Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät

Department
MATHEMATIK

Vorlesungsskript

Mathematik für Ingenieure III

Wintersemester 2022/23

Prof. Dr. Serge Kräutle

am

Lehrstuhl für Angewandte Mathematik (Modellierung und Numerik)

AM1

— Version vom 16.02.2023 —

Friedrich–Alexander–Universität Erlangen–Nürnberg

Vorbemerkung

Bei diesem Skript handelt es sich inhaltlich um die Vorlesungsfolien, ohne irgendwelche Ergänzungen oder Ausformulierungen. Nur das Layout der Folien wurde umgewandelt, der Text ist unverändert.

In theory, there is no difference between theory and practice.
In practice, there is.

unknown author

Inhaltsverzeichnis

1 Analysis im $\mathbb{R}^n$

Dies ist eine Fortsetzung des letzten Kapitels aus dem 2. Semester.

1.1 Extremstellen, Extremwertaufgaben

Problemstellung:
Gegeben: $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$ (ggf.: $D = \mathbb{R}^n$), f sei "hinreichend glatt", z.B. $f \in C^2(D)$,
Gesucht: Die lokalen Extremstellen von f

Erinnerung: Def. Extremstellen

- $\vec{x}_0 \in D$ heißt *lokale Maximal-[Minimal-]stelle* von $f : D \to \mathbb{R}$, falls es ein $\epsilon > 0$ gibt, so dass

$$f(\vec{x}) \leq f(\vec{x}_0) \quad \forall\, \vec{x} \in K_\epsilon(\vec{x}_0) \cap D \qquad [\ f(\vec{x}) \geq f(\vec{x}_0) \quad \forall\, \vec{x} \in K_\epsilon(\vec{x}_0) \cap D\ ]$$

  Bem.: Gilt hierbei "<" [bzw. ">"] für alle $\vec{x} \neq \vec{x}_0$, so spricht man von einer *isolierten Max.-[Min.-]stelle.*

- $\vec{x}_0 \in D$ heißt *globale Maximal-[Minimal-]stelle* von $f : D \to \mathbb{R}$, falls

$$f(\vec{x}) \leq f(\vec{x}_0) \quad \forall\, \vec{x} \in D \qquad [\ f(\vec{x}) \geq f(\vec{x}_0) \quad \forall\, \vec{x} \in D\ ]$$

Zur Notation: $K_\epsilon(\vec{x}_0) := \{\vec{x} \in \mathbb{R}^n \mid \|\vec{x} - \vec{x}_0\| \leq \epsilon\}$, "Epsilon-Kreis um $\vec{x}_0$"

Bevor wir (notwendige/hinreichende) Kriterien herleiten, erinnern wir uns an **notwendige/hinreichende Kriterien im Fall** $n = 1$ (s. Schule bzw. 2. Sem., Kap.3.6.2 und 3.6.4):

Erinnerung: Kriterien für Extremstellen, skalarer Fall
Für $f : \mathbb{R} \to \mathbb{R}$, $f \in C^2(\mathbb{R})$, $\vec{x}_0 \in \mathbb{R}$ gilt:

$$\left.\begin{array}{lllll} (1) & f'(x_0) = 0 \ \wedge\ f''(x_0) < 0 & \Rightarrow & x_0 \text{ ist lok. Max.-stelle} \\ (2) & f'(x_0) = 0 \ \wedge\ f''(x_0) > 0 & \Rightarrow & x_0 \text{ ist lok. Min.-stelle} \end{array}\right\} \text{hinreichende Kriterien}$$

$$\left.\begin{array}{lllll} (3) & x_0 \text{ ist lok. Extremstelle} & \Rightarrow & f'(x_0) = 0 \\ (4) & x_0 \text{ ist lok. Max.-stelle} & \Rightarrow & f'(x_0) = 0 \ \wedge\ f''(x_0) \leq 0 \\ (5) & x_0 \text{ ist lok. Min.-stelle} & \Rightarrow & f'(x_0) = 0 \ \wedge\ f''(x_0) \geq 0 \end{array}\right\} \text{notwendige Kriterien}$$

Beachte:

- In (1)-(2) steht die echte Ungleichheit, in (3)-(5) nicht.
  Dass dies so sein muss, kann man sich am Beispiel $f(x) = x^4$, $\tilde{f}(x) = -x^4$ überlegen.
  D.h. wir haben leider kein "genau-dann-Kriterium" für Max.-/Min.-stellen.
- Für (3) reicht $f \in C^1(\mathbb{R})$.
- Falls D eine *echte* Teilmenge von D ist, und Randpunkte von D zu D gehören (z.B. wenn D abgeschlossen ist), dann ist zu beachten, dass obige Kriterien nur für *innere* Punkte von D gelten, nicht für Randpunkte.

Erinnerung: Wie beweist man (1)-(5)?

Mittels Taylor-Entwicklung:

(I) $f(x) = f(x_0) + f'(\xi)\,(x-x_0)$ (Taylor nullter Ordnung)

(II) $f(x) = f(x_0) + f'(x_0)\,(x-x_0) + \frac{1}{2}f''(\xi)\,(x-x_0)^2$ (Taylor erster Ordnung)

Z.B. kann man mit (I) Aussage (3) beweisen:

Sei x_0 lok. Extremstelle. Angenommen $f'(x_0) \neq 0$.

[z.z.: x_0 ist keine lok. Extremstelle]

Falls $f'(x_0) > 0$, so gibt es (wegen $f' \in C^0$) ein $\epsilon_0 > 0$, so dass in der ϵ_0-Umgebung um x_0 gilt, dass $f'(x) > 0$. Sei nun $0 < \epsilon \leq \epsilon_0$ beliebig.
Wählt man $x := x_0 + \epsilon$, so ist mit (I) $f(x) = f(x_0) + \underbrace{f'(\xi)}_{>0}\underbrace{(x-x_0)}_{=\epsilon>0} > f(x_0)$.
Wählt man $x := x_0 - \epsilon$, so ist mit (I) $f(x) = f(x_0) + \underbrace{f'(\xi)}_{>0}\underbrace{(x-x_0)}_{=-\epsilon<0} < f(x_0)$.
x_0 ist also nicht Extremstelle, Widerspruch!
Analog im Fall $f'(x_0) < 0$. □
Mit (II) kann man (1), (2), (4), (5) beweisen.

Z.B. Beweis*idee* für (4):

Sei x_0 lok. Max.-stelle. Da wir (3) schon bewiesen haben, reicht es, $f''(x_0) \leq 0$ zu zeigen.
Dazu (II):

$$f(x) = f(x_0) + \underbrace{f'(x_0)}_{=0}(x-x_0) + \frac{1}{2}\,\underbrace{f''(\xi)}_{\approx f''(x_0)}\,\underbrace{(x-x_0)^2}_{\geq 0}$$

Wegen der Max.-Eigenschaft von x_0 (d.h. $f(x) \leq f(x_0)$) muss $f''(\xi) \leq 0$ sei.
Die Annahme, dass $f''(x_0) > 0$ sei, führt wegen $f'' \in C^0$ für x aus einer hinr. kleinen Umg. von x_0 zu Widerspruch. □

Nun: **Übertragung auf den mehrdimensionalen Fall:**

Taylor-Entwicklung von f (mehrdimensional):

$$\begin{aligned}&\text{(I)} \quad f(\vec{x}) = f(\vec{x}_0) + \langle \nabla f(\vec{\xi}), \vec{x}-\vec{x}_0 \rangle \\ &\text{(II)} \quad f(\vec{x}) = f(\vec{x}_0) + \langle \nabla f(x_0), \vec{x}-\vec{x}_0 \rangle + \tfrac{1}{2} \langle Hf(\vec{\xi})\,(\vec{x}-\vec{x}_0), \vec{x}-\vec{x}_0 \rangle\end{aligned}$$

wobei $\langle \cdot, \cdot \rangle$ das *Euklidische* Skalarprodukt ist, und wobei (Erinnerung)
$\nabla f(\vec{x}) = \begin{pmatrix} \partial_1 f(\vec{x}) \\ \vdots \\ \partial_n f(\vec{x}) \end{pmatrix}$ der Gradient und $Hf(\vec{x}) = \begin{pmatrix} \partial_1\partial_1 f(\vec{x}) & \dots & \partial_n\partial_1 f(\vec{x}) \\ \vdots & & \vdots \\ \partial_1\partial_n f(\vec{x}) & \dots & \partial_n\partial_n f(\vec{x}) \end{pmatrix}$ die Hesse-Matrix (symmetrisch!) ist,
und wobei $\vec{\xi} = \vec{x}_0 + \tau\,(\vec{x}-\vec{x}_0)$ mit $\tau \in (0,1)$.

Wir zeigen nun, analog zum 1-dim. Fall, dass $\nabla f(\vec{x}_0) = \vec{0}$ notwendiges Kriterium ist für Extremstellen $\vec{x}_0 \in \overset{\circ}{D}$:Sei also $\vec{x}_0 \in \overset{\circ}{D}$ Extremstelle von f.
[z.z.: $\nabla f(\vec{x}_0) = \vec{0}$]

Angenommen, $\nabla f(\vec{x}_0) \neq \vec{0}$.
Dann gibt es ein $i \in \{1, \dots, n\}$ mit $\partial_i f(\vec{x}_0) \neq 0$.

Falls $\partial_i f(\vec{x}_0) > 0$, so gibt es eine ϵ_0-Umgebung um $\vec{x}_0$ mit $\partial_i f(\vec{\xi}) > 0 \;\forall\, \vec{\xi} \in K_{\epsilon_0}(\vec{x}_0)$.
Sei $0 < \epsilon \leq \epsilon_0$ beliebig.

Wähle nun zum einen $\vec{x} := \vec{x}_0 + (0, \dots, 0, \epsilon, 0, \dots, 0)^T$, zum anderen
$\vec{x} := \vec{x}_0 - (0, \dots, 0, \epsilon, 0, \dots, 0)^T$. Man erhält einmal $f(\vec{x}) - f(\vec{x}_0) \overset{(I)}{=} \langle \nabla f(\vec{\xi}), \underbrace{\vec{x}-\vec{x}_0}_{=\epsilon\vec{e}_i} \rangle =$
$\underbrace{\partial_i f(\xi)}_{>0} \cdot \underbrace{\epsilon}_{>0} > 0$,
und einmal $f(\vec{x}) - f(\vec{x}_0) \overset{(I)}{=} \langle \nabla f(\vec{\xi}), \underbrace{\vec{x}-\vec{x}_0}_{=-\epsilon\vec{e}_i} \rangle = -\partial_i f(\xi) \cdot \epsilon < 0$.

Somit kann $\vec{x}_0$ kein Extrempunkt sein. Widerspruch!
Analog im Fall $\partial_i f(\vec{x}_0) < 0$.
Somit ist also $\nabla f(\vec{x}_0) = \vec{0}$. □

An Extremstellen gilt also

$$\begin{aligned} f(\vec{x}) &\overset{(II)}{=} f(\vec{x}_0) + \langle \overbrace{\nabla f(\vec{x}_0)}^{=\vec{0}}, \vec{x}-\vec{x}_0 \rangle + \frac{1}{2} \langle Hf(\vec{\xi})\,(\vec{x}-\vec{x}_0), \vec{x}-\vec{x}_0 \rangle \\ &= f(\vec{x}_0) + \frac{1}{2} \langle Hf(\vec{\xi})\,(\vec{x}-\vec{x}_0), \vec{x}-\vec{x}_0 \rangle \end{aligned}$$

Also: Über die Frage, ob $\vec{x}_0$ lok. Min.- oder Max.-Stelle ist, entscheidet alleine das Vorzeichen des quadratischen Terms.

Zur Beschreibung von Termen der Form $\langle A\vec{v}, \vec{v}\rangle$ führen wir daher Begriffe ein:

Def. (Definitheit)
Eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ heißt

(symmetrisch) positiv definit, falls $\langle A\vec{v}, \vec{v}\rangle > 0 \;\forall\; \vec{v} \in \mathbb{R}^n \setminus \{\vec{0}\}$,

(symmetrisch) negativ definit, falls $\langle A\vec{v}, \vec{v}\rangle < 0 \;\forall\; \vec{v} \in \mathbb{R}^n \setminus \{\vec{0}\}$,

(symmetrisch) positiv semidefinit, falls $\langle A\vec{v}, \vec{v}\rangle \geq 0 \;\forall\; \vec{v} \in \mathbb{R}^n$,

(symmetrisch) negativ semidefinit, falls $\langle A\vec{v}, \vec{v}\rangle \leq 0 \;\forall\; \vec{v} \in \mathbb{R}^n$.

Falls A weder positiv semidefinit noch negativ semidefinit ist, heißt A *indefinit*.
Dabei sei $\langle \cdot, \cdot \rangle$ das Euklidische Skalarprodukt.

$$\langle A\vec{v}, \vec{v}\rangle = \sum_{i,j=1}^{n} a_{ij} v_i v_j$$

Bemerkung: Eine Abbildung $\mathbb{R}^n \to \mathbb{R}$, $\vec{x} \mapsto \langle A\vec{x}, \vec{x}\rangle$ nennt man *quadratische Form*.

Bemerkung:

- (Nur) Bei der *Semi*definitheit ist es egal, ob man $\vec{x} = \vec{0}$ ausschließt oder nicht.
- Manche Autoren definieren die Definitheit auch für *nicht-symmetrische* Matrizen; dies wird aber selten gebraucht, d.h. in dem Fall kann man aus der Definitheit nicht auf die Symmetrie einer Matrix schließen. Um Missverständnisse zu vermeiden, ist es daher sinnvoll, genauer *"symmetrisch positiv/negativ (semi-)definit"* zu sagen.

Bemerkung:

- In obiger Definition von positiver/negativer (Semi-)Definitheit kann man "$\forall \vec{v} \in \mathbb{R}^n \setminus \{\vec{0}\}$" ersetzen durch "für alle $\vec{v} \in \mathbb{R}^n$ mit $\|\vec{v}\| = 1$".
  Grund:
  Für beliebiges $\vec{v} \in \mathbb{R}^n \setminus \{\vec{0}\}$ kann man $\vec{v}_0 := \frac{1}{\|\vec{v}\|}\vec{v}$ setzen; es ist dann $\|\vec{v}_0\| = 1$. Und es ist $\langle A\vec{v}, \vec{v}\rangle = \langle \|\vec{v}\|\, A\vec{v}_0, \|\vec{v}\|\, \vec{v}_0\rangle = \underbrace{\|\vec{v}\|^2}_{>0} \langle A\vec{v}_0, \vec{v}_0\rangle$, woran man sieht, dass $\langle A\vec{v}, \vec{v}\rangle$ und $\langle A\vec{v}_0, \vec{v}_0\rangle$ immer das gleiche Vorzeichen haben.
- "Symmetrisch positiv definit" kann als "s.p.d." abgekürzt werden.

Satz (notwendiges und hinreichendes Kriterium für Extremstellen)

Für $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$, $f \in C^2(D)$, $\vec{x}_0 \in \overset{\circ}{D}$ gilt:

(1)	$\nabla f(\vec{x}_0) = \vec{0} \wedge Hf(\vec{x}_0)$ neg. definit	$\Rightarrow$	$\vec{x}_0$ ist lok. Max.-stelle	}	hinr. Krit.
(2)	$\nabla f(\vec{x}_0) = \vec{0} \wedge Hf(\vec{x}_0)$ pos. definit	$\Rightarrow$	$\vec{x}_0$ ist lok. Min.-stelle		
(*)	$\nabla f(\vec{x}_0) = \vec{0} \wedge Hf(\vec{x}_0)$ indefinit	$\Rightarrow$	$\vec{x}_0$ ist keine lok. Extr.stelle		
(3)	$\vec{x}_0$ ist lok. Extremstelle	$\Rightarrow$	$\nabla f(\vec{x}_0) = \vec{0}$	}	notw. Krit.
(4)	$\vec{x}_0$ ist lok. Max.-stelle	$\Rightarrow$	$\nabla f(\vec{x}_0) = \vec{0} \wedge Hf(\vec{x}_0)$ ist neg. semidef.		
(5)	$\vec{x}_0$ ist lok. Min.-stelle	$\Rightarrow$	$\nabla f(\vec{x}_0) = \vec{0} \wedge Hf(\vec{x}_0)$ ist pos. semidef.		

In den Fällen (1), (2) ist $\vec{x}_0$ sogar *isolierte* Extremstelle.

Bemerkung:

- Wie schon im 1-D Fall haben wir auch hier keine "genau-dann"-Kriterien.
- Ist insbesondere $\nabla f(\vec{x}_0) = \vec{0}$ und $Hf(\vec{x}_0)$ nur (pos. oder neg.) semidefinit, so folgt daraus nichts interessantes. Ausweg für diesen Fall: s. Tafel, z.B. $f(x, y) := x^2 + 5(x-y)^6$.
- (3) gilt bereits für $f \in C^1$.
- Ist Kriterium (*) erfüllt, spricht man von einem *Sattelpunkt*.
- Erinnerung: Ein $\vec{x}_0$ mit $\nabla f(\vec{x}_0) = \vec{0}$ heißt *kritische Stelle* oder *stationärer Punkt* von f.

Zum Beweis:

(3) haben wir bereits erwiesen.

zu (4): Im Fall dass $\vec{x}_0$ lok. Max.-stelle ist, also $0 \geq f(\vec{x}) - f(\vec{x}_0) \forall \vec{x} \in K_\epsilon(\vec{x}_0)$ für ein $\epsilon > 0$, liefert die Taylor-Entwicklung (II), dass

$$\begin{aligned} 0 \geq f(\vec{x}) - f(\vec{x}_0) &\overset{(II)}{=} \langle \overbrace{\nabla f(\vec{x}_0)}^{=\vec{0} \text{ (s.o.)}}, \vec{x} - \vec{x}_0 \rangle + \frac{1}{2} \langle Hf(\vec{\xi})\,(\vec{x} - \vec{x}_0), \vec{x} - \vec{x}_0 \rangle \\ &= \frac{1}{2} \langle Hf(\vec{\xi})\,(\vec{x} - \vec{x}_0), \vec{x} - \vec{x}_0 \rangle \qquad \forall \vec{x} \in K_\epsilon(\vec{x}_0) \qquad (**) \end{aligned}$$

wobei $\vec{\xi}$ auf der geraden Verbindung von $\vec{x}$ und $\vec{x}_0$ liegt: $\vec{\xi} = \vec{x}_0 + \tau\,(\vec{x} - \vec{x}_0)$, $\tau \in (0, 1)$.

Es folgt: $Hf(\vec{\xi})$ ist negativ definit.

Unser Ziel war aber: $Hf(\vec{x}_0)$ ist negativ definit.

Dazu benötigt man nun die *Stetigkeit* von $\vec{x} \mapsto Hf(\vec{x})$ (d.h. die Stetigkeit der einzelnen Komponenten der Matrix $Hf(\vec{x})$). Statt eines festen $\vec{x}$ betrachtet man eine Folge $(\vec{x}_n)$, die gegen $\vec{x}_0$ konvergiert. Da das zugehörige $\vec{\xi}$, besser $\vec{\xi}_n$, immer auf der Verbindungslinie von $\vec{x}_n$ und $\vec{x}_0$ liegt, konvergieren dann auch die ξ_n gegen $\vec{x}_0$. Nun benötigen wir, dass $f \in C^2$, dass also die Einträge der Matrix $Hf(\vec{x})$ stetig von $\vec{x}$ abhängen. Aus dieser Stetigkeit folgt, dass $Hf(\vec{\xi}_n) \overset{(n\to\infty)}{\longrightarrow} Hf(\vec{x}_0)$ und somit letztendlich $\overbrace{\langle Hf(\vec{\xi}_n)\,(\vec{x}-\vec{x}_0), \vec{x}-\vec{x}_0\rangle}^{\leq 0} \overset{(n\to\infty)}{\longrightarrow} \langle Hf(\vec{x}_0)\,(\vec{x}-\vec{x}_0), \vec{x}-\vec{x}_0\rangle \quad \forall \vec{x} \in K_\epsilon(\vec{x}_0)$
Und da die Folgenglieder (links) alle ≤ 0 sind, muss auch der Grenzwert (rechts) ≤ 0 sein.
Damit ist (4) bewiesen.

Der Beweis von (5) ist analog. Und die Beweis von (1), (2), (*) sind ähnlich. □

Beispiele zur Anwendung (z.T. Scheitern) der Kriterien:
(i) $f(x,y) := x^2 - y^2$,
(ii) $f(x,y) := (x^2+y^2)\,e^x$,
(iii) $f(x,y) := x^2+y^2+5(x-y)^6$:
(iv) $f(x,y) := x^2+5(x-y)^6$:
(v) $f(x,y) := x^2+5(x-y)^5$
(iii-v): siehe Tafel

Was momentan noch unbefriedigend ist: Die Überprüfung von $Hf(\vec{x}_0)$ auf pos./neg. (Semi-)Definitheit anhand der Def. erscheint schwierig!
(In obigen Beispielen hatten wir Glück: $Hf(\vec{x}_0)$ enthielt viele Nullen. Einer voll besetzten Matrix sieht man jedoch nur schwer an, ob sie pos./neg. (semi-)definit ist!)

Wir suchen einfach handhabbares Kriterium für (Semi-)Definitheit:
Erinnerung: $Hf(\vec{x}_0)$ ist symmetrisch, und jede symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ ist *diagonalisierbar* ($\to$ Ende des 1. Semesters); man kann sie als

$$A = QDQ^{-1}$$

schreiben, wobei $D = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ ist, die λ_i die Eigenwerte von A sind, und Q eine Orthogonalmatrix ist, d.h. $Q^{-1} = Q^T$.

Wir setzen ein:

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

Es ist also

$$\begin{aligned} A \text{ pos. def.} &\Leftrightarrow \langle A\vec{v}, \vec{v}\rangle > 0 \quad \forall \vec{v} \in \mathbb{R}^n \setminus \{\vec{0}\} \\ &\Leftrightarrow \langle DQ^T\vec{v}, Q^T\vec{v}\rangle > 0 \quad \forall \vec{v} \in \mathbb{R}^n \setminus \{\vec{0}\} \\ &\overset{(*)}{\Leftrightarrow} \underbrace{\langle D\vec{y}, \vec{y}\rangle}_{=\sum_{i=1}^{n} \lambda_i y_i^2} > 0 \quad \forall \vec{y} \in \mathbb{R}^n \setminus \{\vec{0}\} \end{aligned}$$

wobei für die Äquivalenz (*) verwendet wurde, dass die lin. Abb. $\vec{x} \mapsto \vec{y} := Q^T\vec{x}$, als Abb. $\mathbb{R}^n \to \mathbb{R}^n$, damit aber auch als Abb. $\mathbb{R}^n \setminus \{\vec{0}\} \to \mathbb{R}^n \setminus \{\vec{0}\}$, bijektiv ist, denn $\det(Q^T) = \pm 1 \neq 0$.

Daraus ergibt sich:

$$A \text{ pos. def.} \overset{(s.o.)}{\Longleftrightarrow} \sum_{i=1}^{n} \lambda_i \, y_i^2 > 0 \, \forall \vec{y} \in \mathbb{R}^n \setminus \{\vec{0}\} \overset{(**)}{\Longleftrightarrow} \lambda_1, ..., \lambda_n > 0$$

wobei die λ_i die Diagonaleinträge von D sind.

Begründung für (**):
"$\Leftarrow$": Aus $\lambda_i > 0$ folgt zunächst $\lambda_i y_i^2 \geq 0$ und somit $\sum \lambda_i y_i^2 \geq 0$. Da mindestens ein $y_i^2 > 0$ folgt $\sum \lambda_i y_i^2 > 0$.

"$\Rightarrow$": Angenommen ein λ_j ist ≤ 0. Wähle dazu passend $y_j := 1$ und $y_k := 0 \; \forall \, k \neq j$. Dann ist $\sum_{i=1}^{n} \lambda_i \, y_i^2 = \lambda_j \leq 0$, Widerspruch. □

Wir haben somit für die positive Definitheit von A ein notwendiges und hinreichendes Kriterium hergeleitet, das die Einträge der Diagonalmatrix D verwendet. Beachte: Diese Einträge von D sind trivialerweise gleichzeitig die *Eigenwerte* von D, und damit auch die Eigenwerte von A (nach dem Satz: Sind zwei Matrizen ähnlich, dann haben sie die gleichen Eigenwerte).

Analog folgen Kriterien für negative Definitheit, Semidefinitheit und Indefinitheit:

Satz (Kriterium für Definitheit)
Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch, und seien $\lambda_1, ..., \lambda_n [\in \mathbb{R}!]$ die Eigenwerte von A. Dann gilt:

$$\begin{aligned} A \text{ positiv definit} &\iff \lambda_1, ..., \lambda_n > 0 \\ A \text{ negativ definit} &\iff \lambda_1, ..., \lambda_n < 0 \\ A \text{ positiv semidefinit} &\iff \lambda_1, ..., \lambda_n \geq 0 \\ A \text{ negativ semidefinit} &\iff \lambda_1, ..., \lambda_n \leq 0 \\ A \text{ indefinit} &\iff \exists \, i, j : \; \lambda_i > 0 \; \wedge \; \lambda_j < 0 \end{aligned}$$

Hier laufen die Fäden aus 1. und 2. Semester zusammen: Ein Problem der Analysis (Extremwerte bestimmen) wird mit Mitteln der Linearen Algebra (Eigenwerte) gelöst.

Später erleben wir Ähnliches bei Lösen von linearen Differentialgleichungen.

Noch ein Trick für den Fall $n=2$, mit dem man etwas Arbeit einsparen kann:

Im Fall $n=2$ reicht es, Determinante und Spur der Matrix zu kennen, um auf Definitheit zu schließen; die EW sind nicht explizit nötig:

siehe Tafel

Beachte noch: Die genannten Kriterien für lokale Extremstellen gelten für Punkte aus dem *Inneren* des Definitionsbereichs. Falls D Randpunkte hat (die zu D gehören), können auch dort lokale Extremstellen liegen. (Für $f : [a, b] \to \mathbb{R}$ also an den Stellen $x=a$ und $x=b$. Wie man im mehrdimensionalen Fall in Frage kommende Randpunkte von D ausfindig macht: siehe Kap. I.2!)

Schließen von lokalen auf *globale* Extrema:

Es ist mühsam dazu ein *allgemeines* Rezept anzugeben, insbesondere falls D eine *echte* Teilmenge von $\mathbb{R}^n$ ist.
Was man weiß: Für *innere* Punkte gilt:

- Eine globale Max.-Stelle ist immer auch lokale Max.-Stelle. Also: Das globale Max. wird immer an einer der lokalen Max.-Stellen angenommen oder aber es existiert nicht. Insbes.: Falls es keine lok. Max.-Stelle gibt, dann auch keine globale.

- Eine lokale Max-Stelle ist immer auch kritische Stelle. Also: Das globale Maximum, falls es existiert, wird an einer *kritischen* Stelle angenommen – oder an einem Randpunkt. Mann muss also nicht zwingend die Hesse-Matrix bemühen.

- Und: Falls D kompakt ist (=abgeschlossen und beschränkt), dann muss es nach Satz aus IngMath2 ein globales Max. geben. Und dieses muss an einer lok. Max.-Stelle liegen, d.h. an einer kritischen Stelle liegen - oder am Rand. Man muss also nur die kritischen Stellen und die Randpunkte nach dem größten dort vorkommenden Funktionswert durchsuchen. Analog für Minima. (Hesse-Matrix somit gar nicht erforderlich.)

Vorgehensweise also:
Zunächst einmal muss man also alle kritischen Stellen im Inneren von D, aber, falls es Randpunkte von D gibt, die zu D gehören, dann leider auch diese (dazu siehe Kap. I.2) betrachten. Dann die Funktionswerte an allen diesen Stellen berechnen, und darunter den größten Wert $=: M$ und den kleinsten Wert $=: m$ ermitteln.

Falls D kompakt ist, dann muss M das globale Max. und m das globale Minimum sein. Falls D nicht kompakt ist, kann man (durch geeignetes Abschätzen) versuchen zu

zeigen, dass $f(\vec{x}) \geq m \, \forall \vec{x} \in D$ (bzw. dass $f(\vec{x}) \leq M \, \forall \vec{x} \in D$), woraus dach Def. des glob. Min./Max. m *globales* Minimum (bzw. M *globales* Maximum) ist. Siehe Beispiel unten auf nächster Seite.

Falls man stattdessen $\vec{x} \in D$ findet mit $f(\vec{x}) > M$ (bzw. mit $f(\vec{x}) < m$), so handelt es sich bei m bzw. M nicht um globale Extrema, und dann gibt es kein globales Max. bzw. Min.

Falls D beschränkt und abgeschlossen (also kompakt) ist und der Rand von D 'glatt' ist: Diesen Fall werden wir gleich noch systematisch untersuchen! (siehe Kap. I.2)
Ein weiterer Spezialfall wird in Kap. I.5 behandelt: "f konvex"

Beispiel, wo das Abschätzen sehr einfach ist:
Für $f(x,y) := (x-3)^4 + (x+y)^6$, $D := \mathbb{R}^2$, gibt es keine Randpunkte, und unser ∇f-basiertes Kriterium liefert, dass nur $\vec{x}_0 = (3,-3)^T$ einzige potenzielle lokale Extremstelle im Innern von D ist. Es ist $f(3,-3) = 0$. Dann weiß man:
Entweder ist 0 das globale Maximum, oder das globale Maximum existiert nicht. Und:
Entweder ist 0 das globale Minimum, oder das globale Minimum existiert nicht.
Eine triviale Abschätzung zeigt: $f(x,y) \geq 0 \forall (x,y)^T$, somit $f(3,-3) \leq f(x,y) \forall (x,y)^T \in D$, d.h. es
handelt sich um ein globales Minimum. (Dass $f(3,-3)$ nicht globales Max. sein kann (das globale Max. also nicht existiert; es ist $\sup_D f(x) = +\infty$), sieht man z.B. an $f(0,0) = 3^4 > f(3,-3) = 0$.)

1.2 Extremwertaufgaben mit Nebenbedingungen

Motivation, geometrisch:

1. Gegeben seien zwei Flächen F, G im $\mathbb{R}^3$ in der Form $f(x,y,z) = 0$ bzw. $g(x,y,z) = 0$, d.h. $F = \{(x,y,z) \in \mathbb{R}^3 \mid f(x,y,z) = 0\}$, $G = \{(x,y,z) \in \mathbb{R}^3 \mid g(x,y,z) = 0\}$
   Gesucht ist der Abstand der Flächen, d.h. der Abstand der Punkte $(x_1,y_1,z_1) \in F$ und $(x_2,y_2,z_2) \in G$, für die $\|(x_1,y_1,z_1) - (x_2,y_2,z_2)\|$ minimal wird.
   Dies kann formuliert werden als: Finde Minimum der Funktion $h : \mathbb{R}^6 \to \mathbb{R}$, $h(x_1,y_1,z_1,x_2,y_2,z_2) := \|(x_1,y_1,z_1) - (x_2,y_2,z_2)\|$ unter den Nebenbedingungen $f(x_1,y_1,z_1) = 0$, $g(x_2,y_2,z_2) = 0$.

2. Gegeben seien zwei Kurven F, G im $\mathbb{R}^3$, und zwar F in der Form $y = f_1(x)$, $z = f_2(x)$ und G als $x = g_1(y)$, $z = g_2(y)$. Gesucht ist der Abstand der Kurven, d.h. der Abstand der Punkte $(x_1,y_1,z_1) \in F$ und $(x_2,y_2,z_2) \in G$, für die $\|(x_1,y_1,z_1) - (x_2,y_2,z_2)\|$ minimal wird.
   Dies kann formuliert werden als: Finde Minimum der Funktion $h : \mathbb{R}^6 \to \mathbb{R}$, $h(x_1,y_1,z_1,x_2,y_2,z_2) := \|(x_1,y_1,z_1) - (x_2,y_2,z_2)\|$ unter den Nebenbedingungen $y_1 = f_1(x_1)$, $z_1 = f_2(x_1)$, $x_2 = g_1(y_2)$, $z_2 = g_2(y_2)$.

Manchmal(!) kann man Probleme mit Nebenbedingungen auch *ohne* Nebenbedingungen formulieren. Dazu die **Nebenbedingungen** in die zu minimierende Funktion *einsetzen* ('**eliminieren**'). Z.B. falls die Flächen in Beispiel 1 in 'expliziter' Form $z = \tilde{f}(x,y)$ und $z = \tilde{g}(x,y)$ geschrieben werden können, dann kann man die Variablen z_1 und z_2 sowie die Nebenbedingung eliminieren und erhält das 'nicht-restringierte' Optimierungsproblem
Finde Minimum der Funktion
$h : \mathbb{R}^4 \to \mathbb{R}$, $h(x_1,y_1,x_2,y_2) := \|(x_1,y_1,\tilde{f}(x_1,y_1)) - (x_2,y_2,\tilde{g}(x_2,y_2))\|$.
Zu diesem Eliminieren der Nebenbedingung durch Einsetzen später noch mehr.

3. Probleme der Bauart: Finde unter allen Dreiecken (oder Rechtecken) mit fest vorgegebenem Umfang dasjenige, dessen Flächeninhalt (oder Umkreisradius oder Inkreisradius etc.) maximal (bzw. minimal) ist.

Motivation, analytisch-mathematisch:

4. Gesucht sind die *globalen* Extrema einer Funktion $f : D \to \mathbb{R}$, wobei $D \subset \mathbb{R}^n$ *kompakt* ist.
   Man kann die Suche nach potenziellen Extremstellen $\vec{x}_0$ unterteilen nach (1.) $\vec{x}_0 \in \overset{\circ}{D}$ und (2.) $\vec{x}_0 \in \partial D$.
   Zu (1.): Wie man (lok.) Extremstellen in $\overset{\circ}{D}$ findet, haben wir in I.1. gesehen.
   Zu (2.): Die Suche nach Extremstellen in ∂D kann man als
   Optimierungsproblem mit der Nebenbedingungen "$\vec{x}_0 \in \partial D$" formulieren.

Beispiel:
Gesucht seien die globalen Extrema der Funktion $f(x,y) := y - x^2$ unter der Nebenbedingung (NB) $x^2+y^2=1$.

1. Schritt: Schreibe die NB um in die Form $g(x,y)=0$.
Hier (z.B.): $g(x,y) := x^2+y^2-1$

Skizze der Niveaulinien von f und von g: Siehe Tafel.

Man erkennt: Potenzielle Extremstellen sind dort auf der Kurve $g(x,y) = 0$, wo die Niveaulinien von f und von g tangential verlaufen. Da der Gradient immer orthogonal auf Niveaulinien steht, können wir solche potenziellen Extremstellen also dadurch charakterisieren, dass dort $\nabla f(\vec{x})$ und $\nabla g(\vec{x})$ 'kollinear' sind, d.h. dass

$$\exists\, \lambda_1, \lambda_2 \in \mathbb{R} : \quad \lambda_1 \nabla f(\vec{x}) + \lambda_2 \nabla g(\vec{x}) = \vec{0}, \quad (\lambda_1,\lambda_2) \neq (0,0). \qquad (A)$$

Nebenbemerkung: Man sollte bei der Wahl von g ein wenig darauf achten, dass auf der Kurve $g(x,y)=0$ überall $\nabla g(\vec{x}) \neq \vec{0}$ (*) ist, denn andernfalls lässt sich die Bedingung für beliebiges f immer erfüllen per $\lambda_1 := 0$. Im obigen Bsp. sollte man also z.B. nicht $g(x,y) := (x^2+y^2-1)^2$ wählen.

Wir führen hier also zusätzliche Unbekannte λ_1, λ_2 in unser Problem ein.
Wir überzeugen uns nun davor, dass es möglich ist, lediglich *eine* neue Unbekannte λ anstelle von λ_1, λ_2 einzuführen und zu fordern:

$$\exists\,\lambda\in\mathbb{R}: \quad \nabla f(\vec{x}) = \lambda\,\nabla g(\vec{x}) \qquad (B)$$

Begründung:
Aus (B) folgt offensichtlich (A), indem man $\lambda_1 := 1$, $\lambda_2 := -\lambda$ wählt.

Aus (A) folgt (B), falls wir die Bedingung (*) beachten, denn aus (A) und (*) folgt, dass $\lambda_1 \neq 0$ sein muss; man kann also (A) durch λ_1 dividieren und erhält (B) mit $\lambda := -\lambda_2/\lambda_1$.

Ergebnis (Joseph-Louis Lagrange, 1736-1813):

Satz (notw. Kriterium für (lok./glob.) Extremwerte unter Nebenbed.; Lagrange-Formalismus)
Seien $f, g : \mathbb{R}^n \to \mathbb{R}$, $f, g \in C^1(\mathbb{R}^n)$, und es sei $\nabla g \neq \vec{0}$ dort wo $g = 0$ ist.
Notwendige Bedingung dafür, dass $\vec{x} \in \mathbb{R}^n$ (lok./glob.) Extremstelle von f unter der NB $g(\vec{x}) = 0$ ist, ist: Es gibt $\lambda \in \mathbb{R}$ mit

$$\begin{array}{lllll} \text{(I)} & \nabla f(\vec{x}) - \lambda\,\nabla g(\vec{x}) & = & \vec{0} & (n \text{ Gleichungen}) \\ \text{(II)} & g(\vec{x}) & = & 0 & (1 \text{ Gleichung}) \end{array}$$

Formaler Beweis: s. z.B. Heuser, Analysis, Band 2, Kap. 174

- Wir haben also ein i.a. nichtlineares(!) Gleichungssystem aus $n+1$ Gleichungen für die $n+1$ Unbekannten $\vec{x}, \lambda$. Das Lösen kann beliebig schwierig sein.
- Die Unbekannte λ ist eine Hilfsgröße, die am Ende weggeworfen werden kann.
- λ heißt *Lagrange-Multiplikator*
- Die Funktion $\mathcal{L}(\vec{x}, \lambda) := f(\vec{x}) - \lambda\, g(\vec{x})$ heißt das zum obigen Optimierungsproblem gehörende *Lagrange-Funktional*.
- Man kann (I) als $\nabla_x \mathcal{L}(\vec{x}, \lambda) = \vec{0}$ schreiben und (II) als $\partial_\lambda \mathcal{L}(\vec{x}, \lambda) = 0$. (I) und (II) zusammen können somit knapp als $\nabla \mathcal{L}(\vec{x}, \lambda) = \vec{0}$ geschrieben werden. Es werden also kritische Punkte von $\mathcal{L}$ gesucht.
- Indem man λ durch $-\lambda$ ersetzt, kann man genau so gut das $-\lambda$ in (I) und in der Definition von $\mathcal{L}$ ersetzen durch $+\lambda$.
- **Zur Existenz von Lösungen:**
  Man kann zeigen, dass die Menge aller $\vec{x} \in \mathbb{R}^n$, für die $g(\vec{x}) = 0$ ist, bei stetigem g, immer abgeschlossen ist. Ist diese Menge außerdem beschränkt, so ist sie kompakt, und, da f stetig, muss f auf dieser Menge (glob., somit auch lok.) Extrema annehmen.
  Ist die Menge unbeschränkt, so brauchen keine (glob./lok.) Extrema zu existieren.

– **Praktische Vorgehensweise zur Bestimmung von glob. Extrema von f auf kompakter Menge $D \subset \mathbb{R}^n$:**
1. kritische Stellen $\vec{x} \in \overset{\circ}{D}$ suchen per $\nabla f(\vec{x}) = \vec{0}$
2. kritische Stellen $\vec{x} \in \partial D$ suchen per $\nabla \mathcal{L}(\vec{x}, \lambda) = \vec{0}$. Dann den größten/kleinsten der Funktionswerte aller krit. Stellen ermitteln.
(Zumindest in ∂D muss es (mindestens 2) kritische Stellen geben.)
(Bem.: Hesse-Matrix nicht nötig)

Zurück zu obigem **Beispiel**:

Bestimme das Maximum und Minimum von $f(x, y) := y - x^2$
(a) unter der NB $x^2 + y^2 = 1$,
(b) unter der NB $x^2 + y^2 \leq 1$.

Rechnung: Siehe Tafel.

Ein Optimierungsproblem mit Nebenbedingung(en) wird auch als *restringiertes Optimierungsproblem* bezeichnet. Die Menge der Punkte, die die NB erfüllen, wird auch als *Menge der zulässigen Punkte* older kurz *zulässige Menge* bezeichnet (engl.: *constrained optimization problem, admissible points, admissible set*).

Weiterführende Bemerkung: Eine Verallgemeinerung des Lagrange-Systems auf Optimierungsprobleme, die neben Gleichungs- zusätzlich Ungleichungsnebenbedingungen ('$\leq$') beinhaltet, ist die *Karush-Kuhn-Tucker-Bedingung*, kurz *KKT-Bedingung*; siehe Literatur/Optimierungsvorlesungen.

Das Lösen des Lagrange-Systems kann schwierig sein; es kann kein allgemeines Rezept gegeben werden. (Bei komplizierten Systemen muss man damit rechnen, dass es sogar unmöglich ist, die Lösungen exakt zu berechnen; dann können numerische Näherungsverfahren helfen, s. Kap. I.7.)

Einige Tipps, die manchmal helfen:

- Bei $n = 2$: Welche der 3 Variablen lässt sich eliminieren? Brauchen 2 Gleichungen mit nur 2 Variablen. Oft kann man λ leichter als x, y eliminieren, da es nur linear eingeht.
- Oft kann es helfen, Gleichungen auf die Form "Produkt= 0" zu bringen. Ein Produkt ist genau dann null, wenn einer der Faktoren null ist. Dies führt dazu, dass man verschiedene *Fälle* "Faktor1= 0", "Faktor2= 0",... unterscheiden muss. (Manchmal ist es geschickter, die Fälle "Faktor1= 0", "Faktor1$\neq$ 0" zu betrachten.)
- Häufiger Fehler: Gewisse Fälle werden vergessen. Oft geschieht das so: Man dividiert durch Größen, von denen man nicht sicher weiß, dass sie $\neq 0$ sind. Wenn man das tun will, muss man den Fall "... $= 0$" separat abhandeln!

- Multiplikation mit Größen, die $=0$ sein können, ist nicht streng verboten; solche Aktionen führen 'nur' dazu, dass man möglicherweise die Zahl der in Frage kommenden potenziellen Extremstellen vergrößert. (Kurz: Alle Umformungen müssen zumindest "$\Rightarrow$" erfüllen; "$\Leftrightarrow$" ist wünschenswert, aber nicht zwingend erforderlich.)
- Ist $f = g \circ \tilde{f}$, wobei $g : \mathbb{R} \to \mathbb{R}$ *monoton* ist, so haben f und $\tilde{f}$ die gleichen Extremstellen (nicht unbedingt die gleichen Extrema). Es reicht dann also, nach Extremstellen von $\tilde{f}$ zu suchen.
  Bsp.: Für $f(x,y) := e^{(x-y)^2 y}$ suche krit. Stellen von $\tilde{f}(x,y) := (x-y)^2 y$.

Verallgemeinerung der Lagrange-Multiplikator-Methode bei *mehreren* NB

Satz (Lagrange-Formalismus bei *mehreren* Nebenbedingungen)
Problemstellung:
Sei $f, g_1, ..., g_m : \mathbb{R}^n \to \mathbb{R}$, $f, g_1, ..., g_m \in \mathcal{C}^1(D)$.
Suche Extrema von f unter den NB $g_1(\vec{x}) = 0, ..., g_m(\vec{x}) = 0$.
Falls ferner $\nabla g_1(\vec{x}), ..., \nabla g_m(\vec{x})$ lin. unabh. sind an allen Stellen $\vec{x}$, die die NB erfüllen, ist eine notwendige Bedingung für Extremstellen:

$$\begin{array}{lll} \text{(I)} & \nabla f(\vec{x}) = \lambda_1 \nabla g_1(\vec{x}) + ... + \lambda_m \nabla g_m(\vec{x}) & (n \text{ Gleichungen}) \\ & \left.\begin{array}{rcl} g_1(\vec{x}) & = & 0 \\ & \vdots & \\ g_m(\vec{x}) & = & 0 \end{array}\right\} & \\ \text{(II)} & & (m \text{ Gleichungen}) \end{array}$$

I.a. hat man $m < n$ (Großes m erhöht die Wahrscheinlichkeit, dass die zulässige Menge, d.h. die Menge aller Punkte, die die Nebenbedingungen erfüllen, leer ist).
Für jede NB bekommt man einen eigenen Lagrange-Multiplikator.

Das zugehörige Lagrange-Funktional lautet

$$\begin{aligned} \mathcal{L}(\vec{x}, \vec{\lambda}) &:= f(\vec{x}) - \lambda_1 g_1(\vec{x}) - ... - \lambda_m g_m(\vec{x}) \\ &= f(\vec{x}) - \langle \vec{\lambda}, \vec{g}(\vec{x}) \rangle \end{aligned}$$

damit kann das Langrange-System knapp als $\nabla \mathcal{L}(\vec{x}, \vec{\lambda}) = \vec{0}$ geschrieben werden, wobei sich das '∇' auf $\vec{x}$ und $\vec{\lambda}$ bezieht.

Alternative Strategie für Optimierungsprobleme mit NB

Idee: Eliminiere die NB

- Warnung: Ist i.a. eher nicht zu empfehlen, es sei denn ggf. wenn
  - die NB affin-linear ist/sind
  - oder sich zumindest "global" nach einer Variable auflösen lässt, z.B. als $x_1 = \tilde{g}(x_2, ..., x_n)$.
- Die Methode wird hier vorgestellt, da sie als Motivation/Überleitung zum nächsten Kapitel I.3 herhalten kann.

Verdeutlichung am obigen Beispiel: $f(x,y) = y - x^2$ mit NB $x^2 + y^2 = 1$.
Löse NB nach einer der Variablen auf.
Achtung: Leider geht das hier nur *abschnittsweise ("lokal")*, z.B.

$$x^2+y^2=1 \quad \Longleftrightarrow x = \sqrt{1-y^2} =: x_1(y) \ \vee \ x = -\sqrt{1-y^2} =: x_2(y), \qquad x_1, x_2 : [-1,+1] \to \mathbb{R}$$

d.h. x_1 stellt den rechten und x_2 den linken Halbkreis dar.
Betrachte nun f auf linkem und rechtem Halbkreis separat:
$\tilde{f}_1(y) := f(x_1(y), y) = y - x_1(y)^2 = y - (1-y^2) = y^2 + y - 1$
$\tilde{f}_2(y) := f(x_2(y), y) = y - x_2(y)^2 = y - (1-y^2) = y^2 + y - 1$
($\tilde{f}_1, \tilde{f}_2$ sind nur 'zufällig' gleich.)
$\tilde{f}_1, \tilde{f}_2 : [-1,1] \to \mathbb{R}$. Bestimme kritische Stellen von $\tilde{f}_1, \tilde{f}_2$:
$\tilde{f}_1'(y) = \tilde{f}_2'(y) = 2y + 1 \stackrel{!}{=} 0 \ \Leftrightarrow \ y = -\frac{1}{2}$,
$x_1(-\frac{1}{2}) = \sqrt{1 - \frac{1}{4}} = \frac{1}{2}\sqrt{3}, \ x_2(-\frac{1}{2}) = -\frac{1}{2}\sqrt{3}$
Also sind $(\pm\frac{1}{2}\sqrt{3}, \frac{1}{2})$ kritische Stellen von f unter der NB.

Doch Vorsicht! $\tilde{f}_1, \tilde{f}_2$ können auch an den Rändern ihrer Def.-bereiche, also bei ± 1, Extremstellen haben! Es müssen also auch $x_1(\pm 1)$, $x_2(\pm 1)$, die hier alle $= 0$ sind, berechnet, und $(0, \pm 1)$ betrachtet werden:
Kritische Stellen sind also $(\pm\frac{1}{2}\sqrt{3}, \frac{1}{2}), (0, \pm 1)$.

Evaluation von f an diesen Stellen liefert globales Minimum/Maximum.

Bemerkungen:

- Variante: Wenn man *überlappende* abschnittsweise Darstellungen der zulässigen Menge wählt, also z.B. $x_1(y), x_2(y)$ wie oben und zusätzlich $y_1(x) := \sqrt{1-x^2}$, $y_2(x) := -\sqrt{1-x^2}$, alle vier $[-1,+1] \to \mathbb{R}$, dann kann man auf die Untersuchung an den Rändern der Def.-bereiche verzichten.
- Sobald $n > 2$ oder $m > 1$ wird die Eliminationsmethode deutlich unübersichtlicher als die Lagrange-Methode; es sei denn, die NB lässt sich *global* mittels einer Auflösungsfunktion darstellen, was z.B. bei affin-linearen NB der Fall ist.

– Für komplizierte NB lassen sich die (lokalen oder die globale) Auflösungsfunktion(en) oft gar nicht explizit angeben (selbst wenn man weiß, dass sie existieren). Siehe dazu auch das folgende Kapitel 1.3, in dem es um die Existenz von (lokalen) Auflösungsfunktionen geht.

1.3 Satz über implizite Funktionen

Problemstellung:

Wir haben eine Abbildung $f : \mathbb{R}^2 \to \mathbb{R}$ (später auch: $\mathbb{R}^{n+m} \to \mathbb{R}^m$) gegeben und betrachten die Gleichung $f(x,y)=0$. Eine solche Gleichung beschreibt 'meistens' eine *Kurve* im $\mathbb{R}^2$ (jedoch: betrachte auch Bsp. (b) unten!).

Frage: Kann man diese Gleichung nach x oder nach y (oder beides) *auflösen*?
Das heißt: Kann man die Menge (Kurve?) $\{(x,y) \in \mathbb{R}^2 \mid f(x,y)=0\}$ als Graph einer Funktion $y = y(x)$ oder als $x = x(y)$ schreiben?

Beispiele:

(a) $x^2+y^2=1$ (setze $f(x,y) := x^2+y^2-1$)

(b) $x^3y+3xy^4=0$

(c) $x^2+y^2=0$

Skizzen/Auflösungsfunktionen siehe Tafel.

Man sieht: Nicht immer ist die Menge eine Kurve; es können Verzweigungen vorkommen, die Menge kann sogar aus nur einem Punkt bestehen (oder natürlich auch leer sein). Bei Verzweigungspunkten kann an nicht erwarten, die Menge in der Form $x \mapsto y=y(x)$ oder $y \mapsto x=x(y)$ darzustellen, sondern bestenfalls *Teile* der Menge. Sogar im Fall (a), in dem die Menge eine einfache glatte Kurve ist, kann man durch solche Auflosungsfunktionen nur *Teile* der Kurve, nicht die gesamte Kurve, darstellen.

Wir wollen uns mit 'lokalen' Auflösungsfunktionen begnügen, d.h. wir wollen fragen, welche Teilbereiche einer Menge $\{(x,y) \in \mathbb{R}^2 \mid f(x,y)=0\}$ wir als $x \mapsto y = y(x)$ oder $y \mapsto x=x(y)$ schreiben können; wir suchen ein Kriterium für die Existenz einer 'lokalen Auflösungsfunktion'.

Wozu man das brauchen könnte:

- Zur **Eliminierung von Nebenbedingungen** in Optimierungsproblemen (s. Kap. I.2) oder zur Eliminierung einzelnen Gleichungen/Unbekannter in (nichtlinearen) Gleichungssystemen

- Um zu klären, ob sich die Menge, die durch $g(x,y)=0$ wirklich als 'glatte Kurve' darstellen lässt, bzw. an welchen Stellen dies nicht der Fall ist, d.h. Knicke oder **Verzweigungspunkte** vorliegen.

- Dies lässt sich natürlich auch auf **Niveaumengen** von Funktionen, d.h. Mengen die durch $g(x,y)=const$ beschrieben werden, übertragen, indem man die Konstante auf die linke Seite bringt: $\tilde{g}(x,y) := g(x,y) - const = 0$.

- Weitere Anwendung: Zur Klärung, ob eine Funktion $f:\mathbb{R}\to\mathbb{R}$ oder $f:\mathbb{R}^n\to\mathbb{R}^n$ (zumindest "lokal", d.h. eingeschränkt auf eine Umgebung eines vorgegebenen festen Punktes) eine **Umkehrfunktion** hat.
  Bsp.: Hat $\underbrace{f(x)}_{=:y} = x\,e^{-x}$ (lokal um $x_0\in\mathbb{R}$) eine Umkehrfunktion? Schreibe dies als $\underbrace{y - xe^{-x}}_{=:\tilde{f}(x,y)} = 0$, und frage, ob sich $\tilde{f}$ nach x (lokal) auflösen lässt.
  (In diesem Sinne kann man die Suche nach Auflösungsfunktionen als Verallgemeinerung der Suche nach Umkehrfunktionen auffassen.)

