz 2.24 (*Lineare Unabhängigkeit und de-facto Disjunktheit*)

Es sei $A \in \mathbb{K}^{n\times n}$ gegeben.

(a) Ist v ein Hauptvektor der Stufe $j+1$ zum Eigenwert λ, so ist das Vektorsystem bestehend aus $d_0 = v$, $d_1 = \ker(A-\lambda E_n)v$, …, $d_j = \ker(A-\lambda E_n)^j v$ linear unabhängig.

(b) Für je zwei Eigenwerte $\lambda \neq \mu$ der Matrix A mit Indizes f und g gilt $H_f(\lambda) \cap H_g(\mu) = \{0\}$.

Beweis:

(a) Angenommen die Vektoren wären nicht linear unabhängig, so gäbe es mindestens ein $c_i \neq 0$, so dass

$$\sum_{k=0}^{j} c_k v_k = 0.$$

Sei in dieser Gleichung ℓ der größte Index mit $c_\ell \neq 0$. Dann lässt sich der Hauptvektor d_ℓ der Stufe ℓ als Linearkombination von Hauptvektoren kleinerer Stufen schreiben. Gleichzeitig wäre dann aber $\ker(A-\lambda E_n)^{\ell-1} = 0$ im Widerspruch zur Definition eines Hauptvektors der Stufe ℓ. Folglich kann der Nullvektor nur als triviale Linearkombination geschrieben werden, das Vektorsystem ist linear unabhängig.

(b) Sei oBdA. $f \geqslant g$. Wäre nun $0 \neq v \in (\ker(H_f(\lambda)) \cap \ker(H_g(\mu))$, so gäbe es eine Zahl $0 \leqslant j \leqslant f-1$ mit $H_{j+1}(\lambda)v - 0$ und $H_j(\lambda)v \neq 0$. Dann wäre aber auch

$$0 = H_g(\mu)v = ((\lambda-\mu)E_n + (A-\lambda E_n))^g v = \sum_{i=0}^{g} \binom{g}{i} \cdot (\lambda-\mu)^{g-i} H_i(\lambda)v = \sum_{i=0}^{\min\{g,j\}} \xi_i H_i(\lambda)v,$$

wobei $\xi_i := \binom{g}{i} \cdot (\lambda-\mu)^{g-i} \neq 0$ gilt, was in Widerspruch zu Aussage (a) steht.

□

Wir erhalten damit dann schlussendlich die Aussage:

Satz 2.25 (*Aufspannen des* $\mathbb{C}^n$)

Sind $\lambda_1, \ldots, \lambda_m$ die paarweise verschiedenen Eigenwerte der Matrix $A \in \mathbb{C}^{n\times n}$ mit Vielfachheit $k_1, \ldots, k_m$ und bezeichnen $E(\lambda_j) := \ker(A-\lambda_j E_n)^{f(\lambda_j)}$ die Haupträume von A, so gilt die direkte Zerlegung

$$\mathbb{C}^n = \bigoplus_{i=1}^{m} E(\lambda_i),$$

das heißt der Vektorraum $\mathbb{C}^n$ besitzt eine Basis aus Hauptvektoren von A.

Beweis: Satz 2.24(b) stellt zunächst die Direktheit der Summe $E(\lambda_1) \oplus E(\lambda_2) =: U$ sicher. Wir zeigen nun, dass auch $U \cap E(\lambda_3) = \{0\}$ gilt, woraus sich die Direktheit der Summe $E(\lambda_1) \oplus E(\lambda_2) \oplus E(\lambda_3)$ ergibt und per Induktion somit die Behauptung, da ja $\dim(E_j) = k_j$ und $\sum_{j=1}^{m} k_j = n$.
Wäre also nun $U \cap E(\lambda_3) \neq \{0\}$, so gäbe es ein $0 \neq v \in E(\lambda_3)$, sowie $w_1 \in E(\lambda_1)$ und $w_2 \in E(\lambda_2)$ mit $v = w_1 + w_2 \in U$. Sei $g := \max\{f(\lambda_1), f(\lambda_2), f(\lambda_3)\}$. Dann wäre

$$\begin{aligned} 0 &= H_g(\lambda_3)v = ((\lambda_1-\lambda_3)E_n + H_1(\lambda_1))^g w_1 + ((\lambda_2-\lambda_3)E_n + H_1(\lambda_2))^g w_2 \\ &= \sum_{i=0}^{g} \xi_i H_i(\lambda_1)w_1 + \sum_{i=0}^{g} \eta_i H_i(\lambda_2)w_2 =: z_1 + z_2 \end{aligned}$$

mit $\xi_i := \binom{g}{i} \cdot (\lambda_1-\lambda_3)^{g-i} \neq 0 \neq \binom{g}{i} \cdot (\lambda_2-\lambda_3)^{g-i} =: \eta_i$ und $z_1 \in E(\lambda_1)$, sowie $z_2 \in E(\lambda_2)$. Wegen der Direktheit der Summe von $E(\lambda_1)$ und $E(\lambda_2)$ folgt aber hieraus, dass $z_1 = 0 = z_2$, was allerdings im Widerspruch zu Satz 2.24(a) steht. □

Der abschließende Satz soll nun noch den Zusammenhang zwischen Hauptvektoren und dem Finden eines Fundamentalsystems klarstellen.

Satz 2.26 (*Fundamentalsystem im nicht-diagonalisierbaren Fall*)

Sei $A \in \mathbb{K}^{n\times n}$ und $\lambda \in \mathbb{C}$ ein Eigenwert von A. Seien $v_1, \ldots, v_k$ eine „Kette“ von Hauptvektoren zum Eigenwert λ, heißt v_j ist der Hauptvektor der Stufe j zum Eigenwert λ, es gilt (nach Korollar 2.13.2):

$$(A - \lambda E_n)v_{j+1} = v_j \tag{2.40}$$

(a) Dann ist

$$y(x) := \exp(\lambda x) \cdot \left(\sum_{i=0}^{k-1} \frac{x^i}{i!} v_{k-i} \right) \tag{2.41}$$

eine Lösung des Differentialgleichungssystems

$$y' = Ay \tag{2.42}$$

(b) Es gibt dann ein Fundamentalsystem für (2.42), welches aus Funktion der Bauart (2.40)/(2.41) besteht. Jeder Eigenwert $\lambda \in \mathbb{C}$ trägt dabei genau $k := \text{algVfht}(\lambda)$ Fundamentallösungen bei.

Beweis:

(a) Wir setzen dazu (2.41) in (2.42) ein:

$$Ay - y' = \exp(\lambda x) \cdot \left(\sum_{i=0}^{k-1} \frac{x^i}{i!} A v_{k-i} \right) - \lambda \exp(\lambda x) \cdot \left(\sum_{i=0}^{k-1} \frac{x^i}{i!} v_{k-i} \right) + \exp(\lambda x) \cdot \left(\sum_{i=0}^{k-1} \frac{i \cdot x^{i-1}}{i!} v_{k-i} \right)$$

Nach Potenzen sortiert ergibt sich:

$$y' - Ay = \exp(\lambda x) \cdot \left(\sum_{i=0}^{k-2} \frac{x^i}{i!} \cdot \underbrace{(Av_{k-i} - \lambda v_{k-1} - v_{k-i-1})}_{\substack{=(A-\lambda E_n)v_{k-i}-v_{k-i-1} \\ = 0 \text{ nach (2.40)}}} + \frac{x^{m-1}}{(m-1)!} \overbrace{(Av_1 - \lambda v_1)}^{= 0, \text{ da Eigenvektor}} \right) = 0$$

(b) Folgt nahezu unmittelbar aus Korollar 2.13.2 und den Sätzen 2.19 – 2.25

□

Fazit

Mit der Hilfe von Hauptvektoren können wir immer Fundamentalsysteme berechnen. Wir brauchen die Hauptvektoren allerdings (nur) dann, wenn es nicht „genug“ linear unabhängige Eigenvektoren gibt.

2.4.2.4 Fundamentalsysteme bei Jordanmatrizen

Neben den eben vorgestellten Fällen gibt es auch andere Matrixformen, bei der die Hauptvektorbestimmung einfacher sein kann. Wir wollen deswegen hier nur kurz auf Jordanmatrizen und die Jardan'schen Normalformen eingehen. Wir definieren deswegen an dieser Stelle:

Definition 2.17 (*Jordanblock, Jordan'sche Normalform*)

Für ein $r \geqslant 1$ und ein $\lambda \in \mathbb{K}$ heiße die Matrix

$$J_r(\lambda) = J_{r,\lambda} = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \lambda & 1 \\ & & & \lambda \end{pmatrix} = (\iota_{i,j}) = \begin{cases} \lambda & \text{, wenn } i = j \\ 1 & \text{, wenn } j = i+1 \wedge 1 \leqslant i \leqslant r-1 \\ 0 & \text{, sonst} \end{cases} \tag{2.43}$$

mit $J_r(\lambda) \in \mathbb{K}^{r\times r}$ ein **Jordanblock** der Größe r zum Eigenwert λ. Wir bezeichnen ein J, welches zu einer Matrix $A \in \mathbb{K}^{n\times n}$ mit den paarweise verschiedenen Eigenwerten λ_i, $1 \leqslant i \leqslant m \leqslant n$ und deren Vielfachheiten k_i, $1 \leqslant i \leqslant m$ mit $k_1 + \cdots + k_m = n$ ähnlich ist un dessen Form

$$J := \begin{pmatrix} J_{d_1}(\lambda_1) & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_{d_r}(\lambda_r) \end{pmatrix} \in \mathbb{K}^{n\times n} \tag{2.44}$$

mit $\lambda_1, \dots, \lambda_r \in \mathbb{K}$, $d_1, \dots, d_r \in \mathbb{N}$ und $r \geqslant m$, sowie $\lambda_1, \dots, \lambda_r$ sind nicht unbedingt paarweise verschieden und $J_{d_i}(\lambda_i)$ ist definiert nach (2.43), entspricht als in **Jordannormalform** stehend. Die geometrische Vielfachheit eines Eigenwertes λ_i entspricht genau der Anzahl an Jordanblöcken zum Eigenwert λ_i in J. Das charakteristische Polynom von J sei dann durch

$$p_J(\lambda) = (-1)^n \cdot \prod_{i=1}^{r} (\lambda - \lambda_i)^{d_i}$$

definiert.

Anmerkung

Ohne Beweis sei an dieser Stelle angemerkt, dass eine **jede** Matrix $A \in \mathbb{K}^{n\times n}$ eine solche Jordannormalform besitzt, welche bis auf das Tauschen der Jordanblöcke eindeutig ist.

Die einzige Frage, die es jetzt noch zu beantworten gilt, ist warum das Finden von Fundamentalsystemen bei bekannten Jordanformen leichter ist. Wir können nämlich nun die Fundamentalmatrix aus den einzelenen Teilen zusammensetzen, sprich für einen jeden Jordanblock $J_r(\lambda)$ ergibt sich eine Fundamentalmatrix

$$W(J_r(\lambda)) = (w_{ij}) = \begin{pmatrix} \exp(\lambda x) & x\exp(\lambda x) & \cdots & \frac{x^{r-1}}{(r-1)!}\exp(\lambda x) \\ 0 & \exp(\lambda x) & \cdots & \frac{x^{r-2}}{(r-2)!}\exp(\lambda x) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \exp(\lambda x) \end{pmatrix} = \begin{cases} 0 & \text{, falls } i > j \\ \frac{x^{j-i}}{(j-i)!} \cdot \exp(\lambda x) & \text{, falls } i \leqslant j \leqslant r \end{cases}.$$

Die Fundamentalmatrix $W(J)$ ergibt sich dann durch das Zusammenstecken der einzelnen Fundamentalblöcke:

$$W(J) = \begin{pmatrix} W(J_{d_1}(\lambda_1)) & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & W(J_{d_r}(\lambda_r)) \end{pmatrix}$$

Um eine Fundamentallösung **für A** auszurechnen, muss nun die Fundamentallösung von J mit der Ähnlichkeitsmatrix multipliziert werden, es ergibt sich also

$$y_i(x) = C \cdot z_i(x), \text{ wobei } 1 \leqslant i \leqslant d_r,\ z_i \text{ der Spaltenvektor von } W(J) \text{ und } J = C^{-1}AC.$$

Wir behandeln hier allerdings Jordanmatrizen nur am Rande aufgrund ihrer besonders schlechten Kondition, weswegen sie in der numerischen Mathematik meistens vermieden werden.

2.4.3 Teilaufgabe (P) — Bestimmung einer partikulären Lösung

Wir beschäftigen uns nun wieder mit einem inhomogenen System an Differentialgleichungen erster Ordnung (wie in (2.33)). Nach Satz 2.13 brauchen wir nun noch **eine** Lösung y_P des inhomogenen Systems, um die Lösungsgesamtheit zu kennen. Wir wollen nun voraussetzen, dass ein Fundamentalsystem bereis gefunden wurde, wir bezeichnen dies als $\{y_1,\dots,y_n\}$.

Kleine Notiz am Rande

Ab hier darf die Systemmatrix A auch durchaus wieder von der Veränderlichen abhängen, da wir voraussetzen ein Fundamentalsystem bereits gefunden zu haben. Zugegeben könnte diese Aufgabe dann wieder schwierig sein bei einer Abhängigkeit von der Veränderlichen.

Den Ansatz um nun auf eine partikuläre Lösung von (2.33) zu kommen kennen wir bereits aus Kapitel 2.2, wir **variieren der Konstanten** *(VdK)*. Es kommt also zu folgendem Ansatz:

$$y_P(x) = \sum_{i=1}^{n} c_i(x) y_i(x) = W(x)c(x) \tag{2.45}$$

Wir setzen dies nun – zur Bestimmung der c_i – in (2.33) ein und erhalten:

$$y_P'(x) = A(x)y_P(x) + b(x) \overset{(2.45)}{\Leftrightarrow} \sum_{i=1}^{n} c_i'(x)y_i(x) + \sum_{i=1}^{n} c_i(x)\underline{y_i'(x)} = \sum_{i=1}^{n} c_i(x)\underline{Ay_i(x)} + b(x)$$

Da die y_i Lösungen der homogenen Differentialgleichung sind, fallen die unterstrichenen Summen weg (sie sind nach (2.34) identisch), man erhält zum Schluss

$$\underbrace{\sum_{i=1}^{n} c_i'(x)y_i(x)}_{= W(x)c'(x)} = b(x)$$

Wir erhalten damit folgenden Satz:

Satz 2.27 (*Partikuläre Lösung eines inhomogenen Differentialgleichungssytems erster Ordnung*)

Seien die Matrixfunktion $A \in \mathrm{Abb}(\mathbb{R}, \mathbb{R}^{n\times n})$, sowie die Vektorfunktion $b \in \mathrm{Abb}(\mathbb{R}, \mathbb{R}^n)$ mit $b \not\equiv 0$ gegeben und stetig, und sei daraus das Differentialgleichungssystem

$$y'(x) = A(x)y(x) + b(x) \tag{2.33 wiederholt}$$

definiert. Beschreibe $\{y_1,\dots,y_n\}$ ein Fundamentalsystem der Systemmatrix, so ist eine partikuläre Lösung des Differentialgleichungssystems gegeben durch

$$y_P(x) := W(x)c(x), \tag{2.46}$$

wobei sich $c(x)$ aus der trivialen Differentialgleichung

$$c'(x) := W\text{^}{-1}(x)b(x) \tag{2.47}$$

bestimmt.

Beweis: Dieser Satz ergibt sich genau so aus der obigen Herleitung. □

Beispiel 2.14: Bestimmen Sie die allgemeine Lösung von $z' = \begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix} = Az + \zeta = \begin{pmatrix}1 & 1\\ 4 & 1\end{pmatrix}\begin{pmatrix}x(t)\\ y(t)\end{pmatrix} + \begin{pmatrix}2\\ -1\end{pmatrix}\exp(t)$. Wir lösen dazu wieder die zwei Probleme (H) und (P):

(H): Wir bestimmen also das Fundamentalsystem zur Matrix A und damit zuerst deren Eigenwerte und -vektoren. Wir stellen dazu wiederum das charakteristische Polynom auf und bestimmen davon die Nullstellen:

$$p(\lambda) = \det \begin{vmatrix} 1-\lambda & 1 \\ 4 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - 4 = \lambda^2 - 2\lambda - 3$$

Die Nullstellen $\lambda_{1,2}$ bestimmen sich nun über die Mitternachtsformel, wir erhalten

$$\lambda_{1,2} = \frac{2 \pm \sqrt{4 - 4 \cdot 1 \cdot (-3)}}{2 \cdot 1} = \frac{2 \pm 4}{2} = 1 \pm 2 = \begin{cases} 3 = \lambda_1 \\ -1 = \lambda_2 \end{cases}$$

Wir bestimmen daraufhin die Eigenvektoren zu den beide Eigenwerten:

$\lambda_1 = 3$: Wir lösen also das Lineare Gleichungssystem

$$\begin{pmatrix} -2 & 1 \\ 4 & -2 \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

$$\left(\begin{array}{cc|c} -2 & 1 & 0 \\ 4 & -2 & 0 \end{array}\right) \begin{array}{l} \\ \cdot 2 \\ \leftarrow + \end{array} \quad \rightsquigarrow \quad \left(\begin{array}{cc|c} -2 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right) \quad \rightsquigarrow \quad v_{\lambda_1} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$\lambda_2 = -1$: Wir lösen also das Lineare Gleichungssystem

$$\begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

$$\left(\begin{array}{cc|c} 2 & 1 & 0 \\ 4 & 2 & 0 \end{array}\right) \begin{array}{l} \\ \cdot(-2) \\ \leftarrow + \end{array} \quad \rightsquigarrow \quad \left(\begin{array}{cc|c} 2 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right) \quad \rightsquigarrow \quad v_{\lambda_2} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

Wir erkennen, dass die Eigenvektoren für eine Bestimmung des Fundamentalsystems ausreichen, eine weitere Bestimmung von Hauptvektoren ist somit nicht notwendig. Wir stellen das Fundamentalsystem und damit dann die Wronskideterminante mit

$$FS = \left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \exp(3 \cdot t), \begin{pmatrix} 1 \\ -2 \end{pmatrix} \cdot \exp(-t) \right\}, W(t) = \begin{pmatrix} \exp(3 \cdot t) & \exp(-t) \\ 2\exp(3 \cdot t) & -2\exp(-t) \end{pmatrix}$$

auf.

(P): Nun bestimmen wir eine partikuläre Lösung der Differentialgleichung. Wir lösen dazu eine Differentialgleichung, wie sie in (2.47) gegeben ist. Dazu müssen wir an dieser Stelle zuerst W^{-1} berechnen.[3]

$$W^{-1}(t) = \frac{1}{-4 \cdot \exp(2t)} \cdot \begin{pmatrix} -2\exp(-t) & -\exp(-t) \\ -2\exp(3 \cdot t) & \exp(3 \cdot t) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\exp(-3t) & \frac{1}{4}\exp(-3t) \\ \frac{1}{2}\exp(t) & -\frac{1}{4}\exp(t) \end{pmatrix}$$

Eingesetzt in (2.47) ergibt sich somit

$$c'(t) = W^{-1}(t) \cdot b(t) = \begin{pmatrix} \frac{3}{4}\exp(-2t) \\ \frac{5}{4}\exp(2t) \end{pmatrix},$$

[3] Selbstverständlich ist es einfacher und vor allen bei größeren Matrizen numerisch gesehen besser, wenn wir den Zusammenhang (2.47) als zu lösendes lineares Gleichungssystem auffassen. Da aber W hier nur eine 2×2-Matrix ist, brauchen wir auf diese Lösungsart vorerst nicht eingehen.

lösen wir diese Differentialgleichung durch einfaches Integrieren, erhalten wir

$$c(t) = \begin{pmatrix} -\frac{3}{8}\exp(-2t) \\ \frac{5}{8}\exp(2t) \end{pmatrix}.$$

Die partikuläre Lösung erhalten wir dann durch das Einsetzen in (2.46):

$$x_P(t) = W(t) \cdot c(t) = \begin{pmatrix} \frac{1}{4}\exp(t) \\ -2\exp(t) \end{pmatrix}.$$

Durch das Zusammensetzen der beiden Lösungen erhalten wir den Raum der inhomogenen Lösungsmenge mit

$$L_{inhom} = \left\{ c_1, c_2 \in \mathbb{R} \,:\, x(t) = c_1 \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \exp(3 \cdot t) + c_2 \cdot \begin{pmatrix} 1 \\ -2 \end{pmatrix} \cdot \exp(-t) + \frac{1}{4} \cdot \begin{pmatrix} 1 \\ -8 \end{pmatrix} \exp(t) \right\}.$$

※

2.5 Lineare skalare Differenzialgleichungen höherer Ordnung

Wir wollen nun noch einmal kurz folgende Art von Differentialgleichungen betrachten:

Definition 2.9 (*Lineare Differentialgleichungen*)

Sei $I \subset \mathbb{R}$ ein Intervall und seien $a_0, \ldots, a_n, b \in \mathrm{Abb}(I, \mathbb{R})$ stetig. Dann heiße eine Differentialgleichung der Form

$$\sum_{i=0}^{n} a_i(x) y^{(i)} = b(x) \tag{2.1}$$

eine **lineare Differentialgleichung n-ter Ordnung**. Ist das sogenannte **Störglied** $b(x)$ null – sprich für alle $x \in I$ gelte, dass $b(x) = 0$ – so nennt man die Differentialgleichung **homogen**, andernfalls **inhomogen**.
Man nennt die Differentialgleichung *normiert* genau dann, wenn $a_n(x) :\equiv 1$.

Auch wenn wir solche speziellen Differentialgleichungen mit Lemma 2.1 und Kapitel 2.4 bereits lösen können, wollen wir uns diesen Prozess einmal genauer anschauen. Denn wir werden feststellen, dass viele rechenaufwändige Schritte ausgelassen werden können und damit sogar leicht exakte Lösungen von Differentialgleichungen mit von der Veränderlichen abhängigen Koeffizienten möglich ist.
Lemma 2.1 liefert dann für die Systemmatrix A des zugehörigen Systems

$$A(x) = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 & \cdots & \cdots & \cdots & -a_{n-1} \end{pmatrix} \text{ und für die Vektorfunktion } b(x) := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b(x) \end{pmatrix}.$$

Wir wissen allerdings ebenso, dass von den Fundamentallösungen – wegen des skalaren Problems – nur die erste Komponente relevant ist. Sei nun also angenommen, dass die a_i konstant sind, so brauchen wir um Eigenwerte und -vektoren zu berechnen das charakteristische Polynom von A. Interessanterweise gilt der folgende Satz, der eine Berechnung desselbigen stark vereinfacht:

Satz 2.28 (*Charakteristisches Polynom von Systemmatrizen lin. skalarer DGL'en n-ter Ord.*)

Sei eine Matrix $A \in \mathbb{R}^{n \times n}$ mit $n \geqslant 2$ der Form

$$A(x) = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ \end{pmatrix}$$

gegeben, so hat sie das charakteristische Polynom

$$p_n(\lambda) = (-1)^n \cdot \left(\lambda^n + \sum_{i=0}^{n-1} a_i \lambda^i\right) \tag{2.48}$$

Beweis: Wir zeigen dies durch vollständige Induktion.

IA $(n = 2)$:

$$A_2 = \begin{pmatrix} 0 & 1 \\ -a_0 & -a_1 \end{pmatrix} \Rightarrow p_2(\lambda) = \lambda(a_1 + \lambda) + a_0 = \lambda^2 + a_1\lambda + a_0 \qquad \checkmark$$

IS $(n \to n+1)$:
Gelte nun für p_n einer Systemmatrix A der obigen Form bereits der Zusammenhang (2.48). Dann gelte für p_{n+1}:

$$\begin{aligned} p_{n+1}(\lambda) &= \det(A_{n+1} - \lambda E_{n+1}) = \det \begin{vmatrix} -\lambda & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\lambda & 1 \\ -a_0 & & \dots & -a_{n-1} & -a_n - \lambda \end{vmatrix} \\ &\overset{(*)}{=} (-\lambda) \cdot \det \begin{vmatrix} -\lambda & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\lambda & 1 \\ -a_1 & & \dots & -a_{n-1} & -a_n - \lambda \end{vmatrix} - (-1)^{n+2} \cdot a_0 \cdot \det \begin{vmatrix} 1 & & & & \\ -\lambda & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ddots & 1 & \\ & & & -\lambda & 1 \end{vmatrix} \\ &\overset{IV}{=} (-\lambda) \cdot (-1)^n \cdot \left(\lambda^n + \sum_{i=1}^{n} a_i \lambda^{i-1}\right) + (-1)^n \cdot (-a_0) = (-1)^{n+1} \cdot \left(\lambda^{n+1} + \sum_{i=0}^{n} a_i \lambda^i\right), \end{aligned}$$

wobei an Stelle $(*)$ der Laplace'sche Entwicklungssatz bzgl. der ersten Spalte angewendet wurde.

□

Als nächstes müsste man eigentlich die Eigenräume (und eventuell Haupträume) von A bestimmen. In allgemeiner Form ist dies nachgewiesen schwierig. Aber wie eingangs bereits erwähnt benötigen wir nur die erste Komponente. Nimmt man nun von $y = p(x) * \exp(\lambda x)$ die erste Komponente, so erhält man eine skalare Funktion der Form $y_s(x) = p_s(x)\exp(\lambda x)$, wobei $\lambda \in \mathbb{C}$ Eigenwert von A und p ein Polynom von Grad $g \leqslant \text{algVfht}(\lambda)$ ist. Betrachtet man nun noch einmal (2.41) genauer, so kann man feststellen, dass eine Wahl von p als nacheinander $1, x, \dots, x^{r-1}$ mit $r = \text{algVfht}(\lambda)$ günstig ist, da sich einerseits alle algebraischen Vielfachheiten aller Eigenwerte in den komplexen Zahlen immer zu n aufaddieren und andererseits die so entstehenden Funktionen immer linear **un**abhängig sind. Wir erhalten somit n linear unabhängige Lösungen des homogenen Problems. Wir fassen diese Erkenntnisse in einem Satz zusammen:

Satz 2.29 (*Lösungsraum der homogenen skalaren linearen Differentialgleichungen n-ter Ord.*)

Sei g eine Differentialgleichung der Form (2.1) mit konstanten Koeffizienten a_i, welche in homogener Version vorliegt, so hat g folgende Basis $\mathcal{B}$ des homogenen Lösungsraumes L_{hom}:
Für jede Nullstelle des charakteristischen Polynoms (2.48) $\lambda_i \in \mathbb{C}$, mit Vielfachheit r_i, nehme die Funktionen

$$\exp(\lambda_i x), x \cdot \exp(\lambda_i x), \dots, x^{r_i - 1} \cdot \exp(\lambda_i x)$$

zur Basis mit auf.