Präzisierung der Fragestellung:
Sei $f:\mathbb{R}^2\to\mathbb{R}$ eine hinreichend glatte Funktion, und sei $(x_0,y_0)\in\mathbb{R}^2$ ein Punkt mit $f(x_0,y_0)=0$.
Gibt es "lokal" bei (x_0,y_0) eine Auflösungsfunktion $x\mapsto y(x)$, die die durch $f(x,y)=0$ charakterisierte Menge darstellt? Das heißt, es soll $y(x_0)=y_0$ sein und

$$f(x,y(x))=0 \quad \forall\, x\in(x_0-\epsilon, x_0+\epsilon)$$

gelten für ein $\epsilon>0$. Und analog:
Sei (x_0,y_0) ein Punkt mit $f(x_0,y_0)=0$. Gibt es "lokal" bei (x_0,y_0) eine Auflösungsfunktion $y\mapsto x(y)$, die die durch $f(x,y)=0$ charakterisierte Menge darstellt? Das heißt, es soll $x(y_0)=x_0$ sein und $f(x(y),y)=0$ gelten für alle $y\in(y_0-\epsilon,y_0+\epsilon)$, für ein $\epsilon>0$.

Welche Bedingung an f ist zu stellen, damit diese Auflösungsfunktionen existieren?

Einschränkung:

- Leider kann man häufig die (lok.) Auflösungsfunktion (so wie auch Umkehrfunktionen) – selbst wenn man weiß, dass sie existieren – nicht explizit ausrechnen/angeben.
- Der – gleich folgende – Satz über implizite Funktionen sagt leider nichts darüber, ob und wie man die Auflösungsfunktionen *berechnen/angeben* kann, sondern nur, ob eine *existiert*.
- Jedoch: Ein genauerer Blick in den Beweis des Satzes zeigt, wie ein numerisches Rechenverfahren aussehen kann, mit dem man die Auflösungsfunktion näherungsweise berechnen kann!

Bevor wir zum eigentlichen Satz kommen, machen wir folgende **Plausibilitätsüberlegung**:

Teil 1:
Wenn eine glatte Funktion $x \mapsto y(x)$, die $f(x, y(x)) = 0$ erfüllt, existieren soll, dann folgt per Differenzieren nach der Kettenregel, dass $\partial_1 f(x, y(x)) + \partial_2 f(x, y(x)) \cdot y'(x) = 0$ sein muss für $x \in (x_0 - \epsilon, x_0 + \epsilon)$, insbesondere, für $x := x_0$: $\partial_1 f(x_0, y_0) + \partial_2 f(x_0, y_0) \cdot y'(x_0) = 0$.
Falls $\partial_2 f(x_0, y_0) = 0$ wäre, so müsste somit auch $\partial_1 f(x_0, y_0) = 0$ sein. Wir haben somit die Bedigung $[\partial_2 f(x_0, y_0) = 0 \wedge \partial_1 f(x_0, y_0) = 0)]$ oder $\partial_2 f(x_0, y_0) \neq 0$. Der erste der beiden Fälle ist jedoch in den Beispielen (b) und (c) gerade dort erfüllt, wo offensichtlich *keine* lokale Auflosungsfunktion existiert.
Man kann also vermuten, dass die Bedingung $\partial_2 f(x_0, y_0) \neq 0$ ein Kriterium sein könnte für die Existenz der lokalen Auflösungsfunktion $x \mapsto y(x)$
(und analog, aus Symmetriegründen, $\partial_1 f(x_0, y_0) \neq 0$ für $y \mapsto x(y)$).

Teil 2:
Wir können die obige Vermutung mal am Beispiel $f(x, y) := x^2 + y^2 - 1$ testen.
Eine lokale Auflösungsfunktion $x \mapsto y(x)$ (nämlich $y(x) = \sqrt{1 - x^2}$ falls $y_0 > 0$ und $y(x) = -\sqrt{1 - x^2}$ falls $y_0 < 0$) existiert offensichtlich für alle Punkte (x_0, y_0) des Einheitskreises mit Ausnahme von $(1, 0), (-1, 0)$, also für alle $(x_0, y_0) \in$ Einheitskreis mit $y_0 \neq 0$.
Um obige Vermutung zu testen, berechnen wir:
$\partial_2 f(x, y) = 2y$, somit: $\partial_2 f(x_0, y_0) \neq 0 \quad \Leftrightarrow \quad x_0 \neq 0$, was zur Vermutung aus 1. passt!
Analog:
Eine lokale Auflösungsfunktion $y \mapsto x(y)$ (nämlich $x(y) = \sqrt{1 - y^2}$ falls $x_0 > 0$ und $x(y) = -\sqrt{1 - y^2}$ falls $x_0 < 0$) existiert offensichtlich für alle Punkte (x_0, y_0) des Einheitskreises mit Ausnahme von $(0, 1), (0, -1)$, d.h. für alle $(x_0, y_0) \in$ Einheitskreis mit $x_0 \neq 0$.
Um obige Vermutung zu testen, berechnen wir:
$\partial_1 f(x, y) = 2x$, somit: $\partial_1 f(x_0, y_0) \neq 0 \quad \Leftrightarrow \quad x_0 \neq 0$, was zur Vermutung passt!

In der Tat lautet der Satz über implizite Funktionen:

Satz (Satz über implizite Funktionen)
Sei $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^2$ offen, $f \in C^1(D)$, $(x_0, y_0) \in D$ mit $f(x_0, y_0) = 0$.
Es gelte ferner

$$\partial_2 f(x_0, y_0) \neq 0.$$

Dann existiert eine Umgebung $U_\epsilon = (x_0 - \epsilon, x_0 + \epsilon)$ von x_0 und genau eine Abbildung $x \mapsto y(x)$, $U \to \mathbb{R}$ (genannt lok. Auflösungsfunktion) derart, dass $f(x, y(x)) = 0$ und $y_0 = y(x_0)$ ist.

Die lokale Auflösungsfunktion ist stetig diff'bar, und es gilt $y'(x) = -\frac{\partial_1 f(x, y(x))}{\partial_2 f(x, y(x))}$.

Bemerkungen:

- Die Formel für $y'(x)$ ist offensichtlich, siehe Plausibilitätsüberlegung Teil 1.
- Obiger Satz betrachtet das "Auflösen nach y". Natürlich gibt es eine Variante für Auflösung nach dem anderen Argument, x:

Wenn $\partial_1 f(x_0, y_0) \neq 0$, dann existiert eine lok. Auflösungsfunktion $y \mapsto x(y)$ derart, dass $f(x(y), y) = 0$ und $x_0 = x(y_0)$, und es ist $x'(y) = -\frac{\partial_2 f(x(y),y)}{\partial_1 f(x(y),y)}$.

- Der obige Satz sagt nichts darüber aus, wie groß ϵ ist, oder wie man die Auflösungsfunktion findet.
  Jedoch: Ein Blick in den Beweis liefert ein *Iterationsverfahren* zur näherungsweisen numerischen Berechnung der Auflösungsfunktion, s.u.

Beweisidee für Satz über implizite Funktionen

Man zeigt, dass die rekursiv definierte Folge von Funktionen

$$y_{n+1}(x) := y_n(x) - \frac{f(x, y_n(x))}{\partial_2 f(x_0, y_0)}, \quad y_0(x) \equiv y_0,$$

auf einem Intervall $(x_0 - \epsilon, x_0 + \epsilon)$ konvergent ist.
Gegen welche Funktion $x \mapsto y(x)$?
Wende dazu $\lim\limits_{n\to\infty}$ auf die Rekursionsvorschrift an: Die Grenzfunktion $x \mapsto y(x)$ erfüllt offensichtlich $y(x) = y(x) - \frac{f(x,y(x))}{\partial_2 f(x_0,y_0)}$, also $f(x, y(x)) = 0$. Fertig!

Hilfsmittel zum Beweis der Konvergenz der obigen Funktionenfolge ist der sog. *Fixpunktsatz von Banach* (s. später, Kap. I.7). □

Bemerkung: Diese Iterationsvorschrift kann man benutzen, um die Auflösungsfunktion numerisch näherungsweise zu berechnen. Der Computer kann natürlich nicht für *alle* $x \in (x_0 - \epsilon, x_0 + \epsilon)$ die Berechnung durchführen; man nimmt stattdessen endlich viele Punkte $x_1, ..., x_n$ aus dem Intervall, und 'interpoliert' das Ergebnis z.B. stückweise linear; siehe Übung. Das $\epsilon > 0$ kann man raten/ausprobieren.

Beispiel für die Anwendung des Satzes über implizite Funktionen: Hat die Funktion $f(x, y) = x^3 y - 3xy^3$ an der Stelle
(i) $(x_0, y_0) = (1, 1)$
(ii) $(x_0, y_0) = (0, 0)$
(iii) $(x_0, y_0) = (3, 1)$
lokal eine Niveaulinie, die sich als $y = y(x)$ schreiben lässt?
Rechnung siehe Tafel.

Verallgemeinerung des Satzes auf mehr als 2 Argumente

Beispiele:

(a) Sei $f : \mathbb{R}^3 \to \mathbb{R}$, $\quad f(x, y, z) = 0$, z.B. $x^2 + y^2 + z^2 - 1 = 0$ beschreibt eine *"Fläche"* (Kugeloberfläche="Sphäre") im $\mathbb{R}^3$, eine sog. 2-dimensionale Mannigfaltigkeit. Frage nach lok. Auflösungsfunktionen $x = x(y, z)$, $y = y(x, z)$, $z = z(x, y)$?

(b) Sei $f : \mathbb{R}^3 \to \mathbb{R}^2$, $\begin{cases} f_1(x,y,z)=0 \\ f_2(x,y,z)=0 \end{cases}$, ist (meist) der Schnitt zweier Flächen, was 'meist' eine *Kurve* im $\mathbb{R}^3$ ist.
Bsp.: $f_1(x,y,z) := z+5-x^2-y^2$, $f_2(x,y,z) := z^2+y^2-1$.
Das zugehörige Gleichungssystem $\vec{f}(x,y,z) = \vec{0}$ hat 2 Gleichungen und 3 Unbekannte.
Kann man, für vorgegebenes x, die Gleichungen (ggf. lokal) nach y und z auflösen?
D.h. existieren Auflösungsfunktionen $y = y(x)$, $z = z(x)$, kurz
$(y,z) = (y(x), z(x))$?
Ebenso kann man nach lok. Auflösungsfunktionen
$(x,z) = (x(y), z(y))$, bzw.
$(x,y) = (x(z), y(z))$ fragen.

Allgemeiner:
Obige Beispiele verdeutlichen: Man löst immer nach so vielen Variablen auf, wie wir Gleichungen ' $f_i(\vec{x})=0$ ' haben, also welche Dimension der Bildbereichs von f hat. Im Fall $f : \mathbb{R}^{m+n} \to \mathbb{R}^m$ lösen wir also nach m Variablen auf.
Wir teilen die Komponenten des Argument (das sich im $\mathbb{R}^{m+n}$ befindet) auf in einen Vektor $\vec{y} \in \mathbb{R}^m$, nach dem wir auflösen wollen, und einen Vektor $\vec{x} \in \mathbb{R}^n$ (d.h. die Argumente von f werden ggf. umsortiert/umbenannt).

Satz (Satz über implizite Funktionen, mehrdimensionaler Fall)
Sei $\vec{f} : D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^{n+m}$ offen, $n, m \in \mathbb{N}$, $\vec{f} \in C^1(D)$, $(\vec{x}_0, \vec{y}_0) \in D$, wobei $\vec{x}_0 \in \mathbb{R}^n$, $\vec{y}_0 \in \mathbb{R}^m$, mit $\vec{f}(\vec{x}_0, \vec{y}_0) = \vec{0}$.
Falls die $m \times m$-Matrix $\frac{\partial \vec{f}}{\partial \vec{y}}(\vec{x}_0, \vec{y}_0)$ invertierbar ist, existiert eine Umgebung $U = K_\epsilon(\vec{x}_0) \subset \mathbb{R}^n$ von $\vec{x}_0$ und eine Auflösungsfunktion $\vec{x} \mapsto \vec{y}(\vec{x}), U \to \mathbb{R}^m$, d.h. $\vec{f}(\vec{x}, \vec{y}(\vec{x})) = \vec{0} \; \forall \, \vec{x} \in U$, $\vec{y}(\vec{x}_0) = \vec{y}$. Die Auflösungsfunktion ist stetig diff'bar und es gilt $J\vec{y}(\vec{x}) = \frac{d\vec{y}}{d\vec{x}}(\vec{x}) = -\left[\frac{\partial \vec{f}}{\partial \vec{y}}(\vec{x}, \vec{y}(\vec{x}))\right]^{-1} \frac{\partial \vec{f}}{\partial \vec{x}}(\vec{x}, \vec{y}(\vec{x}))$.

Bemerkung: die Iterationsvorschrift wird im Mehrdimensionalen zu

$$\vec{y}_{n+1}(\vec{x}) := \vec{y}_n(\vec{x}) - \left[\frac{\partial \vec{f}}{\partial \vec{y}}(\vec{x}_0, \vec{y}_0)\right]^{-1} \vec{f}(\vec{x}, \vec{y}_n(\vec{x})), \qquad \vec{y}_0(\vec{x}) \equiv \vec{y}_0,$$

Eine Anwendung/Folgerung des Satzes über Implikationen:
Der Satz von der inversen Abbildung

Wir hatten bereits zu Beginn des Kapitels festgehalten, dass ein Kriterium, welches die Existenz von Auflösungsfunktionen sichert, verwendet werden kann, um die Existenz einer lokalen Umkehrfunktion zu sichern. Hier die Überlegung im Mehrdimensionalen:

Sei $\vec{f}: \mathbb{R}^n \to \mathbb{R}^n$ stetig diff'bar und $\vec{x}_0 \in \mathbb{R}^n$.
Frage: Gibt es eine Umgebung U von $\vec{x}_0$, so dass $\vec{f}|_U : U \to f(U)$ eine Umkehrfunktion hat?

Zurückführung auf den Satz über implizite Funktionen:
Frage ist, ob sich die Gleichung $\vec{f}(\vec{x}) = \vec{y}$, bzw. $\vec{y} - \vec{f}(\vec{x}) = \vec{0}$, nach $\vec{x}$ auflösen lässt. Setze also $\tilde{f}: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, $\vec{\tilde{f}}(\vec{x}, \vec{y}) := \vec{y} - \vec{f}(\vec{x})$ und frage, ob die Gleichung $\tilde{f}(\vec{x}, \vec{y}) = \vec{0}$ eine Auflösungsfunktion $\vec{x} = \vec{x}(\vec{y})$ hat. Nach dem Satz über implizite Funktionen brauchen wir dazu, dass die Matrix $\frac{\partial \tilde{f}(\vec{x}, \vec{y})}{\partial \vec{x}}(\vec{x}_0, \vec{y}_0)$ invertierbar ist, $\vec{y}_0 := f(\vec{x}_0)$.
Nach Def. von $\vec{\tilde{f}}$ ist diese Matrix (bis aufs Vorzeichen) nichts anderes als die Jacobi-Matrix von $\vec{f}$ an der Stelle $\vec{x}_0$. Wir erhalten somit:

Satz (von der inversen Abbildung)
Sei $\vec{f}: D \to \mathbb{R}^n$, $D \subseteq \mathbb{R}^n$ offen, $\vec{f} \in C^1(D)$, $\vec{x}_0 \in D$.
Es sei die Jacobi-Matrix $Jf(\vec{x}_0) = \frac{d\vec{f}}{d\vec{x}}(\vec{x}_0)$ invertierbar.
Dann gibt es eine Umgebung $U \subseteq D$ von $\vec{x}_0$, so dass $\vec{f}|_U : U \to \vec{f}(U)$ umkehrbar (d.h. bijektiv) ist.

Erinnerung: $\frac{d\vec{f}}{d\vec{x}}(\vec{x}_0)$ invertierbar $\iff$ $\det \frac{d\vec{f}}{d\vec{x}}(\vec{x}_0) \neq 0$

Spezialfall $n = 1$: Das Kriterium lautet dann $f'(x_0) \neq 0$.

Das leuchtet ein, denn:

Für $f \in C^1$ ist mit $f'(x_0) \neq 0$ auch die Existenz einer Umgebung von x_0 gesichert, auf der $f'(x) \neq 0$ ist.
Daraus folgt, dass f streng monoton auf dieser Umgebung von x_0 ist, und somit dort injektiv. Die Surjektivität ergibt sich trivialerweise für $f: U \to f(U)$.

Zur *Konstruktion* der Umkehrfunktion:

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

Iterationsverfahren zur (näherungsweisen) Berechnung einer lok. Umkehrfunktion

$$\begin{aligned} \vec{x}_{n+1}(\vec{y}) &:= \vec{x}_n(\vec{y}) - \left[\frac{\partial \vec{\tilde{f}}}{\partial \vec{x}}(\vec{x}_0, \vec{y}_0)\right]^{-1} \vec{\tilde{f}}(\vec{x}_n(\vec{y}), \vec{y}) \\ &= \vec{x}_n(\vec{y}) + \left[J\vec{f}(\vec{x}_0, \vec{y}_0)\right]^{-1} (\vec{y} - \vec{f}(\vec{x}_n(\vec{y})), \qquad \vec{x}_0(\vec{y}) \equiv \vec{x}_0, \end{aligned}$$

Bemerkung zur Existenz einer *globalen* Umkehrfunktion:
Wenn die Bedingung, dass die Jacobi-Matrix von $\vec{f}$ an *jeder* Stelle $\vec{x}$ invertierbar ist, dann gibt es **um jeden Punkt** $\vec{x}$ eine **lokale** Umkehrfunktion.
Man könnte vermuten, dass dies hinreichend ist für die Existenz einer **globalen** Umkehrfunktion von $\vec{f} : D \to \vec{f}(D)$ (etwa durch "Zusammenstückeln" der lokalen Umkehrfunktionen).
Das ist jedoch **nicht** der Fall; man kann Beispiele finden, in denen, obwohl in jedem Punkt eine lokale Umkehrfunktion existiert, dennoch keine globale Umkehrfunktion existiert:

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \quad f(x,y) = \begin{pmatrix} e^x \cos y \\ e^x \sin y \end{pmatrix}$$

erfüllt die Voraussetzungen des Satzes von der inversen Abbildung an *jeder* Stelle $(x_0, y_0) \in \mathbb{R}^2$, hat also an jeder Stelle lokal eine Umkehrfunktion, $f : \mathbb{R}^2 \to \mathbb{R}^2$ ist jedoch (offensichtlich!) nicht injektiv, hat somit keine globale Umkehrfunktion. (siehe Übung?)

Beachte, dass dies im *Eindimensionalen* nicht passieren kann: Eine C^1-Funktion $f : (a,b) \to \mathbb{R}$, deren Ableitung überall $\neq 0$ ist, ist (global) streng monoton, und hat deswegen als Abb. $f : (a,b) \to f(a,b)$ immer eine globale Umkehrfunktion.

1.4 Parameterdarstellungen von Kurven, Kurvenintegrale, Flächen

Motivation:

Wir haben in Kap. I.3 gesehen:

Die Darstellung von Kurven (im $\mathbb{R}^2$) in der Form $y = y(x)$ oder $x = x(y)$

- kann meist nur *abschnittsweise (=lokal)* geschehen
- kann schon in recht einfachen Fällen nicht mehr explizit angegeben, sondern nur näherungsweise (iterativ) berechnet werden

Ausweg: Parameterdarstellung ("Parametrisierung") der Kurve

Die Parameterdarstellung einer Kurve hat ferner den Vorteil, diverse Berechnungen, z.B. der (Bogen-)Länge einer Kurve, einfach zu ermöglichen.

Def. (Parameterdarstellung einer Kurve)

Eine stetige Abbildung $\vec{\gamma} : I \to \mathbb{R}^n$, wobei $I \subseteq \mathbb{R}$ ein abgeschlossenes Intervall ist, heißt *Parameterdarstellung* der Kurve $\Gamma := \vec{\gamma}(I) \subset \mathbb{R}^n$.

Ist $\vec{\gamma}$ stetig diff'bar, so so bezeichnet man die Kurve als *C^1-Kurve*.

Falls $I = [a, b]$ und $\vec{\gamma}(a) = \vec{\gamma}(b)$, so bezeichnet man die Kurve als *geschlossen*.

Skizze: siehe Tafel

Beispiel: Kreis: $\vec{\gamma} : [0, 2\pi] \to \mathbb{R}^2, \quad \vec{\gamma}(t) = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$

Weitere Beispiele: Gerade, Schraubenlinie, Zykloide, Graph einer reellen Funktion: s. Tafel

Berechnung der Tangentenrichtung an einem Punkt $\vec{\gamma}(t_0)$:
Betrachte zunächst *Sekante* durch die Kurvenpunkte $\vec{\gamma}(t_0)$ und einem $\vec{\gamma}(t)$:
Diese hat die Richtung $\vec{\gamma}(t) - \vec{\gamma}(t_0)$ bzw. $\frac{\vec{\gamma}(t) - \vec{\gamma}(t_0)}{t - t_0}$
Im Grenzwert $t \to t_0$ bekommen wir die Richtung der Tangente (s. Skizze) als

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

falls $\vec{\gamma}$ diff'bar. Also:

Satz (Tangente an eine Kurve in Parameterdarstellung)

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

Weiter zur Motivation/Deutung/Anwendung von Parameterdarstellungen:

Die Parametrisierung einer Kurve kann als **Beschreibung der Bewegung eines punktförmigen Objekts im physikalischen Raum** aufgefasst werden:

t ist die Zeit, $\vec{\gamma}(t)$ der Ort zur Zeit t, und $\vec{\gamma}'(t) = \frac{d\vec{\gamma}}{dt}(t)$ ist die Geschwindigkeit.

$\vec{\gamma}''(t)$ ist dann übrigens die Beschleunigung, d.h. das zweite Newton'sche Gesetz "Kraft gleich Masse mal Beschleunigung", das die Bewegung eines Körpers mit Masse m, der der Kraft $\vec{F}$ ausgesetzt ist, beschreibt, lautet

$$m\,\vec{\gamma}''(t) = \vec{F}(\vec{\gamma}(t))$$

(Hier wurde angenommen, dass die Kraft eine Funktion des Ortes ist (wie z.B. Gravitationskraft oder elektrostatische Kraft); manchmal ist die Kraft auch eine Funktion der Geschwindigkeit (z.B. Reibungskraft in einem Fluid), d.h. $\vec{F}(\vec{\gamma}\,'(t))$.)

Umgekehrt liefert die Formel für den Fall, dass ein Objekt zu einer Bewegung $\vec{\gamma}$ *gezwungen* wird ($\to$"Achterbahnfahrt") die sich ergebenden Trägheitskräfte.

"Glattheit" von Kurven

Frage: Wann kann man sicher sein, dass eine Kurve keine "Knicke" hat?
Erste Vermutung: Diff'barkeit oder C^1-Eigenschaft ist hinreichend?

Dazu ein **Gegenbeispiel**: Die *Neil'sche Parabel* $\vec{\gamma}(t) = \begin{pmatrix} t^2 \\ t^3 \end{pmatrix}$ hat die C^1-Eigenschaft (sogar C^∞), hat aber an der Stelle $\vec{\gamma}(0) = \vec{0}$ einen Knick.
Veranschaulichung: Dies ist möglich, da $\vec{\gamma}'(0) = \vec{0}$ ist;
Erläuterungen/Skizze s. Tafel.
Anschaulich: Um abrupt die Richtung zu wechseln, ist es nicht zwingend erforderlich, abrupte 'Lenkbewegungen' zu machen; man kann die Geschwindigkeit auf $\vec{0}$ abbremsen und dann moderate ('stetige') Lenkbewegungen machen.
Dies führt auf die Begriffsbildung:

Def. (Regularität einer Kurve)
Eine Parameterdarstellung $\vec{\gamma} : I \to \mathbb{R}^n$ heißt *glatt* oder *regulär*, wenn $\vec{\gamma} \in C^1(I)$ und $\vec{\gamma}'(t) \neq \vec{0}\ \forall\, t \in I$.

Beachte: Die obige Forderung sichert außerdem, dass Kurvenstücke nicht mittels 'Wenden auf der Stelle' mehrfach durchlaufen werden.
Beispiel: $\vec{\gamma}(t) := \begin{pmatrix} t^3 - t \\ t^3 - t \end{pmatrix}$, $t \in [-2, 2]$, ist nicht regulär. Hier wird ein Kurvenstück mehrfach durchlaufen.

Zur fehlenden Eindeutigkeit von Parameterdarstellungen, "Umparametrisierungen":

Es ist offensichtlich, dass Parameterdarstellungen einer Kurve nicht eindeutig sind.
So wird der Einheitskreis nicht nur durch

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

parametrisieriert, sondern z.B. auch durch

$$\tilde{\vec{\gamma}}(t) := \begin{pmatrix} \cos 5t \\ \sin 5t \end{pmatrix}, \qquad t \in \tilde{I} = [0, \tfrac{2\pi}{5}]$$

$$\tilde{\vec{\gamma}}(t) := \begin{pmatrix} \cos(t^2) \\ \sin(t^2) \end{pmatrix}, \qquad t \in \tilde{I} = [0, \sqrt{2\pi}]$$

$$\tilde{\vec{\gamma}}(t) := \begin{pmatrix} \cos(t+1) \\ \sin(t+1) \end{pmatrix}, \qquad t \in \tilde{I} = [-1, 2\pi - 1]$$

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

Was haben diese $\vec{\tilde{\gamma}}$ mit $\vec{\gamma}$ gemeinsam?
Sie lassen sich als $\vec{\tilde{\gamma}} = \vec{\gamma} \circ u$ schreiben, wobei $u : \tilde{I} \to I$.

Einen solchen Wechsel der Parametrisierung bezeichnet man als *Umparametrisierung*.

z.Ü.: Ermittle in den obigen 4 Umparametrisierungen das verwendete u.

Satz (Umparametrisierung)
Ist $\vec{\gamma} : I \to \mathbb{R}^n$ Parametrisierung einer Kurve, und $u : \tilde{I} \to I$ stetig und surjektiv, $\tilde{I} \subset \mathbb{R}$ ein Intervall, so ist $\vec{\gamma} \circ u : \tilde{I} \to \mathbb{R}^n$ ebenfalls eine Parametrisierung dieser Kurve.
Umgekehrt hat *jede* Parametrisierung der Kurve die Form $\vec{\gamma} \circ u$, wobei $u : \tilde{I} \to I$ stetig und surjektiv, $\tilde{I} \subset \mathbb{R}$ ein Intervall.
Umparametrisierung von *regulären* Kurven erfordert, dass $u \in C^1$ und $u'(t) \neq 0 \, \forall \, t \in \tilde{I}$ (wegen $\vec{\tilde{\gamma}}'(t) = \vec{\gamma}'(u(t))\, u'(t)$).
Somit ist u dann global entweder streng monoton wachsend oder streng monoton fallend. Im zweiten Fall ändert sich bei Umparametrisierung der Durchlaufsinn der Kurve.

Die Bogenlänge einer Kurve

Sei Parametrisierung $\vec{\gamma} : [a, b] \to \mathbb{R}^n$ gegeben.
Idee: Wähle *Zerlegung* $Z = \{t_0, t_1, ..., t_{m(Z)}\}$ des Intervalls $[a, b]$, also $a = t_0 < t_1 < ... < t_{m(Z)} = b$; die Feinheit der Zerlegung ist $|Z| := \max\limits_{i=1,..,m} t_i - t_{i-1}$.

Gradlinie Verbindung der Punkte $\vec{\gamma}(t_i)$ ergibt einen *Polygonzug*.
Die Länge des Polygonzugs

- ist leicht zu berechnen als $L_Z = \sum\limits_{i=1}^{m(Z)} \|\vec{\gamma}(t_i) - \vec{\gamma}(t_{i-1})\|$
  (wobei $\| \cdot \|$ die Euklidische Norm ist)
- ist eine Näherung für die Bogenlänge der Kurve

Mittels $|Z| \to 0$ ist zu erwarten, dass die Länge des Polygonzugs gegen die Bogenlänge konvergiert. Wir definieren:

Def. (Bogenlänge)
Sei $\vec{\gamma} : [a, b] \to \mathbb{R}^n$ Parametrisierung einer Kurve.
Gilt für jede Folge von Zerlegungen Z_k mit $|Z_k| \to 0$, dass die Folge der Längen der Polygonzüge L_{Z_k} konvergent ist, und zwar immer gegen den gleichen Wert, dann wird dieser Grenzwert

$$|\Gamma| := \lim_{|Z_k| \to 0} L_{Z_k}$$

als *Bogenlänge der Kurve* Γ bezeichnet. In dem Fall nennt man die Kurve *rektifizierbar*.

Herleitung einer leicht zu handhabenden Formel für die Bogenlänge von regulären Kurven:

Rechnung siehe Tafel.

Ergebnis:

Satz (Bogenlänge einer Kurve)
Sei $\vec{\gamma} : I \to \mathbb{R}^n$ eine Parametrisierung einer regulären Kurve Γ. Dann berechnet sich die Bogenlänge $|\Gamma|$ als

$$|\Gamma| = \int_I \|\vec{\gamma}\,'(t)\|\, dt.$$

Bemerkung: Die Bogenlänge von *stückweise* regulären Kurven kann man durch "Zusammenstückeln" berechnen.

Beispiel: Berechnung der Bogenlänge der Zykloide $\vec{\gamma}(t) := r \begin{pmatrix} t - \sin t \\ 1 - \cos t \end{pmatrix}$, $t \in [0, 2\pi]$.

(Bonaventura Cavalieri 1629, Blaise Pascal 1658)
Rechnung siehe Tafel.

Parametrisierung nach der Bogenlänge

Aufgabe: Gegeben sei eine Parametrisierung $\vec{\gamma} : I \to \mathbb{R}^n$ einer Kurve.
Gesucht sei eine Parametrisierung $\vec{\gamma}_0 : \tilde{I} \to \mathbb{R}^n$ dieser Kurve mit der Eigenschaft

$$\|\vec{\gamma}_0'(t)\| = 1 \quad \forall t \in \tilde{I} \qquad (*)$$

Diese spezielle Parametrisierung der Kurve heißt *Parametrisierung nach der Bogenlänge*.

Grund für diese Bezeichnung: Für beliebige Kurvenpunkte $\vec{\gamma}(t_1), \vec{\gamma}(t_2)$ gilt nach der Bogenlängenformel

$$\int_{t_1}^{t_2} \underbrace{\|\vec{\gamma}_0'(t)\|}_{=1}\, dt = t_2 - t_1,$$

d.h. $t_2 - t_1$ gibt genau den Abstand von $\vec{\gamma}_0(t_1), \vec{\gamma}_0(t_2)$ 'entlang der Kurve' an.

Wie findet man $\vec{\gamma}_0$, wenn lediglich irgend eine Parametrisierung $\vec{\gamma}$ gegeben ist?
Wir wissen (s. Satz über Umparametrisierung), dass es eine Darstellung $\vec{\gamma}_0 = \vec{\gamma} \circ u$ mit $u : \tilde{I} \to I$ gibt. Unsere Forderung (*) ergibt somit eine Forderung an u:

$$1 \overset{!}{=} \|\vec{\gamma}_0'(t)\| = \|(\vec{\gamma} \circ u)'(t)\| = \|(\vec{\gamma}\,'(u(t)) \cdot u'(t)\| = |u'(t)|\, \|\vec{\gamma}\,'(u(t)\|,$$

Falls wir zusätzlich die Forderung stellen, dass sich der Durchlaufsinn bei der Umparametrisierung nicht ändern soll, also $u'(t) > 0$, so stellt dies, da $\vec{\gamma}\,'$ bekannt ist, eine

Differentialgleichung für die Funktion u dar:

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

Bemerkung: Leider kann man diese Differentialgleichung nur in seltenen Fällen exakt lösen.
(Lösungsverfahren für Differentialgleichungen, auch Näherungsverfahren, siehe Kap. 2)

Mögliche Anwendung der Parametrisierung nach der Bogenlänge:

Ist die Parametrisierung nach der Bogenlänge bekannt, so können einige Dinge besonders leicht berechnet werden, z.B.:

- Der Abstand von Punkten der Kurve entlang der Kurve, s.o.
- Man kann die *Krümmung* der Kurve im Punkt $\vec{\gamma}_0(t)$ leicht berechnen:
  Die *Krümmung* κ ist ist definiert als der Kehrwert des Radius R des sich an die Kurve anschmiegenden Kreises, und bei Parametrisierung nach der Bogenlänge gilt (o.Bew.) die einfache Formel

  $$\kappa(t) = \frac{1}{R(t)} = \|\vec{\gamma}_0\,''(t)\|.$$

  z.Ü.: Prüfen Sie obige Formel an folgendem Beispiel: Zeigen Sie, dass der Kreis mit Radius $r \in \mathbb{R}^+$, nach der Bogenlänge parametrisiert, also $\vec{\gamma}_0(t) = (r\cos(t/r), r\sin(t/r))^T$, tatsächlich den Krümmungsradius $R(t) = \text{const} = r$ hat.

  Man kann diese auf der Bogenlängenparametrisierung beruhende Formel auch umrechnen auf *andere* Parametrisierungen (s. Übung); sie wird dann jedoch deutlich komplizierter.

Kurvenintegrale: Das Kurvenintegral erster Art

Motivation:

Aufgabe: Betrachtet werde ein Draht oder Faden mit variabler Stärke. Der Draht/Faden wird als Kurve mit Parametrisierung $\vec{\gamma}$ betrachtet; die Dichte (in Masse pro Strecke(!)) am Punkt $\vec{\gamma}(t)$ sei $f(\vec{\gamma}(t))$.
Wie berechnet man die Gesamtmasse?

Idee: Verwende, wie bei der Berechnung der Bogenlänge, wieder den Polygonzug:
Siehe Tafel.

Man erhält folgende Formel:

Def. (Kurvenintegral erster Art)

Sei $\Gamma \in \mathbb{R}^n$ eine reguläre Kurve mit Parametrisierung $\vec{\gamma} : I \to \mathbb{R}^n$, und sei $f : \Gamma \to \mathbb{R}$ (manchmal auch $f : \mathbb{R}^n \to \mathbb{R}$ oder $f : D \to \mathbb{R}$ mit $\Gamma \subset D \subset \mathbb{R}^n$) eine (skalarwertige!) Funktion. Dann ist

$$\underbrace{\int_\Gamma f \, ds}_{\text{Symbol}} := \underbrace{\int_I f(\vec{\gamma}(t)) \, \|\vec{\gamma}'(t)\| \, dt}_{\text{so rechnet man's aus}}$$

das *Kurvenintegral erster Art* von f über Γ (falls das Integral existiert).

Bemerkung: Im Fall $f \equiv 1$ bekommt man die Bogenlänge der Kurve; das Bogenlängenintegral ist also ein spezielles Kurvenintegral, bzw. das Kurvenintegral kann als Verallgemeinerung der Bogenlängenberechnung ("mit Gewichtsfunktion f") betrachtet werden.
Die obige Schreibweise als $\int_\Gamma f \, ds$ suggeriert, dass der Wert des Kurvenintegrals von f über Γ unabhängig von der gewählten Parametrisierung ist.
Dies ist jedoch nicht ganz offensichtlich; immerhin kommt in der rechten Seite der Definition die Parametrisierung durchaus vor!

Wir sollten also zeigen, dass der Wert des Kurvenintegrals unabhängig von der gewählten Parametrisierung ist.
Das geschieht, indem wir zwei Parametrisierungen $\vec{\gamma}_1, \vec{\gamma}_2$ betrachten, ausnutzen, dass es ein u gibt mit $\vec{\gamma}_1 = \vec{\gamma}_2 \circ u$, und die Substitutionsregel für Integrale verwenden.

Siehe Tafel.

Das Kurvenintegral zweiter Art

Hier geht es, im Gegensatz zum Kurvenintegral erster Art, um *vektorwertige* Funktionen.

Motivation (aus der Physik):

Es wird eine Masse m durch ein Kraftfeld $\vec{x} \mapsto \vec{F}(\vec{x}), \mathbb{R}^3 \to \mathbb{R}^3$ (z.B. Gravitationsfeld) bewegt entlang einer Kurve Γ mit Parametrisierung $\vec{\gamma}$.
(Analog: Eine elektrische Punktladung wird durch ein elektrostatisches Feld bewegt)
Frage: Welche Arbeit wird geleistet (welche Energie wird frei bzw. aufgewendet)?

Erinnerung an sie Schul-Physik: "Arbeit ist Kraft mal Weg"

Jedoch Vorsicht: Kraft und Weg sind i.a. Vektoren, und sobald sie nicht in genau die gleiche Richtung zeigen, gilt: Zur Berechnung der Arbeit darf nur der in Richtung des Weges gerichtete Anteil der Kraft herangezogen werden (eine Bewegung orthogonal zur angreifenden Kraft erfordert keine Energie).

Der in Richtung des Weges, also in Richtung $\vec{\gamma}\,'(t)$ gerichtete Anteil von $\vec{F}(\vec{\gamma}(t))$ ist $\|\vec{F}(\vec{\gamma}(t))\| \cdot \cos(\text{Winkel}(\vec{\gamma}\,'(t), \vec{F}(\vec{\gamma}(t)))) = \frac{\langle \vec{F}(\vec{\gamma}(t)), \vec{\gamma}\,'(t)\rangle}{\|\vec{\gamma}\,'(t)\|}$.

Fazit: Wir können die gesuchte Arbeit als Kurvenintegral erster(!) Art über die skalarwertige Funktion $t \mapsto \frac{\langle \vec{F}(\vec{\gamma}(t)), \vec{\gamma}\,'(t)\rangle}{\|\vec{\gamma}\,'(t)\|}$ berechnen!

Wir erhalten:

$$W = \int_\Gamma \frac{\langle \vec{F}(\vec{\gamma}(\cdot)), \vec{\gamma}\,'(\cdot)\rangle}{\|\vec{\gamma}\,'(\cdot)\|}\, ds = \int_I \frac{\langle \vec{F}(\vec{\gamma}(t)), \vec{\gamma}\,'(t)\rangle}{\|\vec{\gamma}\,'(t)\|} \cdot \|\vec{\gamma}\,'(t)\|\, dt = \int_I \langle \vec{F}(\vec{\gamma}(t)), \vec{\gamma}\,'(t)\rangle\, dt$$

Dies motiviert die folgende Definition:

Def. (Kurvenintegral zweiter Art)

Sei Γ eine reguläre Kurve mit Parametrisierung $\vec{\gamma} : I \to \mathbb{R}^n$, und sei $\vec{F} : \Gamma \to \mathbb{R}^n$ (manchmal auch $\vec{F} : \mathbb{R}^n \to \mathbb{R}^n$ oder $\vec{F} : D \to \mathbb{R}^n$ mit $\Gamma \subset D \subset \mathbb{R}^n$) eine (vektorwertige!) Funktion. Dann ist

$$\int_\Gamma \vec{F} \bullet d\vec{s} := \int_I \langle \vec{F}(\vec{\gamma}(t)), \vec{\gamma}\,'(t)\rangle\, dt$$

das *Kurvenintegral zweiter Art* von $\vec{F}$ über Γ (falls das Integral existiert).

Bemerkungen:

- s.o.: Das Kurvenintegral zweiter Art einer (vektorwertigen) Funktion kann als K'integral erster Art der in Kurvenrichtung gerichteten Komponente der Funktion aufgefasst werden (s.S. 28).
- Das Kurvenintegral zweiter Art ändert sein Vorzeichen, wenn die Kurve in umgekehrter Richtung durchlaufen wird.
  Abgesehen davon ist sein Wert unabhängig von der Wahl der Parametrisierung.

Parametrisierungen von Flächen und Oberflächenintegrale

Erinnerung: Parametrisierung von Kurven $\Gamma \subset \mathbb{R}^n$ erfordert *einen* Parameter.
Zur **Parametrisierung von Flächen** $F \subset \mathbb{R}^n$ braucht man *zwei* Parameter.
Die Parametrisierung ist eine surjektive glatte Abbildung der Form

$$\vec{\gamma} : M \subset \mathbb{R}^2 \to F \subset \mathbb{R}^n, \quad (s,t) \mapsto \vec{\gamma}(s,t)$$

Beispiel: Zur Parametrisierung der Einheitssphäre (=Oberfläche der Einheitskugel) $F = \{(x,y,z) \in \mathbb{R}^3 \mid x^2+y^2+z^2 = 1\}$ im $\mathbb{R}^3$ kann man verwenden

$$\vec{\gamma}(s,t) := \begin{pmatrix} \cos s \cos t \\ \cos s \sin t \\ \sin s \end{pmatrix}, \quad (s,t) \in [-\frac{\pi}{2}, \frac{\pi}{2}] \times [0, 2\pi] =: M$$

(s entspricht der geografischen Breite, t der geografischen Länge; t kann auch aus $[-\pi, \pi]$ genommen werden anstelle von $[0, 2\pi]$.)

Wir beschränken uns von nun an auf den Fall $n=3$.

Man definiert:

Oberflächenintegral erster Art (für skalarwertige Funktionen f):

$$\int_F f \; do := \int_M f(\vec{\gamma}(s,t)) \, \|\vec{\partial_1}\gamma(s,t) \times \vec{\partial_2}\gamma(s,t)\| \, ds \, dt$$

Anwendung:

- z.B. f als Dichte eines Blechs; dann ist $\int_F f \; do$ die Gesamtmasse
- Insbesondere bekommt man für $f \equiv 1$ den *Flächeninhalt* $|F|$!
  (das Analogon für Kurven: Bogenlänge)
- f als lok. Temperatur auf der Erdoberfläche; $\int_F f \; do \; / \int_F 1 \; do$ ist dann die *mittlere* Temperatur

Motivation für den Faktor $\|\vec{\partial_1}\gamma(s,t) \times \vec{\partial_2}\gamma(s,t)\|$: ggf. s. Tafel

Das Integral über M "$ds\,dt$" ist als sog. Doppelintegral gemeint; s. Tafel.

z.Ü.: Unter Verwendung des Oberflächenintegrals

(a) Berechne den Flächeninhalt des Graphen einer Funktion $u : [a,b] \times [c,d] \to \mathbb{R}$

(b) Berechne die Oberfläche der Einheitskugel

(c) Berechne die mittlere Temperatur $\bar{T}$, falls die lokale Temperatur auf der Kugeloberfläche gegeben ist durch $T(s,t) := T_0 + T_1 \cos s \cos^2 t$.

Für vektorwertige(!) Funktionen $\vec{f}$, die auf einer Oberfläche $F \subset \mathbb{R}^3$ 'leben', kommt es in der Physik häufig vor, dass man das Oberflächenintegral erster Art derjenigen Komponente von $\vec{f}$, die senkrecht zur Fläche steht, benötigt, also über $\frac{\langle \vec{f}(\vec{\gamma}(s,t)), \vec{n} \rangle}{\|\vec{n}(s,t)\|}$. Dabei ergibt sich ein Normalenvektor $\vec{n}(s,t)$ als Kreuzprodukt von zwei Tangentialvektoren. Tangentialvektoren sind $\partial_1\vec{\gamma}(s,t)$ und $\partial_2\vec{\gamma}(s,t)$, also z.B. $\vec{n}(s,t) := \partial_1\vec{\gamma}(s,t) \times \partial_2\vec{\gamma}(s,t)$. Verwendet man die Formel für Oberflächenintegrale erster Art, so kürzt sich $\|\partial_1\vec{\gamma}(s,t) \times \partial_2\vec{\gamma}(s,t)\|$ heraus und man erhält die Formel für das **Oberflächenintegral zweiter Art** (für vektorwertige Funktionen):

$$\int_F \vec{f} \bullet \vec{do} := \int_M \langle \vec{f}(\vec{\gamma}(s,t)) \, , \, \vec{\partial_1}\gamma(s,t) \times \vec{\partial_2}\gamma(s,t) \rangle \, ds \, dt$$

Anwendungen in der Physik:

- $\vec{f}$ elektrische Feldstärke und F 'geschlossene' Fläche, dann ist $\int_F \vec{f} \bullet \vec{do}$ ein Maß für die eingeschlossene Ladungsmenge (Gaußsches Gesetz, evtl. aus der Schulphysik bekannt?)
  Ein analoges Gesetz gilt für Gravitationskraft und Masse: Misst und integriert man auf einer geschlossenen Fläche die zur Fläche senkrechte Komponente der Gravitationskraft, dann ist der Wert dieses Oberflächenintegrals ein Mass für die von der Fläche umschlossenen Masse!
- $\vec{f}$ sei die "Flussdichte" eines Fluids (Masse pro Zeit pro Fläche), erhältlich als $\vec{f} = \rho \cdot \vec{v}$, wobei ρ die Dichte (Masse pro Volumen) und $\vec{v}$ die Fluidgeschwindigkeit ist. Dann ist $\int_F \vec{f} \bullet \vec{do}$ die pro Zeit durch die Fläche F fließende Fluidmasse.
- $\vec{f}$ eine Strahlungsdichte, dann ist $\int\limits_F \vec{f} \bullet \vec{do}$ die einfallende Gesamtstrahlung

1.5 Konvexe, quadratische, linear-quadratische Optimierungsprobleme

Motivation:

Wir haben in Kap. I.1 notwendige und hinreichende Kriterien für *lokale* Extremstellen kennengelernt für Funktionen $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$.
(Bisher ist es uns nur im Falle von *kompakten Definitionsbereichen* D gelungen, daraus auf einfache Art und Weise auf *globale* Extrema zu schließen, siehe auch Kap. I.2.) In diesem Kapitel wollen wir gewisse Annahmen an D, f stellen (insbes. "Konvexität"), die gewisse Aussagen über Extremstellen erlauben, z.B. *Globalität* oder *Eindeutigkeit* von lokalen Extremstellen.

Def. (konvexe Menge)
Eine Menge $M \subseteq \mathbb{R}^n$ heißt *konvex*, falls mit beliebigen $\vec{x}, \vec{y} \in M$ auch die Verbindungsstrecke zwischen $\vec{x}$ und $\vec{y}$ in M liegt:

$$\forall\, \vec{x}, \vec{y} \in M \; \forall\, \alpha \in (0,1) : \quad \alpha\, \vec{x} + (1-\alpha)\, \vec{y} \in M$$

Skizze: s. Tafel

Def. (konvexe Funktion)
Sei $f : D \to \mathbb{R}$, wobei $\emptyset \neq D \subseteq \mathbb{R}^n$ eine konvexe(!) Menge sei.
f heißt *konvex*, falls

$$\forall\, \vec{x}, \vec{y} \in D \; \forall\, \alpha \in (0,1) : \quad f(\alpha\, \vec{x} + (1-\alpha)\, \vec{y}) \leq \alpha\, f(\vec{x}) + (1-\alpha)\, f(\vec{y})$$

Gilt diese Aussage sogar mit "$<$" anstelle von "$\leq$", dann heißt f *strikt konvex*.

Skizze: s. Tafel

Bemerkungen:

- (Affin-)Lineare Funktionen $\vec{x} \mapsto f(\vec{x}) := c_1x_1 + ... + c_nx_n\,(+k)$ sind konvex, aber nicht strikt konvex.
- Die Definition der Konvexität von Funktionen erfordert keine Diff'barkeitseigenschaften von f.
  Falls f aber 2mal stetig diff'bar ist, gilt folgendes Kriterium für Konvexität:

Satz (Kriterium für Konvexität von Funktionen)
Sei $D \subseteq \mathbb{R}^n$ konvexe Menge, und sei $f \in C^2(D)$. Dann gilt:

1. f konvex $\iff$ $Hf(\vec{x})$ positiv semidefinit $\forall\, \vec{x} \in D$
2. f strikt konvex $\impliedby$ $Hf(\vec{x})$ positiv definit $\forall\, \vec{x} \in D$
3. Die Umkehrung "$\Rightarrow$" in 2. gilt nicht (s.u.).

Der Beweis von 1.-2. kann in *Rockafellar, Convex Analysis*, S. 26-27, gefunden werden.

Als Beleg für Aussage 3. betrachte das Beispiel $f(x) = x^4$: f ist strikt konvex, hat aber $Hf(0) = f''(0) = 0$.

Konvexität ist nützlich, wenn es um die Charakterisierung der Menge der lok./glob. Minimalstellen geht:

Satz (lokale Minimalstellen konvexer Funktionen)
Sei f konvex.

a) Dann ist jede *lokale* Minimalstelle ist immer auch *globale* Minimalstelle.

b) Insbesondere haben alle lokalen Minima den gleichen Wert.

c) Dann bilden alle lok. Minimalstellen von f eine zusammenhängende, sogar eine konvexe, Menge.

Skizzen: s. Tafel.

Vorsicht: Die Menge aller lok. Min.-stellen in obigem Satz kann *leer* sein!
Nicht jede konvexe (oder strikt konvexe) Funktion hat überhaupt lok./glob. Min.-stelle(n);
siehe Beispiel $f(x)=e^x, \mathbb{R} \to \mathbb{R}$ und die nachfolgenden Sätze.

Beweis des Satzes: s. Tafel.

Satz (Eindeutigkeit des Minimums)
Sei f strikt(!) konvex.
Dann hat f *höchstens* eine lok. Minimalstelle. (Somit höchstens eine glob. Min.-stelle.)

Beweis: s. Tafel

Nun zur *Existenz* von Minimalstellen:

Für die *Existenz* von Minimalstellen reicht weder Konvexität noch strikte Konvexität aus; betrachte dazu $f(x)=e^x$ auf $D=(a,b)$ oder auf $D=\mathbb{R}$.
Ein hinreichendes Kriterium kennen wir bereits aus dem 2. Sem.: Falls der Definitionsbereich kompakt ist (und f stetig), so hat f ein Minimum. Und wenn dieses Kriterium nicht erfüllt ist?
Im folgenden Satz lernen wir eine *Abschwächung* des obigen Kriteriums kennen:

Satz (Existenz des Minimums)
Sei $f : D \to \mathbb{R}$ stetig, $\emptyset \neq D \subseteq \mathbb{R}^n$. Ferner habe f eine nichtleere(!) kompakte(!) Levelmenge

$$\emptyset \neq L_c := \{\vec{x} \in D \mid f(\vec{x}) \leq c\}.$$

Dann hat f mindestens eine globale Minimalstelle. (Somit mindestens eine lok. Min.-stelle.)
(Ist f außerdem strikt konvex, so ist diese (s.o.) eindeutig bestimmt.)

Beweis: s. Tafel

Bemerkung: Für stetiges f und abgeschlossenes D sind die Levelmengen L_c immer abgeschlossen (o. Bew.), d.h. in diesem Fall ist nur die Beschränktheit von L_c zu prüfen.

Anwendungsbeispiele:
Der obige Existenzsatz deckt Fälle ab wie

- f wächst zum Rand von D hin, z.B.
  (i) $f(x) = x^2$ auf $D=(-1,1)$
  (ii) $f(x) = x^2$ auf $D=\mathbb{R}$
  } mit z.B. $c:=\frac{1}{2}$ ist $L_c = \left[-\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right]$ kompakte

und nichtleere Levelmenge.

- f fällt zwar zum Rand von D hin, aber der Rand von D gehört dort zu D, z.B. $f(x)=e^x$, $D=[-1,1)$; mit z.B. $c:=1$ ist $L_c=[-1,0]$ kompakt und nichtleer (obwohl D, anders als in Kap. I.2, nicht kompakt ist!)

Spezialfall (Standard-Beispiel) eines konvexen Optimierungsproblems: Das Quadratische Optimierungsproblem:

Problemstellung: Quadratisches Optimierungsproblem:

$D:=\mathbb{R}^n$, $A\in\mathbb{R}^{n\times n}$ symmetrisch positiv definit, $\vec{b}\in\mathbb{R}^n$ $(, c\in\mathbb{R})$, $\langle\cdot,\cdot\rangle$ sei das Euklidische Skalarprodukt,

$$f(\vec{x}) := \frac{1}{2}\langle A\vec{x},\vec{x}\rangle + \langle\vec{b},\vec{x}\rangle \ \ (+c) \quad \longrightarrow \quad \min \qquad (*)$$

Wir werden das 'c' weglassen, da es keinen Einfluss auf die Minimalstelle (und den Schwierigkeitsgrad der Aufgabe) hat.

Lösen des Problems:
Es ist (ergibt sich leicht durch Rechnen mit Komponenten, siehe Übung)

$$\nabla f(\vec{x}) = A\vec{x}+\vec{b},$$
$$Hf(\vec{x}) = A \quad \forall\vec{x}\in\mathbb{R}^n$$

Nach Voraussetzung an A ist also f strikt konvex (s. Satz 'Kriterium für Konvexität'). Nach Satz gibt es also höchstens eine (lok.) Min.-stelle, und diese ist global.

Wir setzen $\nabla f(\vec{x}_*)\stackrel{!}{=}\vec{0}$ und bekommen

$$\vec{x}_* = -A^{-1}\vec{b} \qquad (**)$$

als kritische Stelle. (Positiv definite Matrizen haben nie den EW 0, sind somit immer invertierbar.)
Da $Hf(\vec{x}_*)$ positiv definit ist, ist $\vec{x}_*$ tatsächlich lokale/globale Minimalstelle.

Veranschaulichung von f:

Falls $A = \mathrm{diag}(\lambda_i)$ (positiv definite) *Diagonalmatrix* ist (die $\lambda_i > 0$ also), so ist $\vec{x} \mapsto \frac{1}{2}\langle A\vec{x},\vec{x}\rangle = \frac{1}{2}\sum_{i=1}^n \lambda_i x_i^2 \geq 0$, und man kann sich überlegen, dass Niveaulinien 'Ellipsoide' mit Mittelpunkt $\vec{0}$ und Achsen in Koordinatenrichtungen sind. $\vec{0}$ ist die einzige Minimalstelle.

Falls A keine Diagonalmatrix ist, sondern nur s.p.d., so kann man jedoch per Ähnlichkeitstransformation A auf Diagonalform bringen (Ende 1. Sem.), und zwar $A = Q^T \text{diag}(\lambda_i) Q$, wobei Q Orthogonalmatrix ('Drehung') ist. Durch Substitution $\vec{y} := Q\vec{x}$ (entspricht Drehung des Koordinatensystems) erkennt man: Die Niveaulinien von $\vec{x} \mapsto \frac{1}{2}\langle A\vec{x}, \vec{x}\rangle$ sind weiterhin Ellipsoide mit Mittelpunkt $\vec{0}$, jedoch mit Achsen in Richtung der Eigenräume von A.

Eine kurze Rechunung zeigt, dass man f schreiben kann als $f(\vec{x}) = \frac{1}{2}\langle A(\vec{x}-\vec{x}_*), \vec{x}-\vec{x}_*\rangle + \frac{1}{2}\langle \vec{b}, \vec{x}_*\rangle$, wobei $\vec{x}_* := -A^{-1}\vec{b}$. Das bedeutet, dass man den Graphen von f aus dem Graphen von $\vec{x} \mapsto \frac{1}{2}\langle A\vec{x}, \vec{x}\rangle$ per Translation (Verschiebung) um $\vec{x}_*$ und Anheben um $\frac{1}{2}\langle \vec{b}, \vec{x}_*\rangle$ bekommt; die Niveaulinien von f sind also ebenfalls Ellipsoide, allerdings mit Mittelpunkt $\vec{x}_*$.

Nochmal zurück zum Lösen des Minimierungsproblems:
Wir haben gesehen: Das Lösen eines quadratischen Optimierungsproblem (*) ist äquivalent zum Lösen eines LGS mit s.p.d. Systemmatrix (**).

Jeder Algorithmus, der eines der beiden Probleme löst, kann auch zur Lösung des anderen verwendet werden.

Man kann quadratische Minimierungsproblem lösen z.B.

- Indem man das LGS z.B. mit der Gauß-Elimination direkt löst.
- Indem man einen anderen direkten Löser, der speziell auf symmetrische pos. def. Systemmatrizen zugeschnitten ist, verwendet, wie die sog. *Cholesky-Zerlegung* ($\rightarrow$ Literatur)

  Bem.: Ein Verfahren heißt *direkt*, wenn es nach endlich vielen Schritten die Lösung liefert. (Im Gegensatz zu *iterativen* Verfahren ($\rightarrow$Kap. I.7), die nur immer besser werdende Näherungslösungen liefern.)
- Sind direkte Verfahren oder iterative Verfahren effizienter zum Lösen von LGS? Das ist schwer pauschal zu beantworten. Als Faustregel gilt, dass für kleines n eher direkte Verfahren und für großes n eher iterative Verfahren zum Einsatz kommen.
  Iterative Löser für Lineare Gleichungssysteme werden wir noch in Kap. I.7 kennenlernen; am Ende des Kapitels wird auch nochmal kurz auf einen Effizienzvergergleich zwischen direkten und iterativen Lösers eingegangen.

  Einen speziellen iterativen Löser für *Minimierungsprobleme* schauen wir uns nun an: Das sog. *Gradientenverfahren*, auch genannt: *Methode des steilsten Abstiegs*.

Ein numerisches Verfahren für Optimierungsprobleme:
Das Gradientenverfahren (="Methode des steilsten Abstiegs")

Aufgabe: Sei $f : \mathbb{R}^n \to \mathbb{R}$; f sei zumindest diff'bar.

Wir wollen eine (lokale?/globale?) Minimalstelle von f bestimmen.
Im allgemeinen muss man dazu das (i.a. nichtlineare(!)) Gleichungssystem $\nabla f(\vec{x}) = \vec{0}$ lösen, was i.a. schwierig/nicht exakt möglich ist.

Daher suchen wir nach einem Verfahren, Minimalstellen näherungsweise/numerisch/iterativ bestimmen.

Idee: Das Negative des Gradienten zeigt immer in die Richtung des steilsten Abstiegs. Falls wir also einen Startwert (eine Näherung an die Minimalstelle) $\vec{x}_0$ haben, können wir von dort aus einen Schritt in Richtung von $-\nabla f(\vec{x}_0)$ machen, und kommen dort (falls die Schrittlänge geeignet gewählt wird) in Bereiche mit kleineren Funktionswerten von f.
Wir verwenden also ein Iterationsverfahren

$$\vec{x}_{m+1} := \vec{x}_m - \alpha_m \nabla f(\vec{x}_m) \qquad (1)$$

Jedoch: Wie soll man α_m wählen?
Betrachte dazu die Werte von f entlang der Geraden $\alpha \mapsto \vec{x}_m - \alpha \nabla f(\vec{x}_m)$, also die Abbildung $h : \mathbb{R} \to \mathbb{R}$, $h(\alpha) := f(\vec{x}_m - \alpha \nabla f(\vec{x}_m))$.
Offensichtlich (Skizze!) ist es sinnvoll, $\alpha = \alpha_m$ als Minimalstelle dieser von-$\mathbb{R}$(!)-nach-$\mathbb{R}$-Funktion zu wählen! Also: Bestimme α_m als Lösung von

$$0 \stackrel{!}{=} h'(\alpha) = -\langle \nabla f(\vec{x}_m - \alpha \nabla f(\vec{x}_m)) \,,\, \nabla f(\vec{x}_m) \rangle \qquad (*)$$

Leider kann man i.a. diese Gleichung nur dann nach α auflösen, wenn man Annahmen an f macht.

Der Einfachheit halber wollen wir momentan den Fall betrachten, dass f eine *quadratische* Funktion (s.o.) ist. Dann ist (s.o.) $\nabla f(\vec{x}) = A\vec{x} + \vec{b}$.
Dies eingesetzt in unsere Bestimmungsgleichung für α_m (*) ergibt:

$$\begin{aligned} 0 = \; & \langle A(\vec{x}_m - \alpha \nabla f(\vec{x}_m)) + \vec{b} \,,\, A\vec{x}_m + \vec{b} \rangle \\ = \; & \langle A\vec{x}_m - \alpha A \nabla f(\vec{x}_m) + \vec{b} \,,\, A\vec{x}_m + \vec{b} \rangle \\ = \; & \langle A\vec{x}_m + \vec{b} \,,\, A\vec{x}_m + \vec{b} \rangle - \alpha \langle A \nabla f(\vec{x}_m) \,,\, A\vec{x}_m + \vec{b} \rangle \end{aligned}$$

Wir erhalten also als optimale Wahl der Schrittweitensteuerung für quadratische Probleme:

$$\alpha_m = \frac{\langle \nabla f(\vec{x}_m) \,,\, \nabla f(\vec{x}_m) \rangle}{\langle A \nabla f(\vec{x}_m) \,,\, \nabla f(\vec{x}_m) \rangle} \quad (2), \qquad \text{mit} \quad \nabla f(\vec{x}_m) = A\vec{x}_m + \vec{b} \quad (3)$$

was in (1) einzusetzen ist.

Das Gradientenverfahren auf quadratische Probleme anzuwenden, erscheint nur begrenzt sinnvoll/interessant, da es ja als Alternative reicht, lediglich das LGS $\nabla f(\vec{x}) = A\vec{x}+\vec{b} \stackrel{!}{=} \vec{0}$ zu lösen.

Es lassen sich jedoch zwei wichtige sinnvolle Folgerungen daraus aufstellen:

1. Obige Überlegung liefert eine Idee, wie man das Gradientenverfahren auch dann anwenden kann, wenn f kompliziert ist, insbesondere also $\nabla f(\vec{x}) \stackrel{!}{=} \vec{0}$ ein *nichtlineares* Gleichungssystem ist, und man also nicht auf das Lösen eines LGS ausweichen kann:

   Im Fall, dass $f : \mathbb{R}^n \to \mathbb{R}$ 'beliebig' ist, verwende einfach dennoch Formeln (1) und (2) (jedoch ohne (3)!; ein "$\vec{b}$" und ein "A" haben wir gar nicht!),
   und anstelle von A verwende $Hf(\vec{x}_m)$ in (2), was man damit begründen kann, dass zumindest im quadratischen Fall $A = Hf(\vec{x}_m)$ ist:

   Gradienten-Verfahren (für allgemeines glattes f)

   $$\vec{x}_{m+1} := \vec{x}_m - \alpha_m \nabla f(\vec{x}_m)\,,$$

   wobei eigentlich α_m als Lsg. der Gleichung $\langle \nabla f(\vec{x}_m - \alpha_m \nabla f(\vec{x}_m))\,, \nabla f(\vec{x}_m)\rangle = 0\,,$
   zu wählen ist; da dieses α_m i.a. nicht explizit zu berechnen ist, nimmt man gerne z.B.

   $$\alpha_m := \frac{\langle \nabla f(\vec{x}_m)\,,\, \nabla f(\vec{x}_m)\rangle}{\langle Hf(\vec{x}_m)\nabla f(\vec{x}_m)\,,\, \nabla f(\vec{x}_m)\rangle}$$

   Man kann das Verfahren auch so verstehen (oder: herleiten), dass man in jedem Iterationsschritt $\vec{x}_m \to \vec{x}_{m+1}$ die zu minimierende Funktion f lokal um $\vec{x}_m$ durch eine quadratische Funktion $\tilde{f}_m(\vec{x}) = \frac{1}{2}\langle A_m\vec{x}, \vec{x}\rangle + \langle \vec{b}_m, \vec{x}\rangle + c_m$ approximiert (also: Taylor-Entw. von f um $\vec{x}_m$), und für dieses $\tilde{f}_m$ einen Schritt des Gradienten-Verfahrens macht, dabei $H\tilde{f}_m(\vec{x}_m) = Hf(x_m) = A_m$, $\nabla\tilde{f}_m(\vec{x}_m) = \nabla f(\vec{x}_m) = \vec{b}_m$, $\tilde{f}_m(\vec{x}_m) = f(\vec{x}_m) = c_m$ ausnutzt.

   Für das so konstruierte Verfahren kann man unter gewissen Annahmen (f strikt konvex mit nichtleerer kompakter Levelmenge) Konvergenz gegen die (dann einzige) Minimalstelle zeigen. Für 'beliebige' f kann man natürlich keine Konvergenz zeigen; ggf. gibt es ja gar keine Min.-stelle.

   In der Praxis taucht bei nicht-konvexen f oft die Schwierigkeit auf, dass das Gradientenverfahren irgendein *lokales* Minimum ansteuert, das nicht das globale ist.

2. Zur Konvergenzgeschwindigkeit (beim quadratischen Modellproblem): Diese wird umso schlechter, je größer das Verhältnis $\kappa := \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$ von größtem zu kleinstem EW von A ist. (Das Gradientenverfahren macht dann Zick-Zack-Wege und kommt nur langsam voran. Dies kann man sich anschaulich leicht erklären, s. Skizze, da

die Verhältnisse der EW die Abplattung der Ellipsoide beschreibt, siehe Übung.)
Es gibt Weiterentwicklungen des Gradientenverfahrens für quadratische Probleme, die schneller konvergieren, und zwar ist das bekannteste die
Methode der konjugierten Gradienten, kurz: das *cg-Verfahren* (→Literatur).
Beim cg-Verfahren hängt die Konvergenzgeschwindigkeit zwar ebenfalls Verhältnis von κ ab, jedoch in weitaus schwächerer Form, so dass das Verfahren für große n häufig effizienter ist als die Gauß-Elimination und andere direkte Verfahren.

Nun noch ein konvexes Prototyp-Problem *mit Nebenbedingungen*:

Gut analysierbar sind die sog. **Linear-Quadratischen Minimierungsprobleme**:

Problemstellung "Linear-Quadratisches Minimierungsproblem"
Die zu minimierende Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei quadratisch, also $f(\vec{x}) = \frac{1}{2}\langle A\vec{x}, \vec{x}\rangle + \langle \vec{b}, \vec{x}\rangle$, mit $A \in \mathbb{R}^{n\times n}$ s.p.d.;
diese soll minimiert werden unter der 'linearen' (affin-linearen) NB $B\vec{x} = \vec{c}$, mit $B \in \mathbb{R}^{m\times n}$, $\vec{c} \in \mathbb{R}^m$ (das sind also m viele 'lineare' skalare NB).

Skizze!

Man bezeichnet die Menge $M := \{\vec{x} \in \mathbb{R}^n \mid B\vec{x} = \vec{c}\}$ als *zulässige Menge* und die Funktion f als *Zielfunktion* oder *Kostenfunktion* (engl.: *objective function, cost function*).
Sinnvoll ist es, $m < n$ zu fordern, denn ansonsten ist zu erwarten, dass die zulässige Menge nur aus einem Punkt besteht (dann ist f irrelevant) oder leer ist.
M ist ein affin-linearer Unterraum von $\mathbb{R}^m$, kann aber auch leer sein.

Analyse: M ist konvex; f ist strikt konvex; somit ist $f|_M$ ebenfalls strikt konvex.

Nach Satz hat $f|_M$ somit höchstens eine (lokale=globale) Minimalstelle.
Im Fall $M \neq \emptyset$ könnte man auch *Existenz* der Lösung abstrakt zeigen (über Levelmengen, siehe Satz); wir verzichten darauf, da wir die Lösung gleich konkret berechnen.

Berechnen der Lösung: Für Optimierungsprobleme mit NB haben wir in Kap. 1.2 zwei Techniken kennengelent: Elimination der NB und Lagrange-Formalismus.

Beachte, dass man die NB $B\vec{x} = \vec{c} \in \mathbb{R}^n$ nicht einfach nach $\vec{x}$ auflösen und somit eliminieren kann, indem man B invertiert, denn B ist eine 'rechteckige' Matrix. (Man müsste stattdessen in B eine invertierbare $n\times n$ Teilmatrix B^* finden, d.h. $B = (B^*|B_{Rest})$, die NB als $B^*\vec{x}^* + B_{Rest}\vec{x}_{Rest} = \vec{c}$ nach $\vec{x}^*$ auflösen und in die Zielfunktion einsetzen, die dadurch zu einem Min.-Problem in $\vec{x}_{Rest}$ wird, ohne NB.)

Wir verwenden zur Handhabung der NB den **Lagrange-Formalismus**:

Lagrange-Formalismus für das Linear-Quadratische Minimierungsproblem

Wir haben m skalare NB $B_i\vec{x}=c_i$, $i=1,...,m$,
wobei $B_i\in\mathbb{R}^{1\times n}$ die i-te Zeile von B ist.

Wir schreiben die m NB als $g_i(\vec{x}):=B_i\vec{x}-c_i=0$.

Wir brauchen m Lagrange-Multiplikatoren $\lambda_1,...,\lambda_m$.
Das Lagrange-System lautet

$$\begin{array}{lcl} \nabla f(\vec{x}) + \lambda_1 \nabla g_1(\vec{x}) + ... + \lambda_m \nabla g_m(\vec{x}) & = & 0 \\ g_1(\vec{x}) & = & 0 \\ & \vdots & \\ g_m(\vec{x}) & = & 0 \end{array}$$

Da $\nabla g_i(\vec{x})=B_i{}^T$ ist, bekommen wir die beiden Vektorgleichungen

$$\begin{array}{lcl} A\vec{x}+\vec{b} + \lambda_1 B_1{}^T + ... + \lambda_m B_m{}^T & = & \vec{0} \\ B\vec{x} - \vec{c} & = & \vec{0} \end{array}$$

Das sind $n+m$ Gleichungen für die $n+m$ Unbekannten $\vec{x}$, $\vec{\lambda}=(\lambda_1,...,\lambda_m)^T$.

Das Gleichungssystem kann geschrieben werden als

$$\begin{array}{lcl} A\vec{x} + B^T\vec{\lambda} & = & -\vec{b} \\ B\vec{x} & = & \vec{c} \end{array} \qquad \text{bzw.} \qquad \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \vec{x} \\ \vec{\lambda} \end{pmatrix} = \begin{pmatrix} -\vec{b} \\ \vec{c} \end{pmatrix} =: \vec{\vec{b}}$$

Dieses LGS ist übrigens – wie A – ebenfalls symmetrisch (aber i.a. nicht positiv definit)!

Man kann es lösen, z.B. mit Gauß-Elimination.

Oder man kann es formal auflösen nach $\vec{x}$, $\vec{\lambda}$, indem man die obere Gleichung nach $\vec{x}$ auflöst, dies in die untere Gleichung einsetzt. Das liefert eine explizite Formel für $\vec{\lambda}$. Diese wiederum kann man in die Gleichung für $\vec{x}$ einsetzen, was eine explizite Lösungsformel für $\vec{x}$ ergibt.

Rechnung: s. Tafel oder Übung.

(Diese Lösungsformel ist allerdings von etwas eingeschränktem Interesse, da ihre Evaluation mehrfach die Berechnung von Inversen erfordert.)

1.6 Lineare Optimierung

Problemstellung:

Gegeben ist eine lineare(!) Zielfunktion $f : \mathbb{R}^n \to \mathbb{R}$, die unter Einhaltung von linearen(!) Gleichungs- und Ungleichungs(!)-Nebenbedingungen minimiert oder maximiert werden soll.

Lineare Optimierung wird (seltsamerweise) häufig auch als *Lineare Programmierung* bezeichnet.
("Programm" vielleicht eher im Sinne von "Planung" als von "Computerprogramm")
Begründer dieses Gebiets: George Dantzig, 1914-2005, in den 1940er Jahren

Beispiel: (Produktionsplanung):

Ein Unternehmen stellt 2 Produkte her,
P_1: Glastür mit Alu-Rahmen
P_2: Fenster mit Holzrahmen

Gewinn pro Tür-Einheit: 3 Geldeinheiten (GE)
Gewinn pro Fenster-Einheit: 5 GE

Die Produktion erfolgt in 3 Fabriken:

F_1 stellt Alu-Rahmen her; Kapazität 4 Alu-Einheiten pro Tag
F_2 stellt Holzrahmen her; Kapazität 6 Rahmen-Einheiten pro Tag
F_3 stellt Glasscheiben (für Türen und Fenster); Kapazität 18 Glas-Flächen-Einheiten pro Tag, wobei pro Tür-Einheit 3 Glas-Einheiten und pro Fenster-Einheit 2 Glas-Einheiten gebraucht werden.

Wir nehmen an, dass alle Einheiten, die produziert werden, verkauft werden.

Ziel: Maximiere Gewinn. (Wie soll das Glas auf Fenster und Türen aufgeteilt werden?)

Formuliere obiges Problem als Lineares Programm (LP):

1. Identifiziere die Entscheidungsvariablen (kurz: Variablen):
   x_1, x_2; das seien die Einheiten, die von P_1 bzw. P_2 pro Tag hergestellt werden.

2. Zielfunktion: Der Gewinn ist $f(x_1, x_2) = 3x_1 + 5x_2 \quad \longrightarrow \quad \max$

3. Nebenbedingungen (NB):
   (i) $x_1 \leq 4$
   (ii) $x_2 \leq 6$
   (iii) $3x_1 + 2x_2 \leq 18$
   (iiii) $x_1, x_2 \geq 0$ (nicht vergessen!)

(Bemerkung: Man hätte auch noch ein x_3 einführen können für die Anzahl der Glas-Einheiten, und $x_3 \leq 18$ und $x_3 = 3x_1 + 2x_2$ an Stelle von (iii). Es sind i.a. mehrere (äquivalente) Modelle möglich.)

Das Problem in Kurzform: $\begin{cases} f(\vec{x}) := \langle \vec{c}, \vec{x} \rangle \ \to \ \max \\ \text{so dass } A\vec{x} \leq \vec{b} \end{cases}$

mit $\vec{c} := \begin{pmatrix} 3 \\ 5 \end{pmatrix}$, $A := \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 3 & 2 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}$, $\vec{b} := \begin{pmatrix} 4 \\ 6 \\ 18 \\ 0 \\ 0 \end{pmatrix}$, und wobei "$\leq$" *komponentenweise* zu verstehen ist.

Gestalt der zulässige Menge, Simplex:

Die Menge aller $\vec{x} \in \mathbb{R}^n$, die alle NB (i.a. sowohl Gleichungs- als auch Ungleichungs-) erfüllt, heißt wieder *zulässige Menge*.

Die zulässige Menge von LP ist ein Schnitt von *Halbräumen*, d.h. von Mengen der Form $\{\vec{x} \in \mathbb{R}^n \mid u(\vec{x}) \leq c\}$, wobei $u : \mathbb{R}^n \to \mathbb{R}$ linear ist und $c \in \mathbb{R}$.

(Bemerkung: Eine *Gleichung* $u(\vec{x}) = c$ kann man in *Ungleichungen* $u(\vec{x}) \leq c \wedge -u(\vec{x}) \leq -c$ zerlegen.)

Def. (Polyeder, Simplex)

(a) Ein nichtleerer Schnitt von endlich vielen Halbräumen heißt *Polyeder*.

(b) Seien $n+1$ viele Punkte $\vec{x}_i \in \mathbb{R}^n$ gegeben. Die von den Punkten 'aufgespannte' Menge

$$\left\{ \sum_{i=1}^{n+1} \alpha_i \vec{x}_i \;\middle|\; \alpha_i \geq 0 \,;\, \sum_{i=1}^{n+1} \alpha_i = 1 \right\}$$

heißt *Simplex*.

Bemerkung: Für $n = 2$ ist ein Simplex ein Dreieck im $\mathbb{R}^2$ (zumindest im Fall linear unabhängiger $\vec{x}_i$), für $n = 3$ ist ein Simplex ein Tetraeder im $\mathbb{R}^3$, allgemein für $n \in \mathbb{N}$ ist ein Simplex ein Polyeder (zum Nachdenken: wie könnte man die zugehörigen Halbräume konstruieren?).

Jedes Polyeder (somit auch: jedes Simplex) ist konvex (z.Ü: zeige dies). Also ist die zulässige Menge konvex.

Da wir hier nur 2 Entscheidungsvariablen haben, können wir die **Lösung *zeichnerisch* finden**:

Wir zeichnen in ein x_1-x_2-Koordinatensystem alle NB ein, somit die zulässige Menge. Dann skizzieren wir die Niveaulinien der Zielfunktion (sind parallele Geraden).

Zeichnung: s. Tafel.

Wir stellen Grundsätzliches fest:

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

2. Es kann passieren, dass es keine Lösung gibt:
   – Wenn die zulässige Menge leer ist, gibt es keine Lösung.
   – Wenn sie unbeschränkt ist, kann(!) es passieren, dass es keine Lösung gibt ("die Lösung liegt im Unendlichen", d.h. $\inf_{x \in M} f = -\infty$ oder $\sup_{x \in M} f = +\infty$)

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

4. Wenn es mehrere Lösungen gibt, dann gibt es mehrere Ecken, die Lösungen sind, und die Menge aller Lösungen ist zusammenhängend.

Fazit: Es reicht, die Ecken zu durchsuchen!

Ein erstes, sehr einfaches Lösungsverfahren wäre:
– Bestimme alle Ecken des zulässigen Bereichs. (Wie geht das?)
– Werte Zielfunktion an allen Ecken aus.
– Zumindest bei Beschränktheit des zulässigen Bereichs findet man so sicher das Minimum.

Einwand: Bei wachsender Anzahl von Entscheidungsvariablen und NB wächst die Anzahl der Ecken derart schnell an, dass der Rechenaufwand zu groß wäre.

Die grundlegende Idee für einen schnelleren Algorithmus (den sog. *Simplex-Algorithmus*):

— Bestimme (wie?) *eine* Ecke des zulässigen Bereichs als Startwert eines iterativen Verfahrens.

- Gehe sukzessive von der aktuellen Ecke entlang einer Kante zu einer Nachbar-Ecke, derart ausgesucht, dass die Zielfunktion dort einen kleineren Wert hat als an der aktuellen Ecke.
  Falls es keine solche Nachbar-Ecke gibt, ist man am Ziel.

Umwandlung des gegebenen LP in 'allgemeiner Form' in die sog. Standard-Form:

Allgemeine Form: Suche $\vec{x} \in \mathbb{R}^n$ so dass

$$
\begin{aligned}
& f(\vec{x}) := \langle \vec{c}, \vec{x} \rangle \longrightarrow \min \text{ oder } \longrightarrow \max \\
\text{so dass } \quad & a_{i1}x_1 + \ldots + a_{in}x_n = b_i, \quad i = 1, \ldots, p \\
& a_{i1}x_1 + \ldots + a_{in}x_n \leq b_i, \quad i = p+1, \ldots, q \\
& a_{i1}x_1 + \ldots + a_{in}x_n \geq b_i, \quad i = q+1, \ldots, m
\end{aligned}
$$

wird umgewandelt in die

Standard-Form: Suche $\vec{x} \in \mathbb{R}^n$ so dass

$$\begin{array}{lll} f(\vec{x}) := \langle \vec{c}, \vec{x} \rangle & \longrightarrow & \min \\ \text{so dass} \quad A\vec{x} & = & \vec{b} \\ \text{und} \quad \vec{x} & \geq & \vec{0} \end{array}$$

Dabei sei $A \in \mathbb{R}^{m \times n}$, die m Zeilen von A seien linear unabhängig, d.h. $\text{rang}(A) = m$.

Umwandlung von allgemeiner Form in Standard-Form:

1. Falls *Maximum* von f gesucht:
   Ersetze die Zeile
   $f(\vec{x}) = \langle \vec{c}, \vec{x} \rangle \to \max$
   durch
   $\tilde{f}(\vec{x}) := -f(\vec{x}) = \langle \underbrace{-\vec{c}}_{=:\tilde{\vec{c}}}, \vec{x} \rangle \to \min$

2. Auch die "$\geq$"-NB werden mit (-1) durchmultipliziert, so dass nur noch "$\leq$"-NB (und Gleichheits-NB) vorhanden sind
   (Ausgenommen davon: Ungleichungen der Form $x_i \geq 0$; die können so bleiben)

3. Die Ungleichungs-NB

   $$a_{i1}x_1 + \ldots + a_{in}x_n \leq b_i \qquad (*)$$

   werden in *Gleichungs*-NB umgeformt durch Einführung sog. *Schlupf-Variablen* $\tilde{x}_i$ (engl.: *slack variables*):

   $$(*) \quad \Longleftrightarrow \quad a_{i1}x_1 + \ldots + a_{in}x_n + \tilde{x}_i = b_i \quad \wedge \quad \tilde{x}_i \geq 0$$

4. Wir sind noch nicht fertig. Wir müssen noch dafür sorgen, dass *jede* Variable x_i eine Bedingung $x_i \geq 0$ hat. Dazu:
   Für jede Variable x_i, für die es noch keine Bedingung "$x_i \geq 0$" gibt, substituiere (d.h. eliminiere x_i)

   $$x_i =: x_i^+ - x_i^-, \quad x_i^+ \geq 0, \quad x_i^- \geq 0$$

   Hintergrund: Jede Zahl $x \in \mathbb{R}$ kann per $x^+ := \max\{x, 0\} \geq 0$, $x^- := \max\{-x, 0\} \geq 0$ als $x = x^+ - x^-$ geschrieben werden.

5. Zur Sicherstellung der linearen Unabhängigkeit der Zeilen von A:
   Lin. Abh. kann sich darin äußern, dass das LGS $A\vec{x} = \vec{b}$ keine Lösung hat ("Zeilen widersprechen sich"), oder darin, dass Zeilen redundant sind.

   Bringe $A\vec{x} = \vec{b}$ auf Stufenform $(\tilde{A}|\tilde{\vec{b}})$.

Ersetze im LP die Gleichungen $A\vec{x}=\vec{b}$ durch die äquivalenten Gleichungen $\tilde{A}\vec{x}=\tilde{\vec{b}}$.
Falls Stufenform 'Typ III' (s. 1. Sem.), ist LGS unlösbar, d.h. die zul. Menge ist leer, d.h. das LP unlösbar.
Falls Stufenform 'Typ II': Streiche Nullzeilen des LGS, die verbleibenden Zeilen der Stufenform sind l.u.
(Der Fall 'Typ I' ist hier der 'Trivialfall', dass die zulässige Menge nur aus 1 Punkt besteht.)

Beispiel: Umwandlung des obigen Beispielproblems "Produktionsplanung": s. Tafel

Bemerkung: Man nimmt in Kauf, dass bei der Umwandlung in 3. und 4. die Anzahl der Variablen (d.h. n, die Dimension des Raumes) i.a. deutlich anwächst.

Visualisierung der zulässigen Menge bei Standard-Form im Fall $n=3$:
s. Tafel (mehrere Varianten)

Die Visualisierung verdeutlicht: Die zulässige Menge ist der Schnitt der affin-linearen Mannigfaltigkeit, die durch $A\vec{x} = \vec{b}$ definiert wird, mit dem positiven 'Rektanten' (:=Verallg. von 'Quadrant' im $\mathbb{R}^2$ auf $\mathbb{R}^n$); sie ist weiterhin ein Polyeder.
Da die Zielfunktion weiterhin linear ist, ist/sind die Lösung(en) weiterhin in den *Ecken* dieses Polyeders zu suchen.

Wie kann man die Ecken charakterisieren/finden?
In den an der Tafel visualisierten Beispielen:
(a) $n=3$, $m=1$: In den Ecken sind jeweils 2 Variablen gleich null, und die dritte wird durch die Gleichung beschrieben.
(b) $n=3$, $m=2$: In den Ecken ist jeweils 1 Variable gleich null, und die beiden übrigen werden durch die beiden Gleichungen beschrieben.
(c) Allgemein für $A\vec{x}=\vec{b}$, $\vec{x}\geq\vec{0}$, mit $A\in\mathbb{R}^{m\times n}$ mit Rang$=m$ und $\vec{x}\in\mathbb{R}^n$, $n\geq m$:
Da man m (l.u.) Gleichungsbedingungen zu erfüllen hat, kann man $n-m$ Variablen zu null setzen ($\dim(\text{Kern}(A))=n-m$), und die übrigen m Variablen dann über die m Gleichungen berechnen.
Formelmäßig: Teile $A = (A_B|A_N)$ auf, wobei $A_B \in \mathbb{R}^{m\times m}$ und $A_N \in \mathbb{R}^{m\times(n-m)}$. Teile analog $\vec{x} = \begin{pmatrix} \vec{x}_B \\ \vec{x}_N \end{pmatrix}$ auf, $\vec{x}_B\in\mathbb{R}^m$, $\vec{x}_N\in\mathbb{R}^{n-m}$.
Wir können das LGS $A\vec{x}=\vec{b}$ schreiben als

$$A_B\vec{x}_B + A_N\vec{x}_N = \vec{b}.$$

Falls A_B (ist quadratisch!) invertierbar ist (falls also die ersten m Spalten von A l.u. waren), kann man das LGS nach $\vec{x}_B$ auflösen: $\vec{x}_B = {A_B}^{-1}(\vec{b} - A_N\vec{x}_N)$

Wir können also $\vec{x}_N := \vec{0}$ setzen (somit $n-m$ Variablen zu null gesetzt), und die übrigen m Variablen $\vec{x}_B$ (s.o.) zu

$$\vec{x}_B = {A_B}^{-1}(\vec{b} - A_N \underbrace{\vec{x}_N}_{=:\vec{0}}) = {A_B}^{-1}\vec{b}$$

setzen.

Jedoch: Wir hatten der Einfachheit halber vorausgesetzt, dass die ersten m Spalten von A l.u. sind.
Um *beliebige* Ecke des Polyeders zu finden, muss man eine *beliebige* Auswahl von m l.u. Spalten von A betrachten. Schreibtechnisch ist es praktisch, die Spalten dann so *umzuordnen*, dass diese m Spalten nach vorne kommen. Diese Umordnung ist erlaubt, wenn man gleichzeitig die Komponenten von $\vec{x}$ und die von $\vec{c}$ analog umordnet.

Dies motiviert die folgende Definition:

Def. (Basis, Basisvariablen, Basislösung eines in Std.-Form geg. LP)
Sei $A = [\vec{a}_1, ..., \vec{a}_n] \in \mathbb{R}^{m \times n}$ mit linear unabhängigen Zeilen (d.h. $\text{Rang}(A) = m$ und $m \leq n$).
Eine Auswahl von m Spalten $\vec{a}_{i_1}, ..., \vec{a}_{i_m}$, die linear unabhängig sind, aus den n Spalten $\vec{a}_1, ..., \vec{a}_n$ bezeichnet man als eine *Basis* des LP $(A, \vec{b}, \vec{c})$.
Zur Vereinfachung der Sprechweise werden wir machmal auch die zug. *Indexmenge* $\{i_1, ..., i_m\} \subseteq \{1, ..., n\}$ als *Basis* des LP bezeichnen.
Die zugehörigen Komponenten von $\vec{x}$, also $\vec{x}_{i_1}, ..., \vec{x}_{i_m}$, sind die zugehörigen *Basisvariablen*, die übrigen $n-m$ Komponenten von $\vec{x}$ sind die zugehörigen *Nichtbasisvariablen*.
Die zugehörige *Basislösung* erhält man, indem man alle Nichtbasisvariablen zu null setzt und die Basisvariablen dann über die m Gleichungen bestimmt; falls man annimmt, die Spalten von A und $\vec{x}$ seien so umsortiert, dass die die Basis aus den *ersten* Spalten $\{1, ..., m\}$ besteht, dann lautet das LGS $A\vec{x} = A_B\vec{x}_B + A_n\vec{x}_N = \vec{b}$; die Basislösung ist

$$\vec{x}_N = \vec{0}\,, \quad \vec{x}_B = {A_B}^{-1}\vec{b}\,.$$

Jede Ecke der zulässigen Menge, also des Schnittgebildes des Raumes "$A\vec{x} = \vec{b}$" mit dem positiven Rektanten, ist, wie oben motiviert, eine Basislösung.
Doch Vorsicht: Die Umkehrung gilt nicht unbedingt: Eine Basislösung kann Komponenten $x_i < 0$ enthalten (s. Skizze im Fall $n = 2$)!
Wir suchen nur solche Basislösungen, die keine negativen Komponenten enthalten.

Dies führt auf die Definition:

Def. (Zulässigkeit einer Basis(-lösung))

Eine Basis bzw. Basislösung heißt *zulässig*, wenn alle Komponenten der Basislösung ≥ 0 sind.

Die Menge der Ecken des Schnittgebildes des Raumes "$A\vec{x} = \vec{b}$" mit dem positiven Rektanten (in der man nach der Lösung des LP sucht), ist identisch zur Menge aller zulässigen Basen des LP.

Die Prüfung einer gegebenen Basis auf Zulässigkeit ist offensichtlich elementar.

Das Finden einer Basis ist elementar (erfordert Tests auf Lin. Unabh. von Spalten).

Es kann bis zu n-über-m-viele Basislösungen geben (und genau so viele Ecken, falls alle zulässig sind)!

Das *schnelle* Finden einer *zulässigen Basis* (also ohne das Durchtesten aller Basen auf Zulässigkeit) ist jedoch nichttrivial (→später)

Vorsicht: Die Zuordnung

$$\text{Basis mit zulässiger Basislösung} \longmapsto \text{Ecke der zulässigen Menge}$$

ist möglicherweise nicht injektiv;
Beispiel: $m=1$, $n=2$, $A = (1\ -1)$, $\vec{b} = (0)$;
wählt man als Basis die 1. Spalte von A, d.h. setzt die Nichtbasisvariable $x_2 := 0$, so ergibt sich x_1 als Lösung von $1 \cdot x_1 - 1 \cdot x_2 = 0$, also $x_1 = 0$.
wählt man als Basis die 2. Spalte von A, d.h. setzt die Nichtbasisvariable $x_1 := 0$, so ergibt sich x_2 als Lösung von $1 \cdot x_1 - 1 \cdot x_2 = 0$, also $x_2 = 0$.
Zwei verschiedene Basen liefern hier also die gleiche Ecke $\vec{x} = (0,0)^T$.

Erläuterung: Das passiert dann, wenn neben den Nichtbasiskomponenten, die zwangsläufig null sind, auch ein oder mehrere Basiskomponenten, die sich als Lösung $\vec{x}_B = {A_B}^{-1}\vec{b}$ ergeben, 'zufällig' null sind. (ggf. Skizze?); in dem Fall gibt es also *mehrere* zulässige Basien, die ein und dieselbe Ecke beschreiben.

Gibt es immer mindestens eine zulässige Basislösung?

Satz (Existenz einer zulässigen Basislösung eines in Std.-Form geg. LP)

Sei $M := \{\vec{x} \in \mathbb{R}^n \mid A\vec{x} = \vec{b},\ \vec{x} \geq 0\}$ die zulässige Menge eines LP.
Falls $M \neq \emptyset$, dann gibt es (mindestens) eine zulässige Basislösung, andernfalls nicht.

Zum Beweis: Die Aussage ist anschaulich klar, wenn man weiß, dass die Menge der zulässigen Basislösungen der Menge der Ecken der zulässigen Menge entspricht.
Einen strengen Beweis findet man in einschlägigen Büchern.

Wir fassen nun noch einmal zusammen, was wir uns zuvor bereits anschaulich überlegt haben:

Falls die zulässige Menge beschränkt ist, muss die Zielfunktion in der zulässigen Menge ein Minimum annehmen (verwendet, dass die zulässige Menge ohnehin abgeschlossen und die Zielfunktion stetig ist); wir hatten uns anschaulich überlegt, dass wenn es überhaupt eine Minimalstelle gibt, es auch eine Minimalstelle in einer Ecke der zulässigen Menge geben muss, und dass die Menge der Ecken mit der Menge der zulässigen Basislösungen übereinstimmt. Somit:

Satz (Fundamentalsatz der Linearen Programmierung)
Sei $(A, \vec{b}, \vec{c})$ ein LP in Standard-Form mit nichtleerer(!) beschränkter(!) zulässiger Menge $M := \{\vec{x} \in \mathbb{R}^n \,|\, A\vec{x} = \vec{b},\ \vec{x} \geq 0\}$. Dann gibt es eine Lösung (d.h. Minimalstelle von $f|_M$), die eine zulässige Basislösung ist.

Beweis: siehe Literatur über Lineare Optimierung.

Obiger Fundamentalsatz rechtfertigt, dass man 'nur' alle Basislösungen durchprobieren muss, jede auf Zulässigkeit prüfen muss, und f dort auswerten muss, um das gesuchte Minimum von f_M zu finden.
Problematisch ist, dass es bis zu $\binom{n}{m}$ viele Basislösungen geben kann.

Eine bessere Strategie: ausgehend von einer gegebenen zulässigen Basislösung/Ecke $\vec{x}$ wandere zu einer "benachbarten" zulässigen Basislösung/Ecke $\tilde{\vec{x}}$, die $f(\tilde{\vec{x}}) < f(\vec{x})$ erfüllt.
"Benachbart" bedeutet, dass man von der alten zur neuen Basis kommt, indem man **nur einen der Basisvektoren ersetzt**. Dieser Vorgang heißt *Basiswechsel* oder *Basisaustauschschritt*.

Dies führt zu einem *iterativen Algorithmus*. Sobald kein solcher Basiswechsel mehr möglich ist, hat man die (besser: eine) Lösung gefunden.

Wir wollen uns im folgenden überlegen:

Eine Basis $\mathcal{B}$ mit Basislösung $\vec{x}$, bestehend aus $\vec{x}_B, \vec{x}_N$, sei gegeben.
Wie testet man *effizient*, ob bei einem Basiswechsel $\mathcal{B} \to \tilde{\mathcal{B}}$, $\vec{x} \to \tilde{\vec{x}}$
$f(\tilde{\vec{x}}) < f(\vec{x})$ gilt und die neue Basis zulässig ist?
Welche Spalte sollte man, um dies zu erreichen, aus der Basis herausnehmen, welche Spalte hineinnehmen?
Wie kann man *effizient* die neue Basislösung $\tilde{\vec{x}}$ ausrechnen, ohne eine Matrix zu invertieren?

Basiswechsel:

Sei $\mathcal{B} := \{b_1, ..., b_m\} \subseteq \{1, ..., n\}$ eine Basis (d.h. die aus den zugeh. Spalten von A bestehende Matrix $A_B \in \mathbb{R}^{m \times m}$ ist invertierbar).
Seien $\mathcal{N} := \{n_1, ..., n_{n-m}\} \subseteq \{1, ..., n\}$ die übrigen, d.h. die Nicht-Basis-Indizes, d.h. $\mathcal{B}$ und $\mathcal{N}$ bilden eine Partition von $\{1, ..., n\}$; und sei $A_N \in \mathbb{R}^{m \times (n-m)}$ der 'Rest' der Matrix A.

Die Komponenten der Vektoren $\vec{x}, \vec{c}$ werden dementsprechend aufgeteilt in $\vec{x}_B, \vec{c}_B \in \mathbb{R}^m$ und $\vec{x}_N, \vec{c}_N \in \mathbb{R}^{n-m}$.

Wir können die Evaluation von f schreiben
(wobei wir an der Stelle (*) $\vec{x}_B$ mittels $A_B\vec{x}_B + A_N\vec{x}_N = \vec{b}$ eliminieren) als:

$$\begin{aligned} f(\vec{x}) &= \langle \vec{c}, \vec{x} \rangle = \langle \vec{c}_B, \vec{x}_B \rangle + \langle \vec{c}_N, \vec{x}_N \rangle \\ &\overset{(*)}{=} \langle \vec{c}_B, {A_B}^{-1}\vec{b} - {A_B}^{-1}A_N\vec{x}_N \rangle + \langle \vec{c}_N, \vec{x}_N \rangle \\ &= \langle \vec{c}_B, {A_B}^{-1}\vec{b} \rangle + \langle \vec{c}_B, -{A_B}^{-1}A_N\vec{x}_N \rangle + \langle \vec{c}_N, \vec{x}_N \rangle \\ &= \langle \vec{c}_B, {A_B}^{-1}\vec{b} \rangle + \langle \vec{c}_N - (A_B{}^{-1}A_N)^T \vec{c}_B, \vec{x}_N \rangle \qquad \forall \vec{x} \in \mathbb{R}^n \qquad (**) \end{aligned}$$

Wir betrachten nun den Basiswechsel $\mathcal{B} \to \tilde{\mathcal{B}}$, $\mathcal{N} \to \tilde{\mathcal{N}}$, $\vec{x} \to \tilde{\vec{x}}$.
Wir benutzten die Formel (**) (mit *festem* $\mathcal{B}$) um sowohl $f(\vec{x})$ vor dem Basiswechsel als auch $f(\tilde{\vec{x}})$ nach dem Basiswechsel zu berechnen:
Sei j_* derjenige Index, der zur Basis hinzugefügt wird, und k_* der Index, der aus der Basis entfernt wird, d.h. $\{j_*\} = \tilde{\mathcal{B}} \setminus \mathcal{B} = \mathcal{N} \setminus \tilde{\mathcal{N}}$, $\{k_*\} = \mathcal{B} \setminus \tilde{\mathcal{B}} = \tilde{\mathcal{N}} \setminus \mathcal{N}$.
D.h. die j_*-te Komp. ist vor dem Basiswechsel null und wird beim Basiswechsel freigegeben ($x_{j_*} = 0$, $\tilde{x}_{j_*}$ kann $\neq 0$ sein), und die k_*-te Komp. kann vor dem Basiswechsel ungleich null sein und wird beim Basiswechsel auf null gesetzt ($\tilde{x}_{k_*} = 0$).
Geometrische Interpretation der Indizes j_*, k_*: Die Wahl der Indizes j_*, k_* entspricht geometrisch der Wahl der Kante, entlang der wir uns von $\vec{x}$ aus bewegen (und an dessen Endpunkt $\tilde{\vec{x}}$ liegen soll).
Für $\vec{x}$ *vor* dem Basiswechsel gilt $\vec{x}_N = \vec{0}$, somit nach (**)

$$f(\vec{x}) = \langle \vec{c}_B, {A_B}^{-1}\vec{b} \rangle.$$

Für $\tilde{\vec{x}}$ *nach* dem Basiswechsel ist (denken!) $\tilde{\vec{x}}_N = (0, ..., 0, \tilde{x}_{j_*}, 0, ..., 0)^T$, somit nach (*)

$$f(\tilde{\vec{x}}) = \langle \vec{c}_B, {A_B}^{-1}\vec{b} \rangle + (\vec{c}_N - (A_B{}^{-1}A_N)^T \vec{c}_B)_{j_*} \tilde{x}_{j_*}$$

Das Anwachsen des Funktionswertes, welches negativ sein soll, ist somit

$$\Delta f_{j_*} := f(\tilde{\vec{x}}) - f(\vec{x}) = (\vec{c}_N - (A_B{}^{-1}A_N)^T \vec{c}_B)_{j_*} \underbrace{\tilde{x}_{j_*}}_{\overset{!}{\geq} 0} \overset{!}{<} 0$$

Da $\tilde{\vec{x}}$ *zulässig* sein soll, muss $\tilde{x}_{j_*} \geq 0$ sein.

Wir erhalten als Kriterium dafür, dass der Wert der Zielfunktion ("Kostenfunktion") bei Basiswechsel fällt (die sog. *reduzierten Kosten*):

$$(\vec{c}_N - (A_B{}^{-1}A_N)^T \vec{c}_B)_{j_*} \overset{!}{<} 0$$

Beim Basiswechsel ist also ein Index $j_* \in \mathcal{N}$ zur Basis hinzuzufügen, so dass obige Bedingung erfüllt ist.
Gibt es kein solches j_*, dann ist $\vec{x}$ bereits Minimalstelle.

– Der Begriff ist ein wenig irreführend, denn die tatsächliche Reduktion der Kosten ist das Produkt aus den 'reduzierten Kosten' und dem $\tilde{x}_{j_*}$.

– Welches j_* wir nehmen, wenn mehrere die Bedingung erfüllen, spezifizieren wir hier nicht. Falls man dasjenige j_* nehmen möchte, für das der Wert von f am stärksten fällt, müsste man den Wert von $\tilde{x}_{j_*}$ zunächst kennen (und dieser wird auch von k_* abhangen, das momentan noch nicht festgelegt ist).

Was uns momentan noch fehlt für einen Basiswechsel:
Der Wert von $\tilde{x}_{j_*}$ und der Index k_*.
Die folgende Überlegung liefert *beides*.

Die j_*-te Komponente von $\vec{x}$ war vor dem Basiswechsel$=0$, und alle Basis-Komponenten waren ≥ 0 (wegen Zulässigkeit von $\vec{x}$). *Nach* dem Basiswechsel soll die j_*-te Komponente ≥ 0 sein (damit Zulässigkeit erfüllt bleibt), und zwar so, dass
1. *alle* Basiskomponenten ≥ 0 sind ($\rightarrow$Zulässigkeit!) und
2. (mindestens) eine genau $= 0$ wird; der Index dieser Komponente wird das k_* sein!
(Die Bedingung 2. stellt sicher, dass (geometrisch:) wir die Kante genau bis zu einer Ecke durchlaufen und nicht weiter laufen bzw (analytisch:) dass wir einen Index k_* finden, den wir aus der Basis streichen können.

Da $\vec{\tilde{x}}_B \overset{(s.S.10)}{=} {A_B}^{-1}\vec{b} - {A_B}^{-1}A_N\vec{\tilde{x}}_N \overset{(s.S.47)}{=} {A_B}^{-1}\vec{b} - {A_B}^{-1}(A_N)_{j_*}\tilde{x}_{j_*} \overset{!}{\geq} \vec{0}$,
wobei $(A_N)_{j_*}$ die j_*-te Spalte von A_N sein soll, lautet unser Kriterium 1.-2. also: Es soll die vektorielle Bedingung gelten

$$ {A_B}^{-1}\vec{b} \geq \tilde{x}_{j_*}\, {A_B}^{-1}\,(A_N)_{j_*}\,; \qquad \text{diese lautet komponentenweise:} $$

$$ (\underbrace{{A_B}^{-1}\vec{b}}_{=\vec{x}_B\geq\vec{0}})_k \geq \underbrace{\tilde{x}_{j_*}}_{\overset{!}{\geq}0} ({A_B}^{-1}\,(A_N)_{j_*})_k, \quad \forall\, k\in\mathcal{B}, $$

wobei (mindestens) eine der Ungleichungen mit "=" erfüllt sein soll.
Die linke Seite ist $=(\vec{x}_B)_k \geq 0$, da $\vec{x}$ als zulässig vorausgesetzt wurde.
Solche k, für die $({A_B}^{-1}\,(A_N)_{j_*})_k \leq 0$, liefern daher keine Beschränkung für $\tilde{x}_{j_*}$. Es reicht, nur die übrigen $k\in\mathcal{B}$ zu betrachten.

Indem wir

$\tilde{x}_{j_*} := \min\limits_{k\in\mathcal{B}} \left\{ \frac{(\vec{x}_B)_k}{({A_B}^{-1}(A_N)_{j_*})_k} \mid ({A_B}^{-1}(A_N)_{j_*})_k > 0 \right\}$ und
$k_* :=$ derjenige Index (oder einer derjenigen Indizes) k, für die das obige Minimum angenommen wird, d.h. so dass $\tilde{x}_{j_*} = (\vec{x}_B)_{k_*}/({A_B}^{-1}(A_N)_{j_*})_{k_*}, ({A_B}^{-1}(A_N)_{j_*})_{k_*} > 0$ und $k_*\in\mathcal{B}$ gilt,

wählen, werden obige Bedingungen 1. und 2. erfüllt. Die obige Regel zur Festlegung von $\tilde{x}_{j_*}$ und k_* heißt *Quotientenregel der linearen Programmierung*.

Falls das Minimum über die leere Menge gebildet wird (d.h. $\forall\, k \in \mathcal{B} :\ ({A_B}^{-1}(A_N)_{j_*})_k \leq 0$), bedeutet dies, dass es keinerlei Einschränkung an $\tilde{x}_{j_*} \geq 0$ gibt, also für beliebig großes $\tilde{x}_{j_*} \geq 0$ die Menge M nicht verlassen wird. Und da j_* so gewählt wurde, dass die Kostenfunktion f in diese Richtung fällt, folgt, dass dann das Ergebnis $\inf f_M = -\infty$ ist.

Wir haben somit den Algorithmus im wesentlichen aufgestellt:
[für II.b-c siehe S. 47, für II.d-e siehe S. 48.]

Simplex-Algorithmus

I. Suche eine zulässige Basis $\mathcal{B}$ und die zugeh. Basislösung $\vec{x}$, oder entscheide, dass es keine solche gibt (dann ist $M = \emptyset$ und es gibt keine Lösung)
(Wie das systematisch geht: Dazu später. Ist nicht trivial!)

II.
a. Bilde aus $A, \vec{x}, \vec{c}$ unter Verwendung der Indexmengen $\mathcal{B}$, $\mathcal{N} = \{1, ..., n\} \setminus \mathcal{B}$ die Größen $A_B, A_N, \vec{c}_B, \vec{c}_N, \vec{x}_B = {A_B}^{-1}\vec{b}, \vec{x}_N = \vec{0}$

b. Falls $\forall\, j \in \mathcal{N} :\ (c_N)_j \geq (({A_B}^{-1}A_N)^T \vec{c}_B)_j$:
Ziel erreicht; Lsg$= f(\vec{x}) = \langle \vec{c}_B, \vec{x}_B \rangle$. Ende.

c. Wähle ein $j_* \in \mathcal{N}$ mit $(\vec{c}_N)_{j_*} < (({A_B}^{-1}A_N)^T \vec{c}_B)_{j_*}$

d. Falls $\forall\, k \in \mathcal{B} :\ ({A_B}^{-1}(A_N)_{j_*})_k \leq 0$:
f ist nach unten unbeschränkt; "Lsg" ist: $\inf_{\vec{x} \in M} = -\infty$, Ende.

e. Setze $\tilde{x}_{j_*} := \min\limits_{k \in \mathcal{B}} \left\{ \frac{(\vec{x}_B)_k}{({A_B}^{-1}(A_N)_{j_*})_k} \mid ({A_B}^{-1}(A_N)_{j_*})_k > 0 \right\}$ und wähle k_* als den Index oder einen der Indizes k, für den die rechte Seite minimal wird.

f. Streiche j_* aus $\mathcal{N}$ und füge es zu $\mathcal{B}$ hinzu.
Streiche k_* aus $\mathcal{B}$ und füge es zu $\mathcal{N}$ hinzu.

g. Gehe zu II.-a.

Bemerkung: Falls es in c. ein j_* gibt, welches die Bedingung d. erfüllt, so sollte man dieses j_* wählen, damit die Rechnung schnell beendet werden kann.