Beweis: Der Satz ergibt sich aus den Vorüberlegungen oben. □

Wir stellen fest, dass das Finden einer Basis von L_{hom} und damit das Finden von homogenen Lösungen von (2.1) ist für lineare **skalare** Differentialgleichungen **n**-ter Ordnung erheblich „einfacher“ als bei linearen *Systemen erster* Ordnung, da die Eigen- und Hauptvektoren der Systemmatrix nicht explizit ausgerechnet werden müssen.

Was passiert bei Eigenwerten $\lambda_i \in \mathbb{C}\backslash\mathbb{R}$?

Wir haben uns bereits gegen Ende von Kapitel 2.4.2.2 dieselbe Frage gestellt und gesehen, dass, sofern eine Matrix A reell ist und sie einen komplexen Eigenwert hat, dieser immer in Verbindung mit dem konjugiert komplexen Partner auftritt. Wir stellen durch gleiche Überlegungen fest, dass auch hier dann Satz 2.18 zutrifft, also sollte $\lambda_i = a + b \cdot i$ komplexer Eigenwert der reellen Matrix A sein, so auch $\overline{\lambda_i} = a - b \cdot i$ und wir erhalten als reelle Fundamentallösungen

$$y_{1/2,\text{reell}} = \frac{1}{2}(y_1(x) \pm y_2(x)) = \begin{cases} \exp(ax) \cdot \cos(bx) = \Re(y_1(x)) \\ \exp(ax) \cdot \sin(bx) = \Im(y_1(x)) \end{cases}$$

zu den komplexen Fundamentallösungen

$$y_{1/2} = \exp((a \pm bi) \cdot x).$$

Beispiel 2.15: Bestimmen Sie ein Fundamentalsystem für $y^{(4)} + 6y''' + 12y'' + 10y' + 3y = 0$.
Wir stellen dazu nach Satz 2.28 das charakteristische Polynom mit

$$p(\lambda) = \lambda^4 + 6\lambda^3 + 12\lambda^2 + 10\lambda + 3$$

auf und suchen die Nullstellen des Polynoms. Durch „cleveres Raten“ erhalten wir $\lambda_1 = -3$ als Nullstelle von $p(\lambda)$; wir erhalten als Restpolynom:

$$\begin{array}{rl}
(\ \lambda^4 + 6\lambda^3 + 12\lambda^2 + 10\lambda + 3) \div (\lambda + 3) = \lambda^3 + 3\lambda^2 + 3\lambda + 1 = (\lambda+1)^3. & \\
\underline{-\lambda^4 - 3\lambda^3} & \\
3\lambda^3 + 12\lambda^2 & \\
\underline{-3\lambda^3 - 9\lambda^2} & \\
3\lambda^2 + 10\lambda & \\
\underline{-3\lambda^2 - 9\lambda} & \\
\lambda + 3 & \\
\underline{-\lambda - 3} & \\
0 &
\end{array}$$

Damit ist die zweite Nullstelle $\lambda_2 = -1$ mit einer algebraischen Vielfachheit von 3. Wir stellen damit das Fundamentalsystem durch

$$\mathcal{FS} = \left\{ \exp(-3x), \exp(-x), x\exp(-x), x^2\exp(-x) \right\}$$

auf. ※

Beispiel 2.16: Bestimmen Sie ein reelles Fundamentalsystem von $y'' + y' + y = 0$.
Wir stellen dazu das charakteristische Polynom durch $p(\lambda) = \lambda^2 + \lambda + 1$ auf und bestimmen die Nullstellen per Mitternachtsformel wie im ersten Semester mit $\lambda_{1/2} = -\frac{1}{2} \pm \sqrt{\frac{1}{4} - 1} = -\frac{1}{2} \pm i \cdot \frac{\sqrt{3}}{2}$.

Dadurch ist es uns möglich ein komplexes Fundamentalsystem durch

$$\mathcal{FS}_{\mathbb{C}} = \left\{ \exp(-1/2x) \cdot \left(\cos\left(\frac{\sqrt{3}}{2}x\right) \pm i \cdot \sin\left(\frac{\sqrt{3}}{2}x\right)\right)\right\}$$

aufzustellen und dieses mit Satz 2.18 in ein reelles mit

$$\mathcal{FS}_{\mathbb{R}} = \left\{ \exp(-1/2x) \cdot \cos\left(\frac{\sqrt{3}}{2}x\right), \exp(-1/2x) \cdot \sin\left(\frac{\sqrt{3}}{2}x\right)\right\}$$

umzuwandeln. ※

Neben Lösungen für rein homogene Differentialgleichungen (Teilaufgabe (H)) wollen wir uns nun wieder mir Lösungsmethoden für inhomogene Differentialgleichungen beschäftigen. Wir verwenden dazu einen ähnlichen Ansatz wie bereits in Kapitel 2.4.3. Dazu betrachten wir wieder das zugehörige System erster Ordnung

$$y'_{neu}(x) = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 & \cdots & \cdots & \cdots & -a_{n-1} \end{pmatrix} y_{neu}(x) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b(x) \end{pmatrix} \tag{2.49}$$

Wir finden nun wieder eine Lösung mit dem Ansatz der **Variation der Konstanten**, also mit $y_P(x) = W(x)c(x)$, wobei $W(x)$ wieder die Wronskimatrix beschreibt. Wir erhalten daraufhin das LGS

$$W(x)c'(x) = b(x). \tag{2.50}$$

Nach dem Lösen integrieren wir komponentenweise und erhalten die $c_i(x)$, welche wir dann in den Ansatz einsetzen um y_P zu bestimmen. Wir fassen die Erkenntnisse in einem abschließenden Satz zusammen:

Satz 2.30 (*„VdK" zum Finden einer partikulären Lösung einer skalaren linearen DGl.*)

Seien $y_1, \dots, y_n$ n linear unabhängige Lösungen der zu (2.1) gehörenden **homogenen** Differentialgleichung. Dann ist (2.50) auch darstellbar als

$$\begin{pmatrix} y_1(x) & \dots & y_n(x) \\ y'_1(x) & \dots & y'_n(x) \\ \vdots & \ddots & \vdots \\ y_1^{(n-1)}(x) & \dots & y_n^{(n-1)}(x) \end{pmatrix} \cdot \begin{pmatrix} c'_1(x) \\ c'_2(x) \\ \vdots \\ c'_n(x) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ b(x) \end{pmatrix}.$$

Ebenfalls ist dann $y_P(x) = \sum_{i=1}^{n} c_i(x) y_i(x)$ eine Lösung der inhomogenen Differentialgleichung, wobei $c_i(x)$ eine beliebige Stammfunktion zu $c'_i(x)$ ist.

Beweis: Dieser Satz wurde bereits in Kapitel 2.4.3 genauer hergeleitet und ergibt sich hier aus den Überlegungen oben. □

Beispiel 2.17: Bestimmen Sie alle Lösungen der Differentialgleichung $y'' - 5y' + 6y = 4\exp(t)$.
Wir lösen dazu wieder die zwei Probleme (H) und (P):

(H): Wir bestimmen das charakteristische Polynom gemäß Satz 2.28 mit

$$p(\lambda) = \lambda^2 - 5\lambda + 6$$

und die Nullstellen desselbigen über die Mitternachtsformel, wir erhalten $\lambda_1 = 2$ und $\lambda_2 = 3$. Damit ergibt sich das Fundamentalsystem $\mathcal{FS} = \left\{\exp(2x), \exp(3x)\right\}$ und der homogene Lösungsraum mit

$$L_{hom.} = \left\{ c_1, c_2 \in \mathbb{R} : \; c_1 \exp(2x) + c_2 \exp(3x) \right\}.$$

(P): Wir verwenden Satz 2.30 und lösen dann das durch (2.50) beschriebene lineare Gleichungssystem:

$$\left(\begin{array}{cc|c} \exp(2x) & \exp(3x) & 0 \\ 2\exp(2x) & 3\exp(3x) & 4\exp(x) \end{array}\right) \begin{array}{l} \cdot(-2) \\ + \end{array} \rightsquigarrow \left(\begin{array}{cc|c} \exp(2x) & \exp(3x) & 0 \\ 0 & \exp(3x) & 4\exp(x) \end{array}\right)$$

Wir bestimmen darauf die Lösungen $c_2'(x) = 4 \cdot \exp(-2x) \rightsquigarrow c_2(x) = -2\exp(-2x)$ und somit $c_1'(x) = -4\exp(-x) \rightsquigarrow c_1(x) = 4\exp(-x)$. Damit lässt sich die partikuläre Lösung darstellen als

$$y_p(x) = 4\exp(-x) \cdot \exp(2x) + (-2\exp(-2x)) \cdot \exp(3x) = 2\exp(x).$$

Wir bestimmen damit den allgemeinen Lösungsraum mit

$$y_a \in \left\{ c_1, c_2 \in \mathbb{R} : \; c_1 \exp(2x) + c_2 \exp(3x) + 2\exp(x) \right\}.$$

※

Sei an dieser Stelle nun noch eine „Alternative“ zu vorherigen Verfahren gegeben. Das Verfahren, welches hier vorgestellt wird, hat besonders, dass man eine partikuläre Losung y_p „**schlau rät**“ und dann restliche Parameter über Koeffizientenvergleiche bestimmt. Es sei also definiert:

Verfahren 2.2 (*Ansatz vom Typ der rechten Seite*)

Wir betrachten eine Differentialgleichung der Form (2.1) mit dem charakteristischen Polynom p:

(1) **Falls** der inhomogene Anteil b von polynomieller Form ist, also $b(x) = \sum_{j=0}^{m} b_j x^j$, wobei $m \in \mathbb{N}_0, b_m \neq 0$ **dann**

(1a) **Falls** $p(0) \neq 0$ **dann** wähle den Ansatz $y_p := \sum_{j=0}^{m} \alpha_j x^j$.

(1b) **Falls** 0 eine r-fache Nullstelle von p ist **dann** wähle den Ansatz $y_p := x^r \cdot \sum_{j=0}^{m} \alpha_j x^j$.

(2) **Falls** der inhomogene Anteil b von reinexponentieller Form ist, also $b(x) = b \cdot \exp(kx)$, wobei $k \in \mathbb{C}, b \neq 0$ **dann**

(2a) **Falls** $p(k) \neq 0$ **dann** wähle den Ansatz $y_p := \alpha \cdot \exp(kx)$.

(2b) **Falls** k eine r-fache Nullstelle von p ist **dann** wähle den Ansatz $y_p := x^r \cdot \alpha \cdot \exp(kx)$.

(3) **Falls** der inhomogene Anteil b von gemischter Form ist, also $b(x) = \exp(kx) \cdot \sum_{j=0}^{m} b_j x^j$, wobei $k \in \mathbb{C}, m \in \mathbb{N}_0, b_m \neq 0$ **dann**

(3a) **Falls** $p(k) \neq 0$ **dann** wähle den Ansatz $y_p := \exp(kx) \cdot \sum_{j=0}^{m} \alpha_j x^j$.

(3b) **Falls** k eine r-fache Nullstelle von p ist **dann** wähle den Ansatz $y_p := x^r \cdot \exp(kx) \cdot \sum_{j=0}^{m} \alpha_j x^j$.

Dabei sind die Parameter α oder α_i über Koeffizientenvergleich zu berechnen. Sollte $b(x)$ an Stelle von oder zusätzlich zum Exponentialterm einen Faktor $\left\{ \begin{array}{c} \cos(\omega x) \\ \sin(\omega x) \end{array} \right\}$ enthält, so wähle

$\widetilde{b}(x) := \exp(i \cdot \omega \cdot x)$, es ist dann zu dieser komplexen Differentialgleichung eine komplexe partikuläre Lösung zu bestimmen, welcher dann der $\left\{ \begin{array}{c} \text{Real-} \\ \text{Imaginär-} \end{array} \right\}$teil entnommen wird.
Sollte $b(x)$ nicht irgendeine der obigen Formen haben, allerdings nur aus Summanden jener bestehen, so ist einfach für jeden Summanden eine partikuläre Lösung zu finden und diese dann anschließend in Gesamtheit wieder aufzusummieren.

2.6 Numerische Verfahren

Wir wollen nun neben den analytischen – *exakten* – Methoden zum Lösen von Anfangswertproblemen auch noch einen kurzen Einblick in die numerischen Verfahren geben.

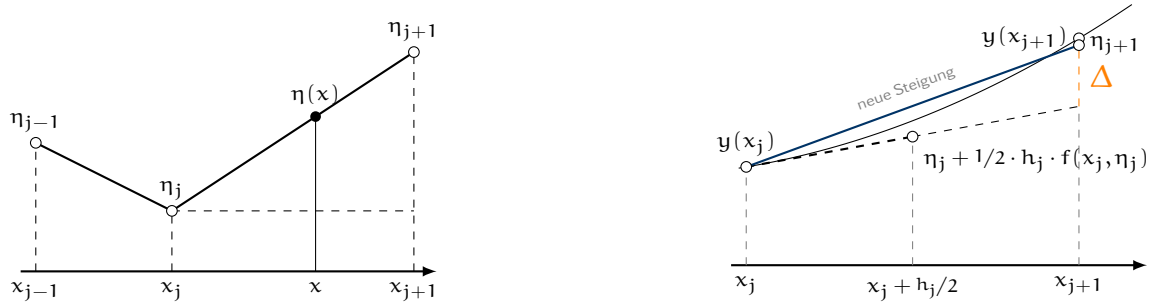
2.6.1 Numerische Lösungsverfahren für gewöhnliche Differentialgleichungen

2.6.1.1 Einschrittverfahren zur Lösung von Anfangswertaufgaben

Sei im Rahmen dieser Veranstaltung nur ein kurzer Ein- und Überblick über die „einfachsten" Verfahren zur numerischen Integration eines Anfangswertproblems

$$y'(x) = f(x, y(x)), y(x_0) = y_0 \tag{2.51}$$

gegeben. Œ sei im Folgenden dieses Anfangswertproblem auf eine gewöhnliche Differentialgleichung erster Ordnung erlaubt, alle Formulierungen lassen aber auch – ohne Zusatz – eine vektorielle Formulierung zu, sprich sie sind ebenso auf Differentialgleichungssysteme und Differentialgleichungen n-ter Ordnung anzuwenden. Wir wollen nun mittels folgender Überlegung auf ein erstes numerisches Verfahren zur Lösung des einfachsten Anfangswertproblems (2.51) kommen:
Ist etwa die Lösung $y(x)$ auf dem Intervall $I := [a, b]$ mit $a := x_0$ zu bestimmen, so wählt man



(a) Das Euler-Cauchy'sche Polygonzugverfahren (b) Verbessertes Eulerverfahren

Abbildung 2.1: Skizzen zu Verfahren 2.3 und 2.4

$n+1$ Stützstellen $x_j \in I$ in der Anordnung $a := x_0 < x_1 < \cdots < x_{n-1} < x_n =: b$. Wir sagen dann $h_j := x_{j+1} - x_j$ heiße **Schrittweite** der obigen Zerlegung von I und n heiße **Schrittzahl**. Im Falle einer äquidistanten Zerlegung ist die Schrittweite $h := \frac{b-a}{n}$ konstant, es gibt somit äquidistante Stützstellen $x_j = a + h \cdot j$. Wie schon in vorherigen Kapiteln immer wieder erwähnt entspricht der Funktionswert $f(x, y)$ ja nach (2.51) genau der Steigung $y'(x)$ der gesuchten exakten Lösung $y(x)$ an der Stelle $x \in I$. Wir wollen diesen Wert der Tangentensteigung nun in einem ersten Versuch durch den Differenzenquotienten[4] $\frac{1}{h}(y(x+h) - y(x))$ annähern. Somit ergibt sich der Zusammenhang

$$y(x+h) \approx y(x) + h \cdot f(x, y(x)). \tag{2.52}$$

[4] geometrisch gesehen ist das ja der Wert der Sekantensteigung

Ausgehend von den Anfangswerten x_0 und $y_0 = y(x_0)$ gelangt man dann auf der endlichen Zerlegung des Intervalls I zu Näherungswerten η_j für die Funktionswerte $y_j := y(x_j)$ der exakten Lösung:

$$\eta_0 := y_0 \text{ und } \eta_{j+1} := \eta_j + h_j \cdot f(x_j, \eta_j), x_{j+1} := x_j + h_j \qquad \forall j = 0, \ldots, n-1 \tag{2.53}$$

Wir verbinden nun die Knotenpunkte (x_j, η_j) und (x_{j+1}, η_{j+1}) durch eine Gerade und erhalten somit als Näherungslösung für unser Anfangswertproblem (2.51) einen **Polygonzug**. Dieser wird dann beschrieben durch

$$\begin{aligned} \eta(x) &:= \eta_j + \frac{(\eta_{j+1} - \eta_j) \cdot (x - x_j)}{h_j} \\ &= \eta_j + f(x_j, \eta_j) \cdot (x - x_j) \qquad \forall x \in [x_j, x_{j+1}], 0 \leqslant j \leqslant n-1 \end{aligned} \tag{2.54}$$

Wir wollen damit dann weiter definieren:

Verfahren 2.3 (Euler-Cauchy'sches *Polygonzugverfahren*)

Sei das Anfangswertproblem (2.51) auf einem Intervall $I := [a, b]$ mit $a := x_0$ zu bestimmen. Sei zudem eine endliche Zerlegung von I mit Schrittweite $h_j := x_{j+1} - x_j$ und $n+1$ Stützstellen $x_j \in I$ bekannt, so nennen wir das Verfahren

$$\eta_0 := y_0 \text{ und } \eta_{j+1} := \eta_j + h_j \cdot f(x_j, \eta_j), x_{j+1} := x_j + h_j \qquad \forall j = 0, \ldots, n-1 \tag{2.53}$$

auch das **Euler-Cauchy'sches Polygonzugverfahren** oder kurz das **(explizite) Eulerverfahren**.

Anmerkung

Die Näherungswerte η_j hängen sicherlich von der Wahl der Schrittweiten h_j ab ($\rightarrow$ *„je kleiner die h_j, desto genauer die Näherungslösung"*). Gelangt man mit anderen Schrittweiten $\widetilde{h}_j$ nun **ebenfalls** zur Stützstelle x_j, so wird der jetzt berechnete Näherungswert $\widetilde{\eta}_j$ aber im Allgemeinen von η_j verschieden sein.

Das eben definierte Euler-Cauchy'sches Polygonzugverfahren (Verfahren 2.3) zählt zu den sogenannten Einschrittverfahren.

Definition 2.18 (*Einschrittverfahren*)

Ein **Einschrittverfahren** zu näherungsweisen Lösung der Aufgabe (2.51) besteht in Vorgeben einer geeigneten Funktion $\Phi = \Phi(x, y; h; f)$ und der Vorgabe der Berechnungsvorschrift

$$\left.\begin{aligned} \eta_0 &:= y_0, \quad \eta_{j+1} := \eta_j + h_j \cdot \Phi(x_j, \eta_j; h_j; f) \\ & \qquad x_{j+1} := x_j + h_j \end{aligned}\right\} \qquad \forall j \in \{0, 1, \ldots, n-1\} \tag{2.55}$$

mit deren Hilfe Näherungen η_j für die Werte $y_j := y(x_j)$ der exakten Lösung bestimmt werden.

Man kann schnell erkennen, dass das Eulerverfahren ein Einschrittverfahren mit $\Phi(x,y;h;f) := f(x,y)$ ist, die Schrittweite hier also keinen Einfluss hat.

Eine ganze Klasse von Einschrittverfahren gewinnt man aus der Taylorentwicklung der exakten Lösung $y(x)$ in einer Umgebung des Startpunktes (x_0,y_0):

$$y(x) = \sum_{k=0}^{p} \frac{1}{k!} y^{(k)}(x_0)(x-x_0)^k + R_{p+1}.$$

Vernachlässigt man das Restglied R_{p+1}, so erhält man mit der Schrittweite $h_j := x_{j+1} - x_j$ den Näherungswert η_{j+1} nach der Rechenvorschrift

$$\eta_{j+1} = \eta_j + h_j \cdot \sum_{k=1}^{p} \frac{1}{k!} h_j^{k-1} y_j^{(k)} \equiv \eta_j + h_j \cdot \Phi(x_j,\eta_j;h_j;f). \tag{2.56}$$

B

Hierbei beschreibt $y_j^{(k)}$ den Wert der k-ten Ableitung im Punkt (x_j,η_j). Wie eingangs bereits angemerkt gilt gemäß (2.51) $y_j' = f(x_j,\eta_j)$. Höhere Ableitungen lassen sich dann im Prinzip durch wiederholte Differentation nach x und Substitution von $y'(x)$ aus (2.51) gewinnen:

$$\left.\begin{array}{rcl} y'(x) &=& f(x,y), \\ y''(x) &=& \frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y} \cdot f(x,y), \\ y'''(x) &=& \frac{\partial^2 f(x,y)}{\partial x^2} + 2 \cdot \frac{\partial^2 f(x,y)}{\partial x \partial y} \cdot f(x,y) + \frac{\partial^2 f(x,y)}{\partial y^2} \cdot f^2(x,y) + \frac{\partial f(x,y)}{\partial y} \cdot y''(x), \\ &\vdots& \end{array}\right\} \tag{2.57}$$

Setzt man hier nun in den rechten Seiten $x = x_j$ und $y = \eta_j$ ein, so hat man formelmäßige Ausdrücke für die Größen $y^{(k)}$ vorliegen. Man beachte aber, dass die Berechnung der höheren Ableitungen nach dem Schema (2.57) sehr rasch kompliziert werden, so dass man sich bei den Verfahren (2.56) nur auf kleine p beschränkt. Mit dem Fall $p = 1$ liegt wieder das Eulerverfahren vor.

Wir wollen nun das kennengelernte Eulerverfahren schrittweise verbessern. Die erste naive Idee ist es $h_j \to 0$ gehen zu lassen. Wir verwerfen diese Idee wieder umgehend aufgrund von zu niedriger Effizienz. Eine weitere Idee ist es, anstelle dem linken Wert des Intervalls zur Steigung einen Wert möglichst in der Mitte des Intervalls zu nehmen. Auch hier stellt sich jedoch die Frage, wie man diese Steigung aus der Mitte erhält. Exakt ist das wie bereits erwähnt sehr schwer, wir erhalten aber zumindest eine Näherung, indem wir zuerst einen „halben“ Schritt von x_j nach $x_j + \frac{h_j}{2}$ mit obigem Eulerverfahren machen, dort dann die Steigung „abgreifen“ und mit dieser „verbesserten“ Steigung den Schritt von x_j nach x_{j+1} durchführen. Wir erhalten somit das Verfahren

Verfahren 2.4 (*Verbessertes* Euler*verfahren*)

Die Idee des Verfahrens bleibt im Vergleich zu Verfahren 2.3 gleich, lediglich ändert sich die Berechnungsvorschrift auf

$$\left.\begin{array}{l} \eta_0 := y_0, \quad \eta_{j+1} := \eta_j + h_j \cdot f\left(x_j + \frac{h_j}{2}, \eta_j + \frac{h_j}{2} \cdot f(x_j,\eta_j)\right) \\ \qquad\qquad\quad x_{j+1} := x_j + h_j \end{array}\right\} \quad \forall j \in \{0,1,\ldots,n-1\}. \tag{2.58}$$

Skizze 2.1b verdeutlicht einen Schritt dieses Verfahrens und stellt in orange auch das Genauigkeitsdelta dar.
Eine andere Idee sei es $y'(x_{j+1})$ durch eine Linearkombination von drei Werten $y(x_{j+1}), y(x_j)$ und

$y(x_{j-1})$ zu approximieren, also suchen wir $\alpha, \beta, \gamma \in \mathbb{C}$, so dass

$$\alpha \cdot y(x_{j+1}) + \beta \cdot y(x_j) + \gamma \cdot y(x_{j-1}) \overset{?}{\approx} y'(x_{j+1}) \tag{2.59}$$

gilt. Sei hier nun eine äquidistante Zerlegung gegeben, so können wir den Zusammenhang (2.59) auch als

$$\alpha \cdot y(x_{j+1}) + \beta \cdot y(x_{j+1} - h) + \gamma \cdot y(x_{j+1} - 2h) \tag{2.60}$$

schreiben. Mit der Taylorentwicklung nach h gilt dann

$$\begin{aligned}
(2.60) \;=\; & \alpha \cdot y(x_{j+1}) + \beta \cdot \left[y(x_{j+1}) + (-h) \cdot y'(x_{j+1}) + \frac{(-h)^2}{2} \cdot y''(x_{j+1}) + \mathcal{O}(h^3) \right] \\
& + \gamma \cdot \left[y(x_{j+1}) + (-2h) \cdot y'(x_{j+1}) + \frac{(-2h)^2}{2} \cdot y''(x_{j+1}) + \mathcal{O}(h^3) \right] \\
= \; & (\alpha + \beta + \gamma) \cdot y(x_{j+1}) + h \cdot (-\beta - 2\gamma) \cdot y'(x_{j+1}) + h^2 \cdot \left(\frac{\beta}{2} + 2\gamma \right) \cdot y''(x_{j+1}) + \mathcal{O}(h^3) \qquad (2.61) \\
\overset{!}{\approx} \; & y'(x_{j+1}).
\end{aligned}$$

Wir erkennen schnell, dass der erste und letzte Faktor jeweils 0 sein müssen, der zweite aber 1 sein muss, damit der Zusammenhang erfüllt ist. Wir erhalten damit ein lineares Gleichungssystem

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 1/2 & 2 \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \text{ gelöst ergibt das } \alpha = \frac{3}{2}, \beta = -2 \text{ und } \gamma = \frac{1}{2}.$$

Damit fassen wir nun (2.59) auf als

$$\frac{3}{2} \cdot y(x_{j+1}) - 2 \cdot y(x_j) + \frac{1}{2} \cdot y(x_{j-1}) = h \cdot y'(x_{j+1} + \mathcal{O}(h^3). \tag{2.62}$$

Unter Ausnutzung von bekanntem Wissen, dass $y'(x_{j+1}) = f(x_{j+1}, y(x_{j+1})$ gilt, sowie leicher Umformung, erhalten wir dann das folgende Verfahren:

Verfahren 2.5 (*Rückwärtsdifferenzenmethode (BDF-2)*)

Sei das Anfangswertproblem (2.51) auf einem Intervall $I := [a, b]$ mit $a := x_0$ zu bestimmen. Sei zudem eine endliche äquidistante Zerlegung von I mit Schrittweite h und $n+1$ Stützstellen $x_j \in I$, sowie zwei Wertepaare (x_0, y_0) und (x_1, y_1) von Beginn an bekannt, so nennen wir das Verfahren mit der Berechnungsvorschrift

$$\eta_0 := y_0, \eta_1 := y_1 \quad \left.\begin{array}{l} \frac{3}{2}\eta_{j+1} - 2\eta_j + \frac{1}{2}\eta_{j-1} := h \cdot f(x_{j+1}, \eta_{j+1}) \\ x_{j+1} := x_j + h \end{array}\right\} \quad \forall j \in \{0, 1, \ldots, n-1\} \tag{2.63}$$

auch **Rückwärtsdifferenzenmethode** oder *Backwards-Difference-Formula* (BDF-2). Die 2 steht dabei für die Anzahl an zusätzlichen Termen, welche für die Berechnung benötigt werden (hier wären das η_j und η_{j-1}). Verallgemeinert man das Verfahren auf k Terme, so ergibt sich ein BDF-k Verfahren.