Wie führt man obigen Algorithmus möglichst effizient durch? In der Praxis geht man wie folgt for:
Man stellt ein "**Tableau**" (eine Erweiterte Matrix) auf, das alle im obigen Algorithmus relevanten Größen beinhaltet. Obigen Schritt II kann man dann mit "Gauß-artigen" elementaren Umformungen des Tableaus durchführen:

Nach Wahl des Startwerts (Schritt I) stellen wir das folgende Tableau auf:

Simplex-Tableau

$$T := \left(\begin{array}{c|c} \vec{c}^T - \vec{c}_B^T A_B{}^{-1} A & -\vec{c}_B^T A_B{}^{-1}\vec{b} \\ \hline A_B{}^{-1} A & A_B{}^{-1}\vec{b} \end{array}\right) \overset{\text{s.S.47}}{\underset{\text{s.S.44}}{=}} \left(\begin{array}{c|c} \vec{c}^T - \vec{c}_B^T A_B{}^{-1} A & -f(\vec{x}) \\ \hline A_B{}^{-1} A & \vec{x}_B \end{array}\right)$$

$$=: (t_{ij})_{i=0..m,\, j=1..n+1} \in \mathbb{R}^{(m+1)\times(n+1)}$$

Interpretation des Simplex-Tableaus, Arbeiten mit dem Simplex-Tableau:

1. Der Bereich $(A_B{}^{-1}A \,|\, A_B{}^{-1}\vec{b})$ entsteht, indem man zunachst das LGS $A\vec{x} = \vec{b}$ einfüllt, und das mit elementaren Gauß-Operationen äquivalent umformt zu $A_B{}^{-1}A\vec{x} = A_B{}^{-1}\vec{b}$; er speichert also die Gleichungs-NB.
   Da für Basisspalten $j \in \mathcal{B}$ $A_B{}^{-1}A_{b_j} = \vec{e}_j$ gilt, enthält der Block $A_B{}^{-1}A \in \mathbb{R}^{m\times n}$ in den Basis-Spalten $b_1, ..., b_m$ die Einheitsvektoren $\vec{e}_{b_1}, ..., \vec{e}_{b_m} \in \mathbb{R}^m$. (Also: Man könnte hier $(A|\vec{b})$ einfüllen und dann Gauss-Umformungen machen bis man in m Spalten die $\vec{e}_i$ hat.)

   Jede Zeile (bis auf die oberste) repräsentiert eine NB; elementare Zeilenumformungen in diesem Bereich sind also erlaubt, weil diese das LGS nur in ein äquivalentes LGS umformen!

   Beachte: Ist der untere Bereich berechnet, kann man auch die oberste Zeile des Schemas leicht ausrechnen.

2. Der Zeilenvektor $\vec{c}^T - \vec{c}_B^T A_B{}^{-1}A$ enthält in Basisspalten (Spaltenindex $j = b_i \in \mathcal{B}$) die Einträge $t_{0,j} = (\vec{c}^T - \vec{c}_B^T A_B{}^{-1}A)_j = c_j - \vec{c}_B^T \underbrace{A_B{}^{-1}A_{b_i}}_{=\vec{e}_{b_i}} = c_j - c_{b_i} = 0$,

   und in Nichtbasisspalten $j = n_i \in \mathcal{N}$ enthält er (s. Def. der reduzierten Kosten S.47) die reduzierten Kosten, die bei der Wahl $j_* := j$ anfallen würden.
   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, d.h. $t_{0,j} < 0$!
   Und falls im oberen Zeilenvektor $(t_{01}, ..., t_{0,n})$ kein Eintrag negativ ist, also keine Kostenreduktion möglich ist, hat man die Lösung gefunden.

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

4. Die Wahl des Index $k_* \in \mathcal{B}$ nach der Quotientenregel mittels Tableau:
   In der zuvor ausgewählten Spalte j_* (die also zuoberst mit einem negativen Eintrag t_{0j_*} beginnt), betrachte die *positiven* Einträge $t_{jj_*} > 0$.
   Falls keiner positiv ist, ist die Lösung $\inf f|_M = -\infty$; Abbruch des Programms.
   Andernfalls: Für Zeilen j, in denen $t_{jj_*} > 0$ ist, dividiere den entsprechenden Eintrag von $\vec{x}_B$ in der rechten Spalte, also $t_{j,n+1}$, durch diesen positiven Wert, und bestimme diejenige Zeile, für die dieser Quotient $t_{j,n+1}/t_{jj_*}$ minimal wird.
   k_* ist der Index derjenigen Basisspalte, die in dieser j-ten Zeile ihre '1' hat.

5. Nachdem man j_* und k_*, für die der Basiswechsel durchgeführt werden soll, ermittelt hat, muss der eigentliche Basiswechsel im Tableau durchgeführt werden: In der j_*-ten Spalte des Matrixblocks muss derjenige Standardbasisvektor entstehen, der bislang in der k_*-ten Spalte stand; dies erreicht man durch elementare Zeilenoperationen. Diese sind erlaubt, da sie lediglich die NB $A\vec{x} = \vec{b}$ äquivalent umformen. Die k_*-te Spalte, in der bisher der ein Standardbasisvektor stand, verändert sich i.a. dabei (wird, wie gewünscht, zu einem Nichtbasisvektor); die übrigen Basisvektoren ändern sich jedoch dabei nicht.
   Was passiert beim Basiswechsel in der obersten Zeile? Basisspalten starten (s.o.) mit einem Null-Eintrag, d.h. da die j_*-te Spalte zur Basisspalte wird, muss in t_{0j_*} eine Null entstehen.
   Man kann sich überlegen, dass man die oberste Zeile *so wie die anderen Zeilen behandeln muss*, d.h. mittels *elem. Zeilenumformungen* bei t_{0j_*} eine Null erzeugt. Die dabei auftretenden Veränderungen der Einträge der obersten Zeile t_{0j} sind genau die richtigen.

   Beachte: Nach dem Basiswechsel muss das Tableau wieder in allen Basis-Spalten Standardbasisvektoren, oberhalb derer eine Null steht, enthalten. Diese Struktur darf nicht verlorengehen. Gauß-Umformungen, die diese Struktur zerstören, sind im Simplex-Algorithmus nicht erlaubt!

Beispiel zum Simplex-Algorithmus: s. Tafel

Anmerkungen zu
– Degeneriertheit des LP und zur
– Konvergenz des Simplex-Algorithmus:

Def. (Degeneriertheit)
Eine zulässige Basislösung $\vec{x}$ heißt *degeneriert*, wenn mehr als $n-m$ der x_i null sind. Ein LP heißt *nicht degeneriert*, wenn es keine degenerierte zulässige Basislösung gibt.

- Für **nicht degenerierte LP** ist der Wert der Zielfunktion während der Simplex-Iteration streng monoton fallend, woraus folgt, dass jede der (maximal n-über-m vielen) Basislösungen höchstens einmal angelaufen werden kann. Der Algorithmus terminiert also nach maximal n-über-m vielen Schritten.

- Es ist in der Tat ("worst case") möglich, dass der Simplex-Algorithmus *alle* diese Basen ansteuert, bis er die Lösung findet.
  In der Praxis kommt sowas "praktisch nicht" vor; der Algorithmus ist i.a. *erheblich* schneller.

- Für **degenerierte LP** *kann(!)* es passieren, dass der Algorithmus "*kreist*", d.h. er zyklisch immer wieder die gleichen Basen durchläuft (die allesamt die gleiche

Polyeder-Ecke repräsentieren, d.h. $\tilde{x}_{j*}=0$ und Wert der Zielfunktion ist konstant). (Degeneriertheit kommt in der Praxis häufig vor, Kreisen eher selten.)
Eine mögliche Strategie, wie man Kreisen ausschließen kann:
– Von allen erlaubten Indizes j_* immer den kleinsten Index auswählen
– Von allen erlaubten Indizes k_* immer den kleinsten Index auswählen
Dann findet der Algorithmus auch für degenerierte LP nach maximal n-über-m vielen Schritten die Lösung.

Was noch fehlt:

Das Finden eines geeigneten Startwerts, also einer beliebigen *zulässigen* Basislösung.

- Obwohl das Finden einer *Basislösung* trivial ist (A auf Stufenform bringen liefert l.u. Spalten), ist das Finden einer *zulässigen Basislösung* i.a. nichttrivial (sofern man nicht alle n-über-m Basislösungen durchprobieren will).
- Eine Ausnahme: Falls die allgemeine Form des LP (d.h. vor Umwandlung in Std.-Form) keine Gleichungs-NB und nur Ungleichungs-NB der Form $A\vec{x}\leq\vec{b}$ mit $\vec{b}\geq\vec{0}$ und $\vec{x}\geq\vec{0}$ enthält, dann bilden immer die Schlupfvariablen eine zulässige Basis (denn für Schlupfvariablen als Basis ist $A_B=E_n$, also $\vec{x}_B=A_B^{-1}\vec{b}=E_m\vec{b}=\vec{b}\geq\vec{0}$), und diese kann man als Startwert nehmen.

Betrachte zum gegebenen LP in Standard-Form

Suche $\vec{x}\in\mathbb{R}^n$ so dass

$$\begin{array}{lllll} f(\vec{x}) := \langle\vec{c},\vec{x}\rangle \longrightarrow \min & & & & \\ \text{so dass} & A\vec{x} & = & \vec{b} & \qquad (LP) \\ \text{und} & \vec{x} & \geq & \vec{0} & \end{array}$$

Dabei sei o.B.d.A. $\vec{b}\geq\vec{0}$ vorausgesetzt (andernfalls Zeilen mit (-1) multiplizieren).

das Hilfsproblem

Suche $\vec{x}\in\mathbb{R}^n$, $\vec{\tilde{x}}\in\mathbb{R}^m$ so dass

$$\begin{array}{lllll} \tilde{f}(\vec{x},\vec{\tilde{x}}) := \sum_{i=1}^{m}\tilde{x}_i \longrightarrow \min & & & & \\ \text{so dass} & A\vec{x}+E_m\vec{\tilde{x}} & = & \vec{b} & \qquad (\tilde{LP}) \\ \text{und} & \vec{x} & \geq & \vec{0}, & \\ \text{und} & \vec{\tilde{x}} & \geq & \vec{0}, & \end{array}$$

Dieses hat ebenfalls Standard-Form, und $\tilde{A}=(A|E_m)\in\mathbb{R}^{m\times(n+m)}$.

Eigenschaften/Nutzen des Hilfsproblems

1. Starten des Hilfsproblems: $(\tilde{LP})$ hat immer eine nichtleere zulässige Menge $\tilde{M}$, und ein Startwert $(\vec{x}, \vec{\tilde{x}}) \in \tilde{M}$ ist trivial zu finden: $\vec{x}_B := \vec{\tilde{x}} = \vec{b} \geq \vec{0}$, $\vec{x}_N := \vec{x} = \vec{0}$.
2. Starten des eigentlichen Problems unter Verwendung der Lösung von $(\tilde{LP})$:
   (a) Falls die Lösung (d.h. das Minimum) von $(\tilde{LP})$ größer als 0 ist, hat (LP) eine leere zulässige Menge und somit keine Lösung.
   (b) Falls die Lösung von $(\tilde{LP}) = 0$ ist, dann ist $M \neq \emptyset$, und eine zulässige Basis von (LP) – die wir also als Startlösung nehmen können – ist dasjenige $\vec{x}$, für das $(\vec{x}, \vec{\tilde{x}})$ die gefundene Lösung von $(\tilde{LP})$ war.

Beweis:

Zu 1: trivial.

Zu 2-a:

Sei also die Lösung von $(\tilde{LP})$ echt positiv. Angenommen $M \neq \emptyset$. Sei also $\vec{x} \in M$. Also ist $\vec{x} \geq \vec{0}$ und $A\vec{x} = \vec{b}$. Dann ist $A\vec{x} + E_m\vec{0} = \vec{b}$, also $(\vec{x}, \vec{\tilde{x}}) \in \tilde{M}$, und offensichtlich $\tilde{f}(\vec{x}, \vec{\tilde{x}}) = 0$. Das ist ein Widerspruch dazu, dass das Minimum von $(\tilde{LP})$ echt positiv ist.

Zu 2-b:

Sei also die Lösung von $(\tilde{LP})$, die an der Stelle $(\vec{x}, \vec{\tilde{x}})$ angenommen werde, gleich null. Es folgt, dass alle $\tilde{x}_i = 0$ sind. Wir können also o.B.d.A. alle $\tilde{x}_i$ als Nichtbasisvariable von der Lösung von $(\tilde{LP})$ auffassen, d.h. alle m Basisvariablen sind unter den x_i zu suchen. Da alle $\tilde{x}_i$ null sind, erfüllt die Lösung von $(\tilde{LP})$ nicht nur die NB von $(\tilde{LP})$, sondern auch die NB von (LP), $A\vec{x} = b$. Fazit: Die zulässige Basis, mit der der Simplexalgorithmus für $(\tilde{LP})$ endete, ist eine zulässige Basis für (LP), mit der wir den Simplexalgorithmus für (LP) starten können. □

Der beschriebene Simplex-Algorithmus wird auch als *Zwei-Phasen-Algorithmus* bezeichnet.
In Phase I wird – mittels Tableau für $(\tilde{LP})$ – ein Startwert für Phase II ermittelt (bzw. entschieden, dass (LP) leere zulässige Menge hat), und Phase II löst (LP) mittels Tableau.

Das Finden eines zulässigen Startwertes ist also i.a. 'ähnlich aufwändig' wie das eigentliche Lösen des LP.

1.7 Fixpunktiterationen

1.7.1 Definition von Fixpunkten und Motivation

Def. (Fixpunkt)
Sei M eine nichtleere Menge und $\Phi : M \to M$ eine Funktion. Ein $x \in M$ mit

$$\Phi(x) = x$$

heißt *Fixpunkt* (FP) (engl: fixed point) von Φ.

Wir betrachten eine Funktion $\Phi : M \to M$, und es sei $M \subseteq V$; V ein normierter Vektorraum (d.h. wir können für Folgen in M über *Konvergenz* reden).
Wir betrachten, für einen Startwert $x_0 \in M$, eine rekursiv definierte Folge

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

Beobachtung: *Falls(!)* diese Folge konvergiert, $\lim\limits_{n\to\infty} x_n =: x_*$, und falls wir annehmen, dass Φ zumindest stetig ist, dann gilt, per $n \to \infty$ auf beiden Seiten von (*):

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

d.h. **der Grenzwert x_* ist dann immer ein Fixpunkt von Φ.**

Das motiviert, die obige Iteration als *Fixpunktiteration* zu bezeichnen:

Def. (Fixpunktiteration)
Sei $\Phi : M \to M$, und es sei $M \subseteq V$; V ein normierter Vektorraum und $x_0 \in M$.
Die rekursive Berechnung

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

heißt *Fixpunktiteration*.

Falls Φ stetig ist, und falls die Fixpunktiteration konvergiert, so ist (s.o.) der Grenzwert immer ein Fixpunkt von Φ.

Wir wollen nun hinreichende Kriterien finden, die die Konvergenz von Fixpunktiterationen garantieren.
Das Ergebnis dieser Überlegung wird im sog. *Fixpunktsatz von Banach* zusammengefasst.

Zuvor brauchen wir einige weitere Begriffe:

Im 1. Semester hatten wir $\mathbb{R}$ als vollständig bezeichnet, da jede Cauchy-Folge in $\mathbb{R}$ einen Grenzwert in $\mathbb{R}$ hat (anders als $\mathbb{Q}$). Diesen Begriff verallgemeinern wir auf normierte Vektorräume:

Def. (Vollständigkeit, Banach-Raum)

Ein normierter $\mathbb{R}$-Vektorraum $(V, \|\cdot\|)$ heißt *vollständig*, falls jede Cauchy-Folge in V konvergiert (gegen ein Element von V).

Ein normierter Vektorraum, der vollständig ist, wird auch als *Banach-Raum* bezeichnet.

Ein normierter Vektorraum, der vollständig ist, und bei dem die Norm durch ein Skalarprodukt erzeugt wird (d.h. es gibt ein Skalarprodukt auf V, so dass $\|x\| = \sqrt{\langle x, x\rangle} \forall x \in V$ gilt), heißt *Hilbert-Raum*. (Stefan Banach 1892-1945, David Hilbert 1862-1943)

Beispiele:

- Der $\mathbb{R}^n$ ist, mit jeder beliebigen Norm, ein Banach-Raum.
  Mit der Euklidischen(!) Norm versehen ist der $\mathbb{R}^n$ außerdem ein Hilbert-Raum.
- Der Raum $V := \{f : [a, b] \to \mathbb{R}, f \text{ stetig}\}$ werde versehen mit
  - der Norm $\|f\|_\infty = \max_{x\in[a,b]} |f(x)|$.
    Ist der Raum vollständig?
    Konvergenz bzgl. der $\|\cdot\|_\infty$-Norm bedeutet *gleichmäßige* Konvergenz. Und wir haben im 2. Sem. gelernt: Für eine gleichmäßig konvergente Folge von stetigen Funktionen ist die Grenzfunktion immer stetig, d.h. in unserem Raum V enthalten. Also: $(V, \|\cdot\|_\infty)$ ist vollständig.
  - V versehen mit der Norm $\|f\|_1 = \int_a^b |f(x)|\, dx$:
    Hierfür haben wir im 2. Semester ein Beispiel einer Folge gefunden (Folge von immer steileren 'Knickfunktionen') deren Grenzfunktion unstetig ist.
    $(V, \|\cdot\|_1)$ ist also kein Banach-Raum.

Weiterführende Bemerkung zum letzten Beispiel: Wenn man unter Verwendung der obigen $\|\cdot\|_1$-Norm einen Banach-Raum konstruieren will, dann muss man eine andere, größere Menge als die Menge der stetigen Funktionen betrachten. Als Vektorraum V muss man die Menge aller Funktionen $f : [a, b] \to \mathbb{R}$, für die $\|f\|_1 = \int_a^b |f(x)|\, dx < \infty$ ist, nehmen, und dabei muss die Norm im Lebesgue'schen anstelle des Riemann'schen Sinne benutzt werden. (s. 2. Sem. "Mangel des Riemann-Integrals"). Der resultierende vollständige normierte Raum wird als $L^1(a, b)$ bezeichnet.
Um für Funktionen $f : [a, b] \to \mathbb{R}$ einen *Hilbert*-Raum zu konstruieren, kann man das Skalarprodukt $\langle f, g\rangle := \int_a^b f(x)g(x)\, dx$ verwenden; als Raum nimmt man alle Funktionen f, für die das Integral $\|f\|_2 := (\int_a^b f(x)^2\, dx)^{\frac{1}{2}} < \infty$ als Lebesgue-Integral existiert. Man nennt diesen Hilbert-Raum $L^2(a, b)$.

Zurück zur Frage, unter welchen Voraussetzungen FP-Iterationen konvergieren:

Einige Skizzen im Fall $\Phi : \mathbb{R} \to \mathbb{R}$ (s. Tafel) legen nahe, dass die "Steilheit" von Φ eine Rolle spielt. Geringe "Steilheit" von Φ kann man nicht nur mittels Φ' erfassen, sondern wir definieren:

Def. (Kontraktion)
Sei $(V, \|\cdot\|)$ ein $\mathbb{R}$-Vektorraum, $D \subseteq V$, und $\Phi : D \to V$.
Φ heißt *Kontraktion*, falls es eine Konstante $k < 1$ gibt, so dass

$$\|\Phi(x) - \Phi(y)\| \leq k \, \|x - y\| \quad \forall x, y \in D$$

(grob gesprochen: Bilder liegen näher beieinander als Urbilder)
Eine Konstante k, die dies erfüllt, heißt *Kontraktionskonstante* von Φ.

Beispiel/Veranschaulichung/Satz (Kontraktionskriterium): Wenn $V = \mathbb{R}$ und $M \subseteq V$ ein Intervall ist, und Φ diff'bar ist mit $\sup_{x \in M} |\Phi'(x)| < 1$, dann ist Φ eine Kontraktion und

$$k := \sup_{x \in M} |\Phi'(x)|$$

ist in dem Fall eine Kontraktionskonstante von Φ.

Beweis: s. Tafel (Mittelwertsatz)

Diese Charakterisierung der Kontraktionseigenschaft ist meist einfacher zu überprüfen als die obige Definition, sie ist jedoch weniger allgemein, da sie $\Phi \in C^1$ voraussetzt.

Anwendungsbeispiel: Ist $\cos : [-1, 1] \to \mathbb{R}$ eine Kontraktion?
Der Kosinus ist diff'bar, berechne also $\sup_{x \in [-1,1]} |\cos'(x)| = \sup_{x \in [-1,1]} |\sin(x)| = \sin 1 < 1$.

Bemerkung: Kontraktionen sind immer stetig.
Begründung: Das ϵ-δ-Kriterium der Stetigkeit

$$\forall \epsilon > 0 \; \exists \delta > 0 : \; \|x - y\| \leq \delta \;\Rightarrow\; \underbrace{\|\Phi(x) - \Phi(y)\|}_{\leq k\delta} \overset{!}{\leq} \epsilon$$

kann offenbar für vorgegebenes $\epsilon > 0$ durch die Wahl $\delta := \epsilon / k$ erfüllt werden.

Weiterführende Bemerkung: Es existieren Verallgemeinerung des obigen Kontraktionskriteriums vom Fall $V = \mathbb{R}$ für diff'bare Funktionen auf den Fall $V = \mathbb{R}^n$, und zwar in der Form $\sup_{x \in M} |||Jf(x)||| < 1$. Allerdings setzt das die Kenntnis von sogenannten *Matrix-Normen* $|||\cdot||| : \mathbb{R}^{n \times n} \to \mathbb{R}$ voraus; $|||M||| := \sup_{\vec{x} \in \mathbb{R}^n} \|M\vec{x}\| / \|\vec{x}\|$.

Nun der wohl wichtigste Satz über Fixpunkte:

Satz (Fixpunktsatz von Banach)

Sei $(V, \|\cdot\|)$ ein Banach-Raum und sei $\emptyset \neq M \subseteq V$ eine abgeschlossene Teilmenge.
Sei $\Phi : M \to V$ mit $\Phi(M) \subseteq M$ (kurz: $\Phi : M \to M$) eine Kontraktion.
Dann hat Φ genau einen Fixpunkt $x_* \in M$, und x_* ist Grenzwert der Folge

$$x_{n+1} := \Phi(x_n), \qquad x_0 \in M \text{ beliebig.}$$

Fehlerabschätzung: Ist k eine Kontraktionskonstante von Φ, so fällt der Approximationsfehler in jedem Iterationsschritt um mindestens den Faktor k:

$$\|x_{n+1} - x_*\| \leq k \, \|x_n - x_*\|, \qquad \text{somit} \qquad \|x_n - x_*\| \leq k^n \, \|x_0 - x_*\|$$

Bemerkungen:

- Der Satz ist *'konstruktiv'*: Er liefert nicht nur Existenz (u. Eindeutigkeit) des FP, sondern er verrät, wie man den FP (näherungsweise, iterativ, mit beliebig großer Genauigkeit) berechnet.
- Falls $M = V$, dann ist die Abgeschlossenheit von M sowie die Selbstabbildungseigenschaft $\Phi(M) \subseteq M$ trivialerweise erfüllt.
- Eine Variante der Fehlerabschätzung, die $\|x_n - x_*\|$ unter Verwendung von $\|x_n - x_{n-1}\|$ (anstelle von $\|x_0 - x_*\|$) abschätzt, ist möglich (ggf. s.Ü.) und nützlicher, da a priori x_* unbekannt ist.

Bevor wir zur Idee des Beweises, dass ein FP existiert, kommen, begründen wir die Fehlerabschätzung, die sehr schön die **Wirkungsweise der Kontraktionseigenschaft** verdeutlicht:

$$\| \overbrace{x_n}^{=\Phi(x_{n-1})} - \overbrace{x_*}^{=\Phi(x_*)} \| = \|\Phi(x_{n-1}) - \Phi(x_*)\| \overset{\text{(Kontr.)}}{\leq} k \, \|x_{n-1} - x_*\|$$

n-malige Anwendung ergibt die Fehlerabschätzung.

Beweisidee für den Fixpunktsatz von Banach:

1. Die Voraussetzung $\Phi : M \to M$ sorgt dafür, dass die Folge $x_{n+1} := \Phi(x_n)$, $x_0 \in M$, wohldefiniert ist.
   Man verwendet die Dreicksungleichung und (wie bei der Fehlerabschätzung, s.o.) die Kontraktionseigenschaft, um zu zeigen, dass (x_n) eine *Cauchy-Folge* ist: Für

$m > n$ ist

$$\begin{aligned}
\|x_m - x_n\| &\leq \|x_m - x_{m-1}\| + ... + \|x_{n+1} - x_n\| \\
&= \|\Phi(x_{m-1}) - \Phi(x_{m-2})\| + ... + \|\Phi(x_n) - \Phi(x_{n-1})\| \\
&\leq k\,\|x_{m-1} - x_{m-2}\| + ... + k\,\|x_n - x_{n-1}\| \\
&\vdots \\
&\leq k^{m-1}\|x_1 - x_0\| + ... + k^n\|x_1 - x_0\| \\
&\leq \sum_{i=n}^{\infty} k^i\,\|x_1 - x_0\| = k^n \sum_{i=0}^{\infty} k^i\,\|x_1 - x_0\| = k^n\,\tfrac{1}{1-k}\,\|x_1 - x_0\| \overset{?}{\leq} \epsilon
\end{aligned}$$

Für vorgegebenes $\epsilon > 0$ kann durch Wahl von hinreichend großem n dies $\leq \epsilon$ gemacht werden (dank $k < 1$). (x_n) ist also eine Cauchy-Folge.

2. Da $(V, \|\cdot\|)$ ein Banach-Raum, d.h. *vollständig* ist, hat die Cauchy-Folge einen *Grenzwert* in V. Den nennen wir x_*.

3. Da alle $x_n \in M$ sind und M abgeschlossen ist, muss auch der Grenzwert $x_* \in M$ sein.

4. Wir hatten uns am Anfang des Kap. bereits überlegt: Da Φ als Kontraktion stetig ist, muss der Grenzwert der FP-Iteration ein FP sein: $\Phi(x_*) = x_*$.

5. Eindeutigkeit des FP: Man führt die Annahme, dass es zwei Fixpunkte $x_* \neq \tilde{x}_*$ zu einem Widerspruch, s.Ü. □

1.7.2 Zusammenhang zu Nullstellenproblemen und Newton-Verfahren

Warum ist das Finden von Fixpunkten interessant?

Nun ja, Fixpunkte sind per se zunächst einmal nicht wirklich interessant.
Aber: Man kann gegeben Probleme häufig umformulieren zu einem Fixpunktproblem! Mit dem FP-Satz findet man den Fixpunkt, und hat damit das ursprüngliche Problem gelöst.

Ein sehr wichtiges Beispiel: Das Nullstellenproblem:

Sei $f : \mathbb{R} \to \mathbb{R}$ gegeben, f zumindest stetig, ggf. stetig diff'bar.
Gesucht ist $x_* \in \mathbb{R}$ mit $f(x_*) = 0$. Wandle das Nullstellenproblem für f in ein FP-Problem um:

Für vorgegebenes $f : \mathbb{R} \to \mathbb{R}$ suche ein $\Phi : \mathbb{R} \to \mathbb{R}$, so dass

$$f(x) = 0 \qquad \Longleftrightarrow \qquad \Phi(x) = x,$$

d.h. x_* ist Nullstelle von f wenn x_* FP von Φ ist.

Zu vorgegebenem f kann man sehr viele Funktionen Φ finden mit dieser Eigenschaft.
Die einfachste Vorgehensweise: Addiere x auf beiden Seiten:
$f(x_*)+x_*=x_*$; dies ist ein FP-Problem für $\Phi(x):=f(x)+x$.
Also: x_* ist Nullstelle von f genau dann wenn x_* Fixpunkt von Φ ist.
Schwierigkeit: Anwendung des FP-Satzes erfordert, dass Φ Kontraktion ist; $\Phi'(x) = f'(x)+1$; dies sollte betragsmäßig <1 sein auf M, d.h. auf $M=\mathbb{R}$ oder zumindest in einer Umgebung der gesuchten Nullstelle von f (auf welcher f auch noch Selbstabbildung ist). Das ist (nur) für einige wenige f der Fall.

Andere Möglichkeit: $-f(x_*)+x_*=x_*$; dies ist ein FP-Problem für $\Phi(x):=x-f(x)$.
Damit Anwendung des FP-Satzes funktioniert, ist notwendig, dass $|1-f'(x)|\le k<1$ auf $\mathbb{R}$ oder zumindest in einer Umgebung der gesuchten Nullstelle gilt.

(Offenbar muss man je nach f 'basteln' um ein solche FP-Formulierung zu finden, *die die Kontraktionseigenschaft erfüllt?*)

Etwas allgemeinerer Ansatz zur Umwandlung Nullstellenproblem→FP-Problem:

Multipliziere $f(x)=0$ zunächt mit einem $\alpha\neq 0$, und addiere dann x:

$$f(x)=0 \qquad \Longleftrightarrow \qquad \underbrace{x+\alpha f(x)}_{=:\Phi(x)} = x\,.$$

Die Erfüllung der Kontraktionseigenschaft von $\Phi(x) := x+\alpha f(x)$ erfordert, dass $|\Phi'(x)| = |1+\alpha f'(x)| \le k < 1$ sein muss, auf $\mathbb{R}$ oder zumindest auf einer Umgebung der Nullstelle von f.
Naheliegende Wahl von α:

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

Dann ist zumindest $\Phi'(x_*)=0$, und, wenn f' und somit Φ' stetig ist, ist $|\Phi'(x)|$ auch in einem Umgebung von x_* kleiner als 1.

Allerdings kennt man ja x_* a priori nicht, d.h. man könnte hier nur eine Näherung x für x_* nehmen.

Dazu ein **Beispiel**: Es soll eine Nullstelle x_* von $f(x):=x^2-8$ berechnet werden.
Das Nullstellenproble für f ist (s.o.) äquivalent zum Fixpunktproblem für $\Phi(x) := x+\alpha\, f(x)$ für $\alpha\neq 0$. Als Wert für α wäre (s.o.) $-\frac{1}{f'(x_*)}$ ideal. Wir wissen, dass $x_*\approx 3$ eine Näherung an eine Nullstelle ist. Also nehmen wir $\alpha := -\frac{1}{f'(3)} = -\frac{1}{6}$:

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

Wir führen die FP-Iteration mit diesem Φ und Startwert $x_0 := 3$ durch:

$x_1 = \Phi(x_0) = 3 - \frac{9-8}{6} = 2.8\bar{3}$
$x_2 = \Phi(x_1) = 2.8\bar{3} - \frac{2.8\bar{3}^2-8}{6} \approx 2.828707$
$x_3 = \Phi(x_2) \approx 2.828442929$
...

Die exakte Lösung wäre $x_* = \sqrt{8} = 2.828427125....$

Dass diese Folge gegen die gesuchte Nullstelle von f konvergieren muss, liegt am FP-Satz von Banach; dessen Voraussetzungen sind z.B. für $M := [2,3]$ in der Tat erfüllt; s. Tafel. Wir bekommen auf diesem M eine Kontraktionskonstante $k = \frac{1}{3}$.

Mit der bekannten exakten Lösung können wir auch die Approximationsfehler ermitteln:

n	x_n	$\lvert x_n - x_* \rvert$
0	3	0.17157
1	2.83333333	0.0049062
2	2.828703703	0.000276578
3	2.828442929	0.000015804
4	2.828428028	0.000000900
⋮	⋮	⋮

Die von uns konstrierte FP-Iteration konvergiert tatsächlich gegen die gesuchte Nullstelle von f (=FP von Φ) !

Bemerkung: Im obigen Fall waren wir in der Lage, rigoros zu überprüfen, dass für die von uns konstruierte PF-Iteration die Voraussetzungen des FP-Satzes erfüllt sind: Wir haben ein $M \subseteq \mathbb{R}$ mit $\Phi : M \to M$ angegeben, so dass Φ Kontraktionskonstante< 1 auf M hat. (Auf $M = \mathbb{R}$ ist Φ ganz sicher keine Kontraktion! In der Praxis kann es aber auch schwierig sein, dies nachzuweisen bzw. ein solches M konkret anzugeben; ggf. verzichtet man dann auf einen Beweis bzw. Angabe von M. Dann allerdings besteht Unsicherheit, wie der Startwert x_0, der im M zu liegen hat, zu wählen ist.)

Zur Konvergenzgeschwindigkeit: Die Theorie sagt voraus, dass der Fehler pro Iterationsschritt um den Faktor k=Kontraktionskonstante fällt; k haben wir oben berechnen können; diese Berechnung kann jedoch i.a. recht schwierig sein (erfordert M). Wir können jedoch (zumindest a posteriori) leicht eine "*asymptotische Kontraktionskonstante*" $k_* := |\Phi'(x_*)|$ ermitteln, die im Grenzwert $x_n \to x_*$ die Fehlerreduktion beschreiben sollte (sofern Φ' stetig ist).

Kontraktionskonstante und asymptotische Kontraktionskonstante

Für stetig diff'bares $\Phi : M \to \mathbb{R}$, $M \subseteq \mathbb{R}$ mit Fixpunkt $x_* \in M$ bezeichnet (s.o.) $k := \sup_{x \in M} |\Phi'(x)|$ die Kontraktionskonstante und $k_* := |\Phi'(x_*)|$ die asymptotische Kontraktionskonstante.

Es ist $k_* \leq k$. k gibt an, um wieviel sich der Fehler pro Iterationsschritt $x_{n+1} := \Phi(x_n)$ mindestens verringert ("worst case"): $|x_{n+1} - x_*| \leq k\,|x_n - x_*|$. Für hinreichend großes n, so dass $x_n \approx x_*$, gibt k_* i.a. einen besseren Schätzwert für die zu erwartende Fehlerreduktion pro Iterationsschritt ab als k: $|x_{n+1} - x_*| \approx k_*\,|x_n - x_*|$.

Für unser Φ hatten wir $k = \frac{1}{3}$ auf $M = [2,3]$, und es ist $k_* = |\Phi'(\sqrt{8})| = 1 - \frac{1}{3}\sqrt{8} \approx 0.057 \approx \frac{1}{17.5}$. Und in der Tat fällt der Fehler von x_1 zu x_2, von x_2 zu x_3, von x_3 zu x_4 um ziemlich genau diesen Faktor k_*! Lediglich der Schritt von x_0 nach x_1 fällt aus dem Rahmen; dort sind wir noch zu weit von x_* entfernt, als dass k_* Gültigkeit hätte.

Eine weitere Verbessung bei der Umwandlung NS-Problem→FP-Problem:
Anstatt mit einer Konstante $\alpha \neq 0$ zu multiplizieren:

Multipliziere $f(x) = 0$ mit einer *Funktion* $h(x) \neq 0$. Nach Addition von x erhalten wir

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

d.h. das Nullstellenproblem von f ist äquivalent zum FP-Problem für $\Phi(x) := x + h(x) f(x)$.

Zur Bestimmung eine geeigneten Funktion h wollen wir die asymptotische Kontraktionskonstante möglichst klein machen:
$k_* = |\Phi'(x_*)| = |1 + h'(x_*) \underbrace{f(x_*)}_{=0} + h(x_*) f'(x_*)| = |1 + h(x_*) f'(x_*)|$

Wir können dies sogar zu Null(!) machen, indem wir h so wählen, dass $h(x_*) \stackrel{!}{=} -\frac{1}{f'(x_*)}$ ist. Die Wahl $h(x) := \text{const} = -\frac{1}{f'(x_*)}$ hat, s.o., das Problem, dass man x_* a priori nicht kennt. Eine andere naheliegende Wahl für h ist deshalb $h(x) := -\frac{1}{f'(x)}$, somit

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

Die zugehörige FP-Iteration lautet

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

Dies ist nicht anderes als das uns wohlbekannte **Newton-Verfahren** (s. 2. Semester)!

Ergebnis:

Das Newton-Verfahren als Fixpunktiteration

Das Newton-Verfahren

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

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

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

Die Asymptotische Kontraktionskonstante dieser FP-Iteration ist 0.

Dass beim Newton-Verfahren, als FP-Iteration betrachtet, $k_* = 0$ ist, äußert sich in einer besonders schnellen Konvergenz des Newton-Verfahrens:

Dazu das folgende numerische Experiment:
$f(x) = x^2 - 8$, $\Phi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^2-8}{2x} = \frac{1}{2}x + \frac{1}{4x}$:

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	2.8284271247493798213	$3.18972 \cdot 10^{-12}$
4	2.8284271247461900976	$1.79859 \cdot 10^{-24}$
⋮	⋮	⋮

Deutung der Tabelle:
Während FP-Iterationen mit $0 < k_* < 1$ den Fehler pro Iterationsschritt in etwa um dem Faktor k_* verringern, d.h. $\|x_{n+1} - x_*\| \approx k_* \|x_n - x_*\|$ (zumindest asymptotisch für großes n), hat das Newton-Verfahren mit $k_* = 0$ eine Fehlerreduktion der Form $\|x_{n+1} - x_*\| \approx c \|x_n - x_*\|^2$ (s. 2.Sem.); zumindest wenn man $f \in C^2$ und $f'(x_*) \neq 0$ voraussetzt.

Man definiert:

Def. (lineare, quadratische Konvergenz)

Sei (x_n) eine Folge mit Grenzwert x_* (in $\mathbb{R}$ oder auch in einem normierten Vektorraum).
Gilt $\|x_n - x_*\| \leq k \|x_{n-1} - x_*\|$ für ein $k < 1$, so heißt das Verfahren *linear konvergent*.
Gilt $\|x_n - x_*\| \leq c \|x_{n-1} - x_*\|^2$ für ein $c > 0$, so heißt das Verfahren *quadratisch konvergent*.

Zusammenfassend:

Konvergenzverhalten von Fixpunkt-Iterationen

Fixpunktiterationen sind, sofern die Voraussetzungen des Fixpunktsatzes von Banach erfüllt werden, i.a. *linear konvergent*; der Fehler reduziert sich pro Iterationsschritt um mindestens den Faktor $k =$Kontraktionskonstante; ist man hinreichend nahe am Fixpunkt, so kann als gute Näherung die (oft leichter zu berechnende) *asymptotische Kontraktionsrate* k_* verwendet werden.

Ist für eine FP-Iteration die asymptotische Kontraktionsrate $k_* = 0$, so ist die Iteration (mindestens) quadratisch konvergent (ohne Beweis hier). Ein Beispiel dafür ist das Newton-Verfahren für $f \in C^2$ sofern $f'(x_*) \neq 0$.

Das Newton-Verfahren, das wir im 2. Semester motiviert hatten als "Nullstellen der *Tangente* nehmen, wenn man Nullstelle der Funktion nicht exakt berechnen kann, und dies iterieren" hat nun also, indem wir es als eine spezielle *Fixpunktiteration* identifiziert haben, eine weitere Motivation/Untermauerung gefunden.

Ein Überblick über **Anwendungen von FP-Iterationen**:

- Ist (wie oben gesehen) *eine* Möglichkeit, das Newton-Verfahren zu motivieren und zu untersuchen;
  z.B.: Um festzustellen, für welchen Anfangswert x_0 das Newton-Verfahren konvergiert, kann man prüfen, ob für ein M das $\Phi : M \to \mathbb{R}$ die Voraussetzungen des Fixpunktsatzes erfüllt, und dann $x_0 \in M$ wählen. Oder man kann prüfen, inwiefern das Newton-Verfahren bei doppelten Nullstellen anwendbar ist (s. Ü.).
- Kann auch auf *lineare* Gleichungssysteme angewendet werden, führt so zu iterativen Lösungsverfahren für LGS ($\to$ Kap. 1.7.4)
- Kap. 1.3: Der Beweis des Satzes über implizite Funktionen (somit der Beweis des Satzes über die inverse Abbildung), sowie die praktische, iterative Berechnung von Auflösungs- und Umkehrfunktionen beruht auf einer Fixpunkt-Iteration.

- Eine weitere Anwendung von FP-Iterationen kommt noch auf uns zu: Beweis von Existenz von Lösungen von Differentialgleichungen (Kap. 2.3)

Zuvor: Verallgemeinerung des Newton-Verfahrens von $\mathbb{R}$ auf $\mathbb{R}^n$, Kap. 1.7.3:

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

In der Praxis tauchen oft nichtlineare Gleichungs*systeme* auf, z.B. wenn wir zur Extremwertberechnung $\nabla F(\vec{x}) \stackrel{!}{=} \vec{0}$ lösen, oder wenn wir Lagrange-Systeme lösen.
Nichtlineare Systeme aus m Gleichungen können allgemein beschrieben werden in der Form

$$\vec{f}(\vec{x}) = \vec{0},$$

wobei $\vec{f} : \mathbb{R}^m \to \mathbb{R}^m$. (Falls auf der rechten Seite etwas $\neq \vec{0}$ steht, bringt man es auf die linke Seite.)
Dies ist ein *Nullstellenproblem* im $\mathbb{R}^m$.
Auch für dieses kann man ein Newton-Verfahren herleiten.
Als Motivation kann man, wie schon im skalaren Fall, entweder
(I.) mittels *Linearisierung* argumentieren (wie im 2. Sem.) oder
(II.) mittels Fixpunktformulierung (wie in Kap. I.7.2).

Zu Motivation I: Ganz analog zum skalaren Fall im 2.Sem.:
Wir verwenden die Taylor-Entwicklung um $x = x_n$, die aktuelle Iterierte

$$\vec{f}(\vec{x}) = \underbrace{\vec{f}(\vec{x}_n) + (Jf)(\vec{x}_n)\,(\vec{x}-\vec{x}_n)}_{=:\vec{T}_n(\vec{x}),\ \text{Linearisierung}} + \text{Restterm}$$

Da wir $\vec{f}(\vec{x}) = \vec{0}$ nicht exakt lösen können, setzen wir stattdessen die Linearisierung null; dies soll die neue Iterierte $\vec{x}_{n+1}$ werden, also $\vec{T}_n(\vec{x}_{n+1}) \stackrel{!}{=} \vec{0}$.
Dies führt, sofern $Jf(\vec{x}_n)$ invertierbar ist, auf die folgende Iteration:

$$\vec{x}_{n+1} := \vec{x}_n - [(Jf)(\vec{x}_n)]^{-1}\,\vec{f}(\vec{x}_n).$$

Zu Motivation II: Motivation mittels FP-Verfahren: Das Nullstellenproblem für $\vec{f}$ ist äquivalent zum FP-Problem für

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

(Existenz von $[(Jf)(\vec{x}_n)]^{-1}$ vorausgesetzt); d.h.

$$\vec{\Phi}(\vec{x}) = \vec{x} \quad \Leftrightarrow \quad \vec{f}(\vec{x}) = \vec{0}\,,$$

und die zugehörige asymptotische Kontraktionskonstante ist $k_* = 0$ (z.Ü.).

Die zu diesem Φ gehörende FP-Iteration ist offenbar

$$\vec{x}_{n+1} := \vec{\Phi}(\vec{x}_n) = \vec{x}_n - [(Jf)(\vec{x}_n)]^{-1}\,\vec{f}(\vec{x}_n)\,.$$

Newton-Verfahren im $\mathbb{R}^m$
Sei $\vec{f} : \mathbb{R}^m \to \mathbb{R}^m$ zweimal stetig differenzierbar, und es sei die Jacobi-Matrix $Jf(\vec{x})$ invertierbar für alle $\vec{x} \in \mathbb{R}^n$.
Dann lautet das Newton-Verfahren zur Bestimmung einer Nullstelle von $\vec{f}$:

$$\vec{x}_{n+1} := \vec{x}_n - [(Jf)(\vec{x}_n)]^{-1}\,\vec{f}(\vec{x}_n)\,.$$

Das Verfahren ist, wie im skalaren Fall, *lokal quadratisch* konvergent (d.h. es gibt eine Umgebung M der Nullstelle $\vec{x}_*$ derart, dass wenn der Startwert $\vec{x}_0$ aus dieser Umgebung gewählt wurde, dann ist das Verfahren quadratisch konvergent).

(Leider kennt man in der Praxis die Größe dieser Umgebung M i.a. nicht bzw. hat nur sehr komplizierte Formeln, die deren Größe angeben.)

Abschwächung der Voraussetzung: Die Anforderung an die Invertierbarkeit der Jacobi-Matrix für *alle* $\vec{x} \in \mathbb{R}^n$ kann man abschwächen: Es reicht, wenn diese für alle $\vec{x}$ aus einer Umgebung M der Nullstelle erfüllt ist, um lokal quadratische Konvergenz zu bekommen. Sogar das lässt sich weiter Abschwächen: Da wir $Jf \in C^1$ vorausgesetzt haben, reicht es, dass $Jf(\vec{x}_*)$ invertierbar ist, um auf Invertierbarkeit von $Jf(\vec{x})$ in einer ganzen Umgebung von $\vec{x}_*$ zu schließen.

Effiziente Durchführung eines Newton-Schritts: (für m groß)
Es ist nicht erforderlich, die Inverse von $Jf(\vec{x}_n)$ explizit zu berechnen; es reicht, ein LGS zu lösen: Dazu schreibt man die Iterationsvorschrift um zu $(Jf)(\vec{x}_n)(\vec{x}_{n+1} - \vec{x}_n) = -\vec{f}(\vec{x}_n)$. Indem man die Hilfsgröße $\vec{\Delta x} := \vec{x}_{n+1} - \vec{x}_n$ einführt, wird ein Newton-Schritt zu:

1. Löse das LGS $(Jf)(\vec{x}_n)\ \vec{\Delta x} = -\vec{f}(\vec{x}_n)$
2. Setze $\vec{x}_{n+1} := \vec{x}_n + \vec{\Delta x}$

Bemerkung: Somit führt das Newton-Verfahren das Lösen eines *nichtlinearen* Gleichungssystems zurück auf das Lösen *linearer* Gleichungssysteme.

1.7.4 Fixpunktverfahren für Lineare Gleichungssysteme

Motivation: Zum Lösen von LGS kennen wir aus dem 1. Semester das Gauß-Verfahren, das nach endlich vielen Schritten die Lösung liefert und offenbar universell einsetzbar ist (keine Anforderungen an LGS).
Weshalb interessiert man sich dann noch für iterative Verfahren zum Lösen von LGS?

Man kann zwei wesentliche Argumente anführen:

1. Der **Rechenaufwand** des Gauß-Verfahrens ist für große n nicht unerheblich.
   Für gewisse LGS (die in der Praxis häufig vorkommen) können iterative Verfahren schneller sein.

2. Für große n muss das Rechnen Computern, die mit Gleitkommaarithmetik arbeiten, überlassen werden. Zahlen werden also nur mit einer gewissen Stellenzahl, d.h. mit einem gewissen relativen Fehler, einem **Rundungsfehler** gespeichert, z.B. mit 16 Ziffern für die Mantisse, das entspricht einem relativen Fehler in der Größenordnung von 10^{-16}.
   Gewisse Rechenoperationen können allerdings den relativen Fehler drastisch erhöhen. Das klassische Beispiel ist die Bildung der Differenz zweier ähnlich großen Zahlen:

   $$0.7948236243749276 - 0.7948236243749241 = 0.0000000000000035 = 0.35 \cdot 10^{-14}$$

Sind also die beiden Zahlen mit 16-ziffriger Mantisse gegeben, dann ist ihre Differenz im obigen Beispiel mit 2 Ziffern bekannt, d.h. der relative Fehler ist von 10^{-16} auf etwa 10^{-2} abgewachsen! Dieser Effekt wird in der Numerik als *Auslöschung* bezeichnet.

Bei großem n sind zum Lösen des LGS derart viele Rechenoperationen erforderlich, dass das Auftreten des obigen Effektes nicht unwahrscheinlich ist.

Es gibt zwar Techniken, mit denen man durch Zeilenvertauschung (=:"Pivotisierung") in vielen Fällen diese Auslöschung vermeiden kann, jedoch sind auch LGS bekannt, in denen Auslöschung unvermeidbar ist beim Verwenden der Gauß-Elimination, und die numerisch berechnete Lösung nicht im entferntesten mit der exakten Lösung übereinstimmt.

Sind denn Fixpunktverfahren weniger anfällig für Auslöschungseffeke?
In der Tat: Selbst wenn Auslöschung auftreten sollte, dann sorgt die Kontraktionseigenschaft dafür, dass in jedem nachfolgende Iterationsschritt der Fehler mit einem Faktor $k<1$ gedämpft wird!

Unser Konzept zur Herleitung iterativer Verfahren für LGS:
Umwandlung eines LGS $A\vec{x}=\vec{b}$, $A\in\mathbb{R}^{n\times n}$, in ein FP-Problem:

Dazu gibt es, wie schon bei skalaren Problemen, unzählige Möglichkeiten; die Schwierigkeit steckt darin, dass die entstehende FP-Funktion die Voraussetzungen des Fixpunktsatzes (insbes.: Kontraktionseigenschaft!) zu erfüllen hat.

Wir können die Vorgehensweise, wie wir im skalaren Fall eine Gleichung auf FP-Form gebracht haben ('Mult. mit Konst $\alpha\neq 0$, Addition von x') übertragen auf LGS:
Multipliziere $\vec{b}-A\vec{x}=\vec{0}$ mit einer Matrix $M\in\mathbb{R}^{n\times n}$, addiere dann $\vec{x}$. Das entstehende FP-Problem lautet $\underbrace{(E_n-MA)\vec{x}+M\vec{b}}_{=:\vec{\Phi}(\vec{x})}=\vec{x}$, somit $\vec{\Phi}(\vec{x}) := (E_n-MA)\vec{x}+M\vec{b}$.
Die Erfüllung der Kontraktionseigenschaft erfordert, dass

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

gelten sollte für alle $\vec{x},\vec{y}\in\mathbb{R}^n$. Ideal wäre die Wahl $M := A^{-1}$; aber man möchte die Berechnung von A^{-1} natürlich vermeiden. Dieser Ansatz ist gangbar, falls man eine Näherung M an A^{-1} kennt.

Beispiel: Falls A die Form „A =Diagonalmatrix D plus kleiner Rest R“ hat, kann man $M := D^{-1}$ setzen (denn $A^{-1}\approx D^{-1}=M$), und dieses $M=D^{-1}$ ist trivial zu berechnen!)

Somit $\Phi(\vec{x}) = (E_n-MA)\vec{x}+M\vec{b} = (E_n-D^{-1}(D+R))\vec{x}+D^{-1}\vec{b} = -D^{-1}R\vec{x}+D^{-1}\vec{b}$, somit lautet die zugehörige FP-Iteration: $\vec{x}_{m+1} := \Phi(\vec{x}_m) = -D^{-1}R\vec{x}_m+D^{-1}\vec{b}$

Dieses Verfahren kann man (vielleicht anschaulicher?) auch mittels *komponentenweiser* Betrachtung herleiten:

Das Jacobi-Verfahren (=Gesamtschrittverfahren):
(Carl Gustav Jacob Jacobi, 1804-51)

Gegeben sei ein LGS $A\vec{x}=\vec{b}$, $A=(a_{ij})\in\mathbb{R}^{n\times n}$, $n\in\mathbb{N}$.
Es seien alle Diagonaleinträge $a_{ii}\neq 0$.

Zur Vereinfachung der Darstellung wollen wir $n=3$ verwenden.
Wir bringen das LGS auf FP-Form, indem wir in jeder der n Gleichungen den Diagonalterm auf die eine Seite und alle anderen Terme auf die andere Seite bringen:

$$\begin{array}{rcll} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &=& b_1 & \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &=& b_2 & \text{(LGS)} \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &=& b_3 & \end{array}$$

wird zu FP-Gleichung $\vec{x}=\Phi(\vec{x})$:

$$\begin{array}{rcll} x_1 &=& \frac{1}{a_{11}}(b_1 - a_{12}x_2 - a_{13}x_3) & \\ x_2 &=& \frac{1}{a_{22}}(b_2 - a_{21}x_1 - a_{23}x_3) & \\ x_3 &=& \frac{1}{a_{33}}(b_3 - a_{31}x_1 - a_{32}x_2) & \text{(FP)} \\ \underbrace{}_{\vec{x}} & & \underbrace{\phantom{\frac{1}{a_{11}}(b_1 - a_{12}x_2 - a_{13}x_3)}}_{=:\vec{\Phi}(\vec{x})} & \end{array}$$

Diese FP-Form kann man auch mittels Matrizen schreiben: Ist $D:=\mathrm{diag}(a_{ii})$ der Diagonalanteil von A, und $R:=A-D$ der 'Rest' von A, dann is obiges $\vec{\Phi}$ offenbar
$\vec{\Phi}(\vec{x}) = D^{-1}\vec{b}-D^{-1}R\vec{x}$
Dies stimmt übrigens mit dem Resultat von der vorigen Folie überein!

Jacobi-Verfahren

Die FP-Iteration $\vec{x}_{m+1} := \vec{\Phi}(\vec{x}_m)$ mit dem oben konstruierten $\vec{\Phi}$ heißt *Jacobi-* oder *Gesamtschrittverfahren*.
Formelmäßig lautet ein Iterationsschritt

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

bzw. $\vec{x}_{m+1} := D^{-1}(\vec{b}-R\vec{x}_m)$, mit D, R wie oben.