Anmerkung

Im Gegensatz zu Einschrittverfahren (nach Definition 2.18) kann die Funktion Φ bei **Mehrschrittverfahren** auch noch von vorherigen Knotenstellen (x_{j-l}, η_{j-l}) abhängen.

Ein mögliches Problem

Man muss mit der Berechnungsvorschrift in (2.63) aufpassen, da diese im Allgemeinen nicht imnmer nach η_{j+1} auflösbar ist. Abhilfe verschaffen dann hier die Verfahren zum Finden von Fixpunkten beispielsweise das Newtonverfahren.

Konsistenz, Diskretisationsfehler und Fehlerordnung
Leicht ist festzustellen, dass die Rechenvorschrift (2.55) sicher nur dann ein brauchbares Näherungsverfahren zur Lösung des Anfangswertproblems (2.51) darstellen kann, wenn $\Phi(x,y;h;f)$ mit $f(x,y)$ in einer bestimmten Relation steht. Diese Relation resultiert aus (2.55) im Limes $h_j \to 0$:

$$\lim_{h_j \to 0} \frac{1}{h_j}(\eta_{j+1} - \eta_j) = y'(x_j) = \Phi(x_j, \eta_j; 0; f)$$

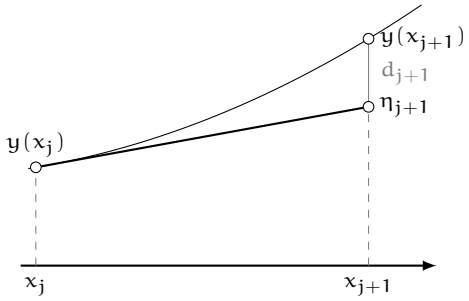


Abbildung 2.2: Zur geometrischen Bedeutung des lokalen Diskretisationsfehlers

Wir definieren mit dieser Erkenntnis:

Definition 2.19 (*Konsistenz von Einschrittverfahren*)

Ein Einschrittverfahren (2.55) heiße mit der Differentialgleichung (2.51) **konsistent** genau dann, wenn

$$\forall x, y.\ \Phi(x,y;0;f) = f(x,y) \tag{2.64}$$

Wegen (2.64) verzichten wir nun bei konsistenten Einschrittverfahren auf das Argument f in der Funktion Φ.
Die Wahl der Funktion Φ nimmt ganz entscheidend Einfluss auf die Güte der Approximation der exakten Lösung. Gilt nämlich $\Phi(x,y;h) \neq f(x,y)$, so wird in jedem Schritt des Verfahrens (2.55) mit einer **falschen** Steigung gerechnet. Die dadurch entstehende Abweichung wird durch den **lokalen Diskretisationsfehler** gemessen:

Definition 2.20 (*Lokaler Diskretisationsfehler*)

Sei $y(X)$ die exakte Lösung des Anfangswertproblems (2.51). Dann heiße die Differenz

$$d_{j+1} := y(x_{j+1}) - y(x_j) - h \cdot \Phi(x_j, y(x_j); h) \tag{2.65}$$

der **lokale Diskretisationsfehler** an der Stelle $x_{j+1} = x_j + h$.

Beispiel 2.18: Als Beispiel wollen wir nun den lokalen Diskretisationsfehler des Eulerverfahrens bestimmen. Hat die Funktion $f(x,y)$ stetig partielle Ableitungen nach x und y bis zur Ordnung $p-1$, so ist die Lösung $y(x)$ des Anfangswertproblems (2.51) sicher p-mal stetig differenzierbar und nach Taylorentwicklung gilt:

$$y(x+h) = \sum_{k=0}^{p-1} \frac{1}{k!} \cdot y^{(k)}(x) \cdot h^k + \frac{1}{p!} \cdot h^p \cdot y^{(p)}(\xi) \text{ mit } \xi \in (x, x+h) \tag{2.66}$$

Unter Verwendung der Formeln (2.57) resultiert

$$y(x+h) - y(x) = h \cdot f(x,y) + \frac{h^2}{2} \cdot \left(\frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y} \cdot f(x,y) \right) + \mathcal{O}(h^3) \text{ für } h \to 0.$$

Mit $x := x_j, x_{j+1} : - = x_j + h$ und $y_j := y(x_j)$ sowie der Beachtung von $\Phi(x,y;h) = f(x,y)$ folgt dann für den lokalen Diskretisationsfehler des Eulerverfahrens an der Stelle x_{j+1}:

$$d_{j+1} = \frac{h^2}{2} \cdot \left(\frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y} \cdot f(x,y) \right) + \mathcal{O}(h^3) \xrightarrow{h \to 0} \mathcal{O}(h^2) \tag{2.67}$$

※

(2.67) veranlasst dann zu folgender Definition:

Definition 2.21 (*Fehlerordnung*)

Ein Einschrittverfahren (nach (2.55)) besitze die **Fehlerordnung** $p \geqslant 0$, falls für seinen lokalen

Diskretisationsfehler d_j die Abschätzung

$$\max_{1\leqslant j\leqslant n} |d_j| \leqslant \text{const.} \cdot h^{p+1} = \mathcal{O}(h^{p+1}) \text{ für } h \to 0^+ \tag{2.68}$$

gilt, wobei $f \in \mathcal{C}^p([a,b]\times\mathbb{R})$ vorausgesetzt wird.

Anmerkung

(2.68) ist äquivalent zu folgender Formulierung:

$$\exists c = c(f) > 0 : \forall h > 0 \forall x_j \in [a,b] : |\eta_j - y(x_j)| \leqslant c \cdot h^p$$

Sei hier nun noch ohne weitere Erklärung postuliert, dass die Fehlerordnung p mit der Konvergenzordnung bzgl. der Schrittweite h eines Einschrittverfahrens übereinstimmt.

Gemäß dieser Definition hat das Eulerverfahren somit die Fehlerordnung $p = 1$.

2.6.1.2 Verwendung von Integralgleichungen zur Lösung von Anfangswertaufgaben — Runge-Kutta-Verfahren

Eine andere Methode Verfahren herzuleiten besteht darin das Anfangswertproblmem in eine Integralgleichung mit Satz 2.10 umzuwandeln und über $[x_j, x_{j+1}]$ zu betrachten:

$$y(x_{j+1}) - y(x_j) = \int_{x_j}^{x_{j+1}} f(x, y(x))dx \tag{2.69}$$

Eine allgemeine Quadraturformel[5] mit den Stützstellen $\xi_1, \ldots, \xi_m \in [x_j, x_{j+1}]$ und zugehörigen Integrationsgewichten $a_1, \ldots, a_m$ führt zu folgendem Ansatz für die Näherung η_{j+1} des Funktionswerts $y(x_{j+1})$:

$$\eta_{j+1} = \eta_j + h_j \cdot \sum_{l=1}^{m} a_l k_l \text{ mit } k_l := f(\xi_l, y(\xi_l)). \tag{2.70}$$

Wählt man nun beispielsweise die **Trapezregel** als Quadraturformel mit $m = 2$, $a_1 = a_2 = \frac{1}{2}$ und $\xi_1 = x_j, \xi_2 = x_{j+1}$ in (2.70), erhält man mit $\eta_j \approx y(x_j), \eta_{j+1} \approx y(x_{j+1})$ das **implizite Integrationsverfahren**:

$$\eta_{j+1} = \eta_j + \frac{h_j}{2} \cdot (f(x_j, \eta_j) + f(x_{j+1}, \eta_{j+1})), \tag{2.71}$$

in welchem der Näherungswert η_{j+1} implizit definiert ist. Jeder Integrationsschritt erfordert im Allgemeinen die Lösung einer *nichtlinearen* Gleichung. Hängt die Funktion $f(x,y)$ nun nichtlinear von y ab, so kann die Fixpunktgleichung (2.71) beispielsweise durch sukzessive Approximation nach η_{j+1} aufgelöst werden. Für einen geeigneten Startwert $\eta_{j+1}^{(0)}$ konvergiert die Fixpunktiteration

$$\eta_{j+1}^{(k+1)} = \eta_j + \frac{h_j}{2}\left(f(x_j, \eta_j) + f(x_{j+1}, \eta_{j+1}^{(k)})\right), \text{ für } k = 0, 1, \ldots \tag{2.72}$$

sofern $f(x,y)$ die übliche Lipschitzbedingung $|f(x,y_1) - f(x,y_2)| \leqslant L \cdot |y_1 - y_2|$ erfüllt und $h_j \cdot \frac{L}{2} < 1$ gilt. Da in (2.72) ohnehin der Funktionswert $f(x_j, \eta_j)$ zu berechnen ist, liegt es nahe als geeigneten Startwert einen Schritt des Eulerverfahrens zu wählen:

$$\eta_{j+1}^{(0)} := \eta_j + h_j f(x_j, \eta_j) \tag{2.73}$$

Wird nun in der Fixpunktiteration (2.72) nun ein **einziger** Iterationsschritt ausgeführt, so resultiert mit dem Startwert (2.73) das folgende Einschrittverfahren:

[5] Formel zur numerischen Integration

Verfahren 2.6 (HEUN*sches Verfahren*)

Das Einschrittverfahren, welches durch die Berechnungsvorschrift

$$\begin{aligned} \eta_{j+1}^{(P)} &= \eta_j + h_j f(x_j, \eta_j), \\ \eta_{j+1} &= \eta_j + \tfrac{h_j}{2}\left(f(x_j, \eta_j) + f(x_{j+1}, \eta_{j+1}^{(P)})\right) \end{aligned} \tag{2.74}$$

beschrieben wird heißt **HEUNsches Verfahren**. Es gehört zu den expliziten **Prädikator-Korrektor-Verfahren**, da zunächst ein **Prädikatorwert** $\eta_{j+1}^{(P)}$ mit dem EULERverfahren bestimmt wird, der dann mit dem impliziten Trapezverfahren **korrigiert** wird.

Lemma 2.31 (*Einschrittverfahren*)

Das HEUNsches Verfahren (Verfahren 2.6) gehört zur Klasse der Einschrittverfahren.

Beweis: Wir geben dazu einfach die Funktion Φ an mit

$$\Phi(x, y; h) := a_1 k_1 + a_2 k_2 = a_1 f(x, y) + a_2 f(x + q_1 h, y + q_2 h f(x, y)), \tag{2.75}$$

worin speziell

$$a_1 = a_2 = \frac{1}{2}, \quad q_1 = q_2 = 1 \tag{2.76}$$

zu setzen sind. □

Wir erhalten damit bezüglich der Fehlerordnung die folgende Aussage:

Satz 2.32 (*Fehlerordnungen von Einschrittverfahren*)

Das mit der Funktion Φ aus (2.75) gebildete Einschrittverfahren hat mindestens die Fehlerordnung $p = 2$, falls a_1, a_2, q_1 und q_2 mit

$$a_1 + a_2 = 1 \text{ und } a_2 q_1 = a_2 q_2 = \frac{1}{2} \tag{2.77}$$

gewählt werden.

Beweis: Entwickelt man (2.75) an der Stelle $h = 0$ in eine Taylorreihe, so erhält man zusammen mit (2.66) für den lokalen Diskretisationsfehler:

$$\begin{aligned} d_{j+1} &= (1 - a_1 - a_2) h_j f(x_j, y_j) \\ &\quad + \frac{1}{2} h_j^2 \left((1 - 2a_2 q_1) \frac{\partial f(x_j, y_j)}{\partial x} + (1 - 2a_2 q_2) \frac{\partial f(x_j, y_j)}{\partial y} \cdot f(x_j, y_j)\right) + \mathcal{O}(h_j^3) \text{ für } h_j \to 0. \end{aligned}$$

Ein Verfahren der Fehlerordnung $p \geqslant 2$ resultiert also bei Wahl von

$$0 = (1 - a_1 - a_2) = (1 - 2a_2 q_1) = (1 - 2a_2 q_2),$$

und diese Bedingungen sind äquivalent mit (2.77). □

Folgerung

Das Heunsche Verfahren erfüllt mit dem Parametersatz (2.76) die Bedingungen (2.77), was es als ein Einschrittverfahren der Fehlerordnung $p = 2$ klassifiziert.

Mit dem Satz folgt auch ebenso die bislang noch unbestätigte Vermutung, dass das verbesserte Eulerverfahren (Verfahren 2.4) besser ist als das „normale“ *Euler*verfahren (Verfahren 2.3). Wir wählen als Parameter dazu $a_1 := 0, a_2 := 1$ und $q_1 = q_2 = \frac{1}{2}$ und erhalten aus dem Ansatz (2.75) die Funktion

$$\Phi(x,y;h) := f\left(x + \frac{h}{2}, y + \frac{h}{2} f(x,y)\right). \tag{2.78}$$

Es ist dann einfach zu erkennen, dass das Verfahren, welches aus (2.78) resultiert, dasselbe ist, wie Verfahren 2.4.

Einschrittverfahren, bei denen die Funktion $f(x,y)$ pro Integrationsschritt k-mal ausgewertet werden muss, heißen k-**stufige Verfahren**. Das Heunverfahren sowie das verbesserte Eulerverfahren sind dementsprechend **zweistufige Verfahren**. Außerdem gehören sie zur Klasse der **Runge-Kutta**-Verfahren.
Allgemeine Runge-Kutta-Verfahren resultieren aus dem Ansatz (2.70), wenn die Integrationsstützstellen ξ_l gemäß

$$\xi_1 := x_j, \xi_l := x_j + q_l h_j \forall 2 \leqslant l \leqslant m \text{ mit } 0 < q_l \leqslant 1 \tag{2.79}$$

fixiert werden und die unbekannten Funktionswerte $y(\xi_l)$ nach der Idee des Prädikatorverfahrens festgelegt werden. Zunächst setzt man $y(\xi_1) \approx \eta_j$, ferner gelte

$$y(\xi_l) \approx y_l^* := \eta_j + h_j b_{l1} f(x_j, \eta_j) + h_j \cdot \sum_{r=2}^{l-1} b_{lr} f(x_j + q_r h_j, y_r^*) \text{ für } 2 \leqslant l \leqslant m. \tag{2.80}$$

Aus (2.70) und (2.80) ergibt sich somit eine rekursive Definition der Funktionswerte $k_l = f(\xi_l, y_l^*)$ gemäß

$$\left.\begin{array}{lcl} k_1 & := & f(x_j, \eta_j), \\ k_l & := & f\left(x_j + q_l h_j, \eta_j + h_j \cdot \sum_{r=1}^{l-1} b_{lr} k_r\right) \text{ für } l = 2, \ldots, m. \end{array}\right\} \tag{2.81}$$

Die Berechnung von (2.70) erfordert also im Allgemeinen in jedem Schritt m Funktionsauswertungen, so dass der dort vorgestellte Algorithmus ein m-stufiges Runge-Kutta-Verfahren darstellt.
Wir wollen jetzt noch ein Runge-Kutta-Verfahren mit Fehlerordnung $p = 4$ kennenlernen, welches vierstufig ist. Gemäß (2.70) und (2.81) hat das allgemeine vierstufige Runge-Kutta-Verfahren die algorithmische Form

$$\left.\begin{array}{lcl} k_1 & := & f(x_j, \eta_j), \\ k_2 & := & f(x_j + q_2 h_j, \eta_h + h_j b_{21} k_1), \\ k_3 & := & f(x_j + q_3 h_j, \eta_h + h_j (b_{31} k_1 + b_{32} k_2)), \\ k_4 & := & f(x_j + q_4 h_j, \eta_h + h_j (b_{41} k_1 + b_{42} k_2 + b_{43} k_3)); \\ \eta_{j+1} & := & \eta_j + h_j \cdot (a_1 k_1 + a_2 k_2 + a_3 k_3 + a_4 k_4). \end{array}\right\} \tag{2.82}$$

Die dreizehn freien Parameter sind dann zunächst nur durch drei Nebenbedingungen, nämlich

$$q_l = \sum_{r=1}^{l-1} b_{lr} \text{ für } 2 \leqslant l \leqslant m,$$

restringiert; es ergeben sich allerdings aus der Bedingung, dass das Verfahren mindestens die Fehlerordnung $p = 4$ hat, weitere acht Bestimmungsgleichungen. Wir geben eine Lösung an, welche das bekannteste Verfahren vierter Ordnung ergeben:

Verfahren 2.7 (*Klassisches* Runge-Kutta-*Verfahren*)

Ein Runge-Kutta-Verfahren mit den Parametern

$$q_2 = q_3 = \frac{1}{2}, q_4 = 1, b_{21} = b_{31} = b_{43} = \frac{1}{2}, b_{32} = b_{41} = b_{42} = 0$$

$$a_1 = a_4 = \frac{1}{6}, a_2 = a_3 = \frac{2}{6}$$

heißt **klassisches vierstufiges Runge-Kutta-Verfahren**. Es hat die algorithmische Form

$$\left.\begin{array}{lcl} k_1 & := & f(x_j, \eta_j), \\ k_2 & := & f(x_j + \frac{1}{2}h_j, \eta_h + \frac{1}{2}h_j k_1), \\ k_3 & := & f(x_j + \frac{1}{2}h_j, \eta_h + \frac{1}{2}h_j k_2), \\ k_4 & := & f(x_j + h_j, \eta_h + h_j k_3); \\ \eta_{j+1} & := & \eta_j + \frac{1}{6}h_j \cdot (k_1 + 2k_2 + 2k_3 + k_4). \end{array}\right\} \tag{2.83}$$

Dies soll es mit der Einführung in numerische Verfahren für gewöhnliche Differentialgleichungen gewesen sein, wir beschäftigen uns nun noch ein bisschen mit partiellen Differentialgleichungen und geben einen kleinen Einblick in deren numerische Verfahren.

2.6.2 Numerische Lösungsverfahren für partielle Differentialgleichungen

Wir haben uns bislang rein der Theorie und Numerik gewöhnlicher Differentialgleichungen gewidmet. Hierbei waren die gesuchten Funktionen *nur* skalar, was auf mache Probleme in physikalisch-technischen, chemischen oder auch wirtschafts-/sozialwissenschaftlichen Anwendungen stößt. Viele dieser Größen hängen von **mehreren unabhängigen Veränderlichen** ab, bei physikalischen Problemen kann dies beispielsweise neben einer zeitlichen Komponente auch eine örtliche sein. Wir in der Motivation eingangs bereits angemerkt wollen wir solche Differentialgleichungen dann **partiell** nennen. Prominente Beispiele hierfür sind beispielsweise die **Navier-Stokes-Gleichungen** der Strömungsmechanik, die **Maxwell-Gleichungen** des Elektromagnetismus oder die **Schrödinger-Gleichung** in der Quantenmechanik.
Als Beispiel wolllen wir uns zunächst die **Wärmeleitungsgleichung** auf einem Gebiet $\Omega \subseteq \mathbb{R}^3$ betrachten:

$$\forall (t,x,y,z) \in [0,T]\times\Omega.\ \frac{\partial u(t,x,y,z)}{\partial t} - k \cdot \underbrace{\left(\frac{\partial^2 u(t,x,y,z)}{\partial x^2} + \frac{\partial^2 u(t,x,y,z)}{\partial y^2} + \frac{\partial^2 u(t,x,y,z)}{\partial z^2}\right)}_{=\Delta u(t,x,y,z) - \frac{\partial^2 u(t,x,y,z)}{\partial t^2}} = f(t,x,y,z) \tag{2.84}$$

In diesem Beispiel beschreibe f eine gegebene Quelle, $k > 0$ den Wärmeleitfähigkeitskoeffizienten und u die gesuchte Temperatur in einem Gebiet $[0,T] \times \Omega$, wobei $\Omega \subseteq \mathbb{R}^3$. Man kann – dies wird hier allerdings ausgelassen – zeigen, dass man, um Existenz und Eindeutigkeit einer Lösung zu garantieren, neben einer Anfangsbedingung „$u(0,x,y,z) = u_0(x,y,z)\ \forall (x,y,z) \in \Omega$“ auch zusätzlich noch **Randbedingungen (RB)** vorgeben muss, beispielsweise mit $\frac{\partial u}{\partial \nu} \stackrel{!}{=} \nu\ \forall (t,x,y,z) \in [0,T] \times \partial\Omega$ mit dem **Normaleneinheitsvektor ν auf** $\partial\Omega$, im Kontext also einen Wärmefluss über den Rand. Sollte man nur nach **stationären Lösungen** $\left(\frac{\partial u}{\partial t} \stackrel{!}{=} 0\right)$ suchen, so fällt in (2.84) der Term $\frac{\partial u(t,x,y,z)}{\partial t}$ weg und wir benötigen fortan nur noch die Randbedingungen.
Man kann sich nun viele interessante **theoretische** Fragen stellen, wie zum Beispiel welche Aussagen über Existenz und Eindeutigkeit getroffen werden können, oder ob und wie „glatt“ die Lösung ist. Aber – anders als *gewöhnliche* Differentialgleichungen – sind **partielle** im Allgemeinen nicht analytisch lösbar, wir müssen also über numerische Verfahren an potentielle Lösungen herankommen.
Wir wollen zunächst aber erstmal den Begriff der partiellen Differentialgleichung näher definieren:

Definition 2.22 (*Multiindex (wiederholt aus C2)*)

Sei $n \in \mathbb{N}$. Ein n-Tupel **nicht**-negativer ganzzahliger Zahlen $\alpha = (\alpha_1, \ldots, \alpha_n), \alpha_i \in \mathbb{N}_0, i = 1, 2, \ldots, n$, heiße **Multiindex**. Der Betrag von α ist dann definiert durch

$$|\alpha| = \sum_{i=1}^{n} \alpha_i.$$

Für eine $|\alpha|$-mal stetig diff'bare Funktion $u \in \mathrm{Abb}(\mathbb{R}^n, \mathbb{R})$ mit $n \in \mathbb{N}$ und α als gegebenen Multiindex gilt dann für eine **mehrfache** Ableitung

$$D^{\alpha} u := \frac{\partial^{|\alpha|} u}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_n} x_n}.$$

Damit lässt sich dann definieren:

Definition 2.23 (*Partielle Differentialgleichung*)

Sei $\Omega \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, ein Gebiet und $F : \Omega \times \mathbb{R}^k \to \mathbb{R}$ mit $k \in \mathbb{N}$, sowie $f : \Omega \to \mathbb{R}$ gegebene Funktionen. Dann heiße eine Gleichung der Form

$$F\left(x, D^{\alpha^1} u, \ldots, D^{\alpha^k} u\right) = f(x) \; \forall x \in \Omega \tag{2.85}$$

partielle Differentialgleichung (pDGl.) in n Veränderlichen für die gesuchte Funktion $u : \Omega \to \mathbb{R}$. α^1, ..., α^k seien dabei Multiindizes, also jeweils Vektoren der Dimension n.
Gelte ferner $\left|\alpha^i\right| \leqslant m$ für alle $i = 1, ..., k$ und existiere ein α^j mit $\left|\alpha^j\right| = m$, so heiße die Differentialgleichung von m-**ter Ordnung**[6]. Wenn $f(x) \equiv 0$, so heiße sie des Weiteren **homogen**, andernfalls **inhomogen**.

Die numerischen Verfahren für partielle Differentialgleichungen lassen sich nun im Wesentlichen in zwei große Klassen[7] aufteilen:

(1) Finite-Differenzen-Methoden (FDM) ($\rightarrow$ 2.6.2.1)

(2) Finite-Elemente-Methoden (FEM) ($\rightarrow$ 2.6.2.2)

Wir wollen an dieser Stelle jeweils nur einen Überblick verschaffen, am genausten aber auf die Finite-Differenzen-Methoden eingehen.

2.6.2.1 Finite-Differenzen-Methoden

Grundlegende Idee: Approximation von Differentialoperatoren durch Differenzenquotienten
Wir betrachten nun – der Einfachheit halber – die stationäre Variante von (2.84) in zwei Raumdimensionen, sprich

$$\forall \begin{pmatrix} x \\ y \end{pmatrix} =: \vec{x} \in \Omega : \; -k\left(\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2}\right) = f(x,y). \tag{2.84.S.2}$$

Das zu betrachtende Gebiet ist hierbei erstmal quadratisch mit $\Omega := [0, L] \times [0, L]$. Um nun die zu suchende Funktion zu approximieren, legen wir ein reguläres Gitter der Größe $(n+1) \in \mathbb{N}$ über Ω. Die Maschenweite entspricht dann $h := \frac{L}{n}$.
Wir suchen nun die Näherungslösung $u_{i,j}$ **nur** an den Gitterpunkten, das heißt es soll $u_{i,j} \approx u(x_i, y_j)$ für $i, j = 0, \ldots, n$ sein. An den *Randpunkten* (für $i, j = 0 \vee i, j = n$) sei $u_{i,j}$ dabei durch eine

[6] Stillschweigend wird hier vorausgesetzt, dass nach $D_x^{\alpha^j} u$ umgestellt werden kann und diese Ableitung auch tatsächlich auftritt.

[7] eventuell auch drei, wenn man Finite-Volumen-Methoden miaufnehmen möchte

Randbedingung bestimmt, was das Problem auf $n-1$ Unbekannte schrumpfen lässt. Zur Lösung konstruieren wir nun eine Lösung ähnlich der Rückwärtsdifferenzenmethode (Herleitung von Verfahren 2.5), wir ersetzen die Ableitung durch geeignete Differenzenquotienten.
Geeignete Differenzenquotienten findet man zum Beispiel mittels Taylor-Entwicklung, so für u':

$$\begin{aligned} u'(x) &= \partial^{+h}u(x) + \mathcal{O}(h) = \tfrac{u(x+h)-u(x)}{h} + \mathcal{O}(h) && \textit{(Vorwärtsdifferenzenquotient)} \\ &= \partial^{-h}u(x) + \mathcal{O}(h) = \tfrac{u(x)-u(x-h)}{h} + \mathcal{O}(h) && \textit{(Rückwärtsdifferenzenquotient)} \\ &= \partial^{c,h}u(x) + \mathcal{O}(h^2) = \tfrac{u(x+h)-u(x-h)}{2h} + \mathcal{O}(h^2) && \textit{(Zentraler Differenzenquotient)} \end{aligned} \tag{2.86}$$

oder u'':

$$u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + \mathcal{O}(h^2) \tag{2.87}$$

unter der jeweiligen Voraussetzung, dass u hinreichend glatt sei. Welche Approximation verwendet werden sollte wird von Funktion zu Funktion bestimmt; wohlgemerkt führt allerdings nicht jede dabei auf ein konvergentes Verfahren.
Wir setzen nun (2.87) in (2.84.S.2) ein und erhalten somit für die exakte Lösung

$$-k \cdot \left[\frac{u(x-h,y) - 2u(x,y) + u(x+h,y)}{h^2} + \frac{u(x,y-h) - 2u(x,y) + u(x,y+h)}{h^2} \right] = f(x,y) + \mathcal{O}(h^2). \tag{2.88}$$

Da die exakte Lösung (2.88) für alle $\vec{x}$, also insbesondere auch für alle Gitterpunkte x_{ij} $(i,j = 1,\dots,n-1)$, erfüllt motiviert dies zur Forderung von

$$-k \cdot \left[\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h^2} \right] = f(x_i, y_i) =: f_{i,j} \quad \forall\, i,j = 1,\dots,n-1 \tag{2.89}$$

an die numerische Lösung.
Wir stellen fest, dass dies genau einem linearen Gleichungssystem mit $(n-1)^2$ Unbekannten und Gleichungen entspricht. Das einzige Problem ist die dünne Besetzung der Matrix, da auf jeder Zeile nur maximal fünf von null verschiedene Einträge stehen.
Für den Fall $n = 5$ ergibt sich folgendes System:

$$\frac{k}{h^2} \cdot \underbrace{\left(\begin{array}{cccc:cccc:cccc} 4 & -1 & & & -1 & & & & & & & \\ -1 & 4 & -1 & & & -1 & & & & & & \\ & -1 & 4 & -1 & & & -1 & & & & & \\ & & -1 & 4 & -1 & & & -1 & & & & \\ \hdashline -1 & & & & 4 & -1 & & & -1 & & & \\ & -1 & & & -1 & 4 & -1 & & & -1 & & \\ & & -1 & & & -1 & 4 & -1 & & & -1 & \\ & & & -1 & & & -1 & 4 & -1 & & & -1 \\ \hdashline & & & & -1 & & & & 4 & -1 & & \\ & & & & & -1 & & & -1 & 4 & -1 & \\ & & & & & & -1 & & & -1 & 4 & -1 \\ & & & & & & & -1 & & & -1 & 4 \end{array}\right)}_{\in \mathbb{R}^{(n-1)^2 \times (n-1)^2}} \cdot \begin{pmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{14} \\ \hdashline u_{21} \\ u_{22} \\ u_{23} \\ u_{24} \\ \hdashline u_{31} \\ u_{32} \\ u_{33} \\ u_{34} \\ \hdashline u_{41} \\ u_{42} \\ u_{43} \\ u_{44} \end{pmatrix} = \begin{pmatrix} f_{11} + \frac{k}{h^2}(u_{01} + u_{10}) \\ f_{12} + \frac{k}{h^2}u_{02} \\ f_{13} + \frac{k}{h^2}u_{03} \\ f_{14} + \frac{k}{h^2}(u_{15} + u_{04}) \\ \hdashline f_{21} + \frac{k}{h^2}u_{20} \\ f_{22} \\ f_{23} \\ f_{24} + \frac{k}{h^2}u_{25} \\ \hdashline f_{31} + \frac{k}{h^2}u_{30} \\ f_{32} \\ f_{33} \\ f_{34} + \frac{k}{h^2}u_{35} \\ \hdashline f_{41} + \frac{k}{h^2}(u_{40} + u_{51}) \\ f_{42} + \frac{k}{h^2}u_{52} \\ f_{43} + \frac{k}{h^2}u_{53} \\ f_{44} + \frac{k}{h^2}(u_{54} + u_{45}) \end{pmatrix}$$

Wie löst man nun ein solch *dünnbesetztes* System?

(a) **Gauß-Elimination:** „normalerweise“ $\in \mathcal{O}\left(((n-1)^2)^3\right) = \mathcal{O}(n^6)$. Da obige Matrix allerdings eine sogenannte Bandmatrix ist, lässt sich der Aufwand durch das Anpassen der Schleifen auf $\mathcal{O}(n^4)$ reduzieren.
Es kommt neben einem Laufzeitproblem bei praxisnahen Größen von $n >> 10^4$ auch zu

Speicherproblemen, da sich während des Eliminationsverfahrens die Werte zwischen den besetzten Diagonalen zu Werten ungleich Null ändern. Man erhält somit einen Speicherbedarf von $\mathcal{O}(n^3)$.

(b) **Fixpunktverfahren:** Diese verändern per se erstmal die gegebene Matrix *nicht*! Speicherprobleme fallen weg und Konvergenz ist durch das Kriterium der Diagonaldominanz gesichert (Matrizen oben sind schwach diagonaldominant, im Falle der zeitabhängigen pDGl. sogar strikt diagonaldominant). Pro Iterationsschritt haben wir so nur $\mathcal{O}(n^2)$ Rechenoperationen unter Ausnutzung der Nullen. Dennoch gilt: **Je größer n desto stärker sinkt die Konvergenzgeschwindigkeit ab.**

(c) **Moderne Löser (Mehrgitterverfahren, ...):** Insgesamt nur $\mathcal{O}(n^2)$ Operationen und Speicherplatzaufwand. (Häufig verwenden Mehrgitterverfahren das Gauß-Seidel-Verfahren als Unteralgorithmus)

Die Diskretisierung der **zeitabhängigen Wärmeleitungsgleichung** (2.84) ergibt sich, indem man zusätzlich eine Zeitschrittweite τ, diskrete Zeitpunkte $t_s = t_0 + s \cdot \tau$ mit $s = 0, 1, ...$ sowie eine Approximation $u_{s,i,j} \approx u(t_s, x_i, y_j)$ einführt. In jedem Zeitschritt $t_s \to t_{s+1}$ erhält man für $\forall i, j = 1, \dots, n-1$ das lineare Gleichungssystem

$$\frac{u_{s,i,j}}{\tau} - k \cdot \left[\frac{u_{s,i-1,j} - 2u_{s,i,j} + u_{s,i+1,j}}{h^2} + \frac{u_{s,i,j-1} - 2u_{s,i,j} + u_{s,i,j+1}}{h^2} \right] = f(t_s, x_i, y_i) + \frac{u_{s-1,i,j}}{\tau} =: \widetilde{f}_{s,i,j}, \tag{2.90}$$

dessen Matrix dann sogar echt diagonaldominant wäre.

2.6.2.2 Finite-Elemente-Methoden

Grundlegende Idee: Energieminimierung in endlichdimensionalen Teilräumen
Sei hierzu wieder die partielle Differentialgleichung

$$\forall \begin{pmatrix} x \\ y \end{pmatrix} =: \vec{x} \in \Omega : \ -k \left(\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} \right) = f(x,y) \tag{2.84.S.2}$$

unter der Randbedingung $u\big|_{\partial\Omega} = 0$ gegeben.
Wir teilen nun den Lösungsraum Ω in endlich viele „Elemente“[8] auf, es soll gelten, dass

$$\Omega = \bigcup_{h=1}^{m} G_h.$$

Statt auf dem Funktionenraum V suchen wir jetzt auf dem **endlich**dimensionalen Teilraum $V_h \subset V$ nach einer Näherungslösung $u_h \in V_h := \left\{ \varphi : \Omega \to \mathbb{R} \;\middle|\; \varphi\big|_{\Omega_h} \text{ sei affin-linear}, \varphi \in \mathcal{C}^0, \varphi\big|_{\partial\Omega} = 0 \right\}$.
Bisheriges Problem: $u_h \in V_h$ ist eben nicht zweimal stetig differenzierbar. Wie kann u_h also Lösung des Problems sein?
Wir multiplizieren dazu Differenzialgleichung mit einer Testfunktion φ und integrieren anschließend über Ω, erhalten somit

$$(2.84.S.2) \Rightarrow -k \int_\Omega \left(\partial_x^2 u\varphi + \partial_y^2 u\varphi \right) \, d\vec{x} = \int_\Omega f \cdot \varphi \, d\vec{x} \quad \forall \varphi \in V_h. \tag{2.91}$$

Mittels partieller Integration ($\to$ Satz von Gauß) im Mehrdimensionalen ergibt sich:

$$(2.91) \xRightarrow[\text{Int.}]{\text{part.}} k \int_\Omega \partial_x u \partial_x \varphi + \partial_y u \partial_y \varphi \, dx - \underbrace{\int_{\partial\Omega} \frac{\partial u}{\partial \nu} \varphi \, ds}_{\substack{= 0, \text{ da gilt, dass} \\ \varphi\big|_{\partial\Omega} = 0}} = \int_\Omega f\varphi \, d\vec{x} \quad \forall \varphi \in V \tag{2.92}$$

[8] Meistens wird eine Triangulierung des Gebiets verwendet.

Sei nun $\varphi_1, \ldots, \varphi_N$ eine Basis von V_h, $N = \dim(V_h) = \#\text{innererGitterpuntkte}$, so gilt:

$$(2.92) \Longleftrightarrow k \int_\Omega \partial_x u \partial_x \varphi_j + \partial_y u \partial_y \varphi_j \, dx = \int_\Omega f \varphi_j \, d\vec{x} \quad \forall j = 1, \ldots, N \tag{2.93}$$

Wir suchen nun ein u_h, dass diese N Gleichungen erfüllt. Wir sehen auch, dass ein u_h diese erfüllen kann, da zweite Ableitungen nicht mehr vorkommen, sondern nur erste und solche sind für obige stückweise-affin-linearen Funktionen (fast überall) wohldefiniert.
Wir stellen u_h also als Linearkombination der φ_1 bis φ_N dar ($\rightarrow$ C1, das dürfen wir Œ):

$$u_h(\vec{x}) = \sum_{i=1}^{N} \alpha_i \varphi_i(\vec{x}) \tag{2.94}$$

Wir setzen (2.94) dann in (2.93) ein und erhalten somit ein lineares Gleichungssystem:

$$k \cdot \int_\Omega \sum_{i=1}^{N} \alpha_i \cdot \left(\partial_x \varphi_i \partial_x \varphi_j + \partial_y \varphi_i \partial_y \varphi_j \right) \, dx = \int_\Omega f \varphi_j \, d\vec{x} \quad \forall j = 1, \ldots, N \tag{2.95}$$

$$\Leftrightarrow \sum_{i=1}^{N} \underbrace{\int_\Omega k \cdot \left\langle \nabla \varphi_i(\vec{x}), \nabla \varphi_j(\vec{x}) \right\rangle \, d\vec{x}}_{a_{ij}\text{: Matrixeinträge}} \cdot \alpha_i = \underbrace{\int_\Omega f \varphi_j \, d\vec{x}}_{\substack{b_j\text{: Komp. der} \\ \text{rechten Seite}}} \quad \forall j = 1, \ldots, N \tag{2.96}$$

Dieses lineare Gleichungssystem ist dann über gleiche Methoden wir die Gleichungssysteme der finiten Differenzen zu lösen.
Eine Basis mit sehr guten Eigenschaften ist $\varphi_j(\vec{x}_i) = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{sonst} \end{cases}$, da die zugehörige Matrix dann dünn besetzt ist, somit zumindest bei Mehrgitterverfahren oder einfachen iterativen Lösern es zu keinem Speicherproblem kommt.
Dies soll es mit der Einführung in Differentialgleichungen gewesen sein, wir beschäftigen uns nun nocheinmal näher mit der Algebra.

KAPITEL 3

EINFÜHRUNG IN DIE ALGEBRA

Wir wollen uns nun mit einem anderen Gebiet der Mathematik beschäftigen, dem der Algebra. Algebren spielen neben „Spezialdisziplinen“ wie der theoretischen Informatik auch im Alltag eine große Rolle. In Kapitel 3.1 wollen wir uns zuerst mit den Grundlagen der Algebra beschäftigen bevor in Kapitel 3.2 die erste große Anwendung mit Prüfziffern und -summen und in Kapitel 3.3 mit RSA kennenlernen werden.
Während hierbei die Grundlagen der Algebra beigebracht werden sollen, wollen wir hier den Körper $\mathbb{R}$ der reellen Zahlen mit seinen Körperaxiomen, den Anordnungsaxiomen und dem Vollständigkeitsaxiom, sowie seinen Erweiterungskörper $\mathbb{C}$ der komplexen Zahlen mit dem Automorphismus $z \mapsto \bar{z}$ der komplexen Konjugation als gegeben voraussetzen. Ebenso wollen wir die ganzen, rationalen sowie natürlichen Zahlen als algebraische Einheiten als gegeben voraussetzen, bei den ganzen Zahlen dann im speziellen noch:

$\boxed{\mathcal{R}}$ $\mathbb{Z}$ ist ein kommutativer Ring mit Einselement 1, der keine Nullteiler besitzt, das heißt in $\mathbb{Z}$ gilt

$$x \cdot y = 0 \Rightarrow x = 0 \vee y = 0$$

$\boxed{\mathcal{O}}$ $\mathbb{Z}$ wird durch $\leqslant$ und $\geqslant$ linear geordnet. Diese Anordnung erfüllt bezüglich Addition und Multiplikation die Beziehungen

$$a, b, c \in \mathbb{Z} \text{ und } a \left\{ \begin{matrix} \leqslant \\ \geqslant \end{matrix} \right\} b \Rightarrow a + c \left\{ \begin{matrix} \leqslant \\ \geqslant \end{matrix} \right\} b + c$$

$$a, b \in \mathbb{Z} \text{ und } 0 \left\{ \begin{matrix} \leqslant \\ \geqslant \end{matrix} \right\} a, b \Rightarrow 0 \leqslant ab$$

$$a, b \in \mathbb{Z} \text{ und } 0 \left\{ \begin{matrix} \leqslant \\ \geqslant \end{matrix} \right\} a, 0 \left\{ \begin{matrix} \geqslant \\ \leqslant \end{matrix} \right\} b \Rightarrow 0 \geqslant ab$$

$\boxed{\mathcal{I}}$ Jede nichtleere und nach unten beschränkte Teilmenge von $\mathbb{Z}$ besitzt ein kleinstes Element, jede nichtleere und nach oben beschränkte Teilmenge von $\mathbb{Z}$ ein größtes Element.

Aus diesen Eigenschaften der ganzen Zahlen $\mathbb{Z}$ lassen sich weitere Gesetze ableiten, einige – bereits seit der Sekundarstufe I bekannten – wie etwa die Regeln für die Benutzung von Exponenten wollen wir ohne weiteres verwenden.

3.1 Grundlagen der Algebra

Eine der einfachsten algebraischen Strukturen besteht aus einer nichtleeren Menge M mit einer einzigen Verknüpfung $*$. Eine solche Algebra $(M, *)$ heiße ein **Gruppoid**. Hat die Verknüpfung $*$ weitere spezielle Eigenschaften, so erhält man spezielle Gruppoide, von welchen die für uns relevantesten hier erwähnt seien:

Definition 3.1 (*Gruppe, Halbgruppe, Monoid*)

Ein Gruppoid $(M, *)$ heiße ...

- ... **Halbgruppe** gdw. $*$ assoziativ ist, also

$$\forall\, a, b, c \in G.\ a * (b * c) = (a * b) * c \tag{G1}$$

  gilt.

- ... **Monoid** gdw. $*$ (G1) erfüllt und zudem noch ein neutrales Element existiert, also

$$\exists\, e \in G. \forall\, a \in G.\ a * e = e * a = a \tag{G2}$$

  gilt.

- ... **Gruppe** gdw. $(G, *)$ ein Monoid ist und zudem noch **jedes** Element ein Inverses bezüglich $*$ besitzt, also

$$\forall\, a \in G.\ \exists\, b \in G.\ a * b = b * a = e, \tag{G3}$$

  wobei e das neutrale Element des Monoids darstellt, gilt.

- ... **abel'sche,kommutative Gruppe** oder auch **Modul** gdw. $(G, *)$ eine Gruppe ist und $*$ kommutativ ist, also

$$\forall\, a, b \in G.\ a * b = b * a \tag{G4}$$

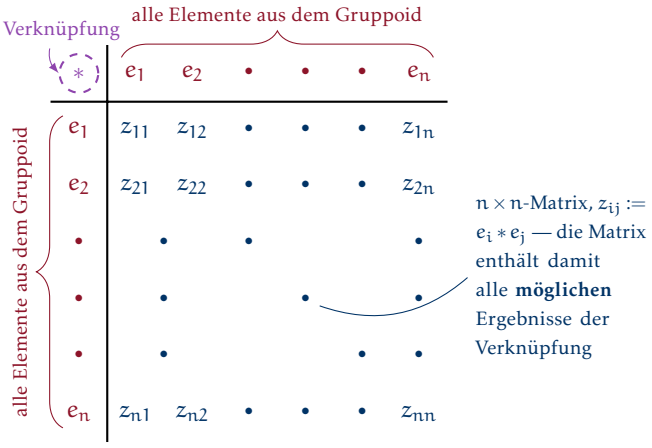
  gilt.

Satz 3.1 (*Eindeutigkeit des neutralen und inversen Elements*)

Sei $(M, *)$ ein Gruppoid, so ...

(a) ... gebe es in M **höchstens** ein neutrales Element bezüglich $*$.

(b) ... gebe es – vorausgesetzt $*$ sei assoziativ und habe ein neutrales Element – zu jedem $x \in M$ **höchstens** ein inverses Element.



Beweis:

(a) Seien sowohl e als auch $\widetilde{e}$ neutrale Elemente in M. Da $\widetilde{e}$ neutrales Element ist gilt $e = e * \widetilde{e}$, da e neutrales Element ist, gilt allerdings auch $\widetilde{e} = e * \widetilde{e}$, also $\widetilde{e} = e * \widetilde{e} = e$.

(b) Sei $x \in M$, e das neutrale Element und seien sowohl y als auch $\widetilde{y}$ inverse Elemente von x, so gilt:

$$y = y * e = y * (x * \widetilde{y}) = (y * x) * \widetilde{y} = e * \widetilde{y} = \widetilde{y}$$

□

Wie aus dem ersten Semester bekannt, werden Verknüpfungen auf **endlichen** Gruppoiden gerne auch durch **Operationstafeln/Verknüpfungstabellen** dargestellt, ein Beispiel für eine solche findet sich in der Skizze oben.

Beispiele für Gruppen (B)

- $(\mathbb{Z},+)$, $(\mathbb{Q},+)$, $(\mathbb{R},+)$, aber **nicht** $(\mathbb{N},+)$
- $(\mathbb{Q}\backslash\{0\},\cdot)$, $(\mathbb{R}\backslash\{0\},\cdot)$, aber **nicht** $(\mathbb{Z}\backslash\{0\},\cdot)$
- $(\mathbb{R}^{n\times n},+)$ und $\left(\left\{A\in\mathbb{R}^{n\times n}\;\middle|\; A \text{ ist invertierbar}\right\},\cdot\right)$, aber **nicht** $(\mathbb{R}^{n\times n},\cdot)$ aufgrund von fehlenden Inversen!
- $(\mathbb{Z}[X],+)$, $(\mathbb{Q}[X],+)$, $(\mathbb{R}[X],+)$

Wir wiederholen weiter den Begriff der Untergruppe:

Definition 3.2 (*Untergruppen*)

Sei $(G,*)$ eine Gruppe, Eine nichtleere Teilmenge $\emptyset \neq U \subseteq G$ heißt **Untergruppe** von G genau dann, wenn

(i) $\forall\, a,b\in U.\ a*b\in U$

(ii) $\forall\, a\in U.\ a^{-1}\in U$

Satz 3.2 (*Untergruppen*)

Eine jede Untergruppe ist mit der ererbten Verknüpfung eine Gruppe.

Beweis: trivial. □

Neben Gruppoiden bilden Mengen mit zwei Verknüpfungen die nächsteinfacheren algebraischen Strukturen. Wir wollen definieren:

Definition 3.3 (*Ring*)

Eine Algebra $(M,+,\cdot)$ heiße ein **Ring**, wenn gilt:

- $(M,+)$ ist ein **Modul**
- $(M,\cdot)$ ist eine **Halbgruppe**
- Es gelte „$\cdot$“ ist **distributiv** bezüglich „$+$“, also

$$\forall\, x,y,z\in M.\ [x(y+z)=xy+xz]\wedge[(x+y)z=xz+yz] \tag{R1}$$

Das neutrale Element des Moduls $(M,+)$ heiße **Nullelement** (0).
Ein Ring $(M,+,\cdot)$ heiße ...

- ... **Ring mit Einselement**, wenn die Halbgruppe $(M,\cdot)$ ein **Monoid** ist, also ein neutrales Element existiert.
- ... **kommutativer Ring**, wenn $(M,\cdot)$ kommutativ ist und somit (G4) erfüllt.

Beispiele für Ringe (B)

- $(\mathbb{Z},+,\cdot)$ ist ein *kommutativer Ring mit Einselement*
- $(\mathbb{R}^{n\times n},+,\cdot)$ und $(\mathrm{Lin}(\mathbb{R}^n,\mathbb{R}^n),+,\cdot)$ sind ***nicht**kommutative Ringe mit Einselement*
- $(\mathbb{Z}[X],+,\cdot)$ ist ein *kommutativer Ring mit Einselement*

Definition 3.4 (*Nullteiler und Integritätsringe*)

Sei $(M,+,\cdot)$ ein Ring mit Nullelement 0. Gilt für $a,b \in M\backslash\{0\}$ die Gleichung

$$a \cdot b = 0,$$

so heißen a,b **Nullteiler**, genauer a ein **Links-** und b ein **Rechtsnullteiler**.
Ein **kommutativer** Ring $(M,+,\cdot)$ **mit** Einselement und **ohne Nullteiler** heiße **Integritätsbereich** oder **Integritätsring**.

i Sei $(M,+,\cdot)$ ein nullteilerfreier Ring, so ist der Schluss

$$a \cdot b = \widetilde{a} \cdot b \wedge b \neq 0 \Rightarrow a = \widetilde{a}$$

gültig.

Lemma 3.3 (*Rechenregeln*)

In einem Ring $R = (M+,\cdot)$ mit Nullelement 0 gilt:

(a) $a0 = 0a = a \quad \forall\, a \in M$

(b) $(-a)b = a(-b) = -(ab)$

(c) $(-a)(-b) = ab$

(d) Ist R ein Ring mit Einselement 1, so gilt $1 = 0$ genau dann, wenn $R = \{0\}$ der Nullring ist.

Beweis:

(a) Es gilt sowohl $a0 = a(0+0) = a0 + a0$ als auch $0a = (0+0)a = 0a + 0a$, was nach Addition des Inversen von $a0$ beziehungsweise $0a$ zu der Beziehung $a0 = 0$ und $0a = 0$ führt.

(b) Es gilt $ab + (-a)b = (a-a)b = 0$ und $ab + a(-b) = a(b-b) = 0$.

(c) Es gilt $ab - (-a)(-b) = ab + a(-b) = 0$.

(d) Ist $R = \{0\}$ so erfüllt 0 offensichtlich alle Eigenschaften des Einselements. Ist umgekehrt $1 = 0$, so folgt für alle $a \in R$, dass $a = 1a = 0a = 0$.

□

Lemma 3.4 (*Spezieller Identitätsring*)

Sei ein Ring $(M,+,\cdot)$ mit Nullelement 0 gegeben. Bildet das Gruppoid $(M\backslash\{0\},\cdot)$ ein Modul, so ist der Ring $(M,+,\cdot)$ kommutativ und es existiert ein Einselement, aber keine Nullteiler.

Beweis: Sei also nun ein Ring $R = (M,+,\cdot)$ mit Nullelement 0 und dem Modul $(M\backslash\{0\},\cdot)$ gegeben. Dann ist trivial zu folgern, dass R kommutativ ist, denn dies folgt unmittelbar aus der Moduleigenschaft. Diese kann einfach auf die 0 fortgesetzt werden, man möchte dazu das Ergebnis aus Lemma 3.3a verwenden.
Ebenso einfach ist es zu zeigen, dass der Monoid $(M,\cdot)$ eine Halbgruppe ist. Wir kennen bereits den Kandidaten für das Einselement, wir verwenden die Eins 1 aus dem Modul. Jetzt bleibt es noch zu zeigen, dass die Eins auf die Null fortgesetzt werden kann, dies entspricht aber genau der Aussage aus Lemma 3.3a.
Wir zeigen nun noch die Nichtexistenz von Nullteilern, angenommen es existierten Nullteiler,

sprich $\exists\, a, b \in M\backslash\{0\}.\ ab = 0$. Sei dann a^{-1} das Inverse von a bezüglich $\cdot$, was aufgrund der Moduleigenschaft von $(M\backslash\{0\}, \cdot)$ existiert und eindeutig ist. Wir folgern dann

$$b = 1b = a^{-1}ab = a^{-1}0 = 0$$

und erhalten einen Widerspruch. Damit kann es keine Nullteiler geben. □
Wir erhalten somit spezielle Integritätsringe, denen wir besondere Namen geben wollen:

Definition 3.5 (*Körper*)

Ein Ring $(M, +, \cdot)$ mit **mehr** als einem Element heiße ein Körper, wenn $(M\backslash\{0\}, \cdot)$ eine abel'sche Gruppe ist.