Beispiel: Für das LGS $\begin{pmatrix} 4 & 1 \\ -1 & 2 \end{pmatrix}\vec{x}=\begin{pmatrix} 6 \\ 4 \end{pmatrix}$, dessen exakte Lösung $\vec{x}_*=(8/9, 22/9)^T = (0.\overline{8}, 2.\overline{4})^T$ ist, führe die ersten 4 Schritte des Jacobi-Verfahrens durch mit Startwert $\vec{x}_0:=\vec{0}$.
Bestimme in jedem Schritt den Fehler in der Maximums-Norm $\|\cdot\|_\infty$.
Versuche auch, eine Kontraktionskonstante bezüglich der ($\|\cdot\|_\infty$-Norm) zu berechnen.
Siehe Tafel.

Konvergenz des Jacobi-Verfahrens:

Kann man für *irgendeine beliebige* Norm $\|\cdot\|$ des $\mathbb{R}^n$ eine Kontraktionskonstante $k<1$ finden (also $\|\vec{\Phi}(\vec{x})-\vec{\Phi}(\vec{y})\|\leq k\|\vec{x}-\vec{y}\|\ \forall\,\vec{x},\vec{y}\in\mathbb{R}^n$), so folgt die Konvergenz der FP-Folge zunächst in der betreffenden Norm; da aber im $\mathbb{R}^n$ alle Normen äquivalent sind, folgt sogar Konvergenz der FP-Folge in *beliebigen* Normen.

Veranschaulichung: Sei $k_1=1.3>1$ die kleinstmögliche Kontraktionskonstante bzgl der $\|\cdot\|_\infty$-Norm und $k_2=0.9<1$ eine Kontraktionskonstante bzgl der $\|\cdot\|_2$-Norm. Dann ist die FP-Folge nicht nur bzgl. der $\|\cdot\|_2$-Norm konvergent, sondern auch bzgl jeder anderen Norm, also auch der $\|\cdot\|_\infty$-Norm!

Man kann folgendes hinreichendes Konvergenzkriterium finden: Obiges $\vec{\Phi}$ ist eine Kontraktion bzgl. der $\|\cdot\|_\infty$-Norm auf ganz $\mathbb{R}^n$, d.h. das Jacobi-Verfahren ist konvergent gegen den FP, also die Lösung des LGS, in beliebiger Norm, falls die Systemmatrix A *diagonaldominant* ist.

Eine Matrix $A\in\mathbb{R}^{n\times n}$ heißt *(zeilenweise) diagonaldominant*, falls

$$|a_{ii}|>\sum_{j\in\{1,\dots,n\}\setminus\{i\}}|a_{ij}|\qquad\forall\,i=1,\dots,n$$

Obiges Konvergenzkriterium kann man noch abschwächen: Falls A die obige Bedingung nur für mindestens ein i erfüllt, und für die übrigen i diese Bedingung mit "$\geq$" statt "$>$" erfüllt (man nennt A dann *schwach diagonaldominant*), und falls das LGS durch eventuelles Umordnen von Zeilen/Spalten nicht zum Zerfallen in kleinere voneinander unabhängige Lineare Gleichungsysteme gebracht werden kann, dann hat das Jacobi-Verfahren weiterhin eine Kontraktionsrate<1 bzgl. der $\|\cdot\|_\infty$-Norm, ist also konvergent.
Man kann analog auch den Begriff der *spaltenweisen* Diagonaldominanz definieren; aus dieser folgt, dass die Kontraktionsrate bzgl. der $\|\cdot\|_1$-Norm <1 ist, ist somit ebenfalls hinreichend für Konvergenz des Jacobi-Verfahrens.

Bemerkung zur Diagonaldominanz:
Obige Anforderung, dass die Systemmatrix diagonaldominant oder schwach diagonaldominant sei, mag auf den ersten Blick seltsam und sehr restriktiv wirken.
In praktischen Anwendungen kommt dieser Fall jedoch sehr häufig vor: Für eine große Klasse von Partiellen Differentialgleichungen (darunter fällt insbesondere die Wärmeleitungsgleichung), führt eine 'Diskretisierung' i.a. auf ein diagonaldominantes oder schwach diagonaldominantes LGS.
Ggf. s. Übung.

Weiterführende Bemerkung zur Effizienz(-steigerung) des Jacobi-Verfahrens:

In den mittels Diskretisierung von Partiellen Differentialgleichungen entstehenden

Linearen Gleichungssystemen ist zwar das Jacobi-Verfahren i.a. konvergent, jedoch ist die Konvergenzgeschwindigkeit meist leider recht niedrig, d.h. die Kontraktionskonstante ist nur ganz knapp unter 1.

Um die Konvergenzgeschwindigkeit zu steigern (die Kontraktionskonstante zu verkleinern) verwendet man eine sog. **Relaxation**:
Man schreibt den Iterationsschritt des Jacobi-Verfahrens als

$$\vec{x}_{m+1} = \vec{\Phi}(\vec{x}_m) =: \vec{x}_m + \vec{\Delta x}_m,$$

und fügt nun einen Parameter $\omega \in \mathbb{R}$ ein, d.h. beim *relaxierten Jacobi-Verfahren, "JOR" (=Jacobi over-relaxation)* iteriert man

$$\vec{x}_{m+1} := \vec{x}_m + \omega\, \vec{\Delta x}_m.$$

einen günstigen Wert für ω kann man theoretisch findet, wenn man die Eigenwerte von A gut kennt. Falls das nicht der Fall ist, kann man gute Werte für ω durch Herumprobieren finden. Die Theorie sagt, dass man grundsätzlich $\omega \in (0, 2)$ zu wählen hat.

Das Gauß-Seidel-Verfahren (=Einzelschrittverfahren)

(Philipp von Seidel, 1821-96; Carl Friedrich Gauß 1777-1855)

Das Gauß-Seidel-Verfahren entsteht aus dem Jacobi-Verfahren durch folgende Überlegung:

Beim Jacobi-Verfahren wird zur Berechnung der i-Komponente des neuen Vektors $\vec{x}_{m+1}$ die Komponenten $j = 1, ..., i-1, i+1, ..., n$ des alten Vektors $\vec{x}_m$ verwendet:

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

Bei der Berechnung der i-ten Komponente des neuen Vektors sind jedoch schon die Komponenten $j = 1, ..., i-1$ des neuen Vektors bekannt. Wenn man davon ausgeht, dass die neuen Komponenten bessere Approximationen als die alten Komponenten sind, dann sollte man, wo immer möglich, *neue* Komponenten anstelle der alten verwenden.

Das führt auf das *Gauß-Seidel-* oder *Einzelschrittverfahren*:

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

Bemerkungen:

- Beachte, dass auf der rechten Seite wirklich nur bereits bekannte Größen stehen!

– Das Gauß-Seidel-Verfahren lässt sich einfacher (speicherplatzsparender) implementieren als das Jacobi-Verfahren:
Bei Berechnung der i-ten Komponente von $\vec{x}_{m+1}$ kann man die i-te Komponente von $\vec{x}_m$ einfach überschreiben; ihr Wert wird nicht mehr gebraucht.

– Man kann also den Iterationsschritt implementieren als

$$\begin{array}{l} \text{für } i = 1, ..., n : \\ x[i] := \frac{1}{a[i,i]} \left(b[i] - \sum_{j=1}^{i-1} a[i,j]\, x[j] - \sum_{j=i+1}^{n} a[i,j]\, x[j] \right) \end{array}$$

wohingegen man beim Jacobi-Verfahren zumindest ein Array x_neu und ein Array x_alt braucht.

– Ist das Gauß-Seidel-Verfahren eine Fixpunktiteration?
In Matrixschreibweise lautet obige Iterationsvorschrift $\vec{x}_{m+1} := D^{-1}(\vec{b} - L\vec{x}_{m+1} - U\vec{x}_m)$, wobei $A = L + D + U$ in den lower/diagonal/upper- Anteil zerlegt wurde. Auch das Gauß-Seidel-Verfahren kann man (zum Zwecke von theoretischen Untersuchungen (Konvergenz?)) als FP-Iteration $\vec{x}_{m+1} := \vec{\Phi}(\vec{x}_m)$ auffassen. Um das $\vec{\Phi}$ zu ermitteln, muss man die Iterationsvorschrift nach $\vec{x}_{m+1}$ auflösen.

– Auch für das Gauß-Seidel-Verfahen gibt es zur Konvergenzverbesserung Relaxation. Das relaxierte Gauß-Seidel-Verfahren wird *SOR-Verfahren* ('successive overrelaxation') genannt.

– In der Praxis zeigt sich, dass – wie in der Motivation angedeutet – **das Gauß-Seidel-Verfahren etwas schneller konvergiert als das Jacobi-Verfahren**. Die theoretische Untersuchung ist jedoch für das Jacobi-Verfahren einfacher als fürs Gauß-Seidel-Verfahren.

– **Zwei verschiedene Klassen von Iterativen Verfahren für LGS, Weiterführendes**

Oben haben wir auf dem *Fixpunkt-Konzept* beruhende iterative Lösern (Jacobi-, Gauß-Seidel-Verfahren) kennengelernt.
Sie finden in der modernen numerischen Mathematik z.B. Anwendung in sogenannten Mehrgitterverfahren (multigrid methods).

In Kap. 1.5 hatten wir zuvor ein ganz anderes iteratives Verfahren, das Gradienten-Verfahren, kennengelernt, zum Lösen von LGS mit s.p.d.-Matrix. Das Verfahren gehört zur Klasse der *Abstiegs-Verfahren*;
in der Literatur findet man weitere 'ähnliche' und modernere Verfahren, z.B. das sog. cg-Verfahren, oder, als Verallgemeinerung, sog. Krylov-Raum-Verfahren, die z.T. sogar auf nicht-s.p.d.-Matrizen anwendbar sind.

– Sind denn nun direkte Verfahren oder iterative Verfahren effizienter zum Lösen von LGS?

Das ist schwer pauschal zu beantworten.

Für "voll-besetzte" Matrizen (d.h. Matrizen, in denen kaum Nullen vorkommen) haben direkte Verfahren den Aufwand $O(n^3)$, und iterative Verfahren meist pro Zeitschritt den Aufwand $O(n^2)$ ($\rightsquigarrow$ Aufwand Matrix-Vektor-Multiplikation); d.h. wenn man ein iteratives Verfahren hat, das weniger als $O(n)$ Iterationsschritte braucht, so ist dieses effizienter als Gauß-Elimination.

In der Praxis (Diskretisierung von Partiellen Differentialgleichungen) hat man es aber meist mit *dünn besetzten* Matrizen zu tun (d.h. die allermeisten Matrixeinträge sind Nullen), die häufig zudem (schwach) diagonaldominant sind.
Bis ca.(!!) $n \approx 10^4$ sind für solche LGS meist direkte Verfahren am effizientesten. Bei der Diskretisierung von Partiellen Differentialgleichungen kommt es jedoch durchaus zu LGS der Größe $n = 10^6$ oder $n = 10^8$. Diese werden niemals mit direkten Verfahren gelöst; sowohl der Rechenaufwand, als auch der Speicherplatzbedarf wären da problematisch (viele Nullen werden bei Gauß-Elimination 'aufgefüllt'!), und iterative Verfahren sind überlegen (hinsichtlich Rechenaufwand, Speicherplatzbedarf, aber auch hinsichtlich Vermeidung/Dämpfung von Rundungsfehlern). Geht man davon aus, dass pro Zeile $O(1)$ viele Nicht-Null-Einträge vorkommen, und dass die sog. Bandbreite b der Matrix $O(n^{\frac{1}{2}})$ ist, so hat die Gauß-Elimination bei schwach besetzten Matrizen den Aufwand $O(nb^2) = O(n^2)$, und iterative Löser haben pro Iterationsschritt(!) den Aufwand $O(n^1)$.

2 Gewöhnliche Differentialgleichungen

2.1 Einführung, Beispiele, grobe Klassifizierung

Motivation: Beispiele für skalare gewöhnliche Differentialgleichungen

Ein Beispiel aus der Mechanik: Federschwinger (Oszillator)

Eine Masse m sei an einer elastischen Feder aufgehängt. Gesucht ist der Ort $x(t)$ der Masse zum Zeitpunkt t. Der Nullpunkt sei so festgelegt, dass die 'Ruhelage' gerade der Nullpunkt der Skala, $x=0$, ist.

Mathematische Modellierung: Nach dem zweiten Newton'schen Gesetz ist die zeitliche Änderung des Impulses I gleich der angreifenden Kraft F: $\frac{d}{dt}I = F$.
Der Impuls ist Masse mal Geschwindigkeit: $I = mv = m\,\frac{dx}{dt}$.
Wir erhalten:

$$\frac{d}{dt}\Big(m\,\frac{dx}{dt}\Big) = F$$

Da die Masse konstant ist:

$$m\frac{d^2x}{dt^2} = F$$

Wir brauchen nun ein Modell für die Kraft F.

Die Feder übt eine Kraft F_F auf die Masse aus. Unter Annahme, dass diese Kraft proportional zur Auslenkung und dieser entgegengesetzt ist (das sog. *Hooke'sche Gesetz*), und dass dies die einzig wirkende Kraft ist (also keine Reibungskräfte), bekommen wir $F = F_F = -kx$, wobei $k>0$ die sog. *Federkonstante* ist.
Wir erhalten das Modell

Mathematisches Modell eines Federschwinger (reibungsfrei)

$$m\,x''(t) + k\,x(t) = 0$$

m und k sind gegebene Zahlen, und gesucht ist eine Funktion(!) $t \mapsto x(t)$, die diese Gleichung erfüllt.

Eine Verfeinerung des Modells: Berücksichtigung einer Reibungskraft F_R.
Es wirkt also eine Gesamtkraft $F = F_F + F_R$, wobei F_F wie oben.
Für die F_R sind verschiedene Modelle möglich, je nach Anwendung.
In jedem Fall sollte die Reibungskraft der Bewegung entgegengesetzt sein, d.h. F_R und x' sollten immer entgegengesetzte Vorzeichen haben.
Ein einfaches Reibungsmodell (das in Fluiden bei moderaten Geschwindigkeiten (näherungsweise) gültig ist): F_R ist proportional zur Geschwindigkeit und dieser entgegengesetzt: $F_R = -cx'$ mit der Reibungskonstante $c>0$.

Wir bekommen das Modell

Mathematisches Modell eines Federschwinger (mit linearer Reibung)

$$m\,x''(t) + c\,x'(t) + k\,x(t) = 0$$

Bemerkung: Ein völlig anderes physikalisches System wird durch exakt die gleiche Gleichung beschrieben: Ein *Schwingkreis* (=Stromkreis mit Kondensator, Spule und Ohm'schem Widerstand); dann entspricht m der sog. Induktivität, c dem Ohm'schen Widerstand, $1/k$ der sog. Kapazität, x der Stromstärke.

Ein Beispiel aus der Biologie oder Soziologie oder Medizin: Sei $x(t)$ die Größe einer Population zum Zeitpunkt t.
Ein einfaches Modell: Die Wachstumsrate (d.h. die Änderung der Populationsgröße pro Zeitintervall $\frac{\Delta x}{\Delta t}$ im Grenzwert Zeitintervall $\Delta t \to 0$) sei proportional zu Populationsgröße ($k \in \mathbb{R}$ sei der Proportionalitätsfaktor):
$\frac{\Delta x}{\Delta t} = k\,x \quad \rightsquigarrow \quad \frac{x(t+\Delta t)-x(t)}{\Delta t} = k\,x(t) \quad \rightsquigarrow \quad x'(t) = k\,x(t)$

Modell des exponentiellen Wachstums

$$x'(t) = k\,x(t)$$

(Bemerkung: Im Fall $k < 0$ hat man ein Modell für den radioaktiven Zerfall eines Isotops.)

Lösen: Man kann erraten, dass $x(t) = x_0\,e^{kt}$ eine Lösung ist; x_0 beliebig. Für $k > 0$ wächst die Population somit über alle Schranken, und zwar recht schnell.

In der realen Welt treten irgendwann Effekte, die man als 'Resourcenknappheit' bezeichnen kann, ein. Die Wachstumsrate k sollte, zumindest für große x, nicht mehr als Konstante modelliert werden, sondern man kann eine Abhängigkeit von k von x postulieren.

Ein mögliches Modell: Es gibt eine Population $\bar{x}$, und für $x(t) < \bar{x}$ kann die Population wachsen, für $x(t) > \bar{x}$ jedoch ist die Wachstumsrate negativ. Genauer: Die Wachstumsrate sei proportional zu $\bar{x} - x(t)$. Dies liefert:

Modell des begrenzten Wachstums, 'logistische Differentialgleichung'

$$x'(t) = k\,(\bar{x} - x(t))\,x(t) = k\bar{x}\,x(t) - k\,x(t)^2$$

Die obigen Gleichungen sind Beispiele für sog. *gewöhnliche Differentialgleichungen*.

eine gDgl beschreibt einen Zusammenhang einer gesuchten Funktion(!) und ihrer/n Ableitung(en). Dabei hängt die gesuchte Funktion nur von *einem* skalaren Argument ab:

Def. (gewöhnliche Differentialgleichung)
Sei $n \in \mathbb{N}$. Eine Gleichung der Form $F(t, y(t), y'(t), ..., y^{(n)}(t)) = 0$, wobei F gegeben ist (definiert auf $\mathbb{R}^{n+2}$ oder auf einer geeigneten Teilmenge davon) und die Funktion $y : \mathbb{R} \to \mathbb{R}$ oder auch $y : [a, b] \to \mathbb{R}$ gesucht ist, heißt *(skalare) gewöhnliche Differentialgleichung*, 'gDgl' (engl.: *ordinary differential equation, ODE*) der Ordnung n.

Bemerkung: Das Wort "gewöhnlich" wird hinzugefügt, um solche Dgln von sog. *partiellen Dgln.* zu unterscheiden:
Falls man eine Funktion y sucht, die von *mehreren* skalaren Argumenten x_i abhängt, und somit partielle(!) Ableitungen $\frac{\partial y}{\partial x_i}$ vorkommen, handelt es sich nicht um eine gewöhnliche, sondern um eine sog. *partielle Differentialgleichung (pDgl, PDE).*

Zu den Beispielen: Das Modell des Federschwingers ist eine Dgl. 2. Ordnung, das Populationsmodell eine Dgl. 1. Ordnung.

Obige Definition ist *sehr* allgemein.
Die 'meisten' Dgln kann man nach der höchsten vorkommenden Ableitungsordnung auflösen, und nur mit solchen werden wir uns beschäftigen.
Solche Dgl. bezeichnet man als *explizit*:

Def. (explizite gewöhnliche Differentialgleichung)
Sei $n \in \mathbb{N}$. Eine Gleichung der Form $y^{(n)}(t) = f(t, y(t), y'(t), ..., y^{(n-1)}(t))$, wobei f gegeben und $t \mapsto y(t)$ gesucht ist, heißt *explizite gewöhnliche Differentialgleichung* der Ordnung n.

Der wohl häufigste Typ von Dgln. (oben $n = 1$ setzen):

explizite Dgl. *erster* Ordnung

$$y'(t) = f(t, y(t))$$

Eine Dgl., in der der das "t" nur als Argument der gesuchten Funktion und ihrer Ableitungen vorkommt, bezeichnet man als *autonom*:

Def. (autonome (explizite, gewöhnliche) Differentialgleichung)
Sei $n \in \mathbb{N}$. Eine Gleichung der Form

$$y^{(n)}(t) = f(y(t), y'(t), ..., y^{(n-1)}(t)),$$

wobei f gegeben und $t \mapsto y(t)$ gesucht ist, heißt *autonome (explizite, gewöhnliche) Differentialgleichung*. Insbesondere im Fall $n = 1$:

$$y'(t) = f(y(t))$$

Obige Beispiele sind allesamt explizit und autonom!

***Systeme* von Dgln., Motivation:**

Bei obigen Beispielen und Definition ging es um "skalare" Dgln, d.h. es wurde *eine* Funktion $t \mapsto y(t)$ gesucht.

In der Praxis sind oft *mehrere* skalare Funktionen gesucht, und mehrere Gleichungen gegeben, die den Zusammenhang zwischen den gesuchten Funktionen und ihren Ableitungen beschreiben.

Ein berühmtes **Beispiel**:

Das Volterra-Lotka-Modell (Räuber–Beute–Modell, engl.: predator–prey model; 1925/26)
Unter gewissen Annahmen gehorchen eine Raubtierpopulation $x(t)$ und eine Beutetierpopulation $y(t)$ den Gleichungen

$$\begin{aligned} x'(t) &= -\alpha\, x(t) + \beta\, x(t)\, y(t) \\ y'(t) &= \gamma\, y(t) - \delta\, x(t)\, y(t) \end{aligned}$$

wobei $\alpha, \beta, \delta, \gamma > 0$ Konstanten sind.

Erläuterungen: s. Tafel, ggf. auch s. Übung

Obiges *System von Dgln* kann man ebenfalls formal als $\vec{y}\,'(t) = \vec{f}(t, \vec{y}(t))$ schreiben, indem man die gesuchten skalaren Funktionen $t \to x(t)$ und $t \to y(t)$ zu einer *vektoriellen* Größe $t \to \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} =: \vec{y}(t)$ zusammenfasst, sowie $\vec{f}(t, \vec{y}) = \vec{f}(\vec{y}) = \begin{pmatrix} -\alpha y_1 + \beta y_1 y_2 \\ \gamma y_2 - \delta y_1 y_2 \end{pmatrix}$.

Weiteres **Beispiel** für ein System von Dgln. aus der **Chemie**:

Drei chemische Stoffe X_1, X_2, X_3 mit Konzentrationen $c_1(t), c_2(t), c_3(t)$ (in mol/l) in einem Reagenzglas reagieren miteinander gemäß der "chemischen Gleichung"

$$X_1 + 2X_2 \longleftrightarrow X_3.$$

Modell für die Reaktionsrate: Die Reaktionsrate sei proportional zur Begegnungswahrscheinlichkeit der reagierenden Moleküle. Die Begegnungswahrscheinlichkeit sei wiederum proportional zum Produkt der Konzentrationen. Wir bekommen für die Reaktion von links nach rechts bzw. für die Reaktion von rechts nach links die beiden Raten $R_{\rightarrow}(c_1, c_2) = k_{\rightarrow} c_1 c_2^2,\ R_{\leftarrow}(c_3) = k_{\leftarrow} c_3$ mit Konstanten $k_{\rightarrow}, k_{\leftarrow} > 0$.
Somit für die Gesamtrate (von links nach rechts):

$$R(c_1, c_2, c_3) := R_{\rightarrow}(c_1, c_2) - R_{\leftarrow}(c_3) = k_{\rightarrow} c_1 c_2^2 - k_{\leftarrow} c_3.$$

Bedenkt man, dass bei Verbrauch von 1 Molekül des Stoffs X_1 ein Molekül des Stoffs X_3 entsteht, und gleichzeitig 2 Moleküle des Stoffs X_2 verbraucht werden, bekommt man die Differentialgleichungen

$$\begin{array}{rcrcl} c_1'(t) &=& -R(c_1(t), c_2(t), c_3(t)) &=& -k_{\rightarrow} c_1 c_2^2 + k_{\leftarrow} c_3 \\ c_2'(t) &=& -2R(c_1(t), c_2(t), c_3(t)) &=& -2k_{\rightarrow} c_1 c_2^2 + 2k_{\leftarrow} c_3 \\ c_3'(t) &=& +R(c_1(t), c_2(t), c_3(t)) &=& k_{\rightarrow} c_1 c_2^2 - k_{\leftarrow} c_3 \end{array}$$

Weiteres **Beispiel** für ein System von Dgln. aus der **Mechanik/Astronomie: Das Dreikörperproblem**:

Szenario: Drei Himmelskörper; jeder ist der Schwerkraft der beiden anderen unterworfen.
Gegeb.: Die Massen $m_S, m_E, m_M > 0$, sowie die Gravitationskonstante G ($=6.67 \cdot 10^{-11} m^3 kg^{-1} s^{-2}$).
Gesucht: Die Positionen $t \mapsto \vec{x}_S(t), \vec{x}_E(t), \vec{x}_M(t)$ als Funktion der Zeit t

Die Gravitationskraft zweier Massen m_1, m_2 im Abstand $\|\vec{d}\|$ ist proportional zu beiden Massen und (betragsmäßig!) umgekehrt proportional zum Quadrat der Entfernungen; die Proportionalitätskonstante ist G, und die Kraft hat die gleiche Richtung wie der vektorielle Abstand:

Gravitationsgesetz, skalar und vektoriell

$$\|\vec{F}\| = \frac{G\, m_1 m_2}{\|\vec{d}\|^2}\,; \qquad \vec{F} = \pm\, G\, m_1 m_2 \frac{\vec{d}}{\|\vec{d}\|^3}$$

Somit bekommen wir, per "Masse mal Beschleunigung = angreifende Kraft", das Dgl-System

$$\begin{array}{rcl} m_S \vec{x}_S'' &=& G\, m_S m_E \frac{\vec{x}_E - \vec{x}_S}{\|\vec{x}_E - \vec{x}_S\|^3} + G\, m_S m_M \frac{\vec{x}_M - \vec{x}_S}{\|\vec{x}_M - \vec{x}_S\|^3} \\ m_E \vec{x}_E'' &=& G\, m_E m_S \frac{\vec{x}_S - \vec{x}_E}{\|\vec{x}_S - \vec{x}_E\|^3} + G\, m_E m_M \frac{\vec{x}_M - \vec{x}_E}{\|\vec{x}_M - \vec{x}_E\|^3} \\ m_M \vec{x}_M'' &=& G\, m_M m_S \frac{\vec{x}_S - \vec{x}_M}{\|\vec{x}_S - \vec{x}_M\|^3} + G\, m_M m_E \frac{\vec{x}_E - \vec{x}_M}{\|\vec{x}_E - \vec{x}_M\|^3} \end{array}$$

Dieses ist von zweiter Ordnung und besteht aus 9(!) skalaren Gleichungen und gesuchten skalaren Funktionen $x_{S,i}, x_{E,i}, x_{M,i}$, $i = 1, 2, 3$.

Bemerkungen zum obigen Beispiel: Außer in wenigen Spezialfällen kann man die Lösung des Systems nicht explizit angeben. Beim *Zweikörperproblem* kann man dagegen die Bahnkurven angeben; diese sind sog. Kegelschnitte (Kreise, Ellipsen, Parabeln, Hyperbeln).

Im Fall von Sonne, Erde, Mond haben wir $m_S = 1.99 \cdot 10^{30} kg$, $m_E = 5.97 \cdot 10^{24} kg$, $m_M = 7.35 \cdot 10^{22} kg$.

Weiteres Beispiel (aus Kap. 1.4):
Gegeben eine Parametrisierung $\vec{\gamma} : I \to \mathbb{R}^n$ einer Kurve. Gesucht: Die Parametrisierung nach der Bogenlänge $\vec{\gamma}_0$. Diese lässt sich als $\vec{\gamma}_0 = \vec{\gamma} \circ u$ schreiben. Für u hatten wir in Kap. 1.4 die Dgl

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

hergeleitet.

Anfangswertprobleme

Frage: Ist die Lösung einer Dgl. i.a. eindeutig bestimmt?

Dazu die heuristischen Überlegungen:
Anschauung: Obige Dgl. beschreiben oft die Bewegung von Objekten. Um die Bahnkurven eindeutig festzulegen, kann es unmöglich reichen, die Kräfte (also die Dgl.), die die Bewegung beeinflussen, zu kennen, sondern man muss in irgend einer Form einen **Anfangszustand** (der Ort z.B.?) festlegen!

Auch bei unserem einfachen Populationsmodell $y'(t) = k\,y(t)$ konnten wir durch Herumprobieren bereits eine *Schar*(!) von Lösungen $y(t) = c\,e^{kt}$, mit $c \in \mathbb{R}$ beliebig, finden.
Um hier zu einer *eindeutigen* Lösung zu kommen, können wir die Dgl. kombinieren mit der Vorgabe eines sog. **Anfangswerts** $y(t_0) \stackrel{!}{=} y_0$; $t_0, y_0 \in \mathbb{R}$ vorgegeben.
Mit Hilfe der Anfangsbedingung kann der freie Parameter c der allgemeinen Lösung bestimmt werden:

$$y_0 \stackrel{!}{=} y(t_0) = c\,e^{kt_0} \quad \Rightarrow \quad c = y_0\,e^{-kt_0} \quad \Rightarrow \quad y(t) = y_0\,e^{k(t-t_0)}$$

Wir bezeichnen die Lösungsschar $y(t) = c\,e^{kt}$, mit $c \in \mathbb{R}$, als **Allgemeine Lösung der Dgl.** und $y(t) = y_0 e^{k(t-t_0)}$ als die (hier zumindest) *eindeutig bestimmte* Lösung des **Anfangswertproblems (AWP)** $y'(t) = k\,y(t),\ y(t_0) \stackrel{!}{=} y_0$.

Frage: Wie kann man das verallgemeinern? Muss man immer genau eine Anfangsbedingung der Form $y(t_0) \stackrel{!}{=} y_0$ zur Dgl. hinzufügen, so dass dann die Lösung eindeutig

bestimmt ist?

Dazu werden später einen Satz kennenlernen ("Satz von Picard-Lindelöf", Kap. 2.3), der besagen wird:
Unter "harmlosen" Anforderungen an f ist eine skalare Dgl. erster(!) Ordnung, aber auch ein Dgl.-System erster(!) Ordnung $y'(t) = f(t, y(t))$ mit einer Anfangsbedingung der Form $y(t_0) = y_0$ bzw. $\vec{y}(t_0) = \vec{y}_0$ auszustatten; das AWP hat dann genau eine Lösung.

Dies motiviert die folgende Begriffsbildung:

Anfangswertproblem erster Ordnung

Unter einem *Anfangswertproblem (AWP) erster Ordnung* verstehen wir eine Dgl. (skalar oder System) erster Ordnung

$$y'(t) = f(t, y(t))$$

zusammen mit einer Anfangsbedingung (einem Anfangswert)

$$y(t_0) = y_0\,,$$

wobei t_0, y_0 gegeben sind.

Dgln und AWP *höherer* Ordnung

Wir haben eben motiviert, dass Dgln und Dgl-Systeme *erster* Ordnung die Vorgabe von genau einem Anfangswert für die gesuchte Funktion $t \to y(t)$ erfordern.

Frage: Gilt dies auch für Dgln. *höherer* Ordnung?

Ein erster Hinweis:
Betrachte die "Dgl" $y''(t) = e^t$ (die rechte Seite ist unabhängig von y; in diesem Sinne ist die Dgl. recht trivial)
Wir können sie durch zweimaliges "Aufintegrieren" lösen:

$$y''(t) = e^t \quad \Leftrightarrow\ y'(t) = \int e^\tau\, d\tau = e^t + c_1 \ \Leftrightarrow\ y(t) = \int (e^\tau + c_1)\, d\tau = e^t + c_1 t + c_2$$

Fazit: Die allgemeine Lösung dieser Dgl enthält *zwei* frei wählbare Parameter, wir können somit *zwei* Anfangsbedingungen stellen, z.B.

$$y(t_0) \stackrel{!}{=} y_0 \quad y'(t_0) \stackrel{!}{=} y_1$$

Bemerkung: Andere Formen von Anfangsbedingungen sind denkbar, z.B. $y(t_0) \stackrel{!}{=} y_0\,,\ y(t_1) \stackrel{!}{=} y_1$, kommen aber in der Praxis seltener vor.

Die Parameter c_1, c_2 der allgemeinen Lösung kann man unter Verwendung der AB bestimmen.

Frage: Gilt über dieses einfache Beispiel hinaus auch für andere, weniger triviale Dgln zweiter Ordnung, dass man zwei AB stellen kann? Gilt für Dgln n-ter Ordnung allgemein, dass man n AB stellen kann?
Dazu vorab das folgende wichtige Konzept:

Skalare Dgln n-ter Ordnung lassen sich äquivalent umwandeln in Systeme erster Ordnung:
Sei $y^{(n)}(t) = f(t, y(t), y'(t), ..., y^{(n-1)}(t))$ eine skalare Dgl n-ter Ordnung. Man setzt

$$\vec{y}_{neu}(t) := \begin{pmatrix} y(t) \\ y'(t) \\ \vdots \\ y^{(n-1)}(t) \end{pmatrix}$$

und leitet daraus ein System erster Ordnung $\vec{y}'_{neu}(t) = \tilde{f}(t, \vec{y}_{neu}(t))$ für $t \to \vec{y}_{neu}(t)$ her.

Details: s. Tafel

Beispiel: Umwandlung der Dgl des Federschwingers in ein System erster Ordnung: s. Tafel

Folgerung aus obiger Umwandlung für *AWPs*:
Das durch die Umwandlung erhaltene System erster Ordnung erfordert einen Anfangswert $\vec{y}(t_0) \stackrel{!}{=} \vec{y}_0$. (Das sind n viele skalare Bedingungen.)
Zurückübersetzt in die skalare Dgl n-ter Ordnung bedeutet dies, n skalare Bedingungen $y(t_0) = y_0$, $y'(t_0) = y_1$,..., $y^{(n-1)}(t_0) = y_{n-1}$ zu fordern!

Bemerkung: Nicht nur *skalare* Dgln n-ter Ordnung, auch *Systeme* n-ter Ordnung können in Systeme erster Ordnung umgewandelt werden; aus einem System n-ter Ordnung, das aus m Gleichungen besteht, wird so ein System erster Ordnung, das aus $n \cdot m$ Gleichungen besteht. (Das Produkt aus Ordnung und Größe bleibt also konstant.
Aus dem Dreikörperproblem wird so ein System erster Ordnung bestehend aus 9 Gleichungen.)

Somit lautet ein skalares AWP n-ter Ordnung:

Anfangswertproblem n-ter Ordnung

Unter einem *Anfangswertproblem (AWP) n-ter Ordnung* verstehen wir eine eine Dgl. (skalar oder System) n-ter Ordnung

$$y^{(n)}(t) = f(t, y(t), y'(t), ..., y^{(n-1)}(t))$$

zusammen mit den n Anfangsbedingungen (Anfangswerten)

$$y(t_0) = y_0, \quad y'(t_0) = y_1, \quad ... \quad y^{(n-1)}(t_0) = y_{n-1}$$

wobei $t_0, y_0, ..., y_{n-1}$ gegeben sind.

Auch hier wäre es denkbar n andersgeartete Anfangswerte vorzugeben; das kommt aber in der Praxis seltener vor, und man würde dann nicht mehr von einem AWP sprechen, wenn Bedingungen zu verschiedenen Zeitpunkten $t_0, t_1(, ...)$ gestellt würden.

2.2 Elementare Lösungsverfahren für skalare Dgln erster Ordnung

2.2.1 Das Verfahren "Trennung der Variablen"

Das Verfahren **"Trennung der Variablen" (T.d.V.)** ist anwendbar auf skalare(!) Dgln erster Ordnung $y'(t) = f(t, y(t))$, wennn f die Produktform

$$f(t, y) = g(y)h(t)$$

hat, und g nicht null wird. Eine solche Dgl wird manchmal auch als *Dgl mit getrennten Variablen* bezeichnet.
Man kann damit sowohl die allgemeine Lösung der Dgl als auch die Lösung eines AWPs bestimmen.

Beispiel: $y'(t) = t\, y(t)^2$, $y(t_0) = y_0$
Idee: Bringe den y-abhängigen Faktor auf die linke Seite, integriere unter Verwendung der Substitution $\eta := y(t)$.

Beispielrechnung: $y'(t) = ty(t)^2$, $y(t_0) = y_0 \neq 0$

Auf Intervallen, auf denen $y(t) \neq 0$ ist, gilt:

$$\begin{aligned}
\text{AWP} \Leftrightarrow\ & \tfrac{y'(t)}{y(t)^2} = t, \quad y(t_0) = y_0 \quad \text{(Variablen sind nun "getrennt")} \\
\Leftrightarrow\ & \int_{t_0}^{t} \tfrac{y'(\tau)}{y(\tau)^2}\, d\tau = \int_{t_0}^{t} \tau\, d\tau, \quad y(t_0) = y_0 \\
\overset{\eta := y(\tau)}{\Longleftrightarrow}\ & \int_{y_0}^{y(t)} \tfrac{d\eta}{\eta^2} = \int_{t_0}^{t} \tau\, d\tau \\
\Leftrightarrow\ & -\tfrac{1}{\eta}\Big|_{y_0}^{y(t)} = \tfrac{1}{2}\,\tau^2\big|_{t_0}^{t} \quad \Leftrightarrow \quad \tfrac{1}{y_0} - \tfrac{1}{y(t)} = \tfrac{1}{2}(t^2 - t_0^2) \\
\Leftrightarrow\ & \tfrac{1}{y(t)} = \tfrac{1}{y_0} - \tfrac{1}{2}t^2 + \tfrac{1}{2}t_0^2 \quad \Leftrightarrow \quad y(t) = \tfrac{1}{\frac{1}{y_0} - \frac{1}{2}t^2 + \frac{1}{2}t_0^2}
\end{aligned}$$

[
Dies gilt für alle t aus einer Umgebung von t_0, auf der der Nenner$\neq 0$ ist.
Wenn man sich die Mühe machen will, diesen Bereich für alle $t_0 \in \mathbb{R}, y_0 \in \mathbb{R}\setminus\{0\}$ zu bestimmen:

Nullsetzen des Nenners ergibt: Nenner$=0 \Leftrightarrow t^2 = t_0^2 + \frac{2}{y_0}$.

Im Fall $t_0^2 + \frac{2}{y_0} < 0$ ist dies nie erfüllt, somit existiert Dgl-Lsg auf ganz $\mathbb{R}$.

Im Fall $t_0^2 + \frac{2}{y_0} \geq 0$ liefert die Bedingung, dass t nicht gleich $\pm\sqrt{t_0^2 + \frac{2}{y_0}}$ werden darf – mit einigem Nachdenken – das t_0 enthaltende Existenzintervall $(\sqrt{t_0^2 + \frac{2}{y_0}}, \infty)$, falls $y_0 < 0$, und $(-\sqrt{t_0^2 + \frac{2}{y_0}}, \sqrt{t_0^2 + \frac{2}{y_0}})$, falls $y_0 > 0$.
]

Obige Art der Berechnung der Integrale mit Integrationsgrenzen ist insbesondere dann sinnvoll, wenn ein Anfangswert gegeben ist. Die Rechnung liefert allerdings *auch* die *Gesamtheit aller Lösungen*, indem man t_0, y_0 als variabel betrachtet.
Ist man an der Gesamtheit aller Lösungen interessiert, dann kann man allerdings auch mit *unbestimmten* Integralen rechnen, was den Vorteil hat, dass die Lösung unter Beinhaltung von nur *einem* Parameter berechnet wird:

$$\begin{aligned}
\text{DGL} \Leftrightarrow\ & \tfrac{y'(t)}{y(t)^2} = t \quad \Leftrightarrow \quad \int \tfrac{y'(t)}{y(t)^2} dt = \int t\, dt \\
\overset{\eta := y(t)}{\Longleftrightarrow}\ & \int \tfrac{1}{\eta^2}\, d\eta = \int t\, dt \quad \Leftrightarrow \quad -\tfrac{1}{\eta} = \tfrac{1}{2}\, t^2 + c \\
\overset{y(t) = \eta}{\Longleftrightarrow}\ & y(t) = -\tfrac{1}{c + \frac{1}{2}t^2}\,, \quad c \in \mathbb{R}
\end{aligned}$$

Beachte, dass man nur *eine* Integrationskonstante braucht; die Integrationskonstante wird zum Parameter der Lösungsschar.

Eine dritte Art der Darstellung der Rechnung: In Leibniz-/Ingenieur-/Physiker-Schreibweise: $y'(t)$ als $\frac{dy}{dt}$ schreiben und mit dt 'multiplizieren':

$$\frac{dy}{dt} = t\, y^2 \quad \Leftrightarrow \quad \frac{dy}{y^2} = t\, dt \quad \Leftrightarrow \quad \int \frac{dy}{y^2} = \int t\, dt \quad \Leftrightarrow \quad -\frac{1}{\eta} = \frac{1}{2}\, t^2 + c \quad \Leftrightarrow \ldots$$

Diese Vorgehensweise ist an sich mathematisch fragwürdig; die mathematisch präzise Begründung, warum man dennoch so rechnen darf, liefert die Rechnung auf der vorangegangenen Folie.

Zum Nachdenken: Obige Rechnung erforderte $y_0 \neq 0$.
Wie lautet eine/die Lösung im Fall $y_0 = 0$?

Weitere Beispiele für T.d.V.: s. Übung

Bemerkungen:

- *Autonome Dgln* sind Spezialfall der Dgl mit getrennten Variablen ($h \equiv 1$); T.d.V. ist also auf autonome Dgln. anwendbar.
- Skalare lineare homogene Dgln $y'(t) = a(t)y(t)$ sind ebenfalls Spezialfall der Dgl mit getrennten Variablen.
- Verfahren scheitert de facto dann, wenn man keine Stammfunktion von $\frac{1}{g}$ oder von h finden kann.

2.2.2 Lineare skalare Dgln erster Ordnung

Def. (lineare skalare Dgl. erster Ordnung)
Eine Dgl. der Form

$$y'(t) = a(t)\, y(t)$$

heißt *lineare homogene skalare Dgl erster Ordnung*.
Eine Dgl. der Form

$$y'(t) = a(t)\, y(t) + b(t)$$

mit $b \not\equiv 0$ heißt *lineare inhomogene skalare Dgl erster Ordnung*.

Bevor wir diese lösen, untersuchen wir die Struktur der Lösungsmenge:

Satz (Struktur der Lösungsmenge der lin. Dgl. erster Ordnung)

(a) homogener Fall: Die Lösungsmenge L_{hom} einer linearen homogenen Dgl. erster Ordnung ist ein *eindimensionaler Vektorraum* (Unterraum eines Funktionenraums). Er wird aufgespannt von $y_{hom} = \exp(\int a\,dt)$, wobei $\int a\,dt$ eine beliebige Stammfunktion von a ist.

$$L_{hom} = \{c\,y_{hom} \,|\, c \in \mathbb{R}\}$$

(b) inhomogener Fall: Die Lösungsmenge L_{inhom} einer linearen inhomogenen Dgl. erster Ordnung hat die Form

$$L_{inhom} = \{y_p\} + L_{hom} = \{y_p + c\,y_{hom} \,|\, c \in \mathbb{R}\}$$

wobei wobei y_p irgend eine beliebige, 'feste' Lösung der inhomogenen Dgl. ist. L_{inhom} ist somit ein *affiner Raum*.

Beweis: s. Tafel

Anmerkungen:

- In obiger Formel für L_{inhom} bezeichnet man die feste, aber *beliebig wählbare(!)* inhomogene Lösung y_p als *partikuläre Lösung*. (Diese Begriffsbildung sorgt erfahrungsgemäß für etwas Verwirrung/Verwunderung.)
- Die obige Formel besagt: **Es reicht L_{hom} sowie *eine einzige* inhomogene Lösung zu kennen, um *alle* inhomogenen Lösungen zu kennen.**
- **Der Zusammenhang der Lösungsmengen L_{hom} und L_{inhom} entspricht genau dem der Lösungsmengen von homogenen bzw. inhomogenen *Linearen Gleichungssystemem!*** Dies motiviert die obige Begriffsbildung "*hom./inhom. lineare* Dgl".

Berechnung einer partikularen Lösung: Variation der Konstanten

Wir zuvor festgestellt, reicht es, L_{hom} zu berechnen (siehe (a)) sowie *eine* ('partikuläre') Lösung des inhomogenen Problems.

Eine inhomogene Lösung bestimmt man, wenn man bereits eine homogene Lösung kennt, mit *Variation der Konstanten ('V.d.K.')*; dies geht zurück auf Lagrange (1736-1813):

Ist $y_h \not\equiv 0$ eine homogene Lösung, so sind auch alle $c\,y_h$, wobei $c \in \mathbb{R}$ konstant, homogene Lösungen (Vektorraumstruktur der Lösungsmenge).
Um eine *inhomogene* Lösung y_p zu finden "lässt man die Konstante variieren", d.h.

man macht den **Ansatz**

$$y_p(t) := c(t)\, y_h(t)\,.$$

Wir leiten ab: $y_p'(t) = c'(t)y_h(t)+c(t)y_h'(t)$. Dann setzen wir y_p und y_p' in die inhomogene Dgl ein:

$$c'(t)y_h(t) + c(t)\,\underline{y_h'(t)} \stackrel{!}{=} \underline{a(t)}\,c(t)\,\underline{y_h(t)} + b(t)$$

Nun nutzt man aus, dass $y_h \in L_{hom}$, d.h. $y_h'(t) = a(t)y_h(t)$, und bekommt

$$c'(t)\,y_h(t) = b(t) \qquad | : y_h(t)$$

$$\Rightarrow \quad c'(t) = \frac{b(t)}{y_h(t)}$$

Das Bilden einer (beliebigen) Stammfunktion von b/y_h liefert also die Funktion c.

Wir fassen zusammen:

Lösungsverfahren „Variation der Konstanten (VdK)“ für inhomogene lineare Dgln

- Um für eine inhomogene lineare skalare Dgl erster Ordnung eine partikuläre Lösung y_p zu finden, macht man den Ansatz („Variation der Konstanten“)

  $$y_p(t) := c(t)\, y_h(t)\,,$$

  wobei $y_h \not\equiv 0$ eine Lösung der zugehörigen *homogenen* Dgl ist.

- Die Koeffizientenfunktion $c(t)$ bekommt man dann, indem man eine beliebige Stammfunktion von

  $$c'(t) = \frac{b(t)}{y_h(t)}$$

  berechnet.

- Die allgemeine Lösung der inhomogenen linearen Dgl ist dann

  $$y(t) = y_p(t) + c\, y_h(t)\,, \quad c \in \mathbb{R}\,.$$

Beachte: Das c im 3. Punkt hat nichts mit dem $c(t)$ im 1. und 2. Punkt zu tun. Hat man eine Anfangsbedingung $y(t_0) = y_0$ gegeben, so kann man das c im 3. Punkt dadurch berechnen und bekommt die Lösung des AWPs.

Man kann den oben angegebenen Algorithmus aus in *allgemeiner* Form durchrechnen und kann so allgemeine **Lösungsformeln** fürs inhomogene lineare AWP herleiten:

T.d.V. ergibt y_h ergibt eine homogene Lösung $y_h(t) = \exp(\int_{t_0}^t a(s)\,ds)$ (1).

$c(t) = \int_{t_0}^t \frac{b(\tau)}{y_h(\tau)}\,d\tau$ in den Ansatz $y_p(t) = c(t)y_h(t)$ einsetzen ergibt: $y_p(t) = \int_{t_0}^t b(\tau)\,\frac{y_h(t)}{y_h(\tau)}\,d\tau$ (2)

Den Bruch in (2) berechnen wir mittels Formel (1), zweimal angewendet:t $\frac{y_h(t)}{y_h(\tau)} = \frac{\exp(\int_{t_0}^t a(s)\,ds)}{\exp(\int_{t_0}^\tau a(s)\,ds)} = \exp(\int_\tau^t a(s)\,ds)$, also $y_p(t) = \int_{t_0}^t b(\tau)\exp(\int_\tau^t a(s)\,ds)\,d\tau$.

Wir haben somit hergeleitet:

Satz: Für homogene lineare Dgln erster Ordnung ist der Lösungsraum

$$L_{hom} = \{c\,\exp(\int_{t_0}^t a(\tau)\,d\tau) \,|\, c\in\mathbb{R}\} = \text{span}\{\,\exp(\int_{t_0}^t a(\tau)\,d\tau)\,\},$$

wobei $t_0\in\mathbb{R}$ beliebig (und fest).
Die allgemeine inhomogene Lösung lautet

$$L_{inhom} = L_{hom} + \{y_p\} = \{\,c\,\exp(\int_{t_0}^t a(\tau)\,d\tau) \;+\; \int_{t_0}^t b(\tau)\exp(\int_\tau^t a(s)\,ds)\,d\tau \mid c\in\mathbb{R}\},$$

wobei $t_0\in\mathbb{R}$ beliebig (und fest).
Beachte, dass hierfür die Existenz der auftretenden Integrale erforderlich ist. Dazu hinreichend ist z.B., dass a, b auf $[t_0, t]$ stetig sind.

Anmerkung zur praktischen Vorgehensweise:
Die *Lösungsformel* ist kompliziert. M.E. ist es im konkreten Anwendungsfall einfacher, die Rechnung gemäß der *zuerst* vorgestellten Vorgehensweise (s.S. 24) auszurechnen, d.h.
(1) hom. Lsg. mittels T.d.V. berechnen,
(2) part. Lsg. mittels V.d.K. berechnen,
(3) allg. Lsg. = part. Lsg. + c mal hom. Lsg.,
als die (schreckliche) Lösungsformel auswendig zu lernen und anzuwenden.
Zu (2): **Merke:** Ansatz $y_p(t) = c(t)\,y_h(t)$ und ggf. noch $c'(t)\,y_h(t) = b(t)$.

Beispiel: Bestimme die allgemeine Lösung der Dgl. $y' = 2ty + te^{t^2}$.
Rechnung: s. Tafel

Bemerkung:

Später (→Kap. 2.4 und 2.5) werden wir obige Vorgehensweise verallgemeinern für **homogene/inhomogene lineare *Systeme* von Dgln.**, und sogar auf **skalare lineare Dgln. n-ter Ordnung**.

2.2.3 Lösen von Dgl mittels Substitution

Bislang haben wir es nur geschafft, folgende Arten von (skalaren) Dgln zu lösen:

– Dgln mit getrennten Variablen, also rechte Seite $f(t,y) = g(y)\,h(t)$ und

– (inhomogene) lineare Dgln, also rechte Seite $t(t,y) = a(t)\,y + b(t)\,.$

Im folgenden wollen wir einigen Situationen betrachten, in denen man mittels einer geeigneten **Substitution** die Dgl. auf eine der obigen Formen zu bringen.

(a) Typus $y'(t) = f(\frac{y(t)}{t})$, kurz: $y' = f(\frac{y}{t})$
Sinnvolle Substitution: $u(t) := \frac{y(t)}{t}$
Daraus kann man eine Dgl für u herleiten, und diese hat getrennte Variablen.
Nachdem man die u-Dgl gelöst hat, substituiert man zurück: $y(t) = t\,u(t)$.
Rechnung: s. Tafel

(b) Typus $y'(t) = f(at+by(t)+c)$, kurz $y' = f(at+by+c)$
mit $b \neq 0$ (der Fall $b=0$ ist trivial)
Sinnvolle Substitution: $u(t) := at+by(t)+c$
Daraus kann man eine Dgl für u herleiten, und diese ist autonom, hat also getrennte Variablen.
Rechnung: s. Tafel

(c) Typus
$$y'(t) = f(t)\,y(t) + g(t)\,y(t)^\alpha, \qquad \alpha \in \mathbb{R}\setminus\{0,1\}$$
Diese Dgl nennt man *Bernoulli'sche Dgl*.
Beachte, dass die Fälle $\alpha=1$ und $\alpha=0$ bereits linear sind und keine Substitution erfordern.

Genialer Trick: Division durch y^α und Anwendung der Kettenregel ('rückwärts') ergibt
$$f(t)\,\underbrace{y(t)^{1-\alpha}}_{u(t)} + g(t) \overset{\text{(Dgl.)}}{=} \frac{y'(t)}{y(t)^\alpha} \overset{\text{(K.R.)}}{=} \frac{1}{1-\alpha}\,\frac{d}{dt}\,\underbrace{y^{1-\alpha}(t)}_{u(t)}$$
Dies legt nahe, $u(t) := y(t)^{1-\alpha}$ zu substituieren. Das ergibt
$$u'(t) = (1-\alpha)\,f(t)u(t) + (1-\alpha)\,g(t)$$
Dies ist eine *inhomogene lineare Dgl. erster Ordnung*; wie man die löst, haben wir in →Kap. 2.2.2 gelernt.

Anwendung: Die logistische Dgl ist eine Bernoulli'sche Dgl. (Allerdings lässt sich die logistische Dgl. auch direkt mit T.d.V. angehen.)

(d) Typus
$$\sum_{k=0}^{n} \alpha_k t^k\, y^{(k)}(t) = 0$$

Eine solche Dgl heißt *Euler'sche Dgl*; es handelt sich um eine *lineare* Dgl $\sum_{k=0}^{n} a_k(t)\, y^{(k)}(t) = 0$, bei der die Koeffizienten in spezieller Weise von t abhängen (als $a_k(t) = \alpha_k t^k$).

Per Substitution

$$u(\tau) := y(e^\tau), \quad \text{d.h. } u(\ln t) = y(t),$$

kann man die t-Abhängigkeit der Koeffizienten verschwinden lassen, d.h. man kann eine Dgl. der Form

$$\sum_{k=0}^{n} \tilde{a}_k\, u^{(k)}(t) = 0$$

herleiten; eine solche ist leichter zu lösen, wie wir später sehen werden (→Kap. 2.5).