Andere Definitionen

Es ist anzumerken, dass der Ring nicht nur ein Element enthalten darf. Dies wird aufgrund von Lemma 3.3d gefordert.

Eine andere äquivalente Definition wäre: „Ein **kommutativer** Ring $(M, +\cdot)$ mit **Eins** heiße ein **Körper**, wenn jedes Element (außer das der Eins verschiedene Nullelement) ein **Inverses** bezüglich »·« hat."

Eine wiederum andere Definition wäre eine Definition über beschreibende Merkmale, ähnlich wie wir es bereits bei der Definition von Ringen und Gruppen gemacht haben. Die Definition lautete dann: „Eine Algebra $(M, +, \cdot)$ heiße ein Körper, wenn gilt:

K1 $(M, +)$ ist eine abel'sche Gruppe mit Nullelement 0 und Inversem $-a$ von a

K2 $(M\backslash\{0\}, \cdot)$ ist eine abel'sche Gruppe mit neutralem Element 1 und Inversem a^{-1} von a

K3 Es gelten die Distributivgesezte (R1)."

Man kann sich hier bei den Distributivgesetzen eine Richtung sparen.

Definition 3.6 (*Schiefkörper*)

Sind alle Bedingungen bis auf die Kommutativität bei **K2** in Variante 3 der Definition 3.5 erfüllt, so sprechen wir von einem **Schiefkörper**. Hier ist es notwendig die Distributivgesetze auch in beide Richtungen zu fordern.

Beispiele für Körper

- $(\mathbb{Q}, +, \cdot)$, $(\mathbb{R}, +, \cdot)$ und $(\mathbb{C}, +, \cdot)$ sind Körper, $(\mathbb{Z}, +, \cdot)$ aber offensichtlich **nicht**.
- $(\mathbb{R}^{n\times n}, +.\cdot)$ und $\left(\left\{A \in \mathbb{R}^{n\times n} \;\middle|\; A \text{ ist invertierbar}\right\}, +, \cdot\right)$ sind **keine** Körper.
- Aus Vorgängerveranstaltungen wissen wir, dass $\mathbb{C} \simeq \mathbb{R}^2$ mit der speziell definierten Addition „$+_{\mathbb{C}}$" und Multiplikation „$\cdot_{\mathbb{C}}$" ein Körper ist, es ist auch eine Isomorphie zu speziell definierten 2×2-Matrizen möglich.
  Tatsächlich ist dies der einzige Körper der Form $\mathbb{R}^n$ mit $n \geqslant 2$.
- Teilmengen von Körpern können wieder Körper sein, mit den vom „Oberkörper" geerbten Eigenschaften und Verknüpfungen.

Beispiel für einen Schiefkörper

B $(\mathbb{H},+,\cdot)$, wobei $\mathbb{H}$ die Quaternionen ($\simeq\mathbb{R}^4$) beschreibt, ist ein **Schiefkörper**. Sie entstehen durch das sogenannte Verdoppelungsverfahren, bei welchem allerdings immer wieder algebraische Eigenschaften wegfallen. Bei den komplexen Zahlen war dies die Ordenbarkeit, bei den Quaternionen ist es die Kommutativität, bei den Oktonionen ($\simeq\mathbb{R}^8$) ist es die Assoziativität (sie bleiben allerdings alternativ) und bei den Sedenionen ($\simeq\mathbb{R}^{16}$) kommen Nullteiler hinzu.

Wir wollen uns im Folgenden mit einer speziellen Art von Ringen und Körpern befassen, den **Restklassenringen** oder **-körpern**. Wir rufen uns dazu noch einmal in Erinnerung:
Für zwei Mengen M,N bezeichnen wir $K\subseteq M\times N$ als **Korrespondenz** zwischen M und N und die Korrespondenz $R\subseteq M\times M$ als **Relation** auf der Menge M. Wir haben daraufhin verschiedene Eigenschaften von Relationen kennengelernt:

Wiederholung 3.7 (*Eigenschaften von Relationen*)

Sei R eine Relation auf einer Menge M. Dann heiße R ...

- ... **reflexiv** gdw. $\forall\, x\in M.\ xRx$.
- ... **transitiv** gdw. $\forall\, x,y,z\in M.\ (xRy\wedge yRz\Rightarrow xRz)$.
- ... **symmetrisch** gdw. $\forall\, x,y\in M.\ xRy\Rightarrow yRx$.
- ... **antisymmetrisch** gdw. $\forall\, x,y\in M.(xRy\wedge yRx)\Rightarrow(x=y)$.
- ... **alternativ** gdw. $\forall\, x,y\in M.(xRy\vee yRx)$.

Anhand dieser Definition haben wir dann den Begriff einer **Äquivalenzrelation** (als eine reflexive, symmetrische sowie transitive Relation) sowie einer **Ordnungsrelation** (als eine reflexive, antisymmetrische und transitive Relation) definiert.
Vor allem im Bereich der Äquivalenzrelationen gibt es dann sogenannte **Äquivalenzklassen** $[x]_R:=\{y\in M\ :\ xRy\}\subset M$, wobei x auch **Vertreter** oder **Repräsentant** der Äquivalenzklasse $[x]_R$ bezüglich der Äquivalenzrelation R genannt wird. Die so gebildeten Äquivalenzklassen bilden dann eine **Partition** von M, also ein System $\{A_i\ :\ i\in I\}$ von nichtleeren, paarweise disjunkten Teilmengen $A_i\subset M$ mit $\bigcup_{i\in M}A_i=M$, das heißt

$$\forall\, x,y\in M.\ x\in[x]_R, M=\bigcup_{x\in M}[x]_R \text{ und } [x]_R\cap[y]_R\neq\emptyset\Leftrightarrow[x]_R=[y]_R.$$

Wir betrachten nun die Quotientenmenge $M/R:=\{[x]_R\mid x\in M\}$ mit einer Verknüpfung $*$ unter der Voraussetzung, dass R mit $*$ strukturverträglich sei. Dann „transportieren“ wir $*$ auf M/R mit

$$\forall\, x,y\in M.\ [x]_R*[y]_R=[x*y]_R=\{z\in M\ :\ (x*y)Rz\}.$$

Damit können wir nun die Kongruenzrelation auf $\mathbb{Z}$ nocheinmal näher betrachten; wir erinnern uns an

$$R_n:=\{(x,y)\in\mathbb{Z}\times\mathbb{Z}\ :\ \overbrace{n\mid(x-y)}^{n\text{ teilt }(x-y)}\}$$

mit der Kurzschreibweise $xR_ny=x\equiv_n y\mod n$ und daran, dass die Faktoralgebra $\mathbb{Z}_n:=\mathbb{Z}/\equiv_n$ aus genau n Äquivalenzklassen, welche wir fortan **Restklassen** nennen wollen, besteht, nämlich:

$$\mathbb{Z}_n:=\{[0]_n,\dots,[n-1]_n\}\text{ mit }[a]_n:=\{x\in\mathbb{Z}:x\equiv_n a\mod n\}=\{x\in\mathbb{Z}\ :\ n\mid(a-x)\},$$

wobei wir $[a]_n$ als **Restklasse von a modulo n** bezeichnen wollen. Dabei gilt $x\in[a]_n$ genau dann, wenn ein $k\in\mathbb{Z}$ existiert mit $k\cdot n=a-x$, sprich a durch n teilbar ist mit Rest x.
Addition und Multiplikation werden nun erklärt durch den „Transport“ von $\mathbb{Z}$, also definieren wir respektive mit

$$[a]_n+[b]_n:=[a+b]_n\qquad\qquad[a]_n\cdot[b]_n:=[a\cdot b]_n.$$

Dass diese Operationen wohldefiniert sind, ist leicht zu zeigen, siehe dazu auch das Skript der Vorgängerveranstaltung. Wir wollen nun Eigenschaften der Strukturen $(\mathbb{Z}_n,+)$, $(\mathbb{Z}_n,\cdot)$ und $(\mathbb{Z}_n,+,\cdot)$ herausfinden:

Behauptung 1

$(\mathbb{Z}_n,+)$ *ist ein Modul mit neutralem Element* $[0]_n$.

Beweis:

→ $(\mathbb{Z}_n,+)$ **ist Halbgruppe**: Zu zeigen ist also die Assoziativität von $+$ (sprich (G1)):

$$[a]_n+([b]_n+[c]_n) = [a]_n+[b+c]_n = [a+(b+c)]_n = [(a+b)+c]_n = [a+b]_n+[c]_n = ([a]_n+[b]_n)+[c]_n \checkmark$$

→ $(\mathbb{Z}_n,+)$ **ist ein Monoid mit neutralem Element** $[0]_n$:

$$[a]_n + [0]_n = [a+0]_n = [a]_n = [0+a]_n = [0]_n + [a]_n \qquad \checkmark$$

→ $(\mathbb{Z}_n,+)$ **ist eine Gruppe**: Zu zeigen ist die Existenz eines inversen Elements zu jedem Element $[a]_n \in \mathbb{Z}_n$. Wir setzen als Kandidaten für das inverse Element $-[a]_n := [-a]_n$, für jedes $[a]_n$ ist dieses wohldefiniert. Wir zeigen, dass es ebenso die Eigenschaft (G3) erfüllt:

$$[a]_n + (-[a]_n) = [a-a]_n = [0]_n = [-a+a]_n = (-[a]_n) + [a]_n \qquad \checkmark$$

→ $(\mathbb{Z}_n,+)$ **ist Modul**: Zu zeigen bleibt Eigenschaft (G4). Seien $[a]_n, [b]_n \in \mathbb{Z}_n$ beliebig, aber fest, so gelte

$$[a]_n + [b]_n = [a+b]_n = [b+a]_n = [b]_n + [a]_n \qquad \checkmark$$

Die Aussagen verwenden allesamt die Eigenschaft, dass $(\mathbb{Z},+)$ ein Modul ist. □

Behauptung 2

$(\mathbb{Z}_n,\cdot)$ *ist ein kommutativer Monoid.*

Beweis:

→ $(\mathbb{Z}_n,\cdot)$ **ist Halbgruppe**: ✓ (funktioniert genauso wie eben)

→ $(\mathbb{Z}_n,\cdot)$ **ist ein Monoid mit neutralem Element** $[1]_n$:

$$[a]_n \cdot [1]_n = [a\cdot 1]_n = [a]_n = [1\cdot a]_n = [1]_n + [a]_n \qquad \checkmark$$

→ $(\mathbb{Z}_n,\cdot)$ **ist kommutativ**: ✓ (funktioniert genauso wie eben)

Die Aussagen verwenden allesamt die Eigenschaft, dass $(\mathbb{Z},\cdot)$ ein kommutativer Monoid ist. □

Behauptung 3

$(\mathbb{Z}_n,+,\cdot)$ *ist ein kommutativer Ring mit Eins.*

Beweis:

→ $(\mathbb{Z}_n,+)$ **ist Modul**: ✓ (siehe Behauptung 1)

→ $(\mathbb{Z}_n,\cdot)$ **ist ein Monoid**: ✓ (siehe Behauptung 2)

→ **Es gelte das Distributivgesetz**: Seien $[a]_n, [b]_n, [c]_n \in \mathbb{Z}_n$ beliebig, aber fest, so gelte:

$$[x]_n\cdot([y]_n+[z]_n) = [x\cdot(y+z)]_n = [xy]_n+[xz]_n \text{ und } ([x]_n+[y]_n)\cdot[z]_n = [(x+y)\cdot z]_n = [xz]_n+[yz]_n \quad \checkmark$$

Die Aussagen verwenden allesamt die Eigenschaft, dass $(\mathbb{Z}, +, \cdot)$ ein kommutativer Ring mit Eins ist. □

Überrascht uns das?

Die mit Behauptung 3 erzielten Ergebnisse sind selbstverständlich nicht überraschend, so das ja $(\mathbb{Z}, +, \cdot)$ bereits ein kommutativer Ring mit Eins ist und die Addition und Multiplikation auf dem Restklassenring die Operationen auf die ganzen Zahlen abbilden. Die klassifizierenden Eigenschaften bleiben hierbei erhalten.

Definition 3.8 (*Restklassenring*)

$(\mathbb{Z}_n, +, \cdot)$ – kurz $\mathbb{Z}_n$ – heißt **Restklassenring** von $\mathbb{Z}$ modulo n.

Wegen der Existenz eines Einselements in $\mathbb{Z}_n$ ist es nun berechtigt zu fragen, ob zu jeder Restklasse $[a]_n \in \mathbb{Z}_n$ auch ein Inverses $[a]^{-1} \in \mathbb{Z}_n$ existiert. Für $[0]_n$ ist die offensichtlich nicht erfüllt, wir schränken unsere Suche also auf $\mathbb{Z}_n \backslash \{[0]_n\}$ ein. Ohne Beweis haben wir im ersten Semester behauptet, dass dies nur dann funktioniert, wenn n prim ist. Wir wollen diese Vermutung nun beweisen:
Als erste Motivation und Idee wollen wir die Begriffe „Abgeschlossenheit“ und „Existenz von Inversen“ miteinander in Beziehung stellen:

$\bullet$	$[1]_4$	$[2]_4$	$[3]_4$
$[1]_4$	$[1]_4$	$[2]_4$	$[3]_4$
$[2]_4$	$[2]_4$	$[0]_4$	$[2]_4$
$[3]_4$	$[3]_4$	$[2]_4$	$[1]_4$

Beispiel der Nichtabgeschlossenheit von $\mathbb{Z}_4 \backslash \{[0]_4\}$ bezüglich „$\cdot$“

Satz 3.5 (*Äqivalenz von Abgeschlossenheit und der Existenz eines Inversen*)

Im Restklassenring $(\mathbb{Z}_n, +, \cdot)$ gilt für jedes $[a]_n \in \mathbb{Z}_n$:

$$[a]_n \text{ ist invertierbar } \Rightarrow \underbrace{\forall\, [b]_n \in \mathbb{Z}_n \backslash \{[0]_n\}.\ [a]_n \cdot [b]_n \neq [0]_n}_{\Leftrightarrow\ \mathbb{Z}_n \text{ ist nullteilerfrei}}$$

Beweis:

„$\Rightarrow$“: Sei $[a]_n$ also invertierbar, das heißt es existiert ein $[a]_n^{-1}$, so dass $[a]_n \cdot [a]_n^{-1} = [1]_n$. Angenommen es existierten $[b]_n \in \mathrm{fieldZ}_n \backslash \{[0]_n\}$, so dass $[a]_n \cdot [b]_n = [0]_n$. Durch die Addition beider Gleichungen und Anwendung des Distributivgesetzes erhalten wir:

$$[a]_n \cdot ([a]_n^{-1} + [b]_n) = [1]_n \Leftrightarrow [a]_n^{-1} + [b]_n = [a]_n^{-1} \Rightarrow [b]_n = 0 \quad \lightning$$

„$\Leftarrow$“: Angenommen es gelte $\forall\, [b]_n \in \mathbb{Z}_n \backslash \{[0]_n\}.\ [a]_n \cdot [b]_n \neq [0]_n$. Wir schließen daraus zuerst, dass damit $[a]_n \neq [0]_n$ gelten muss. Wir wollen nun zeigen, dass die Terme $[a]_n \cdot [1]_n, \ldots, [a]_n \cdot [n-1]_n$ allesamt paarweise verschieden sind, woraus dann die Existenz eines Inversen folgt.
Angenommen es gelte nun

$$[a]_n \cdot [b_1]_n = [a]_n \cdot [b_2]_n \text{ für } [b_1]_n \neq [b_2]_n.$$

Durch Addition des Additivinversen der rechten Seite, das Anwenden des Distributivgesetzes sowie der Definition der Addition auf $\mathbb{Z}_n$ folgern wir:

$$[a]_n \cdot [b_1 - b_2]_n = [0]_n,$$

da aber $[b_1]_n \neq [b_2]_n \Leftrightarrow [b_1 - b_2]_n \neq [0]_n$ gilt, erhalten wir einen Widerspruch $\lightning$. □

Um letztendlich unsere Vermutung zu zeigen, bringen wir nun noch die Begriffe der **Teilerfremdheit** und *Inversenexistenz* in Korrelation. Wir wollen dazu zuerst die Teilbarkeit und danach den **größten gemeinsamen Teiler** einführen.

Definition 3.9 (*Teilbarkeit*)

Seien $a, b \in \mathbb{Z}$ gegeben. Man sagt **a teilt b** (und schreibt $a \mid b$) genau dann, wenn es ein $c \in \mathbb{Z}$ gibt mit der Gleichung $ac = b$. Demnach bedeutet *a teilt b* genau dasselbe wie *a ist ein Faktor von b*.

Wegen $ac = (-a)(-c)$ und $(-a)c = a(-c) = -ac$ sind je zwei der drei Aussagen $a \mid b$, $-a \mid b$ und $a \mid -b$ gleichbedeutend. Es genügt sich daher bei der Teilbarkeitsuntersuchung der ganzen Zahlen sich auf die Untersuchung von $\mathbb{N}_0$ zu beschränken.
Direkt aus der Definition ergibt sich für alle $a, b, c \in \mathbb{N}_0$, dass die Teilbarkeitsrelation reflexiv, transitiv und antisymmetrisch ist. Der Beweis der Reflexivität und Transitivität ist trivial, der der Antisymmetrie beruht auf der Tatsache, dass in $\mathbb{Z}$ außer dem neutralen Element 1 keine multiplikativen Inversen existieren. Damit liefert die Teilbarkeit auf $\mathbb{N}_0$ eine weitere Ordnungsrelation. Ebenso trivial sind die Aussagen, dass für alle $a \in \mathbb{N}_0$ gilt, dass $1 \mid a$ und $a \mid 0$ gilt. Hinsichtlich der Teilbarkeit ist also 1 das kleinste und 0 das größte Element von $\mathbb{N}_0$.
Wir kommen zu einem interessanten Lemma, dass die Teilbarkeit mit der Ordnungsrelation auf den natürlichen Zahlen in Verbindung stellt:

Lemma 3.6 (*Teilbarkeitsanordnung auf den natürlichen Zahlen*)

Auf der Menge der natürlichen Zahlen **ohne** Null ist die Teilbarkeitsrelation vergleichbar mit der Anordnung:

$$a, b \in \mathbb{N} \wedge a \mid b \Rightarrow a \leqslant b$$

Beweis: Seien $a, b \in \mathbb{N}$ beliebig, aber fest, und gelte $a \mid b$. Dann ist dadurch die Existenz eines $c \in \mathbb{Z}$ festgelegt mit $ac = b$. Es gilt nun, dass $c \geqslant 1$ gelten muss, da sonst $-c \geqslant 0$, also $a(-c) = -b \geqslant 0$ gelte, was den Voraussetzungen widerspräche. Es folgt dann die Abschätzung $b - a = a(c-1) \geqslant 0$, also $b \geqslant a$. □
Wir wollen nun bevor wir den größten gemeinsamen Teiler einführen zeigen, dass die Division mit Rest eindeutig ist.

Satz 3.7 (*Division mit Rest*)

Seien $a_0, a_1 \in \mathbb{Z}$ mit $a_1 > 0$ gegeben, so existiert genau ein $q_1 \in \mathbb{Z}$, so dass für den Rest $a_2 = a_0 - a_1 q_1$ gilt, dass $0 \leqslant a_2 < a_1$.

Beweis: Wir zeigen zuerst die **Existenz** eines solchen Paares q_1, a_2, indem wir die Menge

$$R = \left\{ r \in \mathbb{N}_0 \;\middle|\; r = a_0 - a_1 q \quad \text{für ein} \quad q \in \mathbb{Z} \right\}$$

betrachten. Sie ist nicht leer, da falls $a_0 \geqslant 0$ gelte, $a_0 \in \mathbb{R}$ mit $q = 0 \in \mathbb{Z}$ wäre und falls $a_0 < 0$, $-a_0 > 0$ und daher $a_0 - a_1 a_0 = -a_0(a_1 - 1) \geqslant 0$, also $a_0 - a_1 a_0 \in R$, gelte. Die Menge R ist zudem nach unten durch 0 beschränkt, womit es ein kleinstes Element geben muss. Dieses Minimum a_2 hat die Eigenschaft $0 \leqslant a_2 < a_1$, da für jedes $r \in R$ mit $a_1 \leqslant r$ gelte $r - a_1 \in R$ und damit $r - a_1 < r$. Wir wählen dann a_2 als das Minimum der Menge und q_1 als das zu a_2 gehörende q.
Wir beweisen nun noch die Eindeutigkeit der des Paares. Sei nun $a_2 = a_0 - a_1 q_1$ für ein passendes $q_1 \in \mathbb{Z}$. Angenommen es gelte für ein $q' \in \mathbb{Z}$ auch $0 \leqslant a_0 - a_1 q' < a_1$, sprich q' wäre ebenso ein Kandidat für die Division mit Rest. Dann folgt aus der Minimalität von a_2 in R die Abschätzung $a_2 = a_0 - a_1 q \leqslant a_0 - a_1 q'$. Sie ergibt $0 \leqslant a_1(q_1 - q') < a_1$. Hier gilt zunächst $q_1 - q' \geqslant 0$, aber $q_1 - q' > 0$ kann nach Lemma 3.6 nicht gelten. Dementsprechend gilt $q_1 = q'$. □
Wir wollen uns jetzt mit den Untergruppen von $\mathbb{Z}$ und ihrer Struktur beschäftigen, da dieser Zusammenhang spätere Beweise vereinfacht:

Satz 3.8 (*Untergruppen von* $\mathbb{Z}$)

Die Gesamtheit der Untergruppen U von $\mathbb{Z}$ ist gegeben in der Form $U = n\mathbb{Z} := \left\{ n \cdot m \;\middle|\; m \in \mathbb{Z} \right\}$ mit $n \in \mathbb{N}_0$. Ist U eine von $\left\{0\right\}$ verschiedene Untergruppe von $\mathbb{Z}$, so ist n das Minimum des

Durchschnittes $\mathbb{N} \cap U$. Daher ist n durch U stets eindeutig festgelegt.

Beweis: Wir beweisen zunächst die erste Aussage: Sei $n \in \mathbb{N}_0$ beliebig, aber fest. Dann gilt für die Menge $n\mathbb{Z}$, dass ...

... – vorausgesetzt $a, b \in n\mathbb{Z}$ seien beliebig, aber fest – auch $a + b \in n\mathbb{Z}$ aufgrund der Tatsache, dass $\exists\, m_1, m_2 \in \mathbb{Z}$. $a = nm_1 \wedge b = nm_2$ und damit

$$a + b = nm_1 + nm_2 = n\underbrace{(m_1 + m_2)}_{\in\mathbb{Z}} \in n\mathbb{Z}.$$

... – vorausgesetzt $a \in n\mathbb{Z}$ sei beliebig, aber fest – ebenso $-a \in n\mathbb{Z}$, da

$$-a = -nm_1 = n\underbrace{(-m_1)}_{\in\mathbb{Z}} \in n\mathbb{Z}.$$

Damit ist $n\mathbb{Z}$ eine Untergruppe von $(\mathbb{Z}, +)$. Im Fall $n > 0$ ist n nach Lemma 3.6 das Minimum der Menge $\mathbb{N} \cap n\mathbb{Z}$.
Wir wollen nun noch den letzten Teil der Aussage zeigen: Generell gilt, dass jede Untergruppe U von $(\mathbb{Z}, +)$ mit einem Element x auch sein Negativ $-x$ enthält. Daher ist im Fall $U \neq \{0\}$ sicher $\mathbb{N} \cap U \neq \emptyset$. Wir setzen nun $n = \min(\mathbb{N} \cap U)$ und beweisen $U = n\mathbb{Z}$:

$n\mathbb{Z} \subseteq U$: Aus der Tatsache $n \in U$ folgt per Induktion nach k die Tatsache $\forall\, k \in \mathbb{N}.\ nk \in U$.