Durchführung im Detail: Man leitet die Gleichung $u(\ln t) = y(t)$ n-mal ab, und bekommt Gleichungen, die es erlauben, die Terme der Form $t^k y^{(k)}$ in der Dgl durch u-abhängige Terme zu ersetzen. Am Ende substituiert man $\tau := \ln t$.

Beispiel: ggf. s. Tafel/Übung

2.3 Existenztheorie (Existenz und Eindeutigkeit von Lösungen von Anfangswertproblemen)

Motivation:

Anfangswertprobleme (AWPs) entstehen i.a. aus *Anwendungen* heraus mittels *Mathematischer Modellierung* (→Kap 2.1). Häufig (z.B. insbesondere in der klassischen Mechanik, aber auch in der Chemie, in Populationsmodellen,...) erwartet man, dass das Modell genau eine Lösung hat. Falls ein Modell gar keine oder mehrere Lösungen zulässt, so ist die Frage angebracht, ob das mathematische Modell tatsächlich das Anwendungsproblem hinreichend gut beschreibt.

Wir wollen im folgenden Sätze kennenlernen, die die Existenz und Eindeutigkeit von Lösungen von AWPs liefern.

Historische Bemerkung:

Die Sätze und ihre Beweise gehen zurück auf das 19. Jahrhundert. Vor Beginn des 19. Jahrhunderts hat man sich zwar durchaus schon für das Aufstellen und Lösen von AWPs interessiert (Euler, Lagrange, Bernoulli,...), aber man war überhaupt nicht auf die *Idee* gekommen, Existenz/Eindeutigkeit von Lösungen zu beweisen (z.B. in den Fällen, wo man die Lösung nicht explizit ausrechnen kann).
Man hat damals überhaupt nicht zwischen dem Anwendungsszenario und dem mathematischen Modell unterschieden. Und da das Anwendungsszenario "irgend etwas macht", ist "das, was man da sieht", die "Lösung". So die damalige Sichtweise.

Erinnerung: Dgln n-ter Ordnung können in Dgl-Systeme erster Ordnung umgewandelt werden (Kap. 2.1).

Es reicht also, wenn wir uns hinsichtlich Existenz und Eindeutigkeit von Lösungen mit Dgln. bzw. Dgl-Systemen *erster* Ordnung beschäftigen!

Ein Existenzsatz:

Satz (Existenzsatz von Peano, 1886/1890)
Sei $t_0 \in \mathbb{R}$, $\vec{y}_0 \in \mathbb{R}^n$, $T > t_0$, $f : G \to \mathbb{R}^n$, wobei $G \supseteq [t_0, T] \times K_R(\vec{y}_0)$, $R > 0$, f stetig. Dann hat das AWP

$$\vec{y}\,'(t) = \vec{f}(t, \vec{y}(t)), \quad \vec{y}(t_0) = \vec{y}_0$$

auf einem Intervall $[t_0, t_0 + \epsilon]$ (mindestens) eine Lösung $t \mapsto \vec{y}(t)$, und diese ist stetig diff'bar.

Erklärung zu G: Man fordert im obigen Satz, dass f zumindest auf einem 'Zylinder' $[t_0, T] \times K_R(\vec{y}_0) \subseteq \mathbb{R}^{n+1}$ definiert und stetig ist (Skizze!).

Zusatz: Man kann obiges $\epsilon > 0$ immer finden als $\epsilon := \min\{T - t_0, \frac{R}{M}\}$, wobei $M := \max_{(t,\vec{y}) \in [t_0,T] \times [K_R(\vec{y}_0)]} \{\|\vec{f}(t, \vec{y})\|\}$. (Dieses M existiert, da f stetig auf kompakter Menge.) Beachte auch, dass M eine Schranke für die Steigung der Lösung darstellt.)
Es ist also möglich, dass die Lösung nicht bis zum Ende $t = T$ des Zylinders reicht.

Kurzform des Satzes:

$$\text{rechte Seite } \underline{\text{stetig}} \implies \begin{array}{l}\text{(mindestens eine) Lösung } \underline{\text{existiert}} \text{ fürs AWP,} \\ \text{(zumindest 'lokal')}\end{array}$$

Bemerkung: Die Stetigkeitsforderung an f im Existenzsatz kann nicht ersatzlos gestrichen werden; betrachte dazu das AWP mit $f(t, y) := \begin{cases} -1, & t\,y \geq 0, \\ +1, & t\,y < 0 \end{cases}$, $t_0 := 0, y_0 := 0$. Für dieses unstetige f hat das AWP "offensichtlich" keine Lösung (Skizze s. Tafel; ein exakter Beweis für Nichtexistenz geht als Widerspruchsbeweis).

(Dagegen hat man für $f(t, y) := \begin{cases} +1, & y \geq 0, \\ -1, & y < 0 \end{cases}$, obwohl dies ebenfalls unstetig ist, für beliebigen AW immer eine Lösung.)

In den praktischen Anwendungen haben AWPs "fast immer" eine stetige rechte Seite. Beispiel für ein AWP mit unstetiger rechter Seite: Federschwinger mit Rutsch- und Haftreibung, ggf. s. Tafel

Eine Anwendung des Existenzsatzes:
Nicht nur skalare lineare Dgln erster Ordnung $y'(t) = a(t)y(t) + b(t)$ mit stetigem a, b (deren Lösung wir bereits in Kap. 2.2.3 berechnet haben) haben also eine Lösung, sondern auch *Systeme* von linearen Dgln, also $\vec{y}\,'(t) = A(t)\,\vec{y}(t) + \vec{b}(t)$ mit stetigem $A : \mathbb{R} \to \mathbb{R}^{n \times n}, b : \mathbb{R} \to \mathbb{R}^n$ haben nach dem Existenzsatz von Peano eine Lösung, denn

$\vec{f} : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$, definiert durch $\vec{f}(t,\vec{y}) = A(t)\,\vec{y}+\vec{b}(t)$, ist stetig, wenn $t \mapsto A(t)$, $t \mapsto \vec{b}(t)$ stetig sind.

Zur *Eindeutigkeit* von Lösungen von AWPs

Vorüberlegung: Gibt es AWPs, die mehr als eine Lösung haben?

Betrachte dazu das recht berühmte AWP

Gegenbeispiel für Eindeutigkeit von Lösungen

$$y'(t) = \sqrt{|y(t)|}, \quad y(0)=0.$$

Die rechte Seite $f(y) = \sqrt{|y|}$ ist offensichtlich stetig; die *Existenz* einer Lösung ist somit gesichert.

Mittels T.d.V. (oder Raten) kann man eine Lösung $y(t) = \frac{1}{4}t^2$ finden.

Aber auch $y(t) \equiv 0$ ist offensichtlich eine Lösung!

Noch schlimmer: Auch Zusammensetzungen $y(t) = \begin{cases} 0, & t \le c \\ \frac{1}{4}(t-c)^2, & t > c \end{cases}$, wobei $c \ge 0$ beliebig, sind Lösungen!

Unser AWP hat somit unendlich viele Lösungen; wir haben eine "Verzweigung" von Lösungen an allen Stellen $(t,0)$.

Bemerkung: Bei manchen AWP hat man "lokal" zumindest eine *eindeutige* Lösung, die sich dann irgendwann später verzweigt (Bsp.: $y' = -\sqrt{|y|}$ mit AW $y_0 > 0$).

Fazit: Um *Eindeutigkeit* von Lösungen zu bekommen, braucht man also offensichtlich eine stärkere Forderung an f als nur die Stetigkeit. Welche?

Antwort: Die sog. *Lipschitz-Stetigkeit*; siehe folgende Definition:

Def. (Lipschitz-Stetigkeit)

Sei $(V, \|\cdot\|)$ ein normierter Vektorraum und $\emptyset \neq M \subseteq V$ und $\vec{F} : M \to V$.
Wenn es eine Konstante $L > 0$ gibt, so dass

$$\forall\, \vec{x}, \vec{y} \in M : \ \|\vec{F}(\vec{x}) - \vec{F}(\vec{y})\| \le L\, \|\vec{x} - \vec{y}\|,$$

dann heißt $\vec{F}$ *Lipschitz-stetig* auf M, und L heißt *Lipschitz-Konstante* von $\vec{F}$ auf M.

Bemerkung: Die Definition der *Kontraktionseigenschaft* (Kap. I.6) sah sehr ähnlich aus.
Wir stellen fest: Jede Kontraktion ist Lipschitz-stetig. Eine Funktion ist genau dann

eine Kontraktion, wenn sie Lipschitz-stetig mit Lipschitz-Konstante <1 ist.

Zusammenhang zu Stetigkeit und Diff'barkeit (elementare Beweise: s. Tafel)

1. $\vec{F}$ Lipschitz-stetig $\Rightarrow$ $\vec{F}$ stetig.
2. Zumindest im skalaren Fall (lässt sich aber auch auf vektoriellen Fall übertragen):
   (a) F diff'bar mit beschränkter Ableitung
   $\Rightarrow$ F Lipschitz-stetig (mit $L = \sup_{x\in M} |f'(x)|$)
   (b) F stetig diff'bar auf kompaktem M
   $\Rightarrow$ F Lipschitz-stetig (mit $L = \sup_{x\in M} |f'(x)| = \max_{x\in M} |f'(x)|$)
3. Die Umkehrung zu 1. gilt nicht: Gegenbeispiel $F(x) = \sqrt{|x|}$
4. Zur Umkehrbarkeit von 2.: Aus Lipschitz-Stetigkeit kann man nicht auf Differenzierbarkeit schließen (Gegenbeispiel: $F(x) = |x|$).
   Aber: Wenn $F\in C^1(M)$ und $\sup_{x\in M} |f'(x)| = \infty$, dann ist F nicht Lipschitz-stetig auf M.

Der zentrale Satz, der zusätzlich zur Existenz auch die **Eindeutigkeit** von Lösungen von AWPs liefert, und der als wesentliche 'Zutat' die Lipschitz-Stetigkeit hat:

Satz (Satz von Picard–Lindelöf, 1890/94)
Sei $t_0\in\mathbb{R}$, $\vec{y}_0\in\mathbb{R}^n$, $T>t_0$, sei $\vec{f}:[t_0,T]\times\mathbb{R}^n\to\mathbb{R}^n$ stetig, und es gebe ein $L>0$ so dass

$$\|\vec{f}(t,\vec{x}) - \vec{f}(t,\vec{y})\| \le L\,\|\vec{x}-\vec{y}\| \quad \forall\,\vec{x},\vec{y}\in\mathbb{R}^n,\ t\in[t_0,T]$$

(kurz: $\vec{f}$ sei Lipschitz-stetig bzgl. zweitem Argument, gleichmäßig im ersten Argument (d.h. L von t unabhängig))
Dann hat das AWP

$$\vec{y}\,'(t) = \vec{f}(t,\vec{y}(t)), \quad \vec{y}(t_0) = \vec{y}_0$$

eine eindeutig(!) bestimmte Lösung $\vec{y}:[t_0,T]\to\mathbb{R}^n$, und diese ist stetig diff'bar.

Historische Bemerkung: Eine frühere Version dieses Satzes, von Cauchy, 1826, verwendete die stärkere Bedingung "$\vec{f}\in C^1$ und alle $\partial_i f_j$ beschränkt" anstelle der L-Stetigkeit.
Der obige Satz wird daher in Frankreich auch als *Theorème de Cauchy–Lipschitz* bezeichnet.

Zum Beweis des Satzes von Picard-Lindelöf:
Wichtigstes Hilfsmittel: Fixpunktsatz von Banach!
Vorab: Zur Vereinfachung der Darstellung sei $n=1$.

Idee: Schreibe AWP mittels HS der Diff'-u-Int'rechnung in **Integralgleichung** um:

$$y'(t) = f(t, y(t)) \wedge y(t_0) = y_0 \quad \Longrightarrow \quad y(t) \overset{(\mathrm{HS})}{=} y_0 + \int_{t_0}^{t} y'(\tau)\, d\tau \overset{(\mathrm{Dgl})}{=} \underbrace{y_0 + \int_{t_0}^{t} f(\tau, y(\tau))\, d\tau}_{=:\Phi(y)(t)}$$

Das Problem ist nun ein **Fixpunktproblem** $y = \Phi(y)$! Fixpunkt y von Φ entspricht Lösung des AWP!
Untersuchung auf Existenz/Eindeutigkeit des Fixpunktes (der AWP-Lösung also) mittels FP-Satz von Banach!

Dazu vorab die "**Formalien**" klären: Auf welchem normierten VR V ist $\Phi : V \to V$ definiert?
Φ bildet eine Funktion $y : [t_0, T] \to \mathbb{R}$ auf eine Funktion $\Phi(y) : [t_0, T] \to \mathbb{R}$ ab; V ist also ein Funktionenraum.
Es bietet sich

$$V := C^0([t_0, T]) = \{y : [t_0, T] \to \mathbb{R} \,|\, y \text{ ist stetig}\}$$

an, denn für y aus diesem Raum V existiert das Integral, und $\Phi(y) \in V$, d.h. $\Phi : V \to V$ ist wohldefiniert.
Mit welcher Norm statten wir V aus? Der FP-Satz erfordert die *Vollständigkeit* von $(V, \|\cdot\|)$, somit könnte man die ∞-Norm $\|y\|_\infty := \max\limits_{t \in [t_0, T]} |y(t)|$ verwenden.
Technische Details beim Überprüfen der Kontraktionseigenschaft (s. später) bewirken jedoch, dass es sich als sinnvoller herausstellt, eine sog. *gewichtete* Unendlich-Norm zu verwenden, und zwar die Norm

$$\|y\|_{\infty,\alpha} := \max_{t \in [t_0, T]} e^{-\alpha(t - t_0)} |y(t)|;$$

eine solche Gewichtung ändert nichts an der Vollständigkeit. (Welches $\alpha \in \mathbb{R}$ wir fest wählen, können wir später noch festlegen.)

Nach diesen Vorbereitungen bleibt (nur) noch zu prüfen, dass unser $\Phi : V \to V$ die **Kontraktionseigenschaft** erfüllt, dass also für ein $K < 1$

$$\|\Phi(y_1) - \Phi(y_2)\|_{\infty,\alpha} \leq K \, \|y_1 - y_2\|_{\infty,\alpha} \quad \forall\, y_1, y_2 \in V.$$

Die Rechung dazu: siehe Tafel.
Diese Rechnung zeigt, dass der Beweise der Kontraktionseigenschaft gelingt, wenn man $\alpha > L$ wählt! Mit dem FP-Satz folgt also Existenz/Eindeutigkeit der Lösung $y \in C^0$ von $\Phi(y) = y$, also der Integralgleichung. Am Ende zeigt man nur noch, dass der FP in C^1 liegt. □

Bemerkung: Man erinnere sich, dass der FP-Satz von Banach auch ein **Konstruktionsprinzip** zum praktischen Berechnen des FP, also unserer AWP-Lösung, beinhaltet: Für beliebige Startfunktion (z.B.

für die konstante Funktion y_0) konvergiert die Iteration $y_{n+1}(t) := \Phi(y_n)(t)$, also

$$y_{n+1}(t) := y_0 + \int_{t_0}^{t} f(\tau, y_n(\tau))\, d\tau, \qquad t \in [t_0, T],$$

gegen die Lösung des AWPs. In der Praxis jedoch benutzt man andere numerische Verfahren (→Kap. II.6), um AWPs (näherungsweise) zu lösen, wenn man die exakte Lösung nicht finden kann.

Varianten des Satzes von Picard-Lindelöf

Die Anwendbarkeit des Satzes - gegeben in obiger Form - ist in der Praxis häufig nicht direkt möglich, da viele Funktionen f die Lipschitz-Bedingung in obiger Form auf einem "Streifen" $M := [t_0, T] \times \mathbb{R}^n$ nicht erfüllen:
Sogar Polynomfunktionen(!), deren Grad> 1 ist, wie z.B. $f(t, y) = y^2$ oder $f(t, y) = -ty^2$, erfüllen die Lipschitz-Bedingung in obiger Form auf obigem M nicht!
Denn: Für diese f ist $\partial f/\partial y$ auf der unbeschränkten(!) Menge $M := [t_0, T] \times \mathbb{R}^n$ unbeschränkt.

Ein Ausweg: Statt obigem unbeschränktem(!) M ein **beschränktes** abgeschlossenes M betrachten; auf diesem erfüllen polynomielle f (so wie *jedes* $f \in C^1$) eine Lipschitz-Bedingung!

Dies führt dazu, dass die folgende **Variante des Satzes von Picard-Lindelöf** nützlich ist:

> Ersetze in obigem Satz den **Streifen** $M := [t_0, T] \times \mathbb{R}^n$ durch den "**Zylinder**" $M := [t_0, T] \times K_R(\vec{y}_0)$, wobei $R > 0$. Die Aussage des Satzes bleibt weiter erhalten, bis auf die Tatsache, dass die Lösung ggf nicht mehr auf ganz $[t_0, T]$ existiert, sondern auf einem Intervall $[t_0, t_0 + \epsilon]$, $\epsilon := \min\{T - t_0, \frac{R}{m}\}$, welches ggf. kleiner ist (wie beim Satz von Peano); $m := \max_{(t,y)\in M} |f(t, y)|$.
> Ferner liefert der Satz in der neuen Form, dass die Lösung immer bis zum Rand des Zylinders geht.

Auf einen Beweis (der ähnlich wie bei der ersten Version des Satzes verläuft) verzichten wir; Veranschaulichung per Skizze: s. Tafel.

Eine andere Variante: Die bisher betrachtete Forderung der Lipschitz-Stetigkeit kann abgeschwächt werden zu einer sog. ***lokalen* Lipschitz-Bedingung**:

Existenz- und Eindeutigkeitssatz unter *lokaler* Lipschitz-Bedingung an f

Sei $f : U \to \mathbb{R}^n$ mit $U \subseteq \mathbb{R} \times \mathbb{R}^n$, und U sei Umgebung des Anfangswertes von $(t_0, \vec{y}_0)$. Für jedes $(t, \vec{y}) \in U$ gebe es eine Umgebung $V_{t,\vec{y}} \subseteq U$, und eine Zahl $L_{t,\vec{y}} > 0$, so dass (also "lokal" auf $V_{t,\vec{y}}$, nicht auf ganz U) die L-Bedingung

$$\|\vec{f}(\tau, \vec{y}_1) - \vec{f}(\tau, \vec{y}_2)\| \leq L_{t,y} \, \|\vec{y}_1 - \vec{y}_2\| \quad \forall\, (\tau, \vec{y}_1), (\tau, \vec{y}_2) \in V_{t,\vec{y}}$$

gilt. Dann bleibt die Aussage des Satzes, also Existenz und Eindeutigkeit einer Lösung des AWP, erhalten, und die Lösung geht sicher "bis zum Rand" von U (sogar "von Rand zu Rand").

Veranschaulichung/Anwendung:

- **Lokale L-Bedingung im Vergleich zu "globaler" L-Bedingung:**
  Polynomfunktionen f sind, sogar auf ganz $U = \mathbb{R} \times \mathbb{R}^n$, lokal Lipschitz-stetig, aber i.a. nicht "global" Lipschitz-stetig.
  Die Wurzelfunktion jedoch ist, sobald U einen Punkt (t, y) mit $y = 0$ enthält, *nicht* lokal Lipschitz-stetig. (Klar, denn andernfalls stünde obiger Satz im Widerspruch zur uns bereits bekannten Nicht-Eindeutigkeit der Lösung von $y' = \sqrt{y}$, $y(t_0) = 0$!)

- **Präzisierung des Begriffs "bis zum Rand von U"**:
  Falls U unbeschränkt ist, kann dies bedeuten, dass die Lösung auf ganz $(-\infty, \infty)$ existiert oder auch dass die Lösung Blow-Up-Verhalten zeigt.

  Eine Lösung, die "von Rand zu Rand" (des Definitionsbereiches U von f) geht, bezeichnet man auch als *maximale Lösung*; falls sie mindestens auf $[t_0, \infty)$ existiert, als *globale Lösung*. Skizzen siehe Tafel.

Beweisidee des Existenz- und Eindeutigkeitssatzes unter *lokaler* L-Bedingung:
"Zusammenstückeln" der Lösung: Man startet mit $(t_0, \vec{y}_0)$, findet Lösung auf einer Umgebung $V_{t_0,\vec{y}_0}$, diese kann o.B.d.A. zylinderförmig gewählt werden (ggf. verkleinern); Lösung geht nach früherer Version des Satzes bis zum Rand dieses Zylinders. Den Endpunkt der Lsg. am Rand des Zylinders wird als neuer Startpunkt verwendet. Und dies wird iteriert.

Nochmal kurz, und unter Vernachlässigung der mathematischen Präzision, zusammengefasst, was es mit den Varianten des Satzes von Picard–Lindelöf auf sich hat:

- Auf kompakten Mengen U erfüllen "die meisten" f die ("globale") Lipschitz-Bedingung: Stetige Diff'barkeit von f reicht völlig aus. Eines der "wenigen" Beispiele für Funktionen, die nicht L-stetig sind: Die Wurzel-Funktion, wenn $0 \in U$.

- Auf *unbeschränkter* Menge U sind viele "einfache" Funktionen, z.B. alle Polynomfunktionen vom Grad> 1, *nicht* L-stetig (da ihre Steigung unbeschränkt ist).

Mögliche Auswege:
1. Betrachte f bzw. Dgl. eingeschränkt auf *beschränktes* U oder
2. Prüfe ob f auf U die *lokale* L-Bedingung erfüllt. Das tun "die meisten" Funktionen, so z.B. alle Polynomfunktionen, und das sogar auf unbeschränktem U; die Wurzelfunktion erfüllt jedoch weiterhin, d.h. auch in der *lokalen* Version, die L-Bedingung nicht, sofern man $y = 0$ mitbetrachtet.

Beispiel: (Übung oder Tafel): Löse das AWP $y' = f(y) := y^2$, $y(0) = 1$.
Was ist die maximale Lösung? Welche Variante des Satzes von Picard-Lindelöf lässt sich anwenden, bei welchem Def'bereich von f?

Blow-Up: In Anwendungsproblemen (z.B. Populationsmodelle, chemische Reaktionen,...) erwartet man weiterhin häufig, dass die Lösung "für alle Zeiten" existiert, d.h. dass das AWP $y'(t) = f(t, y(t))$, $y(t_0) = y_0$, eine Lösung $y : \mathbb{R} \to \mathbb{R}$ (Systeme: $y : \mathbb{R} \to \mathbb{R}^n$) oder zumindest $y : [t_0, \infty) \to \mathbb{R}$ (Systeme: $y : [t_0, \infty) \to \mathbb{R}^n$) hat.
Es gibt aber auch durchaus AWPs (Beispiel: s.o.), deren Lösung "in endlicher Zeit" gegen unendlich geht, somit die Lösung nur auf $[t_0, T)$ existiert, mit $\lim_{t \to T} y(t) = \infty$. Sowas bezeichnet man als "Blow-Up".

Zeigt das mathematische Modell Blow-Ups, so sollte man sich ggf. die Frage stellen, ob das Modell das Anwendungsproblem hinreichend gut beschreibt.
Häufig liegt dann die Situation vor, dass das Modell nur für kleine Werte von y "gut" ist, aber für große Werte von y sehr ungenau, was zu unterschiedlichen Verhalten des Anwendungsproblems und der Lösung des mathematischen Modells führt.

2.4 Lineare Dgl-Systeme erster Ordnung

Def. (Lineares Dgl.-System erster Ordnung)
Ein Dgl.-System der Form

$$\vec{y}'(t) = A(t)\,\vec{y}(t) + \vec{b}(t) \qquad (*)$$

mit gegebenen $A(t) \in \mathbb{R}^{n \times n}$ und $\vec{b}(t) \in \mathbb{R}^n$ heißt *Lineares Differentialgleichungssystem erster Ordnung*.

Falls $\vec{b}(t) \equiv \vec{0}$, so nennt man das System *homogen*, andernfalls *inhomogen*.

Geht man, unter Streichen von $\vec{b}(t)$, von (*) zum System $\vec{y}'(t) = A(t)\,\vec{y}(t)$, so bezeichnet man das letztere als *das zu (*) zugehörige homogene System*.

Bemerkung zu Existenz und Eindeutigkeit von Lösungen:
Man kann zeigen: Sind $A : \mathbb{R} \to \mathbb{R}^{n\times n}$ und $\vec{b} : \mathbb{R} \to \mathbb{R}^n$ stetige Funktionen, so ist die rechte Seite $f(t, \vec{y}) = A(t)\vec{y}+\vec{b}(t)$ stetig sowie lokal Lipschitz-stetig bzgl. $\vec{y}$, und die Lösung des Systems ist dann, bei vorgegebenem AW $\vec{y}(t_0) = \vec{y}_0$, existent und eindeutig bestimmt, und geht "von Rand zu Rand" des $\mathbb{R}\times\mathbb{R}^n$ (d.h. für jede kompakte Teilmenge $(t_0, \vec{y}_0) \in M \subset \mathbb{R}\times\mathbb{R}^n$ endet die Lösung niemals in M. (Das heißt nicht zwingend, dass das zeitliche Existenzintervall ganz $\mathbb{R}$ ist.)

2.4.1 Die Struktur der Lösungsmenge

Völlig analog zum skalaren Fall (s. Kap. 2.2.2) gilt auch für lineare *Systeme*:

Satz (Struktur der Lösungsmenge der lin. Dgl-Systeme erster Ordnung)

(a) Die Lösungsmenge L_{hom} eines linearen *homogenen* Dgl.-Systems erster Ordnung ist ein *Vektorraum* (ein Unterraum des Raumes $\mathrm{Abb}(I, \mathbb{R}^n)$, wobei $I \subseteq \mathbb{R}$ ein geeignetes Intervall ist).

(b) Die Lösungsmenge L_{inhom} eines linearen *inhomogenen* Dgl.-Systems erster Ordnung hat die Form

$$L_{inhom} = \{\vec{y}_p\} + L_{hom},$$

wobei L_{hom} der Lösungsraum des zugehörigen homogenen Dgl.-Systems ist, und wobei y_p irgend eine beliebige, 'feste' Lösung des inhomogenen Dgl.-Systems ist. L_{inhom} ist somit ein *affiner Raum*.

Beweis: Ist völlig analog zu dem des skalaren Falls in Kap. 2.2.2 □

Anmerkungen (genau wie im skalaren Fall):

- In obiger Formel für L_{inhom} bezeichnet man die feste, aber beliebig wählbare inhomogene Lösung y_p als *partikuläre Lösung*.
- Die obige Formel besagt:
  Es reicht L_{hom} sowie *eine einzige* inhomogene Lösung zu kennen, um *alle* inhomogenen Lösungen zu kennen.
- Die Eigenschaften der Lösungsmengen L_{hom} und L_{inhom} entsprechen genau den Eigenschaften der Lösungsmengen von homogenen bzw. inhomogenen *Linearen Gleichungssystemen!* Dies motiviert die obige Begriffsbildung "hom./inhom. *lineare* Dgl".

Es bleibt noch zu klären, **welche Dimension** der homogene Lösungsraum

$$L_{hom} = \{\vec{y} : I \to \mathbb{R}^n \,|\, \vec{y}\,'(t) = A(t)\,\vec{y}(t),\ \vec{y}(t_0) = \vec{y}_0,\ \vec{y}_0 \in \mathbb{R}^n\}\,,$$

hat ($t_0 \in \mathbb{R}$ fest).

Um Ordnung in diese Menge zu bekommen, betrachten wir eine *Basis* des Raumes der *Anfangswerte*, also des $\mathbb{R}^n$:
Sei also $\vec{b}_1, ..., \vec{b}_n$ eine Basis des $\mathbb{R}^n$.

Sei $t \mapsto \vec{y}_i(t)$ die zu dem Anfangswert $\vec{b}_i$ zugehörige Lösung, $i = 1, ..., n$.

Die Lösung zu einem *beliebigen* Anfangswert $\vec{y}_0 = \sum_{i=1}^{n} \alpha_i \vec{b}_i$ ist offensichtlich die Linearkombination $t \mapsto \vec{y}(t) = \sum_{i=1}^{n} \alpha_i \vec{y}_i(t)$ (kann man elementar nachprüfen).

Zusammengefasst:

Sei $\vec{b}_1, ..., \vec{b}_n$ eine Basis des Raumes $\mathbb{R}^n$ (der Anfangswerte).
Seien $t \to \vec{y}_1(t), ..., t \to \vec{y}_n(t)$ die zu den Anfangswerten $\vec{b}_1, ..., \vec{b}_n$ gehörigen Lösungen.
Zum AW $\vec{y}_0 = \sum_{i=1}^{n} \alpha_i \vec{b}_i$ gehört dann die Lösung $\vec{y}(t) = \sum_{i=1}^{n} \alpha_i \vec{y}_i(t)$.

Man sieht daran: *Jede* Lösung, also jedes $\vec{y} \in L_{hom}$, lässt sich als LK der $\vec{y}_i$, schreiben.
D.h. die $\vec{y}_i$, $i = 1, ..., n$ bilden ein **EZS** von L_{hom}.
Also ist die Dimension von L_{hom} höchstens n.

Zeigen wir nun, dass die $\vec{y}_i$ **linear unabhängig** sind:
Betrachten wir eine LK der $\vec{y}_i$, die die Nullfunktion ergibt: $\sum_{i=1}^{n} \beta_i \vec{y}_i(t) = \vec{0} \, \forall t \in I$.

Durch Einsetzen von $t := t_0$ ergibt sich $\vec{0} = \sum_{i=1}^{n} \beta_i \vec{y}_i(t_0) = \sum_{i=1}^{n} \beta_i \vec{b}_i$.
Wegen der Basiseigenschaft der $\vec{b}_i$ folgt, dass alle $\beta_i = 0$ sind.
Somit sind die $\vec{y}_i$ linear unabhängige Funktionen, somit eines *Basis von* L_{hom}.
Also:

Satz (Struktur der Lösungsmenge, Fortsetzung)
Sei $\vec{b}_1, ..., \vec{b}_n$ eine Basis des Raumes $\mathbb{R}^n$ (der Anfangswerte).
Dann bilden die zu diesen Anfangswerten gehörenden Lösungen $t \mapsto \vec{y}_1(t), ..., t \mapsto \vec{y}_n(t)$ eine *Basis* des Lösungsraumes L_{hom}.
Insbesondere ist also

$$\dim(L_{hom}) = n \, .$$

(Hier versagt also die Analogie zwischen lin. Dgl.-Systemen und lin. Gleichungssystemen: Während der Lösungsraum eines Dgl.-Systems im $\mathbb{R}^n$ immer n-dimensional ist, kann der Lösungsraum eines LGS mit $n \times n$-Matrix auch kleiner als n-dimensional sein.)

Def. (Fundamentalsystem, Fundamentallösungen, etc.)

Eine Basis des Lösungsraumes L_{hom} wird als *Fundamentalsystem (FS)* des Dgl-Systems bezeichnet; seine Mitglieder bezeichnet man als *Fundamentallösungen (FL)*.

Die Matrix $W(t) := [\vec{y}_1(t), ..., \vec{y}_n(t)] \in \mathbb{R}^{n\times n}$, deren Spalten Fundamentallösungen sind, heißt *Fundamentalmatrix*. Die Zahl (eigentlich: Funktion) $\det(W(t))$ heißt *Wronski-Determinante*.

Josef Wronski: 1776-1853, polnischer Philosoph und Mathematiker

Ob n Lösungen ein FS bilden, kann man am Wert der Wronski-Determinante erkennen:

Sei $t_* \in I$ beliebig. Dann gilt für Dgl-Lösungen $t \mapsto \vec{y}_1(t), ..., t \mapsto \vec{y}_n(t)$:

$$\begin{aligned} \vec{y}_1, ..., \vec{y}_n \in \mathrm{Abb}(I, \mathbb{R}^n) \text{ bilden FS} &\iff \vec{y}_1(t_*), ..., \vec{y}_n(t_*) \in \mathbb{R}^n \text{ bilden Basis des } \mathbb{R}^n \\ &\iff W(t_*) = [\vec{y}_1(t_*), ..., \vec{y}_n(t_*)] \text{ ist invertierbar} \\ &\iff \det(W(t_*)) \neq 0, \end{aligned}$$

D.h. man kann die FS-Eigenschaft an einer *beliebigen* Stelle t_* prüfen.

Grund: Die unteren beiden Äquivalenzen sind klar (s. 1. Sem.).
Zur oberen Äquivalenz:
Folgt aus den Überlegungen auf den vorangegangenen beiden Folien (t_* als t_0 betrachten/behandeln).

Im Folgenden wollen wir uns mit dem praktischen Lösen von homogenen linearen Dgl.-Systemen beschäftigen.
Das bedeutet, wir suchen ein Verfahren, um ein FS zu berechnen (denn mit einem FS kennen wir *alle* Lösungen des homogenen Systems).

2.4.2 Berechnung eines Fundamentalsystems für lineare Systeme erster Ordnung mit konstanten Koeffizienten

– Zuerst die schlechte Nachricht:
Im Fall, dass die Systemmatrix $A(t)$ tatsächlich von t abhängt und $n>1$ ist, kann man kein allgemeines Verfahren angeben, das zu einem FS führt.
(Es gibt ein "Verfahren" bekannt als *d'Alembert'sches Reduktionsverfahren*: Falls man eine Lösung raten(!) kann, dann kann man diese Lösung benutzen, um aus dem gegeben System im $\mathbb{R}^n$ ein System im $\mathbb{R}^{n-1}$ zu machen. Damit wollen wir uns nicht beschäftigen.
Im Fall $n=2$ kommt man, sobald man eine Lösung raten kann, immer damit durch: Denn nach der Reduktion ist $n=1$, und *skalare* lineare Dgln erster Ordnung kann man mit T.d.V. immer lösen (→Kap.II.2.2), auch bei t-Abhängigkeit der Daten.)

– **Im Fall, dass die Systemmatrix unabhängig von t ist, gibt es jedoch Verfahren, um ein FS zu berechnen.** Damit wollen wir uns nun beschäftigen. Wir suchen also ein FS für das System

$$\vec{y}\,'(t) = A\,\vec{y}(t)\,. \qquad (*)$$

Idee: Wir tasten uns nach und nach von leichten zu schwierigeren Fällen vor: Wir starten mit dem 'Trivialfall', dass die Matrix A eine **Diagonalmatrix** ist.

Sei also $A = \mathrm{diag}(\lambda_1, ..., \lambda_n)$. Das System (*) lautet dann:

$$\vec{y}\,'(t) = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \vec{y}(t) \qquad \Longleftrightarrow \qquad \begin{cases} y_1'(t) = \lambda_1\, y_1(t) \\ \vdots \\ y_n'(t) = \lambda_n\, y_n(t) \end{cases}$$

Das Dgl-System zerfällt ('*entkoppelt*') in n skalare Probleme, die unabhängig voneinander sind!
Diese n skalaren linearen Dgln können wir leicht lösen (Kap. 2.2.2):
Die Lösungen lauten $y_1(t) = c_1\, e^{\lambda_1 t}$,...,$y_n(t) = c_n\, e^{\lambda_n t}$, wobei $c_1, ..., c_n \in \mathbb{R}$ beliebig.

Unser System hat demnach die Lösungsmenge

$$L_{hom} = \left\{ \vec{y} : I \to \mathbb{R}^n \mid \vec{y}(t) = \begin{pmatrix} c_1\, e^{\lambda_1 t} \\ \vdots \\ c_n\, e^{\lambda_n t} \end{pmatrix} \;,\; c_1, ..., c_n \in \mathbb{R} \right\}$$

Dies kann man schreiben als

$$\begin{aligned} L_{hom} &= \left\{ \vec{y} : I \to \mathbb{R}^n \mid \vec{y}(t) = c_1 \begin{pmatrix} e^{\lambda_1 t} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + ... + c_n \begin{pmatrix} 0 \\ \vdots \\ 0 \\ e^{\lambda_n t} \end{pmatrix} \;,\; c_1, ..., c_n \in \mathbb{R} \right\} \\ &= \mathrm{span} \left\{ \begin{pmatrix} e^{\lambda_1 t} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, ..., \begin{pmatrix} 0 \\ \vdots \\ 0 \\ e^{\lambda_n t} \end{pmatrix} \right\} \end{aligned}$$

Ergebnis: Falls A eine Diagonalmatrix $A = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ ist, ist $e^{\lambda_1 t}\,\vec{e}_1, ..., e^{\lambda_n t}\,\vec{e}_n$ ein Fundamentalsystem.

Erinnerung: Die λ_i sind übrigens die EW von A, die $\vec{e}_i$ sind EV von A.

Nun kommt eine Strategie zum tragen, die wir bereits benutzt haben, um die Definitheit von (symmetrischen) Matrizen (Hesse-Matrix) zu untersuchen (Kap. 1.1):

Falls A keine Diagonalmatrix ist, können wir versuchen, sie zu *diagonalisieren.*

Beachte: in Kap. I.1 ging es nur um *symmetrische* Matrizen; diese sind immer diagonalisierbar.
Bei unserem Problem (*) müssen wir die Diagonalisierbarkeit von A nun explizit voraussetzen.

Bestimmung eines FS im Fall, dass A diagonalisierbar ist:

Sei A diagonalisierbar:

$$A = QDQ^{-1}, \quad D = \mathrm{diag}(\lambda_1, ..., \lambda_n)$$

(Erinnerung: Die λ_i sind EW von A, die Spalten $\vec{q}_i$ von Q sind zugehörige EV von A.)

Wir setzen für A in die Dgl ein:

$$\begin{aligned} & \frac{d}{dt}\vec{y}(t) &=& QDQ^{-1}\vec{y}(t) \qquad |Q^{-1}\cdot \\ \Leftrightarrow \quad & \frac{d}{dt}Q^{-1}\vec{y}(t) &=& DQ^{-1}\vec{y}(t) \qquad |\,\text{Subst. } \vec{u}(t) := Q^{-1}\vec{y}(t) \\ & \Leftrightarrow \quad \vec{u}\,'(t) &=& D\,\vec{u}(t) \qquad (**) \end{aligned}$$

Wir haben somit das Dgl-System auf Diagonalform gebracht ('entkoppelt')!

Ein FS für (**) ist somit (s.o.)

$$\vec{u}_1(t) = e^{\lambda_1 t}\vec{e}_1, \ \dots\ , \vec{u}_n(t) = e^{\lambda_n t}\vec{e}_n\,.$$

Mittels Rücksubstitution $\vec{y}_i(t) := Q\,\vec{u}_i(t)$ bekommen wir daraus ein FS für (*): $\vec{y}_i(t) = Q\vec{u}_i(t) = Q\vec{e}_i\, e^{\lambda_i t} = \vec{q}_i\, e^{\lambda_i t}$, $i = 1, ..., n$.

Ergebnis:

Satz (Fundamentalsystem bei diagonalisierbarer Systemmatrix)
Sei A diagonalisierbar. Seien $\lambda_1, ..., \lambda_n$ die (nicht notwendigerweise verschiedenen) Eigenwerte von A, und sei $\vec{q}_1, ..., \vec{q}_n$ eine Basis des $\mathbb{R}^n$, so dass $\vec{q}_i$ Eigenvektor zum Eigenwert λ_i ist.
Dann ist durch

$$e^{\lambda_1 t}\vec{q}_1, \ \dots\ , e^{\lambda_n t}\vec{q}_n$$

ein Fundamentalsystem für das lineare Dgl-System $\vec{y}\,'(t) = A\vec{y}(t)$ gegeben.

Anmerkungen:

1. Im diagonalisierbaren Fall reicht es also, die EW und EV von A zu kennen, um ein FS aufzustellen.
   Beispielrechnung: s. Tafel.

2. Erinnerung: Eine Matrix ist genau dann diagonalisierbar, wenn für jeden Eigenwert algebraische und geometrische Vielfachheit übereinstimmen. Das wiederum ist genau dann der Fall, wenn es eine Basis des $\mathbb{R}^n$ aus Eigenvektoren gibt.

3. Obiger Satz liefert eine Erklärung dafür, warum bei linearen skalaren Dgln n-ter Ordnung mit konstanten Koeffizienten (z.B. beim Federschwinger mit oder ohne Reibung, Kap. 2.1) ein "Exponentialansatz" $y(t) = c\,e^{\lambda t}$ zum Ziel führt:
   Lineare skalare Dgln n-ter Ordnung können in lineare Systeme erster Ordnung umgewandelt werden. Diese Systeme haben (s.o.) Lösungen der Form $e^{\lambda t}\vec{q}$. Für das ursprüngliche skalare Dgl-Problem braucht man vom Lösungsvektor nur die erste Komponente, diese hat die Form $c\,e^{\lambda t}$.

Der Fall komplexer Eigenwerte

Wir sind bei obigen Überlegungen stillschweigend davon ausgegangen, dass alle Eigenwerte reell sind, dass also A über $\mathbb{R}$ diagonalisierbar ist.

Wie kommt man zu einem FS, wenn es echt-komplexe Eigenwerte gibt?

Zunächst einmal kann man ein Dgl-System, dessen Matrix rein reell ist, problemlos als komplexes Dgl-System per $A \in \mathbb{R}^{n\times n} \subset \mathbb{C}^{n\times n}$ auffassen. Das heißt, man kann durchaus nach *komplexen* Lösungen $t \mapsto \vec{y}(t) \in \mathbb{C}^n$ eines reellen Dgl-Systems suchen.
Obige Überlegungen zur Struktur der Lösungsmenge bleiben korrekt, d.h. L_{hom} ist dann ein n-dimensionaler $\mathbb{C}$-Vektorraum, nämlich Unterraum des $\mathbb{C}$-VR $\mathrm{Abb}(I, \mathbb{C}^n)$. Auch die Herleitung des letzten Satzes unter Verwendung der Diagonalisierung von A bleibt korrekt.
Obiger Satz liefert also im Fall, dass komplexe Eigenwerte von A vorkommen, und A über $\mathbb{C}$ diagonalisierbar ist, ein Fundamentalsystem, d.h. eine Basis des $\mathbb{C}$-Vektorraums L_{hom}.

Jedoch: In der Praxis ist man, bei rein reeller Systemmatrix, an rein reellen Lösungen interessiert, nicht an komplexen! (Und die Theorie besagt ja, dass es auch ein *reelles* FS geben muss!)
Nur unser Satz hat leider ein komplexes FS geliefert!

Es gibt eine Vorgehensweise, wie man aus einem komplexen FS ein reelles FS gewinnen kann:

Wir gehen nun also davon aus, dass die Matrix A nicht nur reelle, sondern auch echt-komplexe EW hat, und sie über $\mathbb{C}$ diagonaliserbar ist.

Erinnerung: Nicht-reelle EW von rellen Matrizen treten immer als konjugiert-komplexe Paare $\lambda, \bar{\lambda}$ auf, und auch die zugehörigen EV sind konjugiert-komplex zueinander.

Man bekommt somit die nicht-reellen Fundamentallösungen immer *paarweise* als

$$\vec{y}_1(t) = e^{\lambda t}\vec{q}, \quad \vec{y}_2(t) = e^{\bar{\lambda} t}\,\overline{\vec{q}}$$

auf. Diese beiden Fundamentallösungen sind dann konjugiert-komplex zueinander:

$$\vec{y}_2(t) = e^{\bar{\lambda} t}\,\overline{\vec{q}} = e^{\overline{\lambda t}}\,\overline{\vec{q}} \overset{(*)}{=} \overline{e^{\lambda t}}\,\overline{\vec{q}} = \overline{e^{\lambda t}\vec{q}} = \overline{\vec{y}_1(t)}\,,$$

dabei kann man (*) mittels Exp.-reihe oder Euler-Formel einsehen.
Man bekommt also durch geeignete LK-Bildung *reelle* Lösungen:

$$\vec{y}_{1,reell} := \frac{1}{2}(\vec{y}_1 + \vec{y}_2) = \mathrm{Re}(\vec{y}_1), \quad \vec{y}_{2,reell} := \frac{1}{2i}(\vec{y}_1 - \vec{y}_2) = \mathrm{Im}(\vec{y}_1)$$

Man kann das noch detaillierter hinschreiben: Wir zerlegen $\lambda \in \mathbb{C}$ und $\vec{q} \in \mathbb{C}^n$ in Real- und Imaginärteil:

$$\lambda = a+bi, \quad \vec{q} = \vec{r}+\vec{s}i, \quad a, b \in \mathbb{R}, \quad \vec{r}, \vec{s} \in \mathbb{R}^n$$

Wir bekommen die Darstellungen

$$\begin{aligned} \vec{y}_1(t) &= (\vec{r}+i\vec{s})\, e^{(a+bi)t} = (\vec{r}+i\vec{s})\, e^{at}e^{bti} = (\vec{r}+i\vec{s})\,(\cos bt + i\sin bt)e^{at} \\ &= \underbrace{(\vec{r}\cos bt - \vec{s}\sin bt)\, e^{at}}_{=Re(\vec{y}_1(t))=\vec{y}_{1,reell}(t)} + i\ \underbrace{(\vec{s}\cos bt + \vec{r}\sin bt)\, e^{at}}_{=Im(\vec{y}_1(t))=\vec{y}_{2,reell}(t)}\,. \end{aligned}$$

Ergebnis:

Umwandlung eines komplexen FS in ein reelles FS

Ist eine Fundamentallösung $y_1(t) = e^{\lambda t}\vec{q} = e^{(a+bi)t}(\vec{r}+\vec{s}i)$ $(a, b \in \mathbb{R}, \vec{r}, \vec{s} \in \mathbb{R}^n)$ eines reellen Dgl-System $\vec{y}' = A\vec{y}$ echt-komplex, dann ist auch $y_2(t) = e^{\bar{\lambda} t}\overline{\vec{q}} = e^{(a-bi)t}(\vec{r}-\vec{s}i)$ eine Fundamentallösung, und es gilt, dass $\vec{y}_1(t)$ und $\vec{y}_2(t)$ konjugiert-komplex zueinander sind.

Die Funktionen

$$\begin{aligned} \vec{y}_{1,reell}(t) := \mathrm{Re}(\vec{y}_1(t)) = \tfrac{1}{2}(\vec{y}_1(t)+\vec{y}_2(t)) &= (\vec{r}\cos bt - \vec{s}\sin bt)\, e^{at} \\ \vec{y}_{2,reell}(t) := \mathrm{Im}(\vec{y}_1(t)) = \tfrac{1}{2i}(\vec{y}_1(t)-\vec{y}_2(t)) &= (\vec{s}\cos bt + \vec{r}\sin bt)\, e^{at} \end{aligned}$$

sind dann *reelle* Lösungen.
Ersetzt man im komplexen FS die komplexen Paare durch die so gefundenen reellen Paare, bekommt man ein *reelles* FS.

Bemerkung: Das Konstruktionsprinzip, also $\mathrm{Re}(\vec{y}_1(t))$, $\mathrm{Im}(\vec{y}_1(t))$, kann man sich leicht merken. Den hinteren Teil dieser Formeln braucht man nicht auswendig zu wissen.

Rechenbeispiel: s. Tafel

Bestimmung eines Fundamentalsystems in dem Fall, dass A (auch über $\mathbb{C}$) nicht diagonalisierbar ist

Erinnerung:

$A \in \mathbb{R}^{n\times n}$ nicht diagonalisierbar (über $\mathbb{C}$) $\iff \sum_{\lambda_i \text{ EW}} \dim(\text{Eig}(\lambda_i)) < n \qquad (*)$

$\iff$ Es gibt EW mit geom.Vfht.<alg.Vfht.

Funktionen der Bauart $e^{\lambda_i t}\vec{q}_i$, wobei $\vec{q}_i$ EV ist, sind auch im nicht-diagonalisierbaren Fall Lösungen (um das zu sehen: einfach in Dgl einsetzen! s.Tafel), aber wegen (*) bekommen wir weniger als n viele (d.h. zu wenige) Fundamentallösungen dieser Bauart!

Beispiel: Sei $A = \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}$. Diese Matrix hat die Eigenwerte $\lambda_1 = 3$ mit algebraischer Vielfachheit 2 und $\lambda_2 = 4$ mit algebraischer Vielfachheit 1. Die geometrische Vielfachheit von λ_1 und λ_2 ist jedoch nur jeweils= 1, denn $\text{Eig}(\lambda_1) = \text{span}\{\vec{e}_1\}$, $\text{Eig}(\lambda_2) = \text{span}\{\vec{e}_3\}$; wir finden mit unserer bisherigen Vorgehensweise nur zwei(!) Fundamentallösungen $\vec{y}_1(t) = \vec{e}_1\, e^{3t}$, $\vec{y}_2(t) = \vec{e}_3\, e^{4t}$; ein FS muss jedoch aus 3 FL bestehen.

Um die "fehlenden" Fundamentallösungen zu finden, brauchen wir einen weiteren Begriff:

Def. (Hauptvektor)
Sei $A \in \mathbb{R}^{n\times n}$ oder $A \in \mathbb{C}^{n\times n}$. Sei $\lambda \in \mathbb{C}$ ein EW von A. Ein $\vec{v} \neq \vec{0}$ heißt *Hauptvektor (HV)* von A zum EW λ, falls es ein $k \in \mathbb{N}$ gibt, so dass $(A-\lambda E_n)^k \vec{v} = \vec{0}$ gilt; d.h. $\vec{v} \in \text{Kern}((A-\lambda E_n)^k)$.
Das kleinste $k \in \mathbb{N}$, fur das obige Gleichung gilt, heißt *Stufe* des Hauptvektors.

Wichtige Eigenschaften von Hauptvektoren:

1. **Hauptvektoren der Stufe 1 sind nichts anderes als die Eigenvektoren.** Der Begriff des Hauptvektors kann in diesem Sinne als Verallgemeinerung des Begriffs des Eigenvektors betrachtet werden.

2. Sei $H_k(\lambda) := \{\vec{v} \in \mathbb{C}^n \mid (A-\lambda E_n)^k \vec{v} = \vec{0}\} = \text{Kern}((A-\lambda E_n)^k)$
   $=$ (Menge aller HV vom Grad $\leq k$ zum EW λ) $\cup \{\vec{0}\}$
   Es ist elementar zu zeigen: Alle $H_k(\lambda)$ sind ineinanderliegende Unterräume (eine "Fahne"):

   $$\text{Eig}(\lambda) = H_1(\lambda) \subseteq H_2(\lambda) \subseteq \ldots \subseteq \mathbb{C}^n$$

   Beispiel: ggf. s. Tafel

3. Sobald in obiger Folge von Räumen $H_k(\lambda) = H_{k+1}(\lambda)$ auftritt, folgt $H_k(\lambda) = H_{k+1}(\lambda) = H_{k+2}(\lambda) = \ldots$.

Beweis: Sei $H_k(\lambda) = H_{k+1}(\lambda)$ (*) und $\vec{v} \in H_{k+2}(\lambda)$. Somit

$$\vec{0} = (A-\lambda E_n)^{k+2}\vec{v} = (A-\lambda E_n)^{k+1} \underbrace{(A-\lambda E_n)\vec{v}}_{=:\vec{w}}$$

Es ist somit $\vec{w} := (A-\lambda E_n)\vec{v} \in \text{Kern}((A-\lambda E_n)^{k+1}) = H_{k+1}(\lambda) = H_k(\lambda) = \text{Kern}((A-\lambda E_n)^k)$, somit $\vec{0} = (A-\lambda E_n)^k \vec{w} = (A-\lambda E_n)^{k+1}\vec{v}$, somit $\vec{v} \in H_{k+1}(\lambda)$ □

4. Aus 3. folgt, dass die Fahne in 2. präzisiert werden kann zu

$$\text{Eig}(\lambda) = H_1(\lambda) \subsetneq H_2(\lambda) \subsetneq ... \subsetneq H_k(\lambda) = H_{k+1}(\lambda) = H_{k+2}(\lambda) = ..., \quad (*)$$

für ein $k = k(\lambda) \in \{1, ..., n\}$.
Insbesondere gibt es zu jedem EW λ eine maximalen Stufe, zu der HV vorliegen (und diese ist höchstens n). Gibt es zu einem EW keine HV der Stufe k, dann gibt es auch keine der Stufen $k+1, k+2, ...$

Anwendung von "dim" auf (*) ergibt

$$\text{geom.Vfht.}(\lambda) = \dim(H_1(\lambda)) \lneq \dim(H_2(\lambda)) \lneq ... \lneq \dim(H_k(\lambda)) = \dim(H_{k+1}(\lambda)) = ...$$

5. Man kann zeigen (hier ohne Beweis): **Die Dimension, bei der die Fahne der Räume $H_k(\lambda)$ in 4. stationär wird, ist gerade die algebraische Vielfachheit von λ.**
Das bedeutet: **Zu jedem EW λ findet man immer genau so viele l.u. Hauptvektoren, wie die algebraische Vielfachheit von λ angibt.**

6. Man kann zeigen (hier ohne Beweis):
Räume von Hauptvektoren zu verschiedenen EW sind immer de-facto-disjunkt:
Für EW $\lambda_1 \neq \lambda_2$ und $k_1, k_2 \in \mathbb{N}$ ist immer $H_{k_1}(\lambda_1) \cap H_{k_2}(\lambda_2) = \{\vec{0}\}$.

7. 5. und 6. ergeben: **Zu jeder Matrix $A \in \mathbb{R}^{n\times n}$ oder $A \in \mathbb{C}^{n\times n}$ gibt es immer n Hauptvektoren, die eine Basis des ganzen $\mathbb{C}^n$ bilden.**
Zum Vergleich: Bei (über $\mathbb{C}$) *diagonalisierbaren* Matrizen gibt es immer eine Basis des $\mathbb{C}^n$ aus *EV*; bei Matrizen, die nicht diagonalisierbar sind, müssen wir noch HV hinzunehmen, um eine Basis zu bekommen.

8. **Zur praktischen Berechnung von HV ist folgender Sachverhalt nützlich:**

Ein $\vec{v} \in \mathbb{C}^n$ ist genau dann HV der Stufe k ($k \geq 2$) zum EW λ, wenn es einen HV $\vec{w}$ der Stufe $k-1$ zum EW λ gibt, so dass

$$(A-\lambda E_n)\,\vec{v} = \vec{w}$$

gilt.

Beweis: s. Tafel

Das bedeutet: Kennt man alle HV der Stufe $k-1$, kann man durch Lösen von LGS HV der Stufe k finden!

Beachte: Die Lösung des obigen LGS ist nicht eindeutig, denn $\dim(\text{Kern}(A-\lambda E_n)) = \dim(\text{Eig}(\lambda)) \geq 1$.

Wir müssen nun noch klären:

Was haben die Hauptvektoren von A mit dem Finden eines Fundamentalsystems der Dgl. $\vec{y}' = A\vec{y}$ zu tun?

(Danach werden wir einige Berechnungen zum Finden von HV (basierend auf 8.) durchführen und damit Fundamentalsysteme aufstellen.)

Satz (Fundamentalsystem im nicht-diagonalisierbaren Fall)

Sei $A \in \mathbb{R}^{n\times n}$ oder $A \in \mathbb{C}^{n\times n}$ und $\lambda \in \mathbb{C}$ ein EW von A. Seien $\vec{v}_1, ..., \vec{v}_m$ eine "Kette" von Hauptvektoren zum EW λ; das soll heißen:
$\vec{v}_k$ ist HV der Stufe k zum EW λ, und es gilt (vgl. 8.):

$$(A-\lambda E_n)\,\vec{v}_{k+1} = \vec{v}_k \qquad (1)$$

(i) Dann ist

$$\vec{y}(t) := e^{\lambda t}\left(\vec{v}_m + t\,\vec{v}_{m-1} + \frac{t^2}{2!}\,\vec{v}_{m-2} + ... + \frac{t^{m-1}}{(m-1)!}\,\vec{v}_1\right) \qquad (2)$$

eine Lösung des Dgl.-Systems

$$\vec{y}\,'(t) = A\,\vec{y}(t)\,. \qquad (3)$$

(ii) Es gibt ein Fundamentalsystem für (3), das aus Funktionen der Bauart (2)/(1) besteht.
Jeder EW $\lambda \in \mathbb{C}$ trägt dabei genau so viele Fundamentallösungen bei, wie seine algebraische Vielfachheit ist.

- Fazit: Mittels HV können wir *immer* Fundamentalsysteme berechnen. Wir brauchen die HV (nur) dann, wenn es nicht 'genug' linear unabhängige EV gibt.
- Beachte: Die HV der höchsten (niedrigsten) Stufe werden mit den niedrigsten (höchsten) t-Potenzen versehen.
- Anforderung (1) legt nahe, die HV mittels 8. zu berechnen und nicht mittels Berechnung von $\text{Kern}((A-\lambda E_n)^k)$, $k = 1, 2, ...$

Beweis.
Beweis von (i): Erfolgt einfach durch Einsetzen von (2) in Dgl. (3):

$$\begin{aligned} A\,\vec{y}(t) - \vec{y}\,'(t) &= e^{\lambda t}\ \left(A\vec{v}_m + t\,A\vec{v}_{m-1} + \frac{t^2}{2!}\,A\vec{v}_{m-2} + ... + \frac{t^{m-2}}{(m-2)!}\,A\vec{v}_2 + \frac{t^{m-1}}{(m-1)!}\,A\vec{v}_1 \right) \\ &\quad -\lambda\, e^{\lambda t}\ \left(\vec{v}_m + t\,\vec{v}_{m-1} + \frac{t^2}{2!}\,\vec{v}_{m-2} + ... + \frac{t^{m-2}}{(m-2)!}\,\vec{v}_2 + \frac{t^{m-1}}{(m-1)!}\,\vec{v}_1 \right) \\ &\quad -e^{\lambda t}\ \left(\qquad \vec{v}_{m-1} + \frac{2t}{2!}\,\vec{v}_{m-2} + ... + \frac{(m-2)t^{m-3}}{(m-2)!}\,\vec{v}_2 + \frac{(m-1)t^{m-2}}{(m-1)!}\,\vec{v}_1 \right) \end{aligned}$$

Wir sortieren nach Potenzen von t:

$$\begin{aligned} A\,\vec{y}(t) - \vec{y}\,'(t) &= e^{\lambda t}\ \Bigg[\underbrace{(A\vec{v}_m - \lambda\vec{v}_m - \vec{v}_{m-1})}_{=\vec{0} \text{ nach (1)}} + \frac{t}{1!}\,\underbrace{(A\vec{v}_{m-1} - \lambda\vec{v}_{m-1} - \vec{v}_{m-2})}_{=\vec{0} \text{ nach (1)}} + ... \\ &\qquad ... + \frac{t^{m-2}}{(m-2)!}\,\underbrace{(A\vec{v}_2 - \lambda\vec{v}_2 - \vec{v}_1)}_{=\vec{0} \text{ nach (1)}} + \frac{t^{m-1}}{(m-1)!}\,\underbrace{(A\vec{v}_1 - \lambda\vec{v}_1)}_{=\vec{0} \text{ da EV}} \Bigg] \\ &= \vec{0} \end{aligned}$$

Beweis von (ii): Kann man aus 5.-8. folgern. □

Wir wollen hier keinen Algorithmus zur FS-Berechnung in voller Allgemeinheit erarbeiten. Stattdessen einige häufig auftretenden Fälle:

Beispiele zur Berechnung von FS unter Verwendung von HV:

(a) Sei A eine 3×3-Matrix mit $p(\lambda) = (-1)^3(\lambda-\lambda_1)(\lambda-\lambda_2)^2$, $\lambda_1 \neq \lambda_2$.;
es seien beide Eigenräume 1-dimensional:
$\text{Eig}(\lambda_1) = \text{span}\{\vec{v}_1\}$, $\text{Eig}(\lambda_2) = \text{span}\{\vec{v}_2\}$.
$\Rightarrow$ Es 'fehlt' genau ein HV; dieser muss (s. 5.) zum EW λ_2 gehören, und er muss (s. 4.) Stufe 2 haben.
Bestimme diesen HV $\vec{v}_{2,2}$ als eine (beliebige) Lösung des LGS $(A-\lambda_2 E_3)\,\vec{v}_{2,2} = \vec{v}_2$.
Dann lautet ein FS: $\{\vec{v}_1\,e^{\lambda_1 t},\ \vec{v}_2\,e^{\lambda_2 t},\ (\vec{v}_{2,2}+t\vec{v}_2)\,e^{\lambda_2 t}\}$

(b) Sei A eine 5×5-Matrix mit $p(\lambda) = (-1)^5(\lambda-\lambda_1)^2(\lambda-\lambda_2)^3$, $\lambda_1 \neq \lambda_2$.;
es seien beide Eigenräume 1-dimensional:
$\text{Eig}(\lambda_1) = \text{span}\{\vec{v}_1\}$, $\text{Eig}(\lambda_2) = \text{span}\{\vec{v}_2\}$.
$\Rightarrow$ Es 'fehlen' ein HV zum EW λ_1 und zwei HV zum EW λ_2.
Bestimme den HV $\vec{v}_{1,2}$ zum EW λ_1 als eine (beliebige) Lösung des LGS $(A-\lambda_1 E_5)\,\vec{v}_{1,2} = \vec{v}_1$.
Bestimme einen HV $\vec{v}_{2,2}$ zum EW λ_2 als eine (beliebige) Lösung des LGS

$(A-\lambda_2 E_5)\,\vec{v}_{2,2} = \vec{v}_2$.
Mehr als einen l.u. HV als Lsg dieses LGS wird man nicht finden, denn $\dim(\mathrm{Kern}(A-\lambda_2 E_5)) = 1$. Der noch fehlende HV muss also Stufe 3 haben (und zum EW λ_2 gehören).
Wir berechnen ihn als eine beliebige Lsg des LGS $(A-\lambda_2 E_5)\,\vec{v}_{2,3} = \vec{v}_{2,2}$
Dann lautet ein FS: $\{\vec{v}_1\,e^{\lambda_1 t},\ (\vec{v}_{1,2}+t\vec{v}_1)\,e^{\lambda_1 t},\ \vec{v}_2\,e^{\lambda_2 t},\ (\vec{v}_{2,2}+t\vec{v}_2)\,e^{\lambda_2 t},\ (\vec{v}_{2,3}+t\vec{v}_{2,2}+\frac{t^2}{2}\vec{v}_2)\,e^{\lambda_2 t}\}$

(c) Ein schwierigerer Fall:
Sei A eine 5×5-Matrix mit $p(\lambda) = (-1)^5(\lambda-\lambda_1)^2(\lambda-\lambda_2)^3$, $\lambda_1 \neq \lambda_2$.;
es seien beide Eigenräume 2-dimensional:
$\mathrm{Eig}(\lambda_1) = \mathrm{span}\{\vec{v}_1, \vec{v}_2\}$, $\mathrm{Eig}(\lambda_2) = \mathrm{span}\{\vec{v}_3, \vec{v}_4\}$.
$\Rightarrow$ Es 'fehlt' ein HV $\vec{v}_5$ zum EW λ_2.
Schwierigkeit: Wir haben hier *zwei* l.u. EV $\vec{v}_3, \vec{v}_4$ zum EW λ_2; sollen wir nun $\vec{v}_5$ als Lgs von $(A-\lambda_2 E_5)\,\vec{v}_5 = \vec{v}_3$ oder als Lsg von $(A-\lambda_2 E_5)\,\vec{v}_5 = \vec{v}_4$ bestimmen?
Antwort: I.a. wird beides nicht funktionieren, sondern auf der rechten Seite muss ein $\alpha\vec{v}_3+\beta\vec{v}_4 \in \mathrm{Bild}(A-\lambda_2 E_5)$ stehen!
(Bem.: Da der Kern von $A-\lambda_2 E_5$ offenbar 2-dim. ist, ist nach dem Dim.-Formel (1.Sem.) das Bild nur $5-2=3$-dim.)
Entweder man 'sieht' geeignete α, β, oder man berechnet zuerst $\mathrm{Bild}(A-\lambda_2 E_5)$ und berechnet α, β, so dass $\alpha\vec{v}_3+\beta\vec{v}_4 \in \mathrm{Bild}(A-\lambda_2 E_5)$, oder man löst allgemein das LGS $(A-\lambda_2 E_5)\,\vec{v}_5 = \alpha\vec{v}_3+\beta\vec{v}_4$ für die Unbekannten $\vec{v}_5, \alpha, \beta$, und wählt eine Lsg mit $(\alpha, \beta) \neq (0, 0)$.
Dann lautet ein FS: $\{\vec{v}_1\,e^{\lambda_1 t},\ \vec{v}_2\,e^{\lambda_1 t},\ \vec{v}_3\,e^{\lambda_2 t},\ \vec{v}_4\,e^{\lambda_2 t},\ (\vec{v}_5+t(\alpha\vec{v}_3+\beta\vec{v}_4))\,e^{\lambda_2 t}\}$

(d) Ein noch schwierigerer Fall:
Sei A eine 4×4-Matrix mit $p(\lambda) = (\lambda-\lambda_1)^4$;
es sei $\mathrm{Eig}(\lambda_1) = \mathrm{span}\{\vec{v}_1, \vec{v}_2\}$.
$\Rightarrow$ Es 'fehlten' zwei HV $\vec{v}_3, \vec{v}_4$ zum EW λ_1.
Schwierigkeit: Es ist a priori unklar, ob zwei HV der Stufe 2 oder ein HV der Stufe 2 und ein HV der Stufe 3 zu finden sind. (Allgemein: Auf jeder Stufe ergeben sich *höchstens* so viele HV, wie auf der um eins niedrigeren Stufe.)
Man stellt also ein LGS wie in (c) auf, d.h. $(A-\lambda_1 E_5)\,\vec{v}_5 = \alpha\vec{v}_1+\beta\vec{v}_2$, und untersucht, welche/wie viele Lösungen $\vec{v}_5, \alpha, \beta$ man findet,...
Solch komplizierte Fälle betrachten wir eher nicht. Für solche Fälle gibt es noch ein anderes Berechnungsverfahren, siehe später.

Siehe Tafel oder Übung: Hauptvektoren für sog. Jordan-Matrizen

Kurze Übersicht, welche Fälle (mit obigem Rechenverfahren) einfach/schwer sind:

Sei λ ein EW von A. Gesucht sind Fundamentallösungen des Dgl-Systems. Fälle:

I. leichtester Fall: $\text{geomVfht}(\lambda) = \text{algVfht}(\lambda) =: k \in \mathbb{N}$:
Kein HV höherer Stufe ist nötig; es genügen k l.u. EV aus $\text{Eig}(\lambda)$
(siehe Satz S. 10)

II. mit obiger Methode recht leicht (aber rechenaufwändiger als I.):
$\text{geomVfht}(\lambda) = 1$ und $\text{algVfht}(\lambda) =: k > 1$:
Suche für jedes $j = 1, ..., k$ genau einen HV der j-ten Stufe (mittels 8.)
$\rightsquigarrow$ Beispiele (a), (b)

III. schwierig mit obiger Methode: Fall $1 < \text{geomVfht}(\lambda) < \text{algVfht}(\lambda)$:
a priori unklar, wie viele HV welcher Stufe zu finden sind, oder welcher Vektor auf die rechte Seite des zur Berechnung von HV aufzustellenden LGS gehört
$\rightsquigarrow$ Beispiele (c), (d)
(Ausnahme, III.-A: Wenn $\text{geomVfht}(\lambda) = \text{algVfht}(\lambda) - 1$ (s. Bsp. (c)): So gerade noch machbar mit obiger Methode.)

Bemerkung:
Im Satz auf S. 104 kann man (2) unter Verwendung von (1) auch als

$$\begin{aligned} \vec{y}(t) &\overset{(1)}{=} e^{\lambda t} \left(\vec{v}_m + t\,(A - \lambda E_n)\,\vec{v}_m + \tfrac{t^2}{2}\,(A - \lambda E_n)^2\,\vec{v}_m + ... + \tfrac{t^{m-1}}{(m-1)!}\,(A - \lambda E_n)^{m-1}\,\vec{v}_m \right) \\ &= e^{\lambda t} \left(\sum_{k=0}^{m-1} \tfrac{t^k}{k!}\,(A - \lambda E_n)^k \right) \vec{v}_m \end{aligned}$$

schreiben.

Bemerkung:
Es existieren, gerade für die 'schwierigeren' Fälle, **alternative Berechnungsmethoden**:
Anstatt Gleichung (1) (also Eigenschaft 8.) benutze die *Definition* von HV, um HV zu berechnen. Also:
Nach dem man zu einem EW λ die EV berechnet hat, also den Kern von $A - \lambda E_n$, berechne HV zweiter Stufe, d.h. ein Element von $H_2(\lambda) \setminus H_1(\lambda)$, indem man $H_2(\lambda) = \text{Kern}(A - \lambda E_n)^2$ berechnet. Doch Vorsicht: Nicht nur HV 2. Stufe, auch die EV sind in dieser Menge($= H_2(\lambda)$) enthalten! Um diese nicht 'nochmal' zu bekommen, suche nur nach solchen Lösungen $\vec{h}$ von $(A - \lambda E_n)^2 \vec{h} = \vec{0}$, die *orthogonal* zu den zuvor berechneten EV sind!

Im Beispiel (c): $\text{Eig}(\lambda_2) = \text{span}\{\vec{v}_3, \vec{v}_4\}$ und es fehlt 1 HV:
Suche HV zweiter Stufe $\vec{h}$ als Lösung des LGS $\begin{pmatrix} (A - \lambda E_n)^2 \\ \vec{v}_3^T \\ \vec{v}_4^T \end{pmatrix} \vec{h} = \begin{pmatrix} \vec{0} \\ 0 \\ 0 \end{pmatrix}$.
Nun muss man a posteriori durch geeignete Wahl des EV $\vec{v} := (A - \lambda E_n)\vec{h}$ dafür sorgen, dass die 'Ketteneigenschaft' (1) erfüllt ist, d.h. nur für dieses $\vec{v}$ ist $(\vec{h} + t\,\vec{v})e^{\lambda t}$ eine FL.

Im Beispiel (d): $\text{Eig}(\lambda_2) = \text{span}\{\vec{v}_1, \vec{v}_2\}$ und es fehlen 2 HV:
Suche HV zweiter Stufe $\vec{h}$ als Lösung(en) des LGs $\begin{pmatrix} (A - \lambda E_n)^2 \\ \vec{v}_1^T \\ \vec{v}_2^T \end{pmatrix} \vec{h} = \begin{pmatrix} \vec{0} \\ 0 \\ 0 \end{pmatrix}$.

Falls die Lösungsmenge 2-dim ist, wähle 2 l.u. Lösungen $\vec{h}_1, \vec{h}_2$ aus, und bilde dann Ketten $\vec{w}_1 := (A-\lambda E_n)\vec{h}_1$, $\vec{w}_2 := (A-\lambda E_n)\vec{h}_2$, d.h. die beiden 2-gliedrigen Ketten $\{\vec{w}_1, \vec{h}_1\}, \{\vec{w}_2, \vec{h}_2\}$ liefern die beiden FL $(\vec{h}_1+\vec{w}_1 t)\, e^{\lambda_1 t}, (\vec{h}_2+\vec{w}_2 t)\, e^{\lambda_1 t}$.
Falls stattdessen die Lösungsmenge nur 1-dim ist, wähle eine Lösung $\vec{h} \neq \vec{0}$ als HV zweiter Stufe aus, und suche dann nach einem HV dritter Stufe $\vec{r} \neq \vec{0}$, indem man das LGS $\begin{pmatrix} (A-\lambda E_n)^3 \\ \vec{v}_1^T \\ \vec{v}_2^T \\ \vec{h} \end{pmatrix} \vec{r} = \begin{pmatrix} \vec{0} \\ 0 \\ 0 \\ 0 \end{pmatrix}$ löst.
Bilde dann eine zweistufige, bei $\vec{h}$ beginnende Kette, sowie eine dreistufige, bei $\vec{r}$ beginnende Kette, also die FL $\{(\vec{h} + (A-\lambda_1 E_4)\,\vec{h}\, t)e^{\lambda_1 t}, (\vec{r} + (A-\lambda_1 E_4)\,\vec{r}\, t + \frac{1}{2}(A-\lambda_1 E_4)^2\,\vec{r}\, t^2)e^{\lambda_1 t}\}$.

Beurteilung dieser Methode: Sie erscheint systematischer, gerade für 'schwierige' Fälle, allerdings für 'einfache' Fälle wie Beispiele (a) und (b) und vielleicht (c) (d.h. für Fälle II. und vielleicht III.-A auf S. 107) vielleicht etwas umständlich, aber ansonsten für Fall III. sinnvoll.

2.4.3 Berechnung einer partikulären Lösung für das *inhomogene* Dgl-System

Für *homogene* Dgl-Systeme zeigten Kap. 2.4.1–2.4.2, wie man eine Basis des Lösungsraumes L_{hom} findet.
Falls das Dgl-System *inhomogen* ist,

$$\vec{y}\,'(t) = A(t)\,\vec{y}(t) + \vec{b}(t),$$

benötigen wir des weiteren *eine* Lösung $\vec{y}_p$ des inhomogenen Systems, um $L_{inhom} = \{\vec{y}_p\} + L_{hom}$ zu haben.

Bemerkung: Von nun an darf die Systemmatrix A durchaus wieder von t abhängen.

Wir setzen voraus, dass wir ein FS $\{\vec{y}_1, ..., \vec{y}_n\}$ bereits gefunden haben (was schwierig ist, falls A von t abhängt).

Idee zum Finden von $\vec{y}_p$: Variation der Konstanten, vgl. Kap. 2.2.2:
Es seien $\vec{y}_1, ..., \vec{y}_n$ ein FS (also l.u. Lösungen der *homogenen* Dgl).

Ansatz:	$\vec{y}_p(t) := \sum_{i=1}^{n} c_i(t)\,\vec{y}_i(t) = W(t)\,\vec{c}(t)$

Zur Bestimmung der c_i setzen wir den Ansatz in die inhomogene Dgl ein:

$$\vec{y}_p'(t) \stackrel{!}{=} A(t)\,\vec{y}_p(t) + \vec{b}(t) \;\Leftrightarrow\; \sum_{i=1}^{n} c_i'(t)\,\vec{y}_i(t) + \sum_{i=1}^{n} c_i(t)\,\underline{\vec{y}_i'(t)} \;\stackrel{!}{=}\; A(t)\left(\sum_{i=1}^{n} c_i(t)\,\vec{y}_i(t)\right) + \vec{b}(t)$$

$$= \sum_{i=1}^{n} c_i(t)\,\underline{A(t)\vec{y}_i(t)} + \vec{b}(t)$$

Da die $\vec{y}_i$ Lösungen der homogenen Dgl sind, fallen Terme weg, und man bekommt:

$$\underbrace{\sum_{i=1}^{n} \vec{y}_i(t)\,c_i'(t)}_{=W(t)\,\vec{c}\,'(t)} = \vec{b}(t)$$

Somit

$$\vec{c}\,'(t) = W(t)^{-1}\vec{b}(t)$$

Indem man jede Komponente dieser Vektorgleichung integriert, bekommt man $\vec{c}(t)$. (Dabei braucht man keine Integrationskonstanten, da man nur *eine* Lösung braucht.) Ist $\vec{c}(t)$ bekannt, setzt man es in den Ansatz ein und hat $\vec{y}_p(t)$.

2.5 Lineare skalare Dgln n-ter Ordnung

Wir betrachten

Lineare skalare Dgl n-ter Ordnung

$$y^{(n)}(t) + a_{n-1}(t)y^{(n-1)}(t) + ... + a_1(t)y'(t) + a_0(t)y(t) = b(t) \qquad (2.1)$$

Idee:
– Wir wissen: Wir können diese **Umwandeln in ein System von Dgln erster Ordnung**.
– Dieses System können wir, zumindest im Fall *konstanter* Koeffizienten a_i, gemäß Kap. 2.4 lösen! Fertig!

Nichtsdestotrotz: **Wir wollen uns diesen Vorgang im Detail anschauen.**
Es wird sich nämlich herausstellen, dass man sich z.B. die mühsame Berechnung von charakteristischem Polynom, Eigenvektoren, Hauptvektoren — anders als man erwarten sollte — komplett ersparen kann!

Bemerkung: Die Umwandlung in ein System ist auch dann möglich, wenn die Koeffizienten a_i von t abhängen. Das Lösen nach Kap. 2.4.2 scheitert jedoch dann meist, da wir dann i.a. kein Fundamentalsystem finden können.

Umwandlung der skalaren Dgl (2.1) in System erster Ordnung:

$$\vec{y}_{neu}(t) := \begin{pmatrix} y(t) \\ y'(t) \\ \vdots \\ y^{(n-1)}(t) \end{pmatrix} \tag{2.2}$$

$$\Rightarrow \quad \vec{y}'_{neu}(t) = \begin{pmatrix} y'(t) \\ y''(t) \\ \vdots \\ y^{(n-1)}(t) \\ y^{(n)}(t) \end{pmatrix} \overset{\text{(Dgl.)}}{=} \begin{pmatrix} y_{neu,2}(t) \\ y_{neu,3}(t) \\ \vdots \\ y_{neu,n}(t) \\ b(t) - [a_{n-1}y_{neu,n}(t) + \ldots + a_0 y_{neu,1}(t)] \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} 0 & 1 & & \\ & \ddots & & \\ & & & 1 \\ -a_0 & \ldots & \ldots & -a_{n-1} \end{pmatrix}}_{=:A} \vec{y}_{neu}(t) + \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \\ b(t) \end{pmatrix}}_{=:\vec{b}(t)} \tag{2.3}$$

Wir wissen: (2.3) hat FS bestehend aus n Lösungen; hinsichtlich des skalaren Ausgangsproblems ist von den Fundamentallösungen (und der partikulären Lösung) nur die 1. Komponente relevant (wegen (2.2))!

Sei angenommen, dass die a_i konstant sind. Um das FS zu berechnen, brauchen wir EW und EV (u. ggf. HV) von A. Daher brauchen wir zuallererst das charakteristische Polynom von A.

Erstaunlicherweise müssen wir das charakteristische Polynom für Matrizen der obigen Form nicht in jedem Fall mühsam berechnen, sondern es gibt eine sehr einfache allgemeine Formel:

Satz (char. Polynom von Matrizen, die sich aus lin. skalaren Dgl. n-ter Ordnung m. konst. Koeff. ergeben)

Die Matrix $A \in \mathbb{R}^{n\times n}$ aus (2.3), $n \geq 2$, hat das charakteristische Polynom

$$p(\lambda) = (-1)^n \left(\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0\right). \tag{2.4}$$

Das bedeutet, dass man das p dem Problem (2.1) direkt ansehen kann, ohne jede Rechnung! Kurz: "Aus Ableitungen werden Potenzen"

Beweis: Per Induktion. Entwickle $p(\lambda)=\det(A-\lambda E_n)$ nach der 1. Spalte. Siehe Tafel. □

Als nächstes müssen wir uns den Eigenräumen von A (sowie ggf HV) zuwenden. Dazu überlegt man sich zunächst, dass alle Eigenräume *ein*dimensional sind. (Denn die oberen $n-1$ Zeilen sind offenbar linear abhängig.)
Um FL zu finden, benötigen wir also zu jedem EW λ_i jeweils einen HV auf den Stufen $2, ..., r_i$, wobei r_i die algebraische Vielfachheit von λ_i sei, und die zum EW λ_i gehörenden r_i vielen FL haben also die Form $\vec{p}_{i,j}(t)e^{\lambda_i t}$, wobei $\vec{p}_{i,j}$ ein vektorielles Polynom vom Grad j ist, $j=0, ..., r_i-1$.

Aber: Wir brauchen, da wir uns eigentlich für (2.1) statt für (2.3) interessieren, nur die erste Komponente $y(t)=y_{neu,1}(t)$ von $\vec{y}_{neu}(t)$, nicht für den gesamten Vektor $\vec{y}_{neu}(t)$.

Diese ersten Komponenten der hergeleiteten FL von (2.3) haben die Form $y_{i,j}(t) = p_{i,j}(t)\, e^{\lambda_i t}$, wobei die $p_{i,j}$ skalare Polynome vom Grad j, $j=0, ..., r_i-1$, sind.

Im Raum der Polynome vom Grad $\leq r_i-1$ machen wir einen Basiswechsel: Wir wechseln zur Standardbasis $t^0, t^1, ..., t^{r_i-1}$ und nehmen somit als FL $y_{i,j}(t)=t^j\, e^{\lambda_i t}$, $j=0, ..., i-1$ nehmen kann.

Satz (Lösungsraum der homogenen skalaren linearen Dgl. n-ter Ordnung)
Die Dgl. (2.1) mit konstanten Koeffizienten a_i hat im homogenen Fall (d.h. $b \equiv 0$) eine Basis des n-dimensionalen Lösungsraumes L_{hom}, die sich wie folgt ergibt:
Für jede Nullstelle $\lambda_i \in \mathbb{C}$ des charakteristischen Polynoms (2.4) nehme die Funktionen

$$e^{\lambda_i t},\ t\, e^{\lambda_i t},\ ...,\ t^{r_i-1}\, e^{\lambda_i t},$$

wobei r_i die Vielfachheit der Nullstelle λ_i ist.

Das Finden einer Basis von L_{hom} ist also für lineare skalare Dgln n-ter Ordnung erheblich einfacher als für beliebige lineare Systeme erster Ordnung: **Man muss die EV und HV der Matrix A aus (3) nicht explizit ausrechnen!**

Beispiel: Bestimme die Lösungmenge der Dgl. $y''''+6y'''+12y''+10y'+3y=0$.
Siehe Tafel.

Fall *komplexer* EW: Falls es, bei *reellen* a_i, nicht-reelle EW gibt, so treten diese immer paarweise als $\lambda = a+bi$, $\bar{\lambda} = a-bi$, $a, b \in \mathbb{R}$, auf, d.h. man hat Paare von komplexen FL

$$y_1(t) = e^{(a+bi)t} = e^{at}(\cos bt + i \sin bt),$$
$$y_2(t) = e^{(a-bi)t} = e^{at}(\cos bt - i \sin bt).$$

Um daraus *reelle* FL zu machen, gehe vor wie in Kap. 2.4 :

$$y_{1,reell}(t) = \tfrac{1}{2}(y_1(t)+y_2(t)) = e^{at}\cos bt = \mathrm{Re}(y_1(t))$$
$$y_{2,reell}(t) = \tfrac{1}{2i}(y_1(t)-y_2(t)) = e^{at}\sin bt = \mathrm{Im}(y_1(t))$$

Beispiel: Berechne ein reelles Fundamentalsystem für $y''+y'+y=0$.
siehe Tafel.

Ein weiteres **Beispiel** ist der **Federschwinger mit Reibung** (Kap. II.1). Dort liefert die obige Vorgehensweise ebenfalls zunächst ein komplexes FS, was in ein reelles FS überführt werden kann, woraus sich Schwingungsterme ergeben.

Hier zeigt sich besonders deutlich die **Nützlichkeit der komplexen Zahlen**: Ein Naturvorgang wird im Reellen beschrieben (durch eine reelle Dgl). Das Rechnen findet im Komplexen statt, liefert aber dennoch am Ende die gesuchte reelle Lösung.

Beispiel z.Ü.: Wir nehmen an, es sei ein Loch gebohrt zentral durch die gesamte Erde. Wir hüpfen in das Loch hinein. Wie lange dauert es, bis wir den Erdmittelpunkt erreichen? Welche Geschwindigkeit haben wir dort? Wie lange dauert es insgesamt, bis wir am entgegengesetzten Punkt der Erdoberfläche auftauchen?
Wir treffen dazu folgende Annahmen: Die Gravitationskraft im Inneren der Erde ist proportional zum Abstand vom Erdmittelpunkt (das ist gültig, sofern man annimmt, dass die Masse im Erdinneren homogen verteilt ist) und beträgt an der Erdoberfläche mg wobei die Erdbeschleunigung $g = 9.81\ m/s^2$ und m unsere Masse ist. Der Erdradius ist in etwa 6370 km.
Wir vernachlässigen jegliche Reibung, sowie jegliche Effekte, die mit der Rotation der Erde (sog. Corioliskräfte würden uns an den Schachtrand drücken, sofern wir das Loch nicht von Pol zu Pol bohren) oder den Temperaturen im Erdinneren zu tun haben.

Was noch aussteht:
Um auch **inhomogene** Probleme der Art (2.1) lösen zu können, brauchen wir zusätzlich zum FS noch eine **partikuläre Lösung** y_p.

Idee: Wir betrachten, wie schon zum Finden des FS, wieder das zugehörige System 1. Ordnung (2.3). Nach Kap. II.4.3 findet man für das System eine partikuläre Lösung mit dem Ansatz **Variation der Konstanten, 'V.d.K.'**, also mit dem Ansatz $\vec{y}_p(t) := W(t)\,\vec{c}(t)$, wobei $W(t)$ die Fundamentalmatrix des Systems ist; dies führt auf ein LGS

$$W(t)\,\vec{c}\,'(t) = \vec{b}(t), \qquad (*)$$

das man löst, anschließend muss komponentenweise integriert werden, um die $c_i(t)$ zu bekommen, die man dann in den Ansatz einsetzt, um $\vec{y}_p(t)$ zu bekommen.

Speziell hier (d.h. wenn das System aus einer skalaren Dgl n-ter Ordnung hervorgeht): Matrix $W(t)$ ergibt sich aus (2.2), Vektor $\vec{b}(t)$ aus (2.3), somit:

'V.d.K.' zum Finden einer partikulären Lösung einer skalaren linearen Dgl n-ter Ordnung

Seien $y_1, ..., y_n$ n linear unabhängige Lösungen der *homogenen* skalaren Dgl. (die wir gemäß dem vorangegangenen Satz berechnet haben).
Das LGS (*) lautet dann

$$\begin{pmatrix} y_1(t) & \dots & y_n(t) \\ y_1'(t) & & y_n'(t) \\ \vdots & & \vdots \\ y_1^{(n-1)}(t) & \dots & y_n^{(n-1)}(t) \end{pmatrix} \begin{pmatrix} c_1'(t) \\ c_2'(t) \\ \vdots \\ c_n'(t) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b(t) \end{pmatrix}$$

Löse dieses LGS, und bilde für jede Komponente $c_i'(t)$ des Lösungsvektors eine Stammfunktion $c_i(t)$. Dann ist

$$y_p(t) = \sum_{i=1}^{n} c_i(t)\, y_i(t)$$

eine Lösung der inhomogenen Dgl.

(Die obige Gleichung für y_p ergibt sich, indem wir aus dem obigen vektoriellen Ansatz $\vec{y}_p(t) = W(t)\,\vec{c}(t) = \sum_{i=1}^{n} c_i(t)\,\vec{y}_i(t)$ die erste Komponente nehmen.)

Bsp.: Bestimme die allgemeine Lösung der (inhomogenen) linearen Dgl. $y'' - 5y' + 6y = 4e^t$
Siehe Tafel.

Eine Alternative zu 'Variation der Konstanten' zum Finden der partikulären Lösung:
"Ansatz vom Typ der Rechten Seite":

(A) Falls die Koeffizienten a_j der Dgl konstant sind und **falls die rechte Seite** $b(t)$ **eine spezielle Form hat**, nämlich $b(t)$ ein Polynom ist oder eine Exponentialfunktion oder ein Produkt aus einem Polynom und einem Exp.-term, kann man mit einem **Ansatz** eine partikuläre Lösung finden:

(1.) Falls $b(t) = \sum_{j=0}^{m} b_j t^j$, $m \in \mathbb{N}_0$, $b_m \neq 0$:
(a) Ansatz $y_p(t) := \sum_{j=0}^{m} \alpha_j t^j$, falls $p(0) \neq 0$, und
(b) Ansatz $y_p(t) := t^r \sum_{j=0}^{m} \alpha_j t^j$, falls 0 eine r-fache Nullstelle von p ist.

(2.) Falls $b(t) = b\, e^{kt}$, $b \neq 0$, $k \in \mathbb{C}$:
(a) Ansatz $y_p(t) := \alpha e^{kt}$, falls $p(k) \neq 0$, und
(b) Ansatz $y_p(t) := \alpha e^{kt} t^r$, falls k eine r-fache Nullstelle von p ist.

(3.) Falls $b(t) = e^{kt} \sum_{j=0}^{m} b_j t^j$, $m \in \mathbb{N}_0$, $b_m \neq 0$, $k \in \mathbb{C}$:
(a) Ansatz $y_p(t) := e^{kt} \sum_{j=0}^{m} \alpha_j t^j$, falls $p(k) \neq 0$, und
(b) Ansatz $y_p(t) := e^{kt} t^r \sum_{j=0}^{m} \alpha_j t^j$, falls k eine r-fache Nullstelle von p ist.

Den Parameter α bzw. die Parameter α_j berechnet man, indem man den Ansatz in die Dgl. einsetzt ($\rightsquigarrow$Koeffizientenvergleich)

Bemerkung: (1.) und (2.) sind Spezialfälle von (3.), mit $k = 0$ bzw. mit $m = 0$, und (a) kann man jeweils als Spezialfall von (b) auffassen mittels Vielfachheit $r = 0$; d.h. es reicht, (3-b) zu kennen.

(B) Erweiterung: Falls $b(t)$ an Stelle des Exponentialterms (auch möglich: *zusätzlich* zum Exponentialterm) einen Faktor $\cos \omega t$ (oder $\sin \omega t$) enthält: "Komplexifizierung der Dgl": Setze $\tilde{b}(t) := e^{i\omega t}$; es ist dann $\tilde{b}(t) = \mathrm{Re}(b(t))$ (bzw. $\tilde{b}(t) = \mathrm{Im}(b(t))$; zu dieser Dgl. mit komplexer rechter Seite $\tilde{b}(t)$ finde komplexe partikuläre Lösung, von dieser nehme den Realteil (bzw. Imaginärteil).

(C) Erweiterung: Falls die rechte Seite nicht die o.g. Form hat, aber eine *Summe* aus Termen der obigen Form ist:

Bestimme mit obiger Maschinerie für jeden der Summanden separat eine partikuläre Lösung, addiere dann diese partikulären Lösungen.

Ggf.:
Physikalisches **Beispiel** für inhomogene lineare Dgl zweiter Ordnung, gelöst mit Ansatz vom Typ der rechten Seite:
Federschwinger mit periodischer Anregung (=äußerer Kraft),
s. Tafel

2.6 Numerische Verfahren

2.6.1 Numerische Verfahren für gewöhnliche Differentialgleichungen und Differentialgleichungssysteme

— Hier nur ein ganz kurzer Einblick —

Oft kann man die Lösung von Anfangswertproblemen nicht exakt berechnen; z.B. bei nichtlinearen Systemen von Dgln oder linearen Systemen mit t-abhängigen Koeffizienten oder auch bei schwierigen nichtlinearen skalaren Dgln.
Dann ist man auf numerische Verfahren (Näherungsverfahren) angewiesen.

Konzept: Sei AWP (skalar oder System)

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0,$$

gegeben.
Wähle eine *Schrittweite* $h > 0$ und die zugehörigen *Gitterpunkte* $t_{n+1} := t_n + h$, $n \in \mathbb{N}$.
Ziel ist es, (nur) zu diesen diskreten t_n-Werten eine Näherung $y_n \approx y(t_n)$ zu finden;

$$\underbrace{y_n}_{\text{Näherung}} \overset{!}{\approx} \underbrace{y(t_n)}_{\text{ex.Lsg.}}, \; n = 1, 2, 3, ...$$

Wir brauchen nun einen Zusammenhang zwischen y_n und y_{n+1}, um sukzessive $y_1, y_2, y_3, ...$ berechnen zu können.

Eine einfache Idee: Ersetze die Ableitung $y'(t)$ in der Dgl durch einen Differenzenquotienten $\frac{y(t+h)-y(t)}{h} \approx y'(t)$:
Die exakte Lösung erfüllt also

$$\frac{y(t+h) - y(t)}{h} \approx f(t, y(t)), \quad y(t_0) = y_0,$$

Das motiviert, für die Näherungslösung zu fordern:

$$\frac{y_{n+1} - y_n}{h} = f(t_n, y_n), \quad \text{wobei } y_0 \text{ gegeben.}$$

Dies ist das sogenante

Euler'sches Polygonzugverfahren, Explizites Euler-Verfahren

$$y_{n+1} = y_n + h\, f(t_n, y_n)$$

Geometrische Interpretation: $f(t_n, y_n)$ entspricht einer Steigung; Skizze!

Man kann vermuten (und liegt damit richtig): Je kleiner h, desto genauer die Näherungslösung.

Anstatt h zu verkleinern, kann man aber auch versuchen, das Verfahren an sich zu verbessern:

Obiges Verfahren verwendet als Steigung der Näherungslösung im Intervall $[t_n, t_{n+1}]$ den Wert von f am linken Endpunkt des Intervalls, t_n. Besser wäre vermutlich eine Steigung aus der Mitte des Intervalls, also zu $\frac{1}{2}(t_n+t_{n+1}) = t_n + \frac{h}{2}$.
Wie bekommt man diese Steigung? Zumindest näherungsweise bekommt man diese Steigung, indem man zunächst einen Schritt von t_n nach $t_n + \frac{h}{2}$ macht mit dem obigen Euler-Verfahren, dort die Steigung 'abgreift', und mit dieser verbesserten Steigung den Schritt von t_n nach t_{n+1} durchführt (Skizze!).

Dies führt auf:

Verbessertes Euler-Verfahren

$$y_{n+1} = y_n + h\, f(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n))$$

Eine andere Art der Verbesserung des Euler-Verfahrens besteht darin, $y'(t_n)$ oder $y'(t_{n+1})$ nicht nur mit einer LK von $y(t_{n+1}), y(t_n)$ zu approximieren (als $y'(t_{n+1}) \approx \frac{1}{h}(y(t_{n+1}) - y(t_n))$, s.o.), sondern mit einer LK von *drei* Werten $y(t_{n+1}), y(t_n), y(t_{n-1})$, also als $h \cdot y'(t_{n+1}) \overset{!}{\approx} \alpha y(t_{n+1}) + \beta y(t_n) + \gamma y(t_{n-1})$.
Indem man die rechte Seite um t_n Taylor-entwickelt, kann man herausfinden, dass die Wahl $\alpha = \frac{3}{2}$, $\beta = -2$, $\gamma = \frac{1}{2}$ die beste Näherung liefert, also

$$\underbrace{\frac{3}{2} y_{n+1} - 2y_n + \frac{1}{2} y_{n-1}}_{\approx h \cdot y'(t_{n+1})} = h\, f(t_{n+1}, y_{n+1})$$

was man, wenn f 'einfach' ist, nach y_{n+1} auflösen kann (sonst: z.B. mit Newton-Verfahren Lösung y_{n+1} bestimmen); diese Methode nennt man *Backward-Difference-Formula, BDF2*.
Das ganze geht auch mit 4 statt 3 Summanden ($\rightsquigarrow$BDF3), usw.

Eine weitere Methode, um Verfahren herzuleiten, besteht darin, das AWP in eine Integralgleichung umzuwandeln und über $[t_n, t_{n+1}]$ zu betrachten:

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt.$$

Nun muss man einen Weg finden, um den Integranden zu approximieren unter Verwendung von $t_n, t_{n+1}, y_n, y_{n+1}, h$; tut man dies unter Verwendung des linken Funktionswertes, also $\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \approx (t_{n+1} - t_n) \cdot f(t_n, y(t_n))$, so liefert dies $y(t_{n+1}) \approx y(t_n) + h\, f(t_n, y(t_n))$, somit das uns bereits bekannte Euler'sche Polygonzugverfahren

$$y_{n+1} = y_n + h\, f(t_n, y_n)\,.$$

Nimmt man anstelle des linken Funktionswertes den Mittelwert aus linkem und rechtem Funktionswert, d.h. ersetzt man das zu berechnende Integral durch ein Trapez, $\int_{t_n}^{t_{n+1}} f(t,y(t))\,dt \approx (t_{n+1}-t_n)\cdot\frac{1}{2}\left[f(t_n,y(t_n))+f(t_{n+1},y(t_{n+1}))\right]$, so bekommt man das sog. (implizite) Trapezverfahren

$$y_{n+1} = y_n + \frac{h}{2}\left[f(t_n,y_n)+f(t_{n+1},y_{n+1})\right].$$

Ausgefeiltere Approximationen für das obige Integral liefern bessere Methoden; all diese sind bekannt unter dem Namen *Runge-Kutta-Verfahren*. Das bekannteste Runge-Kutta-Verfahren, das sog. *Klassische 4-stufige Runge-Kutta-Verfahren* (1902) lautet:

$$\begin{aligned}
k_1 &:= f(t_n, y_n)\\
k_2 &:= f(t_n+\tfrac{h}{2}, y_n+\tfrac{h}{2}k_1)\\
k_3 &:= f(t_n+\tfrac{h}{2}, y_n+\tfrac{h}{2}k_2)\\
k_4 &:= f(t_n+h, y_n+hk_3)\\
y_{n+1} &:= y_n + \tfrac{h}{6}\left(k_1+2k_2+2k_3+k_4\right)
\end{aligned}$$

All diese Verfahren und unzählige weitere funktionieren für skalare Dgln wie auch für Systeme; es wurden Verfahren entwickelt für Dgln, die bestimmt Eigenschaften haben,...
Insgesamt kann die Entwicklung von Näherungsverfahren für gew. Dgl. (anders als für partielle Dgln) heute als weitgehend abgeschlossen betrachtet werden.

Die **Genauigkeit dieser Näherungsverfahren** kann mathematisch exakt untersucht werden; inbesondere interessiert man sich für das Verhalten des Fehlers bei festem hinreichend glattem f für $h \to 0$. So kann man zeigen, dass die o.g. Verfahren eine Fehlerabschätzung der Form

$$\exists\, c=c(f)>0 :\ \forall\, h>0\ \forall\, t_n\in[t_0,t_0+T] :\ |y_n - y(t_n)| \le c\,h^p$$

erfüllen; die Zahl p heißt **Fehlerordnung** des Verfahrens.

Man kann beweisen, dass die Fehlerordnung des Euler'schen Polygonzugverfahrens bzw. fürs Verbesserte Euler-Verfahren bzw. fürs BDF-Verfahren bzw. fürs klassische Runge-Kutta-Verfahren $p=1$ bzw. $p=2$ bzw. $p=2$ bzw. $p=4$ ist, sofern f hinreichend glatt ist.

D.h. bei einer Halbierung der Schrittweite h (also Verdopplung der Anzahl der Schritte) kann man eine Halbierung bzw. Viertelung bzw. Sechzehntelung des Fehlers erwarten.

2.6.2 Numerische Verfahren für partielle Differentialgleichungen

— Hier nur ein ganz kurzer Einblick —

Partielle Differentialgleichungen (pDgl) beschreiben i.a. Vorgänge, bei denen nicht nur zeitliche Entwicklung, sondern auch ein räumlicher Transport eine Rolle spielt.
PDEs kann man nur in ganz seltenen Fällen analytisch lösen; Chancen dazu bestehen grob gesprochen, wenn die pDgl linear ist, alle Koeffizienten konstant sind, und das Gebiet, auf dem die pDgl zu lösen ist, ein Rechteck/Würfel ggf. ein Kreis/Quader ist. In anderen Fällen ist man auf *numerische* Lösungsverfahren angewiesen.

Ein Standard-Beispiel für eine PDE ist die *Wärmeleitungsgleichung* (hier in 2 Raumdimensionen):

$$\frac{\partial}{\partial t}u(t,x,y) - k\left[\frac{\partial^2}{\partial x^2}u(t,x,y) + \frac{\partial^2}{\partial y^2}u(t,x,y)\right] = f(t,x,y) \qquad \forall (t,x,y) \in [0,T]\times\Omega$$

Dabei ist f eine gegebene Quelle ('Heizdichte', ggf. $\equiv 0$), $k > 0$ ein Wärmeleitfähigkeitskoeffizient, und u die gesuchte Temperatur in einem Gebiet $[0,T]\times\Omega$, wobei $\Omega \subset \mathbb{R}^2$. Um Eindeutigkeit von Lösungen zu bekommen, muss man eine AB $u(0,x,y) \stackrel{!}{=} u_0(x,y)\,\forall\,(x,y)\in\Omega$ vorgeben, sowie zusätzlich *Randbedingungen (RB)* vorgeben, z.B. $u(t,x,y) \stackrel{!}{=} v(t,x,y)\,\forall (t,x,y) \in [0,T]\times\partial\Omega$ oder auch $\frac{\partial}{\partial\nu}u \stackrel{!}{=} v\,\forall\,(t,x,y)\in[0,T]\times\partial\Omega$, wobei ν der Normaleneinheitsvektor auf $\partial\Omega$ ist; im letzteren Fall modelliert das v einen Wärmefluss über den Rand; $v\equiv 0$ steht also für thermische Isolation.
Diffusionsprozesse (von Wärme, aber auch von Materie) werden durch obige pDgl beschrieben.

Falls man eine *stationäre* (d.h. zeitlich konstante) Wärmeverteilung $\frac{\partial}{\partial t}u(t,x,y) \stackrel{!}{=} 0$ sucht, so bekommt man die pDgl

$$-k\left[\frac{\partial^2}{\partial x^2}u(x,y) + \frac{\partial^2}{\partial y^2}u(x,y)\right] = f(x,y) \qquad \forall (x,y) \in \Omega$$

zusammen mit einer der obigen Randbedingungen (eine Anfangsbedingung ist hier nicht erforderlich).
Diese pDgl beschreibt übrigens auch (näherungsweise) *Minimalflächen* (z.B. Seifenblasen in einem durch die RB vorgegebenen ggf verbogenen Ring, bei einer ggf durch f gegebenen äußeren Kraftdichte (z.B. der Gravitation oder Wind).
Viele interessante theoretische Fragen kann man untersuchen, z.B. Existenz und Eindeutigkeit von Lösungen (vor allem: in welchem Funktionenraum?!; wie 'glatt' ist die Lösung?)

Numerische Verfahren für pDgl:

Es gibt im wesentlichen zwei-drei große Klassen von **numerischen Verfahren**:

- **Finite-Differenzen-Methoden (FDM)**
- **Finite-Elemente-Methoden (FEM)**
- ggf. kann man noch nenne: Finite-Volumen-Methoden (FVM)

FDM sind einfacher zu verstehen; FEM sind flexibler (z.B. bei krummlinig begrenzten Rechengebieten) aber anspruchsvoller zu verstehen und theoretisch zu analysieren ('Fehlerabschätzung').

Finite-Differenzen-Methoden (FDM):

Betrachte die obige *stationäre* pDgl, und zwar auf einem Gebiet $\Omega := [0,L]\times[0,L]$.
Diskretisierung des Gebietes: Lege ein Gitter über das Gebiet: Maschenweite $h := \frac{L}{n}$, wobei $n \in N$; Gitterpunkte $\vec{x}_{i,j} = (x_i, y_j)$ mit $x_i = ih$, $y_j = yh$, $i,j = 0,...,n$; es gibt also $(n+1)^2$ Gitterpunkte (s. Skizze).
Gesucht wird die Näherungslösung $u_{i,j}$ nur an den Gitterpunkten, d.h. es soll $u_{i,j} \stackrel{!}{\approx} u(x_i,y_j)$ sein, $i,j = 0,...,n$.

An den Randpunkten (d.h. $i=0 \vee i=n \vee j=0 \vee j=n$) sei $u_{i,j}$ durch eine RB vorgegeben.

Wir haben somit $(n-1)^2$ Unbekannte $u_{i,j}$, $i=1,...,n-1$, und brauchen somit $(n-1)^2$ viele Gleichungen.[2mm] Wir konstruieren an jedem Gitterpunkt eine Gleichung, und zwar wie folgt:
Die Ableitungen in der pDgl werden ersetzt (approximiert) durch geeignete "Differenzenquotienten".

Was geeignete Approximationen sind, kann man mittels Taylor-Entwicklung ermitteln, so z.B.: So ist z.B.

$$\begin{aligned} f'(x) &= \tfrac{f(x)-f(x-h)}{h} + O(h) \\ f'(x) &= \tfrac{f(x+h)-f(x-h)}{2h} + O(h^2) \\ f''(x) &= \tfrac{f(x-h)-2\,f(x)+f(x+h)}{h^2} + O(h^2) \end{aligned}$$

falls f hinreichen glatt.
(Man zeigt dies, indem man mit h bzw. h^2 durchmultipliziert und dann die rechte Seite, als Funktion von h aufgefasst, um $h_0=0$ entwickelt.)

Dies in die pDgln eingesetzt, bedeutet, dass die *exakte* Lösung

$$-k \cdot \left(\frac{u(x-h,y) - 2\,u(x,y) + u(x+h,y)}{h^2} + \frac{u(x,y-h) - 2\,u(x,y) + u(x,y+h)}{h^2} \right) = f(x,y) + O(h^2)$$

erfüllt, für alle x, insbes. für alle GP x_{ij}, $i=1,...,n-1$.

Dies motiviert, für die numerische Näherungslösung zu fordern:

$$-k \left(\frac{u_{i-1,j} - 2\,u_{i,j} + u_{i+1,j}}{h^2} + \frac{u_{i,j-1} - 2\,u_{i,j} + u_{i,j+1}}{h^2} \right) = f(x_i,y_i) =: f_{i,j} \quad \forall i,j = 1,...,n-1$$

Dies ist ein LGS mit $(n-1)^2$ Gleichungen und Unbekannten; in jeder Zeile sind aber nur maximal 5 von null verschiedene Einträge vorhanden; eine solche Matrix bezeichnet man als *dünn besetzt.*
Matrix im Fall $n=5$: s. Tafel.