$n\mathbb{Z} \supseteq U$: Man folgert diese Tatsache mit Satz 3.7. Gegeben sei ein beliebig, aber festes, Element $a_0 \in U$. Mit $a_1 := n$ und einem passenden Faktor $q_1 \in \mathbb{Z}$ gilt

$$0 \leqslant \underbrace{a_0 - a_1 q_1}_{\in U} < a_1 = n.$$

Nach Definition von n als Minimum von $\mathbb{N} \cap U$ kann dieses Element nicht positiv sein, es ist deshalb null, also gilt

$$a_0 = a_1 q_1 = n q_1 \in \mathbb{Z}.$$

□

Definition 3.10 (*Größter gemeinsamer Teiler – ggT*)

Eine Zahl $d \in \mathbb{Z}$ heiße **größter gemeinsamer Teiler** der Zahlen $a, b \in \mathbb{Z}$ ($d = \mathrm{ggT}(a, b)$), wenn gleichzeitig gilt:

$$d \geqslant 0, \quad d \mid a, \quad d \mid b \quad \text{und} \quad (t \mid a \wedge t \mid b) \Rightarrow t \mid d.$$

Anmerkungen zum größten gemeinsamen Teiler

Man kann einfach auch den größten gemeinsamen Teiler auf mehr als zwei Zahlen erweitern, man definiert dann die Zahl $d = ggT(a_1, \ldots, a_n)$ über die Eigenschaften

$$d \geqslant 0, \forall 1 \leqslant i \leqslant n.\ d \mid a_i \quad \text{und} \quad \left(\bigwedge_{1 \leqslant i \leqslant n} t \mid a_i \right) \Rightarrow t \mid d.$$

Es gilt dann im Fall $n \geqslant 3$ die Rechenregel

$$ggT(a_1, \ldots, a_n) = ggT(ggT(a_1, \ldots, a_{n-1}), a_n).$$

Beweis: Wir setzen $d = ggT(a_1, \ldots, a_r)$ sowie $d' = ggT(a_1, \ldots, a_{r-1})$ und $d'' = ggT(d', a_r)$. Dann sind per definitionem $d \geqslant 0$ und $d'' \geqslant 0$. Ferner ist d'' ein gemeinsamer Teiler von d' und a_r. Deshalb gilt $d'' \mid a_i$ für $1 \leqslant i \leqslant r-1$ und $d'' \mid a_r$. Dies hat dann wiederum zur Folge, dass $d'' \mid d$. Da andererseits d ein Teiler von $a_1, \ldots, a_{r-1}$ und a_r ist, gilt $d \mid d'$ und $d \mid a_r$, also $d \mid ggT(d', a_r) = d''$. Zusammen ergibt dies mit der Antisymmetrie der Teilbarkeit, dass $d = d''$ gelten muss. □

Satz 3.9 (*Existenzsatz für den größten gemeinsamen Teiler*)

Jedes System von $r \geqslant 1$ ganzen Zahlen $a_1, \ldots, a_r$ besitzt einen größten gemeinsamen Teiler d. Er ist gegeben durch die Gleichung

$$\sum_{i=1}^{r} a_i \mathbb{Z} = d\mathbb{Z}.$$

Beweis: Die Summe $U = \sum_{i=1}^{r} a_i \mathbb{Z}$ der Untergruppen $a_1\mathbb{Z}, \ldots, a_r\mathbb{Z}$ von $\mathbb{Z}$ ist jedenfalls wieder eine Untergruppe von $\mathbb{Z}$.[1] Nach Satz 3.8 hat diese dann die Form $U = d\mathbb{Z}$ mit einem $d \in \mathbb{N}_0$. Da alle $a_i \in U = d\mathbb{Z}$ sind, gilt $d \mid a_i$ mit $1 \leqslant i \leqslant r$. Es gibt nun noch geeignete $c_i \in \mathbb{Z}$ mit $\sum_{i=1}^{r} a_i c_i = d$. Ist t nun ein gemeinsamer Teiler der a_i, etwa $a_i = tb_i$ mit $1 \leqslant i \leqslant r$ und geeigneten b_i, dann folgt unmittelbar

$$d = \sum_{i=1}^{r} a_i c_i = t \cdot \sum_{i=1}^{r} b_i c_i,$$

woraus $t \mid d$ abzulesen ist. Also besitzt d die drei charakteristischen Eigenschaften des größten gemeinsamen Teilers. □

Rechenregeln zum größten gemeinsamen Teiler

Für je drei Zahlen $a, b, q \in \mathbb{Z}$ gilt $ggT(a, b) = ggT(b, a - bq)$.
Beweis: Wir setzen nun $a' := a - bq$. Es reicht nach Satz 3.9 die Gleichung $a\mathbb{Z} + b\mathbb{Z} = b\mathbb{Z} + a'\mathbb{Z}$ zu zeigen. Diese folgt aber daraus, dass a' ein Element ihrer linken Seite und $a = bq + a'$ ein Element ihrer rechten Seite ist. □

[1] Dies folgt aus einem einfach zu beweisenden Satz, dass wenn $(A, +)$ ein additives Modul sei und U_1, U_2 Untergruppen desselbigen sind, die Summe $S := U_1 + U_2 := \{a_1 + a_2 \mid a_i \in U_i \quad 1 \leqslant i \leqslant 2\}$ wiederum eine Untergruppe des Moduls sind. Bedingung (i) wird über die Ausnutzung der Kommutativität bewiesen, so gilt für $a', a'' \in S$ beliebig, aber fest, dass $a_1' + a_2' + a_1'' + a_2'' = a_1' + a_1'' + a_2' + a_2'' = a_1''' + a_2''' \in S$. Bedingung (ii) folgt trivialerweise.

Rechenregeln zum größten gemeinsamen Teiler (fort.)

Sind $a.b.c \in \mathbb{Z}$ und gilt $ggT(a,c) = 1$, so ist $ggT(ab,c) = ggT(b,c)$.
Beweis: Aufgrund der Voraussetzung gilt $a\mathbb{Z} + c\mathbb{Z} = \mathbb{Z}$. Wegen der offensichtlichen Inklusion $bc\mathbb{Z} \subseteq c\mathbb{Z}$ ist darum

$$\begin{aligned} b\mathbb{Z} + c\mathbb{Z} &= b(a\mathbb{Z} + c\mathbb{Z}) + c\mathbb{Z} \\ &= ab\mathbb{Z} + bc\mathbb{Z} + c\mathbb{Z} = ab\mathbb{Z} + c\mathbb{Z}, \end{aligned}$$

womit b und c dieselbe Untergruppe von $\mathbb{Z}$ wie ab und c erzeugen. Nach Satz 3.9 folgt daraus die zu zeigende Aussage. □

Ist $t \in \mathbb{N}_0$, so gilt stets $ggT(ta_1,\ldots,ta_r) = t \cdot ggT(a_1,\ldots,a_r)$.
Beweis: Wir setzen $d' := ggT(ta_1,\ldots,ta_r)$ und $d = ggT(a_1,\ldots,a_r)$. Da d alle a_i teilt, gilt $td \mid ta_i$ für alle $1 \leqslant i \leqslant r$, also $td \mid d'$. Nach Satz 3.9 gibt es aber ganze Zahlen $c_1,\ldots,c_r$ mit $d = \sum_{i=1}^{r} a_i c_i$. Wir folgern daraus

$$td = \sum_{i=1}^{r} ta_i c_i \in \sum_{i=1}^{r} (ta_i)\mathbb{Z} = d'\mathbb{Z},$$

und erhalten damit $d' \mid td$. Aus diesen beiden Aussagen folgt die zu zeigende Gleichheit. □

Sei $d = ggT((a_i)_{1\leqslant i\leqslant r})$ und d' eine Zahl die die Bedingungen an einen größten gemeinsamen Teiler ebenfalls erfüllt, dann gilt $d' = d$ und der größte gemeinsamen Teiler ist damit eindeutig bestimmt. Es gilt ferner $ggT(a_1,\ldots,a_r) = 0$ genau dann, wenn alle $a_i = 0$ sind.
Beweis: Wegen $d' \mid d$ und $d \mid d'$ gelten dann die Gleichungen $d = n'd'$ und $d' = nd$ mit geeigneten $n, n' \in \mathbb{Z}$. Sie ergeben dann $d = 0 \Leftrightarrow d' = 0$, und im Fall $d > 0$ die Abschätzungen $d' \leqslant d \leqslant d'$. Letztere Aussage ergibt sich direkt aus der Definition der Teilbarkeit. □

Definition 3.11 (*Teilerfremdheit*)

Zwei Zahlen $a, b \in \mathbb{Z}$ heißen **teilerfremd** genau dann, wenn $ggT(a,b) = 1$.

Satz 3.10 (*Variante des Lemmas von Bézout*)

$\forall\, a,b \in \mathbb{Z}.\ \exists\, \alpha,\beta \in \mathbb{Z}.\ ggT(a,b) = \alpha a + \beta b.$

Beweis: Wir müssen nur den Fall $a \neq 0$, $b \neq 0$ betrachten, da für die anderen Fälle die obige Zerlegung trivial möglich ist. Wir betrachten hierfür die Menge

$$S := \left\{ ax + by \;\middle|\; x,y \in \mathbb{Z} \text{ und } (ax+by) \in \mathbb{N} \right\}$$

der strikt positiven Linearkombinationen von a und b. Es gilt, dass S nichtleer ist, da entweder a oder $-a$ in S liegen (mit $y = 0$ und $x = \pm 1$). Da S nichtleer und nach unten beschränkt ist, hat es ein kleinstes Element $d = \alpha a + \beta b$. Wir zeigen nun, dass d die Eigenschaften eines größten gemeinsamen Teilers von a und b besitzt.

- $d \mid a$: Mit Satz 3.7 der ganzzahligen Division von a und d existiert ein eindeutiges q und r, so dass

  $$a = d \cdot q + r \quad \text{mit } 0 \leqslant r < d.$$

  Es gilt $r \in S \cup \{0\}$, da

  $$\begin{aligned} r &= a - d \cdot q = a - q(\alpha a - \beta b) \\ &= a \cdot (1 - q\alpha) - b \cdot q\beta. \end{aligned}$$

  Aufgrund der Minimalität von d gilt $r = 0$, woraus $d \mid a$ folgt.

- $d \mid b$ verläuft analog.
- $(t \mid a \wedge t \mid b) \Rightarrow t \mid d$: Sei t als gemeinsamer Teiler von a und b beliebig, aber fest. Dann existieren u, v, so dass $a = cu$ und $b = cv$. Damit gilt

$$d = \alpha a + \beta b = \alpha cu + \beta cv = c \cdot (u\alpha + v\beta),$$

  woraus folgt, dass $c \mid d$.

Mit den drei Eigenschaften folgt dann, dass d der größte gemeinsame Teiler von a und b ist. □
Unmittelbar aus Satz 3.10 folgt folgender Satz:

Satz 3.11 (*Lemma von Bézout*)

Zwei Zahlen $a, b \in \mathbb{Z}$ heißen **teilerfremd** genau dann, wenn $\alpha, \beta \in \mathbb{Z}$ existieren mit

$$\alpha a + \beta b = 1.$$

Beweis: trivial durch Satz 3.10 in Verbindung mit Defintion 3.11. □

Algorithmus zum Finden des ggTs Wir lernen an dieser Stelle den euklidischen Divisionsalgorihtmus kennen:

Verfahren 3.1 (*„erweiterter" euklidischer Divisionsalgorithmus*)

Seien $a, b \in \mathbb{Z}$ und $a \neq 0 \vee b \neq 0$, so berechnet sich der ggT durch:

(1) Setze $r_{-2} \leftarrow |a|$ und $r_{-1} \leftarrow |b|$.

(2) Für $k = 0$ wiederhole solange bis $r_N = 0$

  (2.1) Suche q_k, r_k, so dass $r_{k-2} = q_k r_{k-1} + r_k$

  (2.2) Setze $N \leftarrow k$ und $k \leftarrow k + 1$

(3) $\mathrm{ggT}(a, b) \leftarrow r_{N-1}$

Beweis: Wir zeigen nun, dass Verfahren 3.1 **immer** terminiert und korrekt ist.

Terminierung Wir zeigen dazu zuerst folgende Aussage: In allen Schritten ab $k = 0$ gilt $0 \leqslant r_k < r_{k-1}$. *Beweis:* Schritt (2.1) in Verbindung mit Satz 3.7 ergibt, dass (q_k, r_k) eindeutig existiert und $0 \leqslant r_k < r_{k-1}$ gilt.
Sollte nun $|a| < |b|$ sein, so gilt $r_{-2} < r_{-1}$, aber (2.1) setzt $r_{-2} = |a| = 0 \cdot |b| + \underbrace{|a|}_{=r_0}$,
womit folgt, dass $r_0 < r_{-1}$ gilt, was dann die Aussage ab $k = 0$ erfüllt. □
N kann nun nicht unendlich sein, da es nur eine endliche Anzahl an nichtnegativen Zahlen zwischen r_0 und 0 gibt.

Korrektheit Wir zeigen zuerst folgende Hilfsaussage:

Lemma 3.12 (*Rechenregel für den größten gemeinsamen Teiler*)

$$\mathrm{ggT}(ma_1, ma_2) = |m| \cdot \mathrm{ggT}(a_1, a_2) \quad \text{für } m \in \mathbb{Z}$$

Beweis: Sei $d' = \mathrm{ggT}(ma_1, ma_2)$ und $d = \mathrm{ggT}(a_1, a_2)$, so gilt $md \mid d'$. Wir schlussfolgern daraus:

$$\Rightarrow d' = mdc \mid ma_1, ma_2 \Rightarrow dc \mid a_1, a_2 \Rightarrow dc \mid d \Rightarrow c \mid 1.$$

Daraus folgt dann die zu zeigende Aussage. □
Damit gilt dann auch $ggT(|a|,|b|) = ggT(a,b)$, womit Schritt (1) den Algorithmus nicht verfälscht. Wir zeigen nun, dass $ggT(a,b) = r_{N-1}$ gilt.
Wir stellen leicht fest, dass r_{N-1} r_{N-2} teilt, da $r_{N-2} = q_N r_{N-1} + 0$ gilt. Damit folgt aber auch, dass r_{N-1} r_{N-3} teilt, da $r_{N-3} = q_{N-1} r_{N-2} + r_{N-1}$ gilt und r_{N-1} somit jeden Summanden teilt. Dieselbe Argumentation wird nun wiederholt angewandt und ergibt, dass r_{N-1} alle vorherigen Reste r_k mit $-2 \leqslant k \leqslant N-1$ teilt, insbesondere also auch a und b. Da r_{N-1} gemeinsamer Teiler von a und b ist, gilt $r_{N-1} \mid ggT(a,b)$.
Ebenso schnell wird festgestellt, dass jeder gemeinsame Teiler von a und b alle Reste r_k teilt. Sei nun c ein beliebig, aber fester, gemeinsamer Teiler von a und b. Dann gilt per definitionem, dass $a = c \cdot m$ und $b = c \cdot n$ mit geeigneten $n, m \in \mathbb{Z}$. Daraus können wir schließen, dass $c \mid r_0$, da $r_0 = a - q_0 b = mc - q_0 nc = c(m - q_0 n)$. Analoge Argumente zeigen, dass c auch die Reste $r_1, r_2, \ldots$ teilt, woraus folgt, dass $ggT(a,b) \mid r_{N-1}$ gelten muss. Damit folgt dann unmittelbar, dass $r_{N-1} = ggT(a,b)$ ist. □

Durch die Einführung des ggT schließen wir nun auf den Zusammenhang von Inversen und Teilbarkeit:

Satz 3.13 (*Äquivalenz der Existenz eines multiplikativen Inversen und der Teilerfremdheit*)

$[a]_n \in \mathbb{Z}_n$ hat ein multiplikativ inverses Element $[a]_n^{-1} \in \mathbb{Z}_n$ genau dann, wenn a und n teilerfremd sind.

Beweis:

„⇒“: Ist $[a]_n \in \mathbb{Z}_n$ invertierbar, so muss ein $[b]_n \in \mathbb{Z}_n$ existieren mit $[a]_n \cdot [b]_n = [1]_n$, also mit $ab \equiv_n 1 \mod n$ oder $n \mid (1 - ab)$. Folglich existiert ein $k \in \mathbb{Z}$ mit $kn = 1 - ab$, womit $1 = ab + kn$ gilt. Nach Satz 3.11 sind damit a und n teilerfremd.

„⇐“: Sind nun a und n teilerfremd, so existieren nach Satz 3.11 zwei Zahlen $\alpha, \eta \in \mathbb{Z}$ mit $1 = \alpha a + \eta n$. Für die Restklasse $[\alpha]_n \in \mathbb{Z}_n$ gilt dann $[a]_n \cdot [\alpha]_n = [1]_n$, womit $[\alpha]_n$ zu $[a]_n$ multiplikativ invers ist. □

Für die bereits geäußerte Vermutung sei jetzt noch definiert, was eine Primzahl ist.

Definition 3.12 (*Primzahl*)

Eine Zahl $p \in \mathbb{N}_{\geqslant 2}$ heiße **Primzahl** genau dann, wenn sie **genau zwei** positive Teiler, nämlich 1 und p hat. (Die 1 gehört nicht zu den Primzahlen)
Wir bezeichnen Die Menge aller Primzahlen als $\mathbb{P}$.

Wir formulieren unsere Vermutung jetzt in einem Satz:

Satz 3.14 (*Restklassenring enthält multiplikative Gruppe*)

$\forall n \geqslant 2$. $(\mathbb{Z}_n \backslash \{0\}, \cdot)$ ist ein Modul $\Leftrightarrow$ n ist prim.

Beweis: ergibt sich direkt durch Satz 3.13 i.V.m. Definition 3.12. □

Satz 3.15 (*Restklassenkörper*)

Der Restklassenring $(\mathbb{Z}_n, +, \cdot)$ von $\mathbb{Z}$ modulo n ist genau dann ein Körper, wenn n prim ist.

Beweis: trivial □

Definition 3.13 (*endliche Restklassenkörper*)

Für p prim bezeichnen wir $\mathbb{Z}_p$ mit $\mathbb{F}_p$ und nennen ihn den **Restklassenkörper** von $\mathbb{Z}$ modulo n.

endliche Körper

Eine sehr interessante Frage, die man sich stellen kann, ist, ob es neben den so konstruierten $\mathbb{F}_p$ noch weitere **endliche** Körper gibt.
Per se ist dies erstmal möglich, für p prim und $n \in \mathbb{N}$ kann man einen Körper mit p^n vielen Elementen konstruieren. Dazu sucht man in $\mathbb{F}_p$ ein Element, das nicht die n-te Potenz eines anderen Elements ist und „adjungiert" dieses zu $\mathbb{F}_p$. Sonst lassen sich allerdings keine weiteren endlichen Körper finden.

Wir wollen jetzt noch einmal genauer auf die Multiplikation in $\mathbb{Z}_n$ (mit $n \in \mathbb{N}$ beliebig) eingehen. Wir wissen nach Satz 3.14, dass wenn $n \notin \mathbb{P}$ ist, nicht jedes Element ein Inverses bezüglich der Multiplikation besitzt. Wir suchen jetzt eine Menge, in der das nicht der Fall ist. Wir definieren also:

Definition 3.14 (*multiplikative Gruppe*)

Sei $(\mathbb{Z}_n, +, \cdot)$ ein Restklassenring, dann bezeichne $(\mathbb{Z}_n^*, \cdot)$ mit

$$\mathbb{Z}_n^* := \left\{ [a]_n \in \mathbb{Z}_n \;\middle|\; [a]_n^{-1} \in \mathbb{Z}_n \right\} \subseteq \mathbb{Z}_n \backslash \{[0]_n\}$$

die **multiplikative Gruppe** von $\mathbb{Z}_n$.

Wir wissen aus Satz 3.13, dass $\mathbb{Z}_n^*$ genau die zu n teilerfremden Elemente aus $\mathbb{Z}_n$ enthält. Wir wollen an dieser Stelle unsere Behauptung überprüfen, dass $\mathbb{Z}_n^*$ mit der vom kommutativen Monoid $(\mathbb{Z}_n, \cdot)$ geerbten Multiplikation ein Modul ist.

Behauptung 4

$(\mathbb{Z}_n^*, \cdot)$ ist ein Modul mit neutralem Element $[1]_n$.

Beweis:

→ $(\mathbb{Z}_n^*, \cdot)$ **ist Halbgruppe**: Die Assoziativität folgt unmittelbar aus der Eigenschaft, dass $(\mathbb{Z}_n, \cdot)$ ein kommutativer Monoid ist, zu zeigen bleibt die Wohldefiniertheit der Operation:
Haben $[a]_n, [b]_n \in \mathbb{Z}_n$ inverse Elemente, sind also Elemente von $\mathbb{Z}_n^*$, dann hat auch das Verknüpfungsergebnis $[a]_n \cdot [b]_n$ mit $[b]_n^{-1} \cdot [a]_n^{-1}$ ein Inverses, da

$$[a]_n \cdot [b]_n \cdot [b]_n^{-1} \cdot [a]_n^{-1} = [a]_n \cdot [1]_n \cdot [a]_n^{-1} = [1]_n.$$

Damit ist $[a]_n \cdot [b]_n \in \mathbb{Z}_n^*$. ✓

→ $(\mathbb{Z}_n^*, \cdot)$ **ist ein Monoid mit neutralem Element** $[1]_n$: $[1]_n$ ist neutrales Element der geerbten Operation. Da $\mathrm{ggT}(1, n) = 1$ gilt ebenso $[1]_n \in \mathbb{Z}_n^*$. ✓

→ $(\mathbb{Z}_n^*, \cdot)$ **ist eine Gruppe**: Zu zeigen ist die Existenz eines inversen Elements zu jedem Element $[a]_n \in \mathbb{Z}_n^*$. Sei also $[a]_n \in \mathbb{Z}_n^*$ beliebig, aber fest, womit $[a]_n^{-1} \in \mathbb{Z}_n$ garantiert ist. Da aber auch $[a]_n^{-1}$ mit $[a]_n$ ein Inverses hat, gilt: $[a]_n^{-1} \in \mathbb{Z}_n^*$. ✓

→ $(\mathbb{Z}_n^*, \cdot)$ **ist Modul**: Zu zeigen bleibt Eigenschaft (G4). Diese folgt aber aufgrund der Eigenschaften der Operation. ✓

Wir schlussfolgern, dass $(\mathbb{Z}_n^*, \cdot)$ eine abel'sche Gruppe ist. □
Wir wollen nun nocheinmal Primzahlen und ihre Eigenschaften genauer betrachten.

Lemma 3.16 (*Lemma von Euklid*)

Wenn eine Primzahl p das Produkt ab zweier ganzer Zahlen a, b teilt, dann teilt sie mindestens einen der Faktoren.

Beweis: Angenommen $p \mid ab$, aber p teilt weder a noch b. Damit existieren nach Satz 3.11 Zahlen r, s, m und $n \in \mathbb{Z}$, so dass

$$ra + sp = 1 \quad \text{und} \quad mb + np = 1.$$

Durch Multiplikation der Gleichungen erhalten wir den Zusammenhang

$$1 = (ra + sp)(mb + np) = \underbrace{rm}_{\in\mathbb{Z}} ab + p \underbrace{(smb + ran + snp)}_{\in\mathbb{Z}},$$

aus welchem wiederum mit Satz 3.11 folgt, dass $p \nmid ab$ gelten muss. $\lightning$ □

Satz 3.17 (*Fundamentalsatz der Arithmetik in* $\mathbb{Z}$)

Jede natürliche Zahl $n \in \mathbb{N}$ besitzt eine eindeutige Darstellung

$$n = \prod_{i=1}^{m} p_i^{\alpha_i} \quad \text{mit } m \in \mathbb{N}_0, \alpha_i \in \mathbb{N}, p_i \in \mathbb{P}$$

und einer Anordnung $p_1 < p_2 < \cdots < p_m$. Wir nennen diese Darstellung **Primfaktorzerlegung von n**.

Beweis:

Existenz Wir zeigen die Existenz per Induktion über n.

IA: $n = 1$: Die Darstellung ist durch das leere Produkt mit $k = 0$ erreicht. ✓

IS: $(n \to n + 1)$:
Induktionsvoraussetzung (IV):

Für alle $j \in \mathbb{N}$, so dass $\mathbb{N} \ni j \leqslant n$, besitzt j eine Primfaktorzerlegung.

Wir müssen nun zwei Fälle unterscheiden. Entweder ist $n + 1$ prim, womit mit $k = 1$ und $p_1 = z + 1, \alpha_1 = 1$ die gesuchte Darstellung erreicht wäre, oder $z + 1$ ist nicht prim. Dann muss es allerdings per definitionem eine nichttriviale Zerlegung von $z + 1$ in zwei Faktoren $z + 1 = f_1 \cdot f_2$ geben mit $2 \leqslant f_1, f_2 \leqslant n$, womit beide Faktoren nach Induktionsvoraussetzung eine Primfaktorzerlegung besitzen. Da $(\mathbb{Z}, \cdot)$ ein kommutativer Monoid ist, lässt sich diese gemäß der Vorschrift oben mit Potenzgesetzen umordnen. ✓

Eindeutigkeit Wir zeigen auch die Eindeutigkeit per Induktion über n.

IA: $n = 1$: trivial. ✓

IS: $(n \to n + 1)$:
Induktionsvoraussetzung (IV):

Für alle $j \in \mathbb{N}$, so dass $\mathbb{N} \ni j \leqslant n$, besitzt j eine **eindeutige** Primfaktorzerlegung.

Seien nun im Allgemeinen zwei Zerlegungen $n = p_1 \cdot p_2 \cdot \cdots \cdot p_m = q_1 \cdot q_2 \cdot \cdots \cdot q_m$ in Primfaktoren gegeben. Per definitionem teilt dann p_1 das rechte Produkt, damit nach Lemma 3.16 mindestens einen der Faktoren. Œ sei dies q_1. Da q_1 selbst prim ist, muss automatisch gelten, dass $q_1 = p_1$. Aufgrund der Nullteilerfreiheit von $\mathbb{Z}$ kann man nun eine Kürzungsregel anwenden und erhält

$$n' = p_2 \cdot \cdots \cdot p_m = q_2 \cdot \cdots \cdot q_m,$$

wobei n' aufgrund $n' < n$ per Induktionsvoraussetzung eine eindeutige Zerlegung hat. Damit hat auch n eine eindeutige Zerlegung. □

Anmerkungen

Man kann den Fundamentalsatz der Arithmetik auch auf allgemeine ganze Zahlen erweitern, man erhält sogar die Eindeutigkeit. Dies wird mit einem $\pm$ vor dem Produkt geschafft.

Ebenso lässt sich das Konzept der Primfaktorzerlegung auf Polynomringe übertragen. Das Analogon zur Primzahl ist dann das *irreduzible Polynom*, also Polynome, welche sich nicht als Produkt von Polynomen mit echt niedrigerem Grad schreiben lassen. Abgesehen von konstanten Vorfaktoren hat dann **jedes** Polynom eine (bis auf die Reihenfolge der Faktoren) eindeutige Darstellung als Produkt von irreduziblen Polynomen.

Wir wissen ebenso:

Satz 3.18 (*Anzahl an Primzahlen*)

Es gibt unendlich viele Primzahlen.

Beweis: Angenommen es gäbe nur endlich viele Primzahlen $\mathbb{P}_{\mathbb{F}} = \{p_1, \ldots, p_r\}$[2], so bilden wir die Zahl

$$\eta = \left(\prod_{i=1}^{r} p_i \right) + 1.$$

Entweder η ist prim, dies führt aber unmittelbar zum Widerspruch, da $\eta > p_r$ ist, oder η ist nicht prim. Dann müsste sie allerdings einen Primteiler p_j besitzen, der sowohl $p_1 \cdot \cdots \cdot p_r$ also auch 1 teilt. Da 1 aber keine Primteiler hat, führt dies zum Widerspruch. $\lightning$ □

Wie findet man nun Primzahlen $\leqslant n$?

Verfahren 3.2 (*Sieb des Eratosthenes*)

(1) Stelle eine Liste ℓ der Zahlen $2, \ldots, n$ auf

(2) Bezeichne $\mathfrak{P}$ die Liste der aktuellen Primzahlen bis k

(3) Für $k := 2$ wiederhole solange bis $k^2 > n$

  (3.1) Notiere k als Primzahl ($\mathfrak{P} := \mathfrak{P} \cup \{k\}$)

  (3.2) Streiche **alle *echten* Vielfachen** von k aus der Liste ℓ

  (3.3) Setze $M := \{ j \mid j \notin \ell \}$

  (3.4) Setze $k := \min M$

(4) $\mathbb{P}_{\text{bis } n} := \mathfrak{P} \cup \ell$

Beweis: Wir zeigen nun, dass Verfahren 3.2 **immer** terminiert und korrekt ist.

Korrektheit Wir zeigen nun, dass die Menge $\mathbb{P}_{\text{bis } n}$ am Ende des Verfahrens alle Primzahlen $\leqslant n$ enthält. Wir bezeichnen nun die Menge der natürlichen Zahlen $\leqslant n$ als $\mathbb{N}_n$. Wir wissen, dass die Menge $\mathfrak{P}$ genau die Zahlen aus $\mathbb{N}_n$ enthält, welche nicht gestrichen wurden. Dies ist klar, da nur die *echten* Vielfachen (mit Faktor $\geqslant 2$) gestrichen werden. Da eine Primzahl niemals ein *echtes Vielfaches* einer anderen Zahl ist, werden damit die Primzahlen in $\mathbb{N}_n$ nicht gestrichen. Wir müssen jetzt noch zeigen, dass eine jede Zahl, welche nicht gestrichen wurde, notwendigerweise prim ist.

[2]Es wird hier angenommen, dass $p_{i-1} < p_i$ für alle $1 < i \leqslant r$ gilt.

Sei also m eine beliebige Zahl, die nicht gestrichen wird. Man betrachte den kleinsten Primteiler q von m und sieht dann, dass $m = k \cdot q$ mit $\mathbb{Z} \ni k \geqslant 1$ gelten muss (Folgerung aus Satz 3.17). Gölte nun $k \geqslant 2$, so wäre m allerdings ein *echtes Vielfaches* einer Primzahl $q \leqslant n$. Nach eben getroffener Aussage wird q allerdings nicht gestrichen und ist damit eine Zahl der Menge $\mathfrak{P}$. Bei einem Siebeschritt (Schritt (3.2)) müsste damit aber m gestrichen werden. Damit ist $k = 1$ und m eine Primzahl. Damit gilt, dass alle im Siebverfahren auftretenden Zahlen $\in \mathfrak{P}$ prim sind.
Jetzt müssen wir noch beweisen, dass das Verfahren auch **alle** Primzahlen $\leqslant n$ herausfindet. Mit anderen Worten müssen wir die Schleifenbedingung nachweisen. Wir wollen dazu zuerst zeigen, dass wenn mit allen Primzahlen $\leqslant p$ bereits gesiebt wurde, p^2 die – beim Schritt mit p – nächste zu streichende Zahl ist. Dies folgt eigentlich direkt aus der Tatsache, dass bereits alle Vielfachen $k \cdot p$ mit $2 \leqslant k < p$ bereits gesiebt wurden. Sei nämlich q kleinster Primteiler von einem solchen k, so gilt $k \cdot p = q \cdot (k' \cdot p)$. Da aber $q \leqslant k < p$ gilt, ist $k \cdot p$ als echtes Vielfaches von q bereits beim Schritt mit q gestrichen worden. Beim Schritt mit p wird damit $p \cdot p = p^2$ als nächstes gestrichen.
Jetzt zeigen wir noch, dass wenn mit allen Zahlen $p \leqslant \sqrt{n}$ gesiebt wurde, weitere Schritte keine Streichungen in der Liste ℓ mehr bewirken. Sei $p_m \leqslant \sqrt{n}$ die größte Primzahl $\leqslant \sqrt{n}$. Wie eben festgestellt, ist bei einem Schritt mit p_m die nächste zu streichende Zahl $p_m^2 < n$. Die nachfolgende Zahl, mit der gestrichen wird, ist dann p_{m+1}, welche nach Voraussetzung allerdings größer $\sqrt{n}$ ist. Damit ist die erste zu streichende Zahl ebenfalls größer n, in der Liste der Zahlen $\leqslant n$ wird damit ab jetzt nichts mehr gestrichen. Damit erhalten wir alle Primzahlen, wenn wir das Streichen bis $\sqrt{n}$ beschränken und die übrigen Zahlen als Primzahlen mit dazu nehmen. Der Algorithmus ist damit korrekt.

Terminierung Das Verfahren terminiert immer. Es gibt immer noch Zahlen in ℓ, welche noch nicht gestrichen wurden, dann wählen wir in Schritt (3.4) die kleinste aus. Da diese Menge endlich und nach unten beschränkt ist, gibt es das Minimum in der Menge. Da wir nur Primzahlen **aufsteigend** auswählen und es nur endlich viele Primzahlen $\leqslant \sqrt{n}$ gibt, muss es nach endlich vielen Schritten eine Zahl geben, welche größer als $\sqrt{n}$ und prim ist, wird diese ausgewählt, so terminiert der Algorithmus. □

Wir wollen uns nun Fragen, wie viele Elemente die multiplikative Gruppe $\mathbb{Z}_n^*$ enthält. Wir definieren deswegen:

Definition 3.15 (Euler'sche Phi-Funktion)

Die **Euler'sche Phi-Funktion** $\varphi : \mathbb{N} \to \mathbb{N}$ ist definiert durch

$$\varphi(n) := \operatorname{ord}(\mathbb{Z}_n^*, \cdot) = |\mathbb{Z}_n^*| = \#\left\{1 \leqslant k \leqslant n \;\middle|\; \operatorname{ggT}(k,n) = 1\right\}.$$

Dabei ist die Ordnung $\operatorname{ord}(G, *)$ einer endlichen Gruppe G die Anzahl ihrer Elemente.

Wie man $\varphi(n)$ berechnen kann, entnimmt man am besten folgendem Satz:

Satz 3.19 (Berechnung von $\varphi(n)$)

Sei $p \in \mathbb{P}$ eine Primzahl und $k, l, m \in \mathbb{N}$, dann gilt:

(i) $\varphi(p) = p - 1$

(ii) $\varphi(p^k) = p^k - p^{k-1} = p^{k-1}(p-1)$

(iii) $\operatorname{ggT}(l,m) = 1 \Rightarrow \varphi(l \cdot m) = \varphi(l) \cdot \varphi(m)$

Beweis: An dieser Stelle wird nur (ii) gezeigt, da (i) trivial und (iii) mit dem Homomorphiesatz (Satz 3.25) nachgeholt wird.
Von 1 bis p^k gibt es p^k viele Zahlen, von denen nur diejenigen „entfernt“ werden müssen, die durch p teilbar sind, also

$$p, 2p, \ldots, p^i \cdot p, (p^i + 1) \cdot p, \ldots, (p^{k-1} - 1) \cdot p.$$

Damit bleiben $p^k - 1 - (p^{k-1} - 1) = p^k - p^{k-1}$ Elemente übrig. □
Anwendung findet die φ-Funktion vor allem beim kleinen Fermatschen Satz sowie den Konstruktionen mit Zirkel und Lineal. Gauß zeigte, dass das regelmäßige n-Eck mit Zirkel und Lineal konstruierbar ist genau dann, wenn ein $k \in \mathbb{N}$ existiert, so dass $\varphi(n) = 2^k$. Man schloss daraus dann auch, dass nicht jeder Winkel mit Zirkel und Lineal dreigeteilt werden konnte. Ebenso findet die Funktion Anwendung in der Fehlererkennung (→ Kapitel 3.2).
Wir rufen uns jetzt Defintion 3.2 in Erinnerung und fragen uns wie man am besten Untergruppen finden kann. Wir merken eine Methode hier an, denn sei $(G, \cdot)$ beliebig und $a \in G$, so ist $\langle a \rangle := \{a^k \mid k \in \mathbb{Z}\}$ eine Untergruppe von $(G, \cdot)$. Man nennt sie die von a erzeugte Untergruppe. Wir definieren:

Definition 3.16 (*Zyklische Gruppen und Ordnungen*)

Eine von *einem* Element erzeugte Gruppe heißt **zyklisch**. Falls es ein $k \in \mathbb{N}$ gilt mit $a^k = 1$, so wird das **kleinste** $k \in \mathbb{N}$ mit dieser Eigenschaft als **Ordnung** von a bezeichnet, andernfalls gilt $\mathrm{ord}(a) = \infty$.

Definition 3.17 (*Morphismen*)

Seien $(M, *)$ und (N, Δ) algebraische Strukturen mit Verknüpfungen $*$ und Δ. Eine Abbildung $f \in \mathrm{Abb}(M, N)$ heiße **Homomorphismus** genau dann, wenn

$$\forall x, y \in M.\ f(x * y) = f(x) \Delta f(y) \tag{H}$$

Ein Homomorphismus f heiße …

- … **Monomorphismus** genau dann, wenn f **in**jektiv ist.
- … **Epimorphismus** genau dann, wenn f **sur**jektiv ist.
- … **Isomorphismus** genau dann, wenn f **bi**jektiv ist. (Wir sagen dann, M ist isomorph zu N und schreiben $M \cong N$)
- … **Endomorphismus** genau dann, wenn $(M, *) = (N, \Delta)$
- … **Automorphismus** genau dann, wenn $(M, *) = (N, \Delta)$ und f **bi**jektiv ist.

(Tragen M, N mehrere Verknüpfungen $*_i, \Delta_i$ mit $1 \leqslant i \leqslant k$, so muss (H) für alle Indizes i gelten!)

Satz 3.20 (*Eigenschaften des Gruppenhomomorphismus*)

Seien $(G, *)$ und (H, Δ) Gruppen und beschreibe $f : (G, *) \to (H, \Delta)$ einen Gruppenhomomorphismus, so gilt …

(a) … $f(1_G) = 1_H$, wobei $1_G, 1_H$ die neutralen Elemente in G resp. H sind.

(b) … $\forall\, a \in G.\ f(a^{-1}) = f(a)^{-1}$, wobei x^{-1} das Inverse zu x ist.

(c) … f Isomorphismus $\Rightarrow f^{-1}$ ebenso

(d) … $\mathrm{im}(f)$ ist Untergruppe von H

(e) … $\ker(f) := \{a \in G \mid f(a) = 1_H\}$ ist Untergruppe von G

(f) … f ist injektiv genau dann, wenn der Kern von f trivial ist.

Beweis:

(a) $1_H \Delta f(1_G) = f(1_G) = f(1_G * 1_G) = f(1_G) \Delta f(1_G) \Rightarrow 1_H = (f(1_G) \Delta f(1_G)) \Delta f(1_G)^{-1} = f(1_G)$

(b) Mit (a) folgt $1_H = f(1_G) = f(a * a^{-1}) = f(a)\Delta f(a^{-1})$, damit verhält sich $f(a^{-1})$ genauso wie das Inverse zu $f(a)$.

(c) Als Isomorphismus ist f bijektiv, damit wissen wir aus der Vorgängerveranstaltung, dass f^{-1} existieren muss und ebenfalls bijektiv ist. Seien $b_1, b_2 \in H$ und $a_1 = f^{-1}(b_1)$ und $a_2 = f^{-1}(b_2)$. Aufgrund der Bijektivität von f existieren auch a_1 und a_2. Wir überprüfen nun, ob f^{-1} ein Homomorphismus ist. Aus der Homomorphismuseigenschaft von f folgt unmittelbar, dass $f(a_1 * a_2) = f(a_1)\Delta f(a_2) = b_1 \Delta b_2$. Per definitionem gilt dann aber auch, dass $f^{-1}(b_1 \Delta b_2) = a_1 * a_2 = f^{-1}(b_1) * f^{-1}(b_2)$, was f^{-1} zu einem Homo- und damit auch Isomorphismus macht.

(d) Sei $B := \mathrm{im}(f)$. Dann gilt $1_H = f(1_G) \in B$. Seien nun $h_1, h_2 \in B$, so gibt es $g_1, g_2 \in G$ mit $f(g_1) = h_1$ und $f(g_2) = h_2$. Damit ist $h_1 \Delta h_2 = f(g_1)\Delta f(g_2) = f(g_1 * g_2) \in B$. Ebenso gibt es für $h \in B$ ein $g \in G$ mit $f(g) = h$. Somit ist $h^{-1} = (f(g))^{-1} = f(g^{-1}) \in B$, woraus die zu zeigende Untergruppeneigenschaft von B folgt.

(e) Wegen $f(1_G) = 1_H$ ist $1_G \in \ker(f)$. Seien $g, g' \in \ker(f)$, so ist

$$f(g * g') = f(g)\Delta f(g') = 1_H \Delta 1_H = 1_H$$

und daher auch $g * g' \in \ker(f)$. Sei nun $g \in \ker(f)$. Wir betrachten dann das inverse Element g^{-1}. Es gilt

$$f(g^{-1}) = (f(g))^{-1} = 1_H^{-1} = 1_H,$$

also auch $g^{-1} \in \ker(f)$.

(f)„⇒“: Wenn f injektiv ist, so darf auf jedes Element $h \in H$ höchstens ein Element aus G gehen. Da $f(1_G) = 1_H$, darf kein weiteres $a \in G$ existieren mit $f(a) = 1_H$. Der Kern ist damit trivial.

„⇐“: Sei $\ker(f) = \{1_G\}$. Angenommen es existierte ein $h \in H$ und $g, g' \in G$, so dass $f(g) = f(g') = h$ gelte. Dann ist $f(g * g'^{-1}) = f(g)\Delta f(g')^{-1} = h \Delta h^{-1} = 1_H$, also gelte $g * g^{-1} \in \ker(f)$, dementsprechend muss gelten, dass $g * g'^{-1} = 1_G$, woraus unmittelbar die Gleichheit von g und g' folgt. □

Es sei U nun eine Untergruppe von $(G, *)$. Dann beschreibe die Gleichung

$$a \sim\big|_U\, b :\Leftrightarrow\ a^{-1} * b \in U \tag{3.1}$$

eine Äquivalenzrelation auf G, denn sowohl

(1) **Reflexivität** ist durch

$$a \sim\big|_U\, a, \text{ da } a^{-1} * a = e \in U,$$

(2) als auch **Symmetrie** durch den Zusammenhang

$$a \sim\big|_U\, b \Leftrightarrow a^{-1} * b \in U \Leftrightarrow U \ni (a^{-1} * b)^{-1} = b^{-1} * a \Leftrightarrow b \sim\big|_U\, a$$

(3) sowie **Transitivität** durch den Zusammenhang

$$a \sim\big|_U\, b \wedge b \sim\big|_U\, c \Leftrightarrow a^{-1} * b \in U \ni b^{-1} * c \Leftrightarrow (a^{-1} * b) * (b^{-1} * c) = a^{-1} * c \in U \Leftrightarrow a \sim\big|_U\, c$$

erfüllt. Die Äquivalenzklassen der Relation bezeichnen wir mit $[a]_U$. Es ist

$$[a]_U = \left\{b \in G \;\middle|\; a \sim\big|_U\, b\right\} = \left\{b \in G \;\middle|\; \exists\, u \in U.\ b = a * u\right\} = \left\{a * u \;\middle|\; u \in U\right\} = a * U$$

eine **Linksnebenklasse** von U.
Wir schreiben für die Menge der Linksnebenklassen auch $G/U := \left\{[a]_U \;\middle|\; a \in G\right\}$. Wir wollen uns nun mit Struktureigenschaften ebendieser Strukturen näher beschäftigen, wobei folgender Satz ein hinreichendes Kriterium gibt.

Satz 3.21 (*Struktureigenschaft von* G/U)

Ist $(G,*)$ eine abel'sche Gruppe und $U \subseteq G$ eine Untergruppe, so wird durch die auf G/U wohldefinierte Verküpfung

$$[a]_U * [b]_U := [a * b]_U \tag{3.2}$$

G/U zu einer abel'schen Gruppe mit neutralem Element $[e]_U = U$.

Beweis: Entscheidend ist zunächst einmal die Wohldefiniertheit der Verknüpfung „$*$" auf G/U. Es gelte also $[a]_U = [a']_U$ und $[b]_U = [b']_U$, das heißt

$$\exists u_1, u_2 \in U.\ a' = a * u_1 \text{ und } b' = b * u_2.$$

Wir schließen daraus:

$$\begin{aligned} a' * b' &= (a * u_1) * (b * u_2) = a * (u_1 * b * u_2) \\ &= a * (b * u_1 * u_2) = (a * b) * u, \text{ wobei } u := u_1 * u_2 \in U. \end{aligned}$$

Daraus folgt unmittelbar, dass $a' * b' \in [a * b]_U$, und daher $[a' * b']_U = [a * b]_U$, woraus die Behauptung der Wohldefiniertheit folgt.
Das unter der Verknüpfung $*$ $(G/U,*)$ zu einer abel'schen Gruppe wird, ist trivial gezeigt. □
Wir definieren damit:

Definition 3.18 (*Faktor-/ Quotientengruppe*)

Wir bezeichnen $(G/U,*)$ aus Satz 3.21 als **Quotienten**- oder auch **Faktorgruppe**. Im allgemeinen gilt:
Aus einer algebraischen Struktur A und einer Kongruenzrelation ϑ auf ebendieser kann eine neue algebraische Struktur A/ϑ gewonnen werden, diese heiße dann **Faktorstruktur**, **Faktoralgebra** oder auch **Quotientenstruktur**, **Quotientenalgebra**. Die Grundmenge ist dabei gerade die Faktormenge A/ϑ und für jede n-stellige Operation $f_A : A^n \to A$ von A wird eine neue Operation

$$f_{A/\vartheta} : (A/\vartheta)^n \to A/\vartheta \quad \text{mit} \quad f_{A/\vartheta}([a_1]_\vartheta, \ldots, [a_n]_\vartheta) := [f_A(a_1, \ldots, a_n)]_\vartheta$$

auf A/ϑ definiert.

Für unser U oben brauchen wir allerdings nur die Eigenschaft eines Normalteilers. Wir definieren weiter:

Definition 3.19 (*Normalteiler*)

Sei $(G,*)$ eine Gruppe und $H \subseteq G$ eine Untergruppe. Man nennt H einen **Normalteiler** genau dann, wenn

$$\forall x \in G.\ xH = Hx,$$

also wenn die Linksnebenklasse zu x mit der Rechtsnebenklasse zu x übereinstimmt.

Damit lässt sich Satz 3.21 zu folgender Aussage abschwächen:

Satz 3.22 (*Struktureigenschaften der Faktorgruppe*)

Sei $(G,*)$ eine Gruppe und $U \subseteq G$ ein **Normalteiler** derselbigen, so ist dann die Faktorgruppe G/U mit der in Satz 3.21 definierten Verknüpfung (3.2) eine Gruppe.

Beweis: trivial. □
Wir erweitern nun Satz 3.20 um eine interessante Eigenschaft:

Satz 3.20 (*Eigenschaften des Gruppenhomomorphismus, (fort.)*)

Seien $(G,*)$ und (H,Δ) Gruppen und beschreibe $f : (G,*) \to (H,\Delta)$ einen

Gruppenhomomorphismus, so gilt ...

(g) ... $\ker(f)$ ist Normalteiler.

Beweis:

(g) Sei $a \in G, b \in \ker(f)$ beliebig. Dann ist

$$f(a * b * a^{-1}) = f(a)\Delta f(b)\Delta(f(a))^{-1} = 1_H,$$

woraus unmittelbar folgt, dass $a * b * a^{-1} =: u \in \ker(f)$ und somit, dass $a * b = u * a$ gilt, woraus folgt, dass $\ker(f)$ ein Normalteiler ist. □

Wir wollen nun – bevor wir uns mit Homomorphie-, Struktur- und Fermats Sätzen beschäftigen – einmal näher anschauen, wie die Anzahl der Elemente in einer Untergruppe mit der der Obergruppe korreliert. Wir definieren dazu zuerst noch einmal genauer die Begrifflichkeiten der Nebenklassen:

Definition 3.20 (*Index und Nebenklassen*)

Sei $(G, *)$ eine Gruppe und $H \subseteq G$ eine Untergruppe. Dann bezeichnet man mit G/H (lies: „G nach H“ oder „G modulo H“) die Menge aller **Linksnebenklassen** und mit $H\backslash G$ die Menge aller **Rechtsnebenklassen** in G bezüglich H:

$$G/H := \left\{ g * H \;\middle|\; g \in G \right\} \qquad \text{Linksnebenklassen}$$

$$H\backslash G := \left\{ H * g \;\middle|\; g \in G \right\} \qquad \text{Rechtsnebenklassen}$$

Die Anzahl der Linksnebenklassen wird als der **Index** von H in G bezeichnet, man schreibt auch $G : H$.

In Wirklichkeit sind die Linksnebenklassen gegenüber den Rechtsnebenklassen in keiner Weise ausgezeichnet, man erkennt leicht, dass der Index von H in G auch gleich der Anzahl der Rechtsnebenklassen von H in G ist. Damit ist es uns möglich folgenden Satz genau aufzufassen:

Satz 3.23 (*Satz von Lagrange*)

Es sei G eine endliche Gruppe und H eine Untergrupepe. Dann ist die Ordnung von H ein Teiler der Ordnung von G. Der Quotient ist dann genau der Index von H in G:

$$\#G = \#G/H \cdot \#H = (G : H) \cdot \#H$$

Beweis: Es seien $a_1H, \ldots, a_kH$ die verschiedenen Nebenklassen bezüglich H. Es ist definitionsgemäß also $k = \#G/H$ die Anzahl der Linksnebenklassen. Nach dem Vorkommentar zu Satz 3.21 handelt es sich bei den Mengen $a_1H, \ldots, a_kH$ um Äquivalenzklassen einer Äquivalenzrelation auf G. Nach einer generellen Eigenschaft von Äquivalenzrelationen ist G die disjunkte Vereinigung der Klassen, womit für die Mächtigkeit gilt, dass

$$\#G = \sum_{i=1}^{k} \#(a_iH).$$

Man erkennt schnell, dass alle Mengen a_iH gleich viele Elemente haben, genauer gleichmächtig zu H sind. Dazu betrachte man die Abbildung

$$x \mapsto a_i x, \quad H \to a_i H,$$

welche per definitionem surjektiv und da die Existenz von neutralem und inversen Elementen gesichert ist auch injektiv ist. Zusammengefasst gilt dann:

$$\#G = k \cdot \#H = \#G/H \cdot \#H$$

□

Aus Satz 3.23 folgt dann unmittelbar:

Korrolar 3.23 (*„Korollar" von Lagrange*)

Es sei G eine endliche Gruppe und $a \in G$. Dann ist die Ordnung $\mathrm{ord}(a)$ von a ein Teiler der Ordnung von G.

Mit Definition 3.18 haben wir Faktorgruppen konstruiert. Wir wollen nun ein bisschen mehr über die Struktur von solchen Faktorgruppen herausfinden. Dazu erklären wir folgenden Satz, der den Zusammenhang zu Gruppenhomomorphismen erläutert:

Satz 3.24 (*Homomorphiesatz*)

Es seien $(G,*)$ und (H,Δ) Gruppen mit $f : (G,*) \to (H,\Delta)$ als Gruppenhomomorphismus, so beschreibt

$$\begin{array}{rccl} \widetilde{f}: & {}^{G}/_{\ker(f)} & \longrightarrow & \mathrm{im}(f) \\ & [a]_{\ker(f)} & \longmapsto & \widetilde{F}([a]_{\ker(f)}) := f(a) \end{array}$$

einen Gruppen***iso***morphismus. Damit gilt: ${}^{G}/_{\ker(f)} \cong \mathrm{im}(f)$.

Beweis: Nach Satz 3.20(g) ist ${}^{G}/_{\ker(f)}$ eine Gruppe. Wir zeigen nun die Eigenschaften des Gruppenisomorphismus:

- $\widetilde{f}$ **ist wohldefiniert:** Sei $[a]_{\ker(f)} = [a']_{\ker(f)}$, also $a' = a * u$ mit $u \in \ker(f)$, so folgt direkt

$$\widetilde{f}([a']_{\ker(f)}) = f(a') = f(a*u) = f(a)\Delta f(u) = f(a)\Delta e_H = f(a) = \widetilde{f}([a]_{\ker(f)})$$

- $\widetilde{f}$ **ist ein Gruppenhomomorphismus:**

$$\widetilde{f}([a]_{\ker(f)} * [b]_{\ker(f)}) = \widetilde{f}([a*b]_{\ker(f)}) = f(a*b) = f(a)\Delta f(b) = \widetilde{f}([a]_{\ker(f)})\Delta\widetilde{f}([b]_{\ker(f)})$$

- $\widetilde{f}$ **ist surjektiv:** trivial
- $\widetilde{f}$ **ist injektiv:** Nach Satz 3.20(f) müssen wir nur zeigen, dass der Kern von f trivial ist. Es gilt

$$\widetilde{f}([a]_{\ker(f)}) = f(a) = e_H \Rightarrow a \in \ker(f) \Rightarrow [a]_{\ker(f)} = \ker(f) = [e_G]_{\ker(f)},$$

womit die Injektivität gezeigt ist. □

Korrolar 3.24 (*Elementzahl der Gruppe*)

$$\#G = \#\ker(f) \cdot \#{}^{G}/_{\ker(f)} = \#\ker(f) \cdot \#\mathrm{im}(f)$$

Wir wollen nun zur Euler'schen Phifunktion $\varphi(n)$ zurückkehren. Satz 3.19(iii) gab zur Berechnung von $\varphi(l\cdot m)$ eine Rechenregel an, welche es uns erlaubt den Wert für jede beliebige Zahl auszurechnen. Wir wollen diese nun beweisen und legen unsere Aussagen dafür in einen zusammenfassenden Satz nieder:

Satz 3.25 (*Zusammenfassender Satz*)

Es seien $n, m \in \mathbb{N}\backslash\{1\}$. Dann gilt:

(i) Für $ggT(n,m) = 1$ ist die Abbildung

$$\begin{array}{rccc} \widetilde{f}: & \mathbb{Z}_{m\cdot n} & \longrightarrow & \mathbb{Z}_m \times \mathbb{Z}_n \\ & [x]_{mn} & \longmapsto & ([x]_m, [x]_n) \end{array}$$

einen Gruppenisomorphismus bezüglich „+“, dessen Einschränkung

(ii)

$$\widetilde{f}: \quad \mathbb{Z}^*_{m\cdot n} \quad \longrightarrow \quad \mathbb{Z}^*_m \times \mathbb{Z}^*_n$$

ebenfalls bijektiv und damit wohldefiniert ist.

(iii) Für $ggT(n,m) \neq 1$ gilt diese Isomorphie der Gruppen **nicht**.

Beweis:

(i) Nach Satz 3.24 müssen wir nun zeigen, dass die Abbildung

$$f : \mathbb{Z} \quad \longrightarrow \quad \mathbb{Z}_m \times \mathbb{Z}_n, x \mapsto ([x]_m, [x]_n)$$

gerade den Kern $\ker(f) = m \cdot n\mathbb{Z}$ hat. Dann gilt $\mathrm{im}(\widetilde{f}) = \mathrm{im}(f)$ ist isomorph zu $\mathbb{Z}_{m\cdot n}$ und daher auch

$$\mathrm{ord}(\mathrm{im}(f)) = \mathrm{ord}(\mathbb{Z}_{m\cdot n}) = m \cdot n = \mathrm{ord}(\mathbb{Z}_m \times \mathbb{Z}_n),$$

woraus zu schließen ist, dass f und $\widetilde{f}$ surjektiv und damit $\widetilde{f}$ ein Isomorphismus ist.
Sei also $f(x) = ([0]_m, [0]_n)$, das heißt $x = 0 \mod m$ und $x = 0 \mod n$. Dies gilt für $x = m \cdot n \cdot k$ mit $k \in \mathbb{Z}$, damit ist $m \cdot n\mathbb{Z} \subseteq \ker(f)$.
Umgekehrt ist jedes $x \in \ker(f)$ durch m und n teilbar. Daher gibt es $k_1, k_2 \in \mathbb{Z}$ mit

$$x = k_1 \cdot m = k_2 \cdot n.$$

Da aber nach Voraussetzung m und n **teilerfremd** sin, muss

$$k_1 = l_1 \cdot n \quad \text{und} \quad k_2 = l_2 \cdot n$$

gelten (dies folgt sofort aus Satz 3.11 mit $l \cdot m + k \cdot n = 1$). Damit folgt dann $x = l_1 \cdot n \cdot m = l_2 \cdot m \cdot n \in mn\mathbb{Z}$, also gilt $\ker(f) \subseteq mn\mathbb{Z}$.

(ii) Zunächst zeigen wir die Wohldefiniertheit, indem wir zeigen, dass für ein beliebiges $a \in \mathbb{Z}^*_{mn} \subseteq \mathbb{Z}_{mn}$ gilt, dass $\widetilde{f}(a) \in \mathbb{Z}^*_m \times \mathbb{Z}^*_n$. Wir betrachten dazu ein beliebiges $[a]_{mn} \in \mathbb{Z}^*_{mn}$, womit gilt, dass $ggT(a, mn) = 1$. Wir können nun schließen, dass $ggT(a,m) = 1$ und $ggT(a,n) = 1$ gelten muss, woraus folgt, dass $[a]_m \in \mathbb{Z}^*_m$ und $[a]_n \in \mathbb{Z}^*_n$. Per definitionem ist $\widetilde{f}$ dann wohldefiniert.
Wir zeigen nun noch die Bijektivität: Die Injektivität ist aus (i) bekannt, zu zeigen bleibt die Surjektivität: Sei also $[a]_m \times [a]_n \in \mathbb{Z}^*_m \times \mathbb{Z}^*_n$ beliebig, aber fest. Wir müssen nun zeigen, dass das Urbild $[a]_{mn} \in \mathbb{Z}^*_{mn}$, also dass $ggT(a, mn) = 1$ gilt. Wir wissen, dass $ggT(a,m) = ggT(a,n) = 1$ gilt, wonach mit Satz 3.10 $i, j, k, l \in \mathbb{Z}$ existieren mit $ia + jm = 1$ und $ka + ln = 1$. Wir multiplizieren die erste Gleichung mit n und setzen dies in die zweite Gleichung ein und erhalten somit

$$1 = ka + l(ian + jmn) = (k + iln)a + (lm)nm,$$

was mit erneuter Anwendung von Satz 3.10 liefert, dass $ggT(a, mn) = 1$.

(iii) Sei nun $t > 1$ ein gemeinsamer Teiler von m und n, also gelte $n = n't$ und $m = m't$. Wir wollen zeigen, dass es in $\mathbb{Z}_m \times \mathbb{Z}_n$ kein Element der Ordnung $mn = m'n't^2$ gibt, da daraus die Nichtisomorphie der Gruppen direkt folgt. Wir zeigen also, dass alle Elemente aus $\mathbb{Z}_m \times \mathbb{Z}_n$ eine Ordnung $\leqslant n'm't < n'm't^2 = mn$ haben. Sei dazu $([a]_m, [b]_n) \in \mathbb{Z}_m \times \mathbb{Z}_n$ beliebig, aber fest, so gilt $n'm't([a]_m, [b]_n) = ([n'm'ta]_m, [n'm'tb]_n) = ([0]_m, [0]_n)$ □

Aus Satz 3.25(ii) folgt nun die beabsichtigte Rechenregel:

Korrolar 3.25 (*φ-Rechenregeln*)

Seien $n, m \in \mathbb{N}\backslash\{1\}$ teilerfremd, so ist

$$\varphi(m \cdot n) = \#\mathbb{Z}^*_{m\cdot n} = \#\mathbb{Z}^*_m \cdot \#\mathbb{Z}^*_n = \varphi(m) \cdot \varphi(n)$$

Wir kommen nun zum kleinen Fermat'schen Satz, welcher die Grundlage für das RSA-Verschlüsselungsverfahren (Verfahren **??**) bildet.

Satz 3.26 (*Kleiner Fermat'scher Satz*)

Sei $x \in \mathbb{Z}, n \in \mathbb{N}$ und $p \in \mathbb{P}$, so gilt:

(i) $\mathrm{ggT}(x,n) = 1 \Rightarrow x^{\varphi(n)} \equiv 1 (\mod n)$

(ii) $p \not| x \Rightarrow x^{p-1} \equiv 1 (\mod p)$

(iii) $x^p \equiv x (\mod p)$

Beweis:

(i) Sei also $\mathrm{ggT}(x,n) = 1$, so folgt aus Definition 3.14, dass $[x]_n \in \mathbb{Z}^*_n$ und mit Korollar 3.23, dass $\mathrm{ord}([x]_n) \mid \varphi(n)$, was äquivalent zu der Existenz eines $\alpha \in \mathbb{N}$ ist, womit gilt, dass $\varphi(n) = \alpha \cdot \mathrm{ord}([x]_n)$. Damit gilt dann aber auch:

$$\left[x^{\varphi(n)}\right]_n = \left[x^{\alpha\cdot \mathrm{ord}[x]_n}\right]_n = \left[x^{\mathrm{ord}[x]_n}\right]^\alpha_n = [1]^\alpha_n = [1]_n.$$

(ii) Setze $n = p$ in (i), so gilt mit Satz 3.19 $\varphi(p) = p - 1$.

(iii) Wir unterscheiden zwei Fälle, einerseits kann p kein Teiler von x sein, dann gilt $x^{p-1} \equiv 1$, was auf die Aussage durch beidseitige Multiplikation mit x führt. Andererseits kann p Teiler von x sein, dann gilt $x \equiv 0$ und damit $x^p \equiv 0$, also $x^p \equiv x$. □

Wir wollen zum Abschluss noch einen Struktursatz endlicher abel'scher Gruppen kennenlernen:

Satz 3.27 (*Struktursatz fur abel'sche Gruppen*)

Jede endliche abel'sche Gruppe G ist isomorph zum Produkt

$$\mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2} \times \cdots \times \mathbb{Z}_{m_k}, \quad m_1 \mid m_2 \mid \ldots \mid m_k, m_1 > 1, k \geqslant 1.$$

Dabei sind die Anzahl k und die auftretenden Ordnungen m_i mit ihren Vielfachheiten **eindeutig** bestimmt.

3.2 Anwendung: Kodierungstheorie — Prüfziffern

Bei Datenübertragungen sollen gewisse Fehler erkannt werden. In diesem Kapitel wollen wir uns mit den algebraischen Grundlagen befassen.

Definition 3.21 (*Information, Paket, Wort, Fehler*)

Wir bezeichnen eine Teilmenge an Wissen, die einem Empfänger mittels Nachrichten über einen Informationskanal zugänglich gemacht werden können als **Information**.
Wir bezeichnen denjenigen, welcher die Informationsübermittlung anstößt als **Sender**, denjenigen an den die Information gerichtet ist als **Empfänger**.
Zum Senden zerlegt man die Information in kleinere **Pakete**. Ein solches Paket nennen wir dann **Wort**, welches wiederum aus einer endlichen Anzahl $m \in \mathbb{N}$ an Zeichen – für uns sollen das Ziffern sein – besteht, also $d_1, \ldots, d_m \in Z$, wobei $n := \#Z$ und – für uns – $Z \subseteq \mathbb{N}$ gilt.
Wir sprechen vin einem **Übertragungsfehler**, wenn die ursprüngliche Nachricht von der empfangenen abweicht.
Wir wollen für die Fehlererkennung das Wort um ein **Prüfzeichen** (**Prüfziffer**) erweitern, sprich ein $d_{m+1} \in Z$ anhängen. Das so erhaltene Tupel $(d_1, \ldots, d_{m+1})$ bezeichnen wir dann als **erweitertes Wort**.

Wir definieren dann zur Berechnung der Prüfziffer:

Definition 3.22 (*Prüfziffern*)

Wir definieren eine Funktion $f : Z^m \to Z$ – bei uns gilt weiterhin $Z := \{0, \ldots, n-1\}$ – mit der wir das Prüfzeichen definieren als

$$d_{m+1} := f(d_1, \ldots, d_m).$$

Wir definieren des Weiteren eine Prüffunktion $P : Z^{m+1} \to \mathbb{Z}$, welche bei uns wie folgt arbeitet:

$$P(d_1, \ldots, d_m, d_{m+1}) := d_{m+1} - f(d_1, \ldots, d_m) \begin{cases} \neq 0 & \text{Fehler}\lightning \\ = 0 & \text{vermutlich } \textbf{kein} \text{ Fehler} \end{cases}$$

Dabei ist f Sender **und** Empfänger bekannt.
Recht praktikabel und einfach ist es ein f der folgenden Struktur zu nehmen:

$$f(d_1, \ldots, d_m) :\equiv -\sum_{i=1}^{m} g_i \cdot d_i \equiv \sum_{i=1}^{m} (n - g_i) d_i \mod n,$$

wobei g_i mit $i = 1, \ldots, m$ **Gewichte** heißen, welche geeignet zu wählen sind.
Damit gilt für obiges P, dass

$$P(d_1, \ldots, d_m, d_{m+1}) :\equiv \sum_{i=1}^{m+1} g_i d_i \mod n,$$

wobei $g_{m+1} := 1$.

Wir stellen uns nun die Frage wie die Gewichte zu wählen sind. Ohne Beweis sei angenommen, dass dies von der Art des Fehlers abhängt. Wir definieren deswegen:

Definition 3.23 (*Fehlerarten*)

Wir wollen Übertragungsfehler wie folgt unterscheiden:

- Wenn sich **höchstens** ein d_i unterscheidet, so sprechen wir von **Einzelfehlern**.
- Wenn **nur** ein d_i und ein d_{i+1} permutiert wurden, so sprechen wir von einem

Nachbarvertauschungsfehler.

- Wenn **nur** ein d_i und ein d_j mit $i \neq j$ permutiert wurden, so sprechen wir von einem **Vertauschungsfehler**.

Uns ist an dieser Stelle selbstverständlich bewusst, dass dies nicht alle Fehlerarten abdeckt, für unsere Zwecke sollten diese allerdings genügen.
Im Folgenden werde an des richtigen Wortes $d_1, \ldots, d_m | d_{m+1}$ statt ein möglicherweise falsches Wort $d'_1, \ldots, d'_m | d'_{m+1}$ übertragen. Wir berechnen $P(d'_1, \ldots, d'_{m+1}) = c'$ und kennen damit $[\delta]_n = [c - c']_n$.

3.2.1 Erkennung von Einzelfehlern

Satz 3.28 (*Invertierbarkeit*)

Ein $[g]_q \in \mathbb{Z}/=_q$ ist invertierbar genau dann, wenn

$$\forall x \in \{1, \ldots, q-1\}. \, [g \cdot x]_q = [g]_q \cdot [x]_q \neq [0]_q$$

Beweis:

„⇒“: Sei $[g]_q$ invertierbar und $[g]_q \cdot [x]_q = [0]_q$ für ein $x \in \{1, \ldots, q-1\}$, so folgt unmittelbar, dass

$$[x]_q = [g]_q^{-1} \cdot [0]_q = [0]_q \quad \lightning$$

„⇐“: Ist $[g]_q$ nicht invertierbar, so ist nach Satz 3.13 $\mathrm{ggT}(g, q) \neq 1$. Œ sei $g > 0$, so gilt, dass $l, a, b \in \mathbb{N}$ existieren mit $l \geqslant 2, a < g$ und $b < q$ und $g = l \cdot a$ und $q = l \cdot b$. Daraus folgt allerdings, dass $a \cdot q = b \cdot g$, womit ein Widerspruch folgt.

$$[a \cdot q]_q = [a]_q \cdot [q]_q = [a]_q \cdot [0]_q = [0]_q \overset{\lightning}{=} [b]_q \cdot [g]_q \quad \text{, da } b \in \{1, \ldots, q-1\}$$

□

Satz 3.29 (*Erkennbarkeit von Einzelfehlern*)

Einzelfehler werden sicher erkannt genau dann, wenn

$$\forall i \in \{1, \ldots, m\}. \, [g_i]_m \text{ ist invertierbar.}$$

Beweis: Wir haben einen Fehler bei $j = j_0$, wobei $j_0 \leqslant n+1$ unbekannt ist. Damit gilt dann

$$[\delta]_n = \left[\sum_{j=1}^{m+1} g_j (d_j - d'_j) \right]_n = [g_{j_0}(d_{j_0} - d'_{j_0})]_n = [g_{j_0}]_n \cdot [d_{j_0} - d'_{j_0}]_n.$$

Wir erkennen den Fehler, wenn $[\delta]_n \neq [0]_n$, also muss für $0 < \left| d_{j_0} - d'_{j_0} \right| < n$ gelten, dass $[g_{j_0}(d_{j_0} - d'_{j_0})]_n \neq [0]_n$. Da die Differenz aber unbekannt ist und weil $[-x]_n = [n - x]_n$ gilt, verlangen wir strenger, dass für alle $d \in \{1, \ldots, n-1\}$

$$[g_{j_0}]_n \cdot [d]_n \neq [0]_n \text{ gilt.}$$

Nach Satz 3.28 ist dies genau dann der Fall, wenn $[g_{j_0}]_n$ invertierbar ist. Da aber auch j_0 unbekannt ist, verlangen wir diese Eigenschaft dann einfach von allen g_i. □

3.2.2 Erkennung von Vertauschungsfehlern

Satz 3.30 (*Erkennbarkeit von Vertauschungsfehlern*)

Vertauschungsfehler werden sicher erkannt genau dann, wenn

$$\forall i,j \in \{1,\dots,m+1\}, i \neq j.\ [g_i - g_j]_m \text{ ist invertierbar.}$$

Beweis: Anstatt der korrekten Nachricht $d_1,\dots,d_i,\dots,d_j,\dots,d_m|d_{m+1}$ werde $d_1,\dots,d_j,\dots,d_i,\dots,d_m|d_{m+1}$ übertragen, wobei die Positionen $1 \leqslant i < j \leqslant m+1$ unbekannt seien und $d_i \neq d_j$.
Wir folgern:

$$[\delta]_n = \left[\sum_{j=1}^{m+1} g_j(d_j - d_j')\right]_n = [g_i(d_i - d_j) + g_j(d_j - d_i)]_n$$

Wie schon beim Beweis zu Satz 3.29 ergibt dies dann die Bedingung

$$\mathrm{ggT}(|g_i - g_j|, n) = 1 \quad \text{für alle } i,j \in \{1,\dots,m+1\} \text{ mit } i \neq j,$$

was genau unserem Satz entspricht. □

Anmerkung

Ist n eine Primzahl und gilt zudem noch $0 < g_j \leqslant n-1$ für $j = 1,\dots,m+1$, so sind die obigen Bedingungen äquivalent zu

$$g_j \neq 0 \text{ für alle } j = 1,\dots,m+1 \qquad \text{bzw.} \qquad g_i \neq g_j \text{ für alle } i \neq j.$$

Bei nichtsimultanem Auftreten werden also **alle** Vertauschungs- und Einzelfehler erkannt.

Mit $j = i+1$ sprechen wir von Nachbarvertauschungsfehlern. Man kann die Bedingung abschwächen und erhält somit folgendes Korollar:

Korrolar 3.30 (*Erkennbarkeit von Nachbarvertauschungsfehlern*)

Nachbarvertauschungsfehler werden sicher erkannt genau dann, wenn

$$\forall i \in \{1,\dots,m\}.\ [g_i - g_{i+1}]_m \text{ ist invertierbar.}$$

3.2.3 „Fehlerkorrektur“

Im letzten Teil wollen wir uns kurz die Frage stellen, ob wir einen Einzelfehler auch „rekonstruieren“ können. Im Allgemeinen ist das mit unseren Kodes nicht möglich. Ist allerdings eine Steille/Ziffer d_{i_0} mit **festem** und **bekannten** i_0 **unlesbar**, so kann man die Prüfziffer zur Rekonstruktion von d_{i_0} verwenden, sofern $[g_{i_0}]_n$ invertierbar ist. Zusammengefasst:

Satz 3.31 (*Fehlerkorrektur*)

Sei in einem Wort $w = d_0 \dots d_{i_0} \dots d_m|d_{m+1}$ d_{i_0} unleserlich mit $0 \leqslant i_0 \leqslant m+1$, so ist eine Rekonstruktion möglich genau dann, wenn $[g_{i_0}]_n$ invertierbar ist.

Beweis: Wir wollen diesen Satz nun konstruktiv beweisen. Sei $I := \{1,\dots,m+1\} \setminus \{i_0\}$, so gilt:

$$P(w) = \sum_{i=0}^{m+1} g_i d_i \overset{!}{=} 0 \mod n \Leftrightarrow [g_{i_0}]_n [d_{i_0}]_n + \sum_{i \in I} [g_i]_n [d_i]_n \overset{!}{=} [0]_n$$

Damit leiten wir die Rechenvorschrift für d_{i_0} mit

$$[d_{i_0}]_n = -[g_{i_0}]_n^{-1} \cdot \sum_{i \in I} [g_i]_n [d_i]_n$$

her. □
Man kann diese Fähigkeit natürlich mit anderen Kodes verbessern (→ Reed-Solomon-Kodes).

3.3 RSA-Verschlüsselung

In diesem letzen Kapitel in C3 wollen wir uns den „alltäglichen" Nutzen von der euler'schen Phi-Funktion $\varphi(n)$ und Primzahlen im Sinne der Verschlüsselung kümmern. Dazu betrachten wir insbesondere das asymmetrische RSA-Verfahren, benannt nach den „Erfindern" R. **R**ivest, A. **S**hamir und L. **A**dlemann. Doch zuerst wollen wir einige Grundbegriffe klären.

3.3.1 Grundbegriffe der Kryptographie

Definition 3.24 (*Kryptographie*)

Kryptographie, von altgriechisch κρυπτός („verborgen") und γράφειν („schreiben"), ist im ursprünglichen Sinne die Wissenschaft der Verschlüsselung von Informationen.

Die zu verschlüsselnde Nachricht wir also vorab in „Pakete" $\in \mathbb{Z}_n$, wobei n fest, eingeteilt. Damit gilt:

Definition 3.25 (*Nachricht*)

Eine Aneinanderreihung von Wörtern $w \in \mathbb{Z}_n$, mit festem n, nennen wir eine **Nachricht**.

Jedes Wort einer Nachricht wird dabei für sich verschlüsselt, wir stellen den Prozess der Ver- und Entschlüsselung wie folgt dar:

Definition 3.26 (*Ver-/Entschlüsselung*)

Der Sender einer Nachricht **verschlüsselt** diese durch das Anwenden einer bijektiven Funktion

$$E_n : \mathbb{Z}_n \to \mathbb{Z}_n.$$

Der Empfänger **entschlüsselt** diese Nachricht durch das Anwenden der Umkehrfunktion E_n^{-1}.

Wir wollen nun noch Verfahren im Allgemeinen klassifizieren:

Definition 3.27 (*Symmetrische Verfahren*)

Bei symmetrischen Verfahren gilt $E_n = E_n^{-1}$. Damit kann jeder, der **verschlüsseln** kann, auch **entschlüsseln**.
Beispiele: Caesar, DES, Blowfish, 3DES, AES, Cast-256.

Definition 3.28 (*Asymmetrische Verfahren*)

Bei asymmetrischen Verfahren gilt $E_n \neq E_n^{-1}$. Der Empfänger veröffentlicht nun seinen **Ver**schlüsselungsschlüssel E_n, nicht aber den **Ent**schlüsselungsschlüssel E_n^{-1}. Deswegen ist bei asymmetrischen Verfahren auch der Name *„privater/öffentlicher Schlüssel-Verfahren"* (*private-public-key*) geläufig. Der Sender sollte dem Empfänger mit dessen E_n verschlüsselte Nachrichten zukommen lassen.
Man sollte in diesen Fällen sogenannte **Einwegfunktionen** für E_n verwenden, also Funktionen, bei denen man allein von Kenntnis von E_n nicht leicht auf E_n^{-1} schließen kann. (*Dies schließt eine gewisse Größe von n ein, so dass sich Wertetabellen **nicht zwangsläufig** lohnen.*)

Beispiele: RSA, McEliece, Chor-Rivest oder Elgamal.

3.3.2 RSA-Verfahren

Verfahren 3.3 (*RSA*)

Ziel: Sender A möchte Empfänger J eine verschlüsselte Nachricht übermitteln.

(1) Der Empfänger J bestimmt zwei wirklich **große** Primzahlen p und q mit $p \neq q$ und bestimmt das Produkt n mit

$$n := p \cdot q.$$

(2) Die Anzahl der zu n teilerfremden Zahlen ist (nach Definition 3.15) genau

$$\varphi(n) = \varphi(p) \cdot \varphi(q) = (p-1) \cdot (q-1)$$

(3) J bestimmt mit $1 < v < \varphi(n)$ eine zu $\varphi(n)$ teilerfremde Zahl (bspw. also eine Primzahl mit $v > \max\{p, q\}$).

(4) J bestimmt weiter mit $1 < e < \varphi(n)$ eine natürliche Zahl mit

$$v \cdot e \equiv 1 \mod \varphi(n) \Leftrightarrow e = [v]^{-1}_{\varphi(n)}$$

– Man verwende hierzu Verfahren 3.1 und löse damit dann $e \cdot v + l \cdot \varphi(n) = 1$ –

(5) Im Folgenden werden nur noch die Zahlen e, v und n benötigt, wovon J v und n veröffentlicht (sprich: A kann darauf zugreifen). e ist geheimzuhalten und p, q und $\varphi(n)$ zu zerstören!

(6) A verwandelt die zu sendende Information in eine Ziffernfolge und zerlegt die Ziffernfolge in kleinere, gleich lange Pakete Z_i mit $1 \leqslant i \leqslant k$.

(7) **Verschlüsselung**: A verschlüsselt nun mit den öffentlichen Informationen v, n die Nachricht mit der Beziehung

$$C_i := (Z_i)^v \mod n = [Z_i^v]_n = E_{n,v}(Z_i).$$

A übermittelt dann die Ziffernfolge $C_1, \ldots, C_k$ an J.

(8) **Entschlüsselung**: J kann nun mittels e den Kode entschlüsseln:

$$Z_i \equiv (C_i)^e \mod n.$$

Wir wollen nun die Korrektheit des Verfahrens zeigen. *Beweis:* $\mathbb{Z}$: $Z_i^{v \cdot e} \equiv Z_i \mod n$.
Nach Definition von e ist $e \cdot v = 1 + l \cdot \varphi(n)$, also gilt

$$(Z_i^v)^e = Z_i^{1+l\cdot\varphi(n)} = Z_i \cdot (Z_i^{\varphi(n)})^l.$$

Für $[Z_i]_n = [0]_n$, das heißt $Z_i = k \cdot n$, ist also die Behauptung trivial.
Nach Satz 3.26(i) ist also $(Z_i)^{\varphi(n)} = 1 \mod n$, falls $ggT(Z_i, n) = 1$, womit die Behauptung folgte.
Für $ggT(Z_i, n) > 1$ betrachten wir die Primfaktorzerlegung ($\rightarrow$ Satz 3.17) von Z_i und stellen fest, dass für sie gilt

$$Z_i = p_1 \cdot \cdots \cdot p_k < p \cdot q,$$

sie also **entweder** p **oder** q enthalten muss. *Beides ist allerdings nicht möglich, da sonst* $Z_i \geqslant n$.
Sei dann Œ $p \mid Z_i$, womit unmittelbar $ggT(Z_i, q) = 1$ folgt. Mit Satz 3.26(ii) folgt $Z_i^{q-1} = 1 \mod q$

und somit

$$Z_i^{\varphi(n)} = Z_i^{(p-1)(q-1)} = (Z_i^{q-1})^{p-1} = 1 \mod q,$$

das heißt $(Z_i^{\nu})^e - Z_i = 0 \mod q$, was wiederum äquivalent zur Aussage

$$(Z_i^{\nu})^e - Z_i = l \cdot q \text{ ist.}$$

Nun ist $p \mid Z_i$, womit gilt, dass $[Z_i]_p = [0]_p$ und somit auch $[Z_i]_p^{\nu \cdot e} = [0]_p$. Damit gilt allerdings ebenso, dass ein l' existieren muss mit

$$Z_i^{\nu \cdot e} - Z_i = l'p = lq,$$

woraus abzuleiten ist, dass $q \mid l'$ und somit $l' = l''q$ gilt. Damit gilt aber auch

$$Z_i^{\nu \cdot e} - Z_i = l'' \cdot q \cdot p = l'' \cdot n$$

und somit dann die zu zeigende Aussage

$$(Z_i^{\nu})^e \equiv Z_i \mod n.$$

□

Warum ist RSA so sicher?

Nach heutigem Kenntnisstand ist es sehr schwer eine große Zahl n in das Produkt zweier Primzahlen zu zerlegen und damit $\varphi(n)$ auszurechnen

Nach heutigem Kenntnisstand ist es sehr schwer in $\mathbb{Z}_n$ mit großem n, ν-te Wurzeln zu ziehen. *Interessanterweise ist es das **nicht** in $\mathbb{R}$ oder $\mathbb{Z}$. Hier schafft zum Beispiel das Newton-Verfahren (Verfahren 1.4) Abhilfe.*