Die Diskretisierung der *zeitabhängigen* Wärmeleitungsgleichung ergibt sich, indem man zusätzlich eine Zeitschrittweite τ einführt und diskrete Zeitpunkte $t_s=t_0+s\tau$, $s=0,1,2,...$, und eine Approximation $u_{s,i,j} \overset{!}{\approx} u(t_s,x_i,y_j)$. In jedem Zeitschritt $t_s \to t_{s+1}$ erhält man das LGS

$$\begin{aligned} &\frac{u_{s,i,j}}{\tau} - k \left(\frac{u_{s,i-1,j} - 2\,u_{s,i,j} + u_{s,i+1,j}}{h^2} + \frac{u_{s,i,j-1} - 2\,u_{s,i,j} + u_{s,i,j+1}}{h^2} \right) \\ &\qquad = f(t_s,x_i,y_i) + \frac{u_{s-1,i,j}}{\tau} =: \tilde{f}_{s,i,j} \qquad \forall i,j = 1,...,n-1\,. \end{aligned}$$

Zum Lösen des LGS:

- Per Gauß-Elimination: Für vollbesetzte Matrizen der Größe $(n-1)^2 \times (n-1)^2$ würde man $O((n^2)^3) = O(n^6)$ Rechenoperationen brauchen; unter Ausnutzung, dass nur in den inneren $2n$ vielen "Bändern" Werte ungleich null stehen, kommt man noch auf $O(n^4)$ Rechenoperationen.
  Problem: Während der Gauß-Elimination füllen sich die Nullen zwischen den besetzten Diagonalen auf mit Werten ungleich null, d.h. anstelle der 4 Einträge pro Zeile bekommt man bis zu n Einträge $\neq 0$ pro Zeile, insgesamt $O(n^3)$ viele Einträge.
  Denkt man an richtig große Probleme (z.B. $n=10^4$ oder $n=10^6$), so kann es schwierig sein, diese Matrix überhaupt zu *speichern!* (abgesehen vom riesigen Rechenaufwand).

– Fixpunkt-Verfahren (Jacobi-, Gauß-Seidel-) verändern die gegebene Matrix nicht → keine Speicherplatzprobleme!
Sie sind konvergent, d.h. verwendbar, da die obigen Matrizen *schwach diagonaldominant* bzw. echt *diagonaldominant* sind!
Rechenoperationen pro Iterationsschritt (Matrix-Vektor-Multiplikation): Nur $O(n^2)$ unter Ausnutzung der Nullen.
Die Konvergenzgeschwindigkeit sinkt jedoch leider ab, je größer n ist.

– Für LGS der obigen Bauart gibt es speziell angepasste moderne Löser, z.B. sog. *Mehrgitterverfahren*, die mit insgesamt nur $O(n^2)$ Operationen/Speicherplätzen auskommen (ab ca. 1980). (Mehrgitterverfahren verwenden häufig das Gauß-Seidel-Verfahren als Unteralgorithmus.)

Finite-Elemente-Methode (FEM):

(Ist flexibler, was kompliziertere Gebiete angeht, ist etwas schwieriger zu verstehen und zu analysieren; wurde von Ingenieuren erfunden, ab den 1970er/1980er Jahren mathematisch analysiert, Varianten werden auch heute noch weiterentwickelt.)

Das Rechengebiet wird mit einem *Dreiecksgitter* ("Triangulierung") überzogen. Die Näherungslösung sucht man (anders als bei FDM) nicht nur an den Gitterpunkten (des Dreiecksgitters), sondern an allen Punkten $\vec{x} \in \Omega$. Man macht als Ansatz, dass die Lösung auf jedem der Dreiecke eine affin-lineare Funktion sei, und global stetig (d.h. keine Sprünge an den Dreieckskanten). Diesen Ansatz möchte man die die pDgl einsetzen. Doch das kann man nicht unmittelbar machen, da die o.g. Ansatzfunktion nicht C^2 ist, in der pDgl aber zweite Ableitungen vorkommen. Stattdessen multipliziert man die pDgl. mit einer sog. Testfunktion, dann integriert man über Ω, und dann führt man partielle Integration durch, d.h. man verschiebt eine Ableitungsordnung von der Lösung auf die Testfunktion; es kommen nur noch Ableitungen erster Ordnung vor, und solche sind für obige stückweise-affin-linearen Funktionen (fast überall) wohldefiniert! Der oben beschriebene Ansatzraum ist endlichdimensional (Dimension=Anzahl Gitterpunkte). Man fordert, dass für alle Testfunktionen, die im Ansatzraum liegen, die hergeleitete Integralgleichung gelten soll. Das ist äquivalent dazu, dass, nach Auswahl einer geeigneten Basis des Ansatzraumes, für alle Basisfunktionen des Ansatzraumes die Integralgleichung erfüllt wird. Dies sind so viele Bedingungen, wie man Freiheitsgrade in der numerischen Lösung hat. Die zu berechnende numerische Lösung schreibt man als LK derselben Basisfunktionen des Ansatzraumes. Für die Koeffizienten dieser LK bekommt man so ein LGS. Das LGS ist ebenfalls schwach diagonaldominant und kann mit den gleichen Methoden wie das FDM-LGS gelöst werden.

Diese Erklärung war vermutlich zu knapp, um die FE-Methode zu verstehen; spezielle Bücher oder Vorlesungen über dieses Verfahren sind dazu wohl nötig.

3 Algebra und Anwendungen

Der Inhalt von Kapitel 3 umfasst die folgenden anvisierten Anwendungen:

— In Teil 1: Prüfsummen (Codierungstheorie; erkennen von Fehlern in Datensätzen)

- In Teil 2: RSA-Verschlüsselung (Kryptografie; Verschlüsselung mittels großer Primzahlen)

...sowie die Algebra, die man dazu braucht.

3.1 Algebra und Anwendung in der Codierungstheorie

Folgenes ist z.T. Wiederholung aus dem 1. Semester:

Def. (Gruppe, Monoid)
Eine Menge $G \neq \emptyset$ mit einer Verknüpfung $* : G \times G \to G$ heißt *Gruppe*, falls

(a) $\forall\, a,b,c \in G:\ a*(b*c) = (a*b)*c$ (Assoziativgesetz)

(b) $\forall\, a \in G \exists\, e \in G:\ a*e = e*a = a$ (Existenz eines neutralen Elements)

(c) $\forall\, a \in G\ \exists\, b \in G:\ a*b = b*a = e$ (Existenz eines Inversen)

Sind nur (a) und (b) erfüllt, heißt $(G,*)$ *Monoid*; ist nur (a) erfüllt, spricht man von einer *Halbgruppe*.
Ist zusätzlich zu (a),(b),(c) noch

(d) $a*b = b*a\ \forall\, a,b \in G$ (Kommutativgesetz),

so heißt $(G,*)$ *kommutative* oder *Abel'sche Gruppe*.

Wir haben im 1. Semester gezeigt:

– Neutrales Element ist in Gruppen (sogar in Monoiden) immer *eindeutig*.

– Inverse Elemente sind in Gruppen (sogar in Monoiden) immer eindeutig.

– Falls bekannt ist, dass $a \in G$ ein Inverses hat (also z.B. falls G eine Gruppe ist), so reicht es $a*b=e$ oder $b*a=e$ zu zeigen, um auf $b=a^{-1}$, $a=b^{-1}$ zu schließen; wenn nicht, dann müssen beide Gleichungen überprüft werden.

Beispiele für (unendliche) Gruppen:

- $(\mathbb{Z},+), (\mathbb{Q},+), (\mathbb{R},+), (\mathbb{C},+)$ und $(\mathbb{Q}\setminus\{0\},\cdot), (\mathbb{R}\setminus\{0\},\cdot), (\mathbb{C}\setminus\{0\},\cdot)$,
  jedoch *nicht* $(\mathbb{Z}\setminus\{0\},\cdot)$ oder $(\mathbb{N}_0,+)$, da Inverse fehlen; $\to$Monoid),

- $(\mathbb{R}^{n\times m}, +)$, jedoch nicht $(\mathbb{R}^{n\times n}, \cdot)$ oder $(\mathbb{R}^{n\times n}\setminus\{0\}, \cdot)$,
  sondern $(\{A\in\mathbb{R}^{n\times n} \mid A \text{ invertierbar}\}, \cdot)$ (diese Gruppe ist nicht abelsch)

- $(\mathbb{Z}[X], +), (\mathbb{Q}[X], +), (\mathbb{R}[X], +), (\mathbb{C}[X], +)$ Menge aller Polynome mit Koeffizienten in $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$,
  sowie $(\mathbb{Z}_n[X], +), (\mathbb{Q}_n[X], +), (\mathbb{R}_n[X], +), (\mathbb{C}_n[X], +)$ Menge aller Polynome mit Koeffizienten in $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ mit Polynomgrad $\leq n$.

Beispiele für endliche (abelsche) Gruppen sind

- $(\mathbb{Z}_n, +)$, wobei $\mathbb{Z}_n := \{0, 1, ..., n-1\}$ und die Verknüpfung '+' die Addition aus $\mathbb{Z}$ modulo n sei.
  (Ob auch $(\mathbb{Z}_n\setminus\{0\}, \cdot)$ eine Gruppe ist, hängt interessanterweise von n ab. Dazu später mehr.)

- $(\mathbb{Z}_{n_1}\times\mathbb{Z}_{n_2}\times ... \times\mathbb{Z}_{n_m}, +)$, wobei die Verknüpfung durch *komponentenweise* Anwendung der Verknüpfung aus $(\mathbb{Z}_n, +)$ sei.

Insbesondere bei *endlichen* Gruppen wird die Verknüpfung oft mittels einer Verknüpfungstabelle dargestellt, z.B. für $(\mathbb{Z}_3, +)$

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

In der Verknüpfungstabelle einer Gruppe kommt in jeder Zeile und Spalte jedes Element höchstens einmal vor,
denn angenommen in der Zeile von a sind zwei Einträge gleich, also $a * b_1 = a * b_2$ mit $b_1 \neq b_2$. Dann folgt per Multiplikation mit a^{-1}, dass $b_1 = b_2$, Widerspruch.

Ferner kommt in jeder Zeile und Spalte der Verküpfungstabelle einer Gruppe jedes Element *mindestens* einmal, somit also *genau einmal*, vor;
Grund: Ist die Gruppe *endlich*, so kann man dies einfach durch Abzählen der Elemente folgern. In endlichen und auch unendlichen Gruppen kann man wie folgt argumentieren: Seien $a, b\in G$ beliebig. Um zu zeigen, dass 'b' in der Zeile von a vorkommt, muss man ein $c\in G$ (entspricht einer Spalte) finden, so dass $a*c=b$. In einer Gruppe ist das einfach: $c:=a^{-1}*b$.

Die obige Überlegung kann man auch in *Monoiden* $(M, *)$ anstellen (wo also i.a. nicht *jedes* Element ein Inverses hat):

Hat $a \in M$ ein Inverses, so kommt in der zu a gehörenden Zeile und Spalte der Verknüpfungstabelle jedes Element jeweils genau einmal vor.

Sogar die Umkehrung gilt: Kommt in der Zeile und Spalte von $a \in M$ jedes Element jeweils genau einmal vor, dann ist a invertierbar.

Ringe:

Def. (Ring)
Eine Menge M mit zwei Verknüpfungen $+, \cdot : M \times M \to M$, $(M, +, \cdot)$, heißt *Ring*, falls

1. $(M, +)$ ist abelsche Gruppe
2. $(M, \cdot)$ ist assoziativ (d.h. ist Halbgruppe)
3. Distributivgesetze gelten: $(a+b) \cdot c = a \cdot c + b \cdot c$, $a \cdot (b+c) = a \cdot b + a \cdot c \, \forall \, a, b, c \in M$

Hat $(M, \cdot)$ ein neutrales Element (Bezeichung: "1"), so heißt $(M, +, \cdot)$ *Ring mit Eins(-element)*.
Ist $(M, \cdot)$ kommutativ, heißt der Ring *kommutativer Ring*.

Beispiele für Ringe:

- $(\mathbb{Z}, +, \cdot)$ ist kommutativer Ring mit Eins.
- $(2\mathbb{Z}, +, \cdot)$ ist kommutativer Ring ohne Eins.
- $(\mathbb{R}^{n \times n}, +, \cdot)$ und $(\mathrm{Lin}(\mathbb{R}^n, \mathbb{R}^n), +, \circ)$ sind (nicht kommutative) Ringe mit Eins; sie sind zueinander 'isomorph', d.h. es gibt Bijektion $\Phi : R_1 \to R_2$ mit $\Phi(a+b) = \Phi(a) + \Phi(b)$ und $\Phi(a \cdot b) = \Phi(a) \cdot \Phi(b) \, \forall \, a, b \in R_1$).
- $(\mathbb{Z}[X], +, \cdot)$, $(\mathbb{Q}[X], +, \cdot)$, $(\mathbb{R}[X], +, \cdot)$, $(\mathbb{C}[X], +, \cdot)$ sind kommutative Ringe mit Eins.
- $(\mathbb{Z}_n, +, \cdot)$ ist ein kommutativer Ring mit Eins.
  Warum: Dazu später mehr!

Körper:

Def. (Körper)
Ein kommutativer Ring $(M, +, \cdot)$ mit Eins, in dem jedes Element$\neq 0$ ein Inverses bzgl. '$\cdot$' hat, heißt *Körper*.
Äquivalent: Ein Ring $(M, +, \cdot)$, in dem $(M \setminus \{0\}, \cdot)$ eine abelsche Gruppe ist.
Äquivalent: Eine Menge M mit Verknüpfungen $+, \cdot : M \times M \to M$, wobei $(M, +)$ und $(M \setminus \{0\}, \cdot)$ abelsche Gruppen sind und zwischen $+, \cdot$ Distributivgesetze gelten.

Beispiele für Körper:

- $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, jedoch nicht $\mathbb{Z}$
- Die Menge der Matrizen $\mathbb{R}^{n\times n}$ ist kein(!) Körper (fehlende Inverse), ist nur ein Ring;
  die Menge der *invertierbaren* $n\times n$-Matrizen ist ebenfalls kein(!) Körper (fehlende Abgeschlossenheit bzgl. '+', darüber hinaus fehlende Kommutativität)
  und auch kein Ring (fehlende Abgeschlossenheit bzgl. '+').
- Man kann (s. 1. Sem.) $\mathbb{C}$ als $\mathbb{R}^2$ auffassen, wobei die Addition in $\mathbb{C}$ der Vektoraddition in $\mathbb{R}^2$ entspricht und zusätzlich eine Multiplikation eingeführt wird, derart, dass $\mathbb{C}$ ein Körper ist.
  Frage: Kann man auch für $\mathbb{R}^n$ mit $n > 2$ unter Verwendung der Addition von $\mathbb{R}^n$ und der Hinzunahme einer geeigneten Multiplikation den $\mathbb{R}^n$ zu einem Körper machen?
  Antwort: Nein. Allenfalls im Fall $n = 4$ kann man die sog. *Quaternionen* finden (Hamilton, 1843); diese bilden jedoch nur einen sog. *Schiefkörper*, d.h. es gelten die Körperaxiome nur mit Ausnahme des Kommutativgesetzes der Multiplikation. Nichtsdestotrotz haben die Quaternionen Anwendungen in der theoretischen Physik (Quantenphysik, Relativitätstheorie) und in der Computergrafik (effiziente Berechnung von Hintereinanderausführung von Drehungen (Orthogonalmatrizen) im $\mathbb{R}^3$).
- Teilmengen von Körpern können Köper sein mit den vom 'Oberkörper' 'ererbten' Verknüpfungen:
  - $\mathbb{Q}[\sqrt{2}] := \{q + \sqrt{2}r \,|\, q, r \in \mathbb{Q}\} \subseteq \mathbb{R}$ ist Körper mit den von $\mathbb{R}$ ererbten Verknüpfungen
    (Wie sehen Inverse aus in diesem Körper?)
  - $\{q + \sqrt[3]{2}r \,|\, q, r \in \mathbb{Q}\} \subseteq \mathbb{R}$ ist kein(!) Körper mit den von $\mathbb{R}$ ererbten Verknüpfungen (fehlende Abgeschlossenheit bzgl. '$\cdot$')
  - $\mathbb{Q}[2^{\frac{1}{3}}] := \{q + 2^{\frac{1}{3}}r + 2^{\frac{2}{3}}s \,|\, q, r, s \in \mathbb{Q}\} \subseteq \mathbb{R}$ ist ein Körper mit den von $\mathbb{R}$ ererbten Verknüpfungen
  - $\mathbb{Q}[i\sqrt{2}] := \{q + i\sqrt{2}r \,|\, q, r \in \mathbb{Q}\} \subseteq \mathbb{C}$ ist ein Körper mit den von $\mathbb{C}$ ererbten Verknüpfungen

  (Die Potenzen des 'erzeugenden Elements' müssen wieder in der Menge liegen, anderfalls keine multiplikative Abgeschlossenheit.)
- Vorgriff auf später:
  $(\mathbb{Z}_n, +, \cdot)$ ist ein kommutativer Ring mit Eins.
  Für gewisse n ist $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ eine (abelsche) Gruppe, d.h. dann ist $(\mathbb{Z}_n, +, \cdot)$ ein Körper!
  Für welche $n \in \mathbb{N}$ das so ist: Dazu später mehr!

Weiterhin Wiederholung aus dem 1. Semester:

Konstruktion von $\mathbb{Z}_n$ bzw. der Strukturen $(\mathbb{Z}_n, +)$, $(\mathbb{Z}_n, \cdot)$, $(\mathbb{Z}_n, +, \cdot)$

Sei $\sim$ eine Äquivalenzrelation (d.h. eine reflexive, symmetrische, transitive Relation) auf einer Menge M. Man definiert zu jedem $x \in M$ die zugehörige *Äquivalenzklasse*

$$[x]_\sim := \{y \in M \mid x \sim y\} \subseteq M.$$

Die so gebildeten Äquivalenzklassen bilden eine *Partition* auf M, d.h.

1. $\bigcup_{x \in M} [x]_\sim = M$ und
2. $[x]_\sim \neq [y]_\sim \;\Rightarrow\; [x]_\sim \cap [y]_\sim = \emptyset$ $\quad\big[$Äquiv.: $\forall x, y \in M : \; (\; [x]_\sim = [y]_\sim \; \vee \; [x]_\sim \cap [y]_\sim = \emptyset \;)\big]$.

Anwendung auf $M := \mathbb{Z}$ (s.Ü. 1.Sem.):
Sei $n \in \mathbb{N}$ beliebig vorgegegeben. Die Relation

$$\begin{aligned} a \sim_n b \;&:\Leftrightarrow\; n \text{ teilt } b-a \\ &:\Leftrightarrow\; \exists\, z \in \mathbb{Z} : \; b-a = n \cdot z \end{aligned}$$

ist reflexiv, symmetrisch, transitiv, also eine Äquivalenzrelation auf $\mathbb{Z}$.
Es gibt genau n Klassen ("Restklassen") $[0]_n, ..., [n-1]_n$.

Die **Menge der Restklassen** bezeichnen wir mit $\mathbb{Z}/\sim_n$ oder kurz $\mathbb{Z}_n$.

Beispiel: Die Klassen haben für $n = 3$ die Form

$$\begin{aligned} [0]_3 &= \{..., -6, -3, 0, 3, 6, ...\} = \{z \in \mathbb{Z} \mid z \equiv 0 \,(\text{mod } 3)\} \\ [1]_3 &= \{..., -5, -2, 1, 4, 7, ...\} = \{z \in \mathbb{Z} \mid z \equiv 1 \,(\text{mod } 3)\} \\ [2]_3 &= \{..., -4, -1, 2, 5, 8, ...\} = \{z \in \mathbb{Z} \mid z \equiv 2 \,(\text{mod } 3)\} \end{aligned}$$

und $\mathbb{Z} = [0]_3 \cup [1]_3 \cup [2]_3$ (paarweise disjunkt), und $\mathbb{Z}_3 = \{[0]_3, [1]_3, [2]_3\}$.

Wir statten die Menge der Restklassen $\mathbb{Z}_n$ mit Verknüpfungen aus:

$$\begin{aligned} [a]_n + [b]_n \;&:=\; [a + b]_n \\ [a]_n \cdot [b]_n \;&:=\; [a \cdot b]_n \end{aligned}$$

Beachte: Auf der rechten Seite bezeichnen '+', '·' Operationen in $\mathbb{Z}$, auf der linken Seite (neu definierte) Operationen auf der Restklassenmenge $\mathbb{Z}_n$!
(Eigentlich sollte man neue Symbole verwenden.)

Die oben die Verknüpfung von Klassen unter Verwendung von Repräsentanten erklärt wurde, muss man nachprüfen, dass die rechte Seite unabhängig von der Auswahl der Repräsentanten ist, also: Seien $\tilde{a} \in [a]_n$, $\tilde{b} \in [b]_n$.

Zu zeigen: $[a+b]_n = [\tilde{a}+\tilde{b}]_n$, $[a \cdot b]_n = [\tilde{a} \cdot \tilde{b}]_n$.
Siehe Tafel oder Übung.

Beispiel zum Rechnen in $\mathbb{Z}_n$: $[3]_6 + [5]_6 = [3+5]_6 = [8]_6 = [2]_6$
(bzw. das Gleiche in der 'alten' Notation: $3+5 \equiv 8 \equiv 2 \pmod 6$)

Welche Eigenschaften haben die Strukturen $(\mathbb{Z}_n, +)$, $(\mathbb{Z}_n, \cdot)$, $(\mathbb{Z}_n, +, \cdot)$?
Man rechnet leicht nach (s. Tafel oder Übung), dass $(\mathbb{Z}_n, +)$ eine kommutative Gruppe ist.
(Das neutrale Element ist $[0]_n$, und $-[x]_n = [-x]_n$.)

Und $(\mathbb{Z}_n, \cdot)$? Diese Struktur ist assoziativ und hat ein neutrales Element $[1]_n$, ist somit ein Monoid, ist außerdem kommutativ (leicht nachzurechnen, folgt aus den entspr. Eigenschaften von $(\mathbb{Z}, \cdot)$).

Ferner gelten Distributivgesetze in $(\mathbb{Z}_n, +, \cdot)$ (was aus der Distributivität von $(\mathbb{Z}, +, \cdot)$ folgt);
$(\mathbb{Z}_n, +, \cdot)$ ist also ein **kommutativer Ring mit Eins**, der sog. *Restklassenring*.

Nochmal zu $(\mathbb{Z}_n, \cdot)$: Dies ist (für $n > 1$) keine Gruppe, denn $[0]_n$ hat offensichtlich kein Inverses.

Und wenn wir die $[0]_n$ weglassen? Hat denn wenigstens jedes Element aus $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ ein Inverses? Wenn ja, dann ist $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ eine Gruppe, und dann wäre $(\mathbb{Z}_n, +, \cdot)$ sogar ein Körper! (Bemerkung: Statt $(\mathbb{Z}_n \setminus \{[0]_n\}, \cdot)$ haben wir kurz $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ geschrieben)

Im 1. Semester hatten wir ohne Beweis, durch betrachten von einigen verschiedenen $n \in \mathbb{N}$, vermutet/behauptet, dass dies davon abhängt, ob n eine Primzahl ist.

Beispiele:

$n = 4$ (nicht prim) :

$\cdot$	$[1]_4$	$[2]_4$	$[3]_4$
$[1]_4$	$[1]_4$	$[2]_4$	$[3]_4$
$[2]_4$	$[2]_4$	$''[0]_4''$	$[2]_4$
$[3]_4$	$[3]_4$	$[2]_4$	$[1]_4$

$n = 5$ (prim) :

$\cdot$	$[1]_5$	$[2]_5$	$[3]_5$	$[4]_5$
$[1]_5$	$[1]_5$	$[2]_5$	$[3]_5$	$[4]_5$
$[2]_5$	$[2]_5$	$[4]_5$	$[1]_5$	$[3]_5$
$[3]_5$	$[3]_5$	$[1]_5$	$[4]_5$	$[2]_5$
$[4]_5$	$[4]_5$	$[3]_5$	$[2]_5$	$[1]_5$

- Für $n=4$ ist $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ keine Gruppe.
  Dafür kann man zwei verschiedene Argumente anführen:
  1. Mit $[2]_4 \cdot [2]_4 = [0]_4 \notin \mathbb{Z}_4 \setminus \{0\}$ ist $(\mathbb{Z}_4 \setminus \{0\}, \cdot)$ noch nicht einmal abgeschlossen!
  2. $[2]_4$ hat kein Inverses, denn $[2]_4 \cdot [x]_4 \neq [1]_4 \, \forall x \in \mathbb{Z}_n \setminus \{0\}$
- Für $n=5$ dagegen ist $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ abgeschlossen, und sogar jedes Element hat ein Inverses:

$$
\begin{array}{rcl}
[1]_5 \cdot [1]_5 = [1]_5 & \Rightarrow & [1]_5^{-1} = [1]_5 \\
[2]_5 \cdot [3]_5 = [3]_5 \cdot [2]_5 = [1]_5 & \Rightarrow & [2]_5^{-1} = [3]_5, \quad [3]_5^{-1} = [2]_5 \\
[4]_5 \cdot [4]_5 = [1]_5 & \Rightarrow & [4]_5^{-1} = [4]_5
\end{array}
$$

Man kann zeigen, dass die beiden Eigenschaften 'Existenz von Inversen' und 'Abgeschlossenheit' unmittelbar zusammenhängen für beliebiges $n \in \mathbb{N}$:

Satz 1

Im Restklassenring $(\mathbb{Z}_n, +, \cdot)$ sind für jedes $[a]_n \in \mathbb{Z}_n$ äquivalent:

1. $[a]_n$ ist invertierbar (bzgl. '$\cdot$')
2. Für alle $[b]_n \in \mathbb{Z}_n \setminus \{0\}$ ist $[a]_n \cdot [b]_n \neq [0]_n$, d.h. $[a]_n \cdot [b]_n \in \mathbb{Z}_n \setminus \{0\}$.

Veranschaulichung: Obige Tabelle für $n=4$, oder zur Übung Tabelle für $n=10$.

Nutzen des Satzes: Um zu überprüfen, ob für ein $n \in \mathbb{N}$ die Struktur $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ eine Gruppe ist, reicht es also, die Existenz aller Inversen zu prüfen; die Abgeschlossenheit gilt dann automatisch ebenfalls.

Beweis des Satzes:

(1)⇒(2) Sei $[a]_n$ invertierbar im Monoiden $(\mathbb{Z}_n, \cdot)$. Dann folgt mit der Aussage auf S. 4 unten, dass in der Zeile von $[a]_n$ jedes Element von $\mathbb{Z}_n$ einmal vorkommt, so auch das Element $[0]_n$. Da offensichtlich $[a]_n \cdot [0]_n = [a \cdot 0]_n = [0]_n$ ist, die Null also in der 'nullten' Spalte in der Zeile von $[a]_n$ vorkommt, sind die *übrigen* Einträge in der Zeile von $[a]_n$ keine Nullen.

(2)⇒(1) Es gelte (2). Wir zeigen zunächst, dass für festes $[a]_n$ die Terme $[a]_n \cdot [1]_n, ..., [a]_n \cdot [n-1]_n$ paarweise verschieden sind:
Angenommen $[a]_n \cdot [b_1]_n = [a]_n \cdot [b_2]_n$ für $[b_1]_n \neq [b_2]_n$.
Mit den Rechenregeln im Ring $(\mathbb{Z}_n, +, \cdot)$ kann man dies umstellen zu $[a]_n \cdot [b_1 - b_2]_n = [0]_n$.
Da $[b_1 - b_2]_n \neq [0]_n$, ist dies ein Widerspruch zu (2).
Wir haben somit gezeigt, dass die Werte in der $[a]_n$-Zeile der Multiplikationstabelle von $\mathbb{Z}_n \setminus \{0\}$, also die $[a]_n \cdot [b]_n \forall [b]_n \in \mathbb{Z}_n \setminus \{0\}$, paarweise verschieden sind. Da es sich um $n-1$ viele Werte aus der $n-1$-elementigen Menge $\{[1]_n, ..., [n-1]_n\}$ handelt, muss jedes dieser Elemente genau einmal vorkommen, also auch das Element $[1]_n$.
Somit gibt es ein $[b]_n \in \mathbb{Z}_n \setminus \{0\}$ mit $[a]_n \cdot [b]_n = [1]_n$. □

Kurze Zusammenfassung der bisherigen Ergebnisse über das Rechnen in $(\mathbb{Z}_n, +, \cdot)$:

$(\mathbb{Z}_n, +)$ ist eine (abelsche) Gruppe. $(\mathbb{Z}_n, \cdot)$ ist keine Gruppe, da '0' kein Multiplikativ-Inverses hat.
Um zu prüfen, ob $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ eine Gruppe ist (Bem.: Genau dann ist $(\mathbb{Z}_n, +, \cdot)$ ein

Körper!), müsste man (1.) die Abgeschlossenheit und (2.) die Existenz von Inversen von allen Elementen zeigen. Der obige Satz jedoch besagt, dass (1.) und (2.) äquivalent sind; $(\mathbb{Z}_n\backslash\{0\},\cdot)$ ist also genau dann eine (abelsche) Gruppe, wenn (2.) für jedes Element gilt.

Bevor wir klären können, ob jedes Element ein Inverses hat, benötigen wir den *Euklidischen Divisionsalgorithmus* zur Bestimmung des *größten gemeinsamen Teilers (ggT)* von zwei Zahlen:

Def.: Der *größte gemeinsame Teiler (ggT)* von $a,b\in\mathbb{N}$, $t=ggT(a,b)$, ist die größte Zahl $t\in\mathbb{N}$ (äquivalent: $t\in\{1,...,\min\{a,b\}\}$, die sowohl Teiler von a als auch Teiler von b ist.
Ist $ggT(a,b)=1$, so nennt man a,b *teilerfremd*.
Falls man auch null als Argument von ggT zulassen möchte, so legt man fest, dass $ggT(0,b)=ggT(a,0)=0$.

Der Euklidische Divisionsalgorithmus beruht auf der Beobachtung, dass wenn man von der größeren der beiden Zahlen a,b die kleinere der beiden oder ein Vielfaches der kleineren abzieht, sich der ggT dabei nicht ändert: $ggT(26,8)=ggT(26-8,8)=ggT(26-2\cdot 8,8)=ggT(26-3\cdot 8,8)=ggT(2,8)=2$; dabei haben wir die Zahl 3 per Division von 26 durch 8 mit Rest ermittelt: $26=3\cdot 8+2$

Hier noch ein größeres Zahlenbeispiel: $ggT(99,78)$:

$$\begin{array}{rcl} 99 &=& 1\cdot\underline{78}+\underline{21} \\ & \swarrow & \swarrow \\ 78 &=& 3\cdot\underline{21}+\underline{15} \\ & \swarrow & \swarrow \\ 21 &=& 1\cdot\underline{15}+\underline{6} \\ & \swarrow & \swarrow \\ 15 &=& 2\cdot\underline{6}+\underline{3} \\ & \swarrow & \swarrow \\ 6 &=& 2\cdot\underbrace{\underline{3}}_{ggT}+\underline{0} \end{array} \quad \Rightarrow \quad \begin{array}{rcl} ggT(99,78) &=& ggT(99-1\cdot 78,78) \\ &=& ggT(78-3\cdot 21,21) \\ &=& ggT(21-1\cdot 15,15) \\ &=& ggT(15-2\cdot 6,6) \\ &=& 3 \end{array} \quad = \quad \begin{array}{l} ggT(78,21) \\ ggT(21,15) \\ ggT(15,6) \\ ggT(6,3) \end{array}$$

Neben der Berechnung des ggT liefert der Euklidische Divisionsalgorithmus noch mehr: Liest man ihn rückwärt, so liefert er eine Darstellung des $ggT(a,b)$ als 'Linearkombination' von a,b, also $\alpha,\beta\in\mathbb{Z}$ mit $ggT(a,b)=\alpha a+\beta b$:

$$\begin{array}{rcl} ggT(99,78) &=& \underline{3}=\underline{15}-2\cdot\underline{6} \\ &=& \underline{15}-2\cdot(\underline{21}-1\cdot\underline{15})=3\cdot\underline{15}-2\cdot\underline{21} \\ &=& 3\cdot(\underline{78}-3\cdot\underline{21})-2\cdot\underline{21}=3\cdot\underline{78}-11\cdot\underline{21} \\ &=& 3\cdot\underline{78}-11\cdot(\underline{99}-1\cdot\underline{78})=\underbrace{14}_{\beta}\cdot\underline{78}\ \underbrace{-11}_{\alpha}\cdot\underline{99} \end{array}$$

Erläuterung: In der obigen "Vorwärts-Rechnung" kommen in jeder Zeile 2 unterstrichene Zahlen vor; in jedem Schritt wird die größere der beiden durch eine kleinere ausgetauscht. In der unteren "Rückwärt-Rechnung" kommen die gleichen unterstrichene Zahlen vor; der ggT ist dort jeweils als "Linearkombination" von 2 unterstrichenen Zahlen dargestellt; in jedem Schritt ersetzt man die kleinere der beiden durch größere, und dazu verwendet man die Gleichungszusammenhänge aus der Vorwärts-Rechnung.

Wir halten obige Beobachtung als Satz fest:

Satz 2
Für beliebige $a, b \in \mathbb{N}$ gibt es $\alpha, \beta \in \mathbb{Z}$, so dass

$$ggT(a,b) = \alpha\, a + \beta\, b\,.$$

Bemerkung zur ggT-Berechnung: Alternativ zum Euklidischen Divisionsalgorithmus könnte man auch die Primfaktorzerlegungen bestimmen und darin nach den gemeinsamen Faktoren suchen; das ist jedoch bei großen Zahlen i.a. deutlich aufwändiger; $99 = 3 \cdot 33 = 3 \cdot 3 \cdot 11; 78 = 2 \cdot 39 = 2 \cdot 3 \cdot 13 \Rightarrow ggT(99,78) = 3\,.$

Weiterführende Bemerkung: Nicht nur im Ring $(\mathbb{Z}, +, \cdot)$, auch im Polynomring $(\mathbb{R}[X], +, \cdot)$ bzw $(\mathbb{C}[X], +, \cdot)$ bzw $(\mathbb{Q}[X], +, \cdot)$ kann man eine Primfaktorzerlegung und einen ggT definieren (bei $\mathbb{C}[X]$ sind die Primfaktoren eines Polynoms gerade die Linearfaktoren!):
So hat z.B. das Polynom $p := X^3 - 3X^2 + 4 = (X+1)(X-2)^2$ die Primfaktoren $X+1, X-2, X-2$ und das Polynom $q := X^3+X^2-4X-4 = (X-2)(X+2)(X+1)$ hat die Primfaktoren $X-2, X+2, X+1$. Somit ist $ggT(p,q) = (X+1)(X-2)$.

Bem. 1: Die Primfaktoren (und auch der ggT) von Polynomen sind jedoch nur bis auf konstante Faktoren bestimmt.
Bsp.: $2X^2-2 = (2X-2)(X+1)$, aber auch $2X^2-2 = (X-1)(2X+2)\,.$

Bem. 2: Was die korrekte Primfaktorzerlegung ist, kann davon abhängen, ob man sie in $(\mathbb{R}[X], +, \cdot), (\mathbb{C}[X], +, \cdot)$ oder $(\mathbb{Q}[X], +, \cdot)$ durchführt:
Der Faktor X^2+1 ist 'prim' in $(\mathbb{R}[X], +, \cdot)$, aber nicht in $(\mathbb{C}[X], +, \cdot)$;
Der Faktor X^2-2 ist 'prim' in $(\mathbb{Q}[X], +, \cdot)$, aber nicht in $(\mathbb{R}[X], +, \cdot)$.

Bem. 3: Ringe, in denen es 'Division mit Rest' gibt, heißen *Euklidische Ringe*; Beispiele sind $\mathbb{Z}$ sowie $\mathbb{Q}[X], \mathbb{R}[x], \mathbb{C}[X]$; in solchen Ringen haben alle Elementepaare einen ggT, und diesen kann man per Eukl. Div.algor. finden. Es gibt auch nichteuklidische Ringe, z.B. $\mathbb{R}^{n\times n}$.

Folgerung aus dem Satz 2: Seien $a, b \in \mathbb{N}$ teilerfremd. Dann gibt es eine Darstellung der Eins als $1 = \alpha\, a + \beta\, b$ mit $\alpha, \beta \in \mathbb{Z}$.

Frage: Gilt auch die Umkehrung? — Antwort: Ja. Beweis: s. Übung oder Tafel.

Die obige Folgerung und ihre Umkehrung ergeben:

Satz 3 (Lemma von Bézout, 1730-83)
Zahlen $a, b \in \mathbb{N}$ sind *genau dann* teilerfremd, wenn es $\alpha, \beta \in \mathbb{Z}$ gibt mit

$$\alpha\, a + \beta\, b = 1\,.$$

Wir übertragen die Aussage des Lemmas von Bézout nun auf $\mathbb{Z}_n$:

$$\begin{aligned} ggT(a,n)=1 &\overset{\text{Bezout}}{\iff} \exists\, \alpha,\beta\in\mathbb{Z}:\ \alpha\, a+\beta\, n=1 \\ &\overset{(*)}{\iff} \exists\, \alpha,\beta\in\mathbb{Z}:\ [\alpha\, a+\beta\, n]_n=[1]_n \\ &\iff \exists\, \alpha,\beta\in\mathbb{Z}:\ [\alpha]_n\cdot[a]_n+[\beta]_n\underbrace{\cdot[n]_n}_{=[0]_n}=[1]_n \\ &\iff [a]_n \text{ ist invertierbar} \end{aligned}$$

Vorsicht: An der Stelle (*) ist zunächst nur "⇒" trivial; über "⇐" muss man nachdenken!

Wir haben damit hergeleitet:

Satz 4 (Folgerung aus dem Lemma von Bézout)
$[a]_n \in \mathbb{Z}_n$ hat genau dann ein Multiplikativ-Inverses, wenn a und n teilerfremd sind.
Es ist dann $[a]_n^{-1} = [\alpha]_n$, wobei α aus dem Lemma von Bézout ist (mit $b := n$). Und dieses α kann man unter Verwendung des Euklidischen Divisionsalgorithmus ('rückwärts') finden, s.o..

Beispiel für Anwendung von Satz 4: Bestimmung aller invertierbaren Elemente aus $\mathbb{Z}_{10}$:
Aus $\{1, 2, ..., 9\}$ sind genau die Zahlen $1, 3, 7, 9$ zu 10 teilerfremd.
Also sind $[1]_{10}, [3]_{10}, [7]_{10}, [9]_{10}$ die invertierbaren Elemente aus $\mathbb{Z}_{10}$.
Zur Kontrolle: $[1]_{10}\cdot[1]_{10} = [1]_{10},\ [3]_{10}\cdot[7]_{10} = [21]_{10} = [1]_{10},\ [9]_{10}\cdot[9]_{10} = [81]_{10} = [1]_{10}$,
also $[1]_{10}^{-1} = [1]_{10},\ [3]_{10}^{-1} = [7]_{10},\ [7]_{10}^{-1} = [3]_{10},\ [9]_{10}^{-1} = [9]_{10}$.

Mit Satz 4 haben wir es nun fast geschafft, die bereits geäußerte Vermutung zu beweisen, dass:

Satz

$$\forall n \geq 2:\ (\mathbb{Z}_n \setminus \{0\}, \cdot) \text{ ist Gruppe} \quad \iff \quad n \text{ ist prim}$$

Beweis: Nach Satz 1 ist $(\mathbb{Z}_n \backslash \{0\}, \cdot)$ genau dann eine Gruppe, wenn jedes Element ein Inverses hat. Nach Satz 4 ist das genau dann der Fall, wenn jedes $a \in \{1, ..., n-1\}$ zu n teilerfremd ist. Das wiederum ist äquivalent dazu, dass n prim ist. □

Folgerung:

Satz 5 und Definition (endliche Körper)

$$\forall p \geq 2: \ (\mathbb{Z}_p, +, \cdot) \text{ ist Körper} \quad \Longleftrightarrow \quad p \text{ ist prim}$$

Man bezeichnet diesen Körper als $\mathbb{F}_p$ (engl.: Körper=field). In ihm wird also, sowohl was Addition als auch Multiplikation betrifft, "modulo p" gerechnet.

Auch wenn der Begriff "(endlicher) Körper" erst 1895 geprägt wurde, so hatte bereits Gauß (1777-1855) erkannt, dass man "in $\mathbb{Z}$ modulo Primzahl p genau so rechnen kann wie in $\mathbb{Q}$".

Weiterführende Bemerkung: Man kann sich fragen, ob es außer den so gefundenen $\mathbb{F}_p$ mit p Elementen (p prim) noch **weitere *endliche* Körper** gibt. Man kann zeigen, dass dies in der Tat der Fall ist: Für p prim und $n \in \mathbb{N}$ kann man einen Körper mit p^n vielen Elementen konstruieren. (Darüber hinaus existieren keine weiteren endlichen Körper!)
Wie man Körper $\mathbb{F}_{p^2}$ konstruiert: Man sucht in $\mathbb{F}_p$ ein Element q, das kein Quadrat eines anderen Elements ist, und "adjungiert" dessen "Wurzel" zu $\mathbb{F}_p$, d.h. man adjungiert zu $\mathbb{F}_p$ ein $w \notin \mathbb{F}_p$ mit $w^2 = q$.
Beispiel: In $\mathbb{F}_5$ ist 2 kein Quadrat eines der Elemente, denn $1^2 = 1$, $2^2 = 4$, $3^2 = 9 = 4$, $4^2 = 16 = 1$. Somit kann man $\mathbb{F}_{25} := \{a + b\sqrt{2} \,|\, a, b \in \{0, 1, ..., 4\}\}$ setzen (hat 25 Elemente) und z.B. wie folgt rechnen (modulo 5): $(2+4\sqrt{2}) \cdot (1+3\sqrt{2}) = 2 \cdot 1 + 4 \cdot 3 \cdot \sqrt{2}\sqrt{2} + 2 \cdot 3\sqrt{2} + 4 \cdot 1\sqrt{2} = 2 + 12 \cdot 2 + 10\sqrt{2} = 1 + 0 \cdot \sqrt{2} = 1$.
Noch etwas skuriler: In $\mathbb{F}_7$ ist $6 = -1$ kein Quadrat, wie man leicht herausfindet. Somit kann man zu $\mathbb{F}_7$ wahlweise $\sqrt{6}$ oder auch $\sqrt{-1} = i$ adjungieren, um $\mathbb{F}_{49}$ zu bekommen:
$\mathbb{F}_{49} = \{a + bi \,|\, a, b \in \{0, 1, ..., 6\}\}$, wobei addiert und multipliziert wird modulo 7 und die Rechenregel $i^2 = -1 = 6$ gilt.

Noch eine Bemerkung für Informatiker: Stellt man die Additions- und die Multiplikationstabelle für $\mathbb{F}_2$ auf, so sieht man, dass man, wenn man '0' mit 'falsch' und '1' mit 'wahr' identifiziert, die Operationen 'exklusives Oder' und 'Und' vor sich hat, d.h. $(\{\text{falsch}, \text{wahr}\}, \text{exkl.oder}, \text{und})$ ist ein Körper.

Eine Anwendung des Rechnens im Restklassenring $(\mathbb{Z}_n, +, \cdot)$ ($n \in \mathbb{N}$ beliebig):

1. Mit welcher Ziffer endet die Zahl $z = 9^{123}$?
   Gesucht ist also $[9^{123}]_{10}$.
   Wir rechnen dazu im Ring $(\mathbb{Z}_{10}, +, \cdot)$:

   $$[9^{123}]_{10} = [9]_{10}^{123} = [-1]_{10}^{123} = [(-1)^{123}]_{10} = [-1]_{10} = [9]_{10}$$

   (Das Heraus- und Hereinziehen von Exponenten ist erlaubt nach Definition der Multiplikation in $\mathbb{Z}_{10}$.)

Antwort also: 9

2. Finde eine einfache Regel, wann eine im Dezimalsystem gegebene Zahl $z = d_k d_{k-1} ... d_1 d_0 = \sum_{i=0}^{k} d_i 10^i,\ d_i \in \{0, ..., 9\}$, durch 3 teilbar ist:
   Dazu rechnen wir in $(\mathbb{Z}_3, +, \cdot)$: Wann ist $[z]_3 = [0]_3$?

$$[0]_3 \stackrel{!}{=} [z]_3 = \Big[\sum_{i=0}^{k} d_i 10^i\Big]_3 = \sum_{i=0}^{k} [d_i]_3 \cdot [10^i]_3 = \sum_{i=0}^{k} [d_i]_3 \cdot [10]_3^i$$
$$= \sum_{i=0}^{k} [d_i]_3 \cdot [1]_3^i = \sum_{i=0}^{k} [d_i]_3 \cdot [1^i]_3 = \sum_{i=0}^{k} [d_i]_3 \cdot [1]_3 = \Big[\sum_{i=0}^{k} d_i \cdot 1\Big]_3$$

   Ergebnis: z ist genau dann durch 3 teilbar, wenn die *Quersumme* von z, $\sum_{i=0}^{k} d_i$, durch 3 teilbar ist.

   Ähnlich bekommt man Regeln für Teilbarkeit durch 9,11,101 (im Dezimalsystem), auch für Teilbarkeit einer Binärzahl durch 3, ... : siehe Übung, ggf. Tafel.

Eine weitere Anwendung des Rechnens in $\mathbb{Z}_n$ ist das Thema *Fehlererkennung/Prüfziffern*. Es geht dabei darum, einer Nachricht eine redundante Information (=Prüfziffer) hinzuzufügen, die derart beschaffen sein soll, dass man aus der erweiterten Information ermitteln kann, ob bei der Informationsübermittlung (gewisse) Fehler passiert sind.

Als Vorbereitung müssen wir uns die Multiplikation in $\mathbb{Z}_n$ noch einmal genauer anschauen. Was wir wissen:
Ist n prim, dann ist $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ eine Gruppe, d.h. jedes Element hat ein Inverses. Andernfalls hat nicht jedes Element von $(\mathbb{Z}_n \setminus \{0\}, \cdot)$ ein Inverses.

Wir definieren die Menge $\mathbb{Z}_n^*$:

$$\mathbb{Z}_n^* := \{[a]_n \in \mathbb{Z}_n \mid [a]_n \text{ hat Inverses (bzgl. Multiplikation) in } \mathbb{Z}_n\} \subseteq \mathbb{Z}_n \setminus \{0\}$$

Es ist also $\mathbb{Z}_n^* = \mathbb{Z}_n \setminus \{0\}$ genau dann wenn n prim ist, und andernfalls ist $\mathbb{Z}_n^* \subsetneq \mathbb{Z}_n \setminus \{0\}$.

Wie findet man die Elemente von $\mathbb{Z}_n^*$ (falls n nicht prim ist)?
Man könnte natürlich (jedoch wohl nur für sehr kleines n) die Multiplikationstabelle von $\mathbb{Z}_n$ komplett aufstellen, dort nach Einsen suchen und damit die Inversen finden. (Beispiel $n = 4$ ggf. s. Tafel.)
Besser ist jedoch ein Blick auf Satz 4:
$\mathbb{Z}_n^*$ besteht genau aus den zu n teilerfremden Zahlen aus $\{1, ..., n-1\}$.
Beispiel $n = 12$: $\mathbb{Z}_{12}^* = \{[1]_{12}, [5]_{12}, [7]_{12}, [11]_{12}\}$

Für größere(!) Zahlen n führt man die **Prüfung auf Teilerfremdheit mit dem Euklidischen Divisionsalgorithmus** durch (also $ggT(a,n)$ berechnen für $a=2,...,n–1$).

Zur Übung: Warum reicht es, alle $a \leq \frac{n}{2}$ auf Teilerfremdheit(=Invertierbarkeit) zu testen?
Hinweis: Überlegen Sie: Wenn $[a]_n$ und $[b]_n$ in $(\mathbb{Z}_n, \cdot)$ invers zueinander sind, dann sind auch $[n-a]_n$ und $[???]_n$ invers zueinander.

Multiplikation auf $\mathbb{Z}_n^*$:

Wir können, da $\mathbb{Z}_n^* \subset \mathbb{Z}_n$, Elemente aus $\mathbb{Z}_n^*$ mit der aus $\mathbb{Z}_n$ ererbten Multiplikation $\cdot : \mathbb{Z}_n \times \mathbb{Z}_n \to \mathbb{Z}_n$ verknüpfen, haben also auf $\mathbb{Z}_n^*$ eine ererbte Verknüpfung $\mathbb{Z}_n^* \times \mathbb{Z}_n^* \to \mathbb{Z}_n$; das Ergebnis einer solchen Verknüpfung ist zunächst einmal trivialerweise in $\mathbb{Z}_n$. Liegt es auch in $\mathbb{Z}_n^*$?

Dazu: Haben $[a]_n, [b]_n \in \mathbb{Z}_n$ Inverse, dann hat auch das Verknüpfungsergebnis $[a]_n \cdot [b]_n$ ein Inverses (nämlich $[b]_n^{-1} \cdot [a]_n^{-1}$; dies gilt in jedem Monoid), somit liegt auch das Verknüpfungsergebnis in $\mathbb{Z}_n^*$, d.h. die Menge $\mathbb{Z}_n^*$ ist bezüglich der ererbten Verknüpfung also *abgeschlossen*.

Wir prüfen, dass $(\mathbb{Z}_n^*, \cdot)$ eine Gruppe ist:

Das neutrale Element von $(\mathbb{Z}_n, \cdot)$, also $[1]_n$, liegt ebenfalls in $\mathbb{Z}_n^*$, (denn 1 und n sind teilerfremd (auch: denn $[1]_n$ hat Inverses $[1]_n$)).
Ist $[a]_n \in \mathbb{Z}_n^*$, so hat $[a]_n$ nach Def. von $\mathbb{Z}_n^*$ ein Inverses $[a]_n^{-1}$, welches zunächst einmal in $\mathbb{Z}_n$ ist, weil aber dann auch $[a]_n^{-1}$ ein Inverses hat (nämlich $[a]_n$), liegt $[a]_n^{-1}$ in $\mathbb{Z}_n^*$.

Wir haben somit gezeigt:

> $\mathbb{Z}_n^*$ ist mit der vom Monoid $(\mathbb{Z}_n, \cdot)$ ererbten Multiplikation eine Gruppe.
> (Sie ist abelsch, da in $(\mathbb{Z}_n, \cdot)$ ein Kommutativgesetz gilt.)

Wie berechnet man Inverse in der Gruppe $(\mathbb{Z}_n^*, \cdot)$ (effizient)?

1. Wissen bereits: Mit dem Euklidischen Divisionsalgorithmus findet man $\alpha, \beta \in \mathbb{N}$ mit $\alpha\, a + \beta\, n = 1$; es ist dann $[a]_n^{-1} = [\alpha]_n$.
   Auch bei großem n ist diese Methode recht effizient.

2. Bei kleinem n kommen auch folgende Vorgehensweisen in Betracht:

   – Bilde nacheinander die Potenzen von $[a]_n$, bis eine Potenz $[a]_n^k = [1]_n$ ist.
   Es ist dann $[a]_n^{k-1}[a]_n = [1]_n$, somit $[a]_n^{-1} = [a]_n^{k-1}$.

   Beispiel: Berechne $[3]_{16}^{-1}$ (existiert, da 3,16 teilerfremd):
   $3^2 = 9, 3^3 = 27 \equiv 11 \pmod{16}$,
   $3^4 \equiv 3 \cdot 11 \equiv 33 \equiv 1 \pmod{16}$
   Somit ist $[3]_{16}^{-1} = [3^3]_{16} = [11]_{16}$.

(Die Rechnung ergibt übrigens zusätzlich, dass $3^2 \cdot 3^2 \equiv 1 (\text{mod } 16)$, also $[9]_{16}^{-1} = [9]_{16}$.)

– Fülle die zu $[a]_n$ gehörende Zeile (oder Spalte) der Verknüpfungstabelle aus, bis das neutrale Element $[1]_n$ erscheint. Die Spalte (bzw. Zeile) gibt das inverse Element an.

Es können bis zu $n-1$ 'Schritte' erforderlich sein.

Wir testen diese Methoden in den Übungen.

Anwendung: Prüfziffern

Ziel: Bei der Datenübertragung sollen (gewisse) Fehler detektiert werden.

Die Information werde zerlegt in Pakete, sog. *Wörter*. Ein Wort besteht aus m Ziffern $d_1, ..., d_m \in \{0, ..., n-1\}$; $m, n \in \mathbb{N}$ seien fest.

Der Sender der Nachricht berechnet aus diesen m Ziffern eine $m+1$-te Ziffer, die sog. *Prüfziffer* $d_{m+1} \in \{0, ..., n-1\}$. Die Prüfziffer wird an das Wort angehängt. Das Tupel $d_1, ..., d_m, d_{m+1}$ bezeichnen wir als *erweitertes Wort*. ($\rightarrow$"Redundanz")

Der Empfänger soll aus dem erweiterten Wort berechnen können, ob im Wort ein (gewisser) Fehler enthalten ist.

Anwendungen von Prüfziffern: ISBN (Buchhandel), IBAN (Banking), Dt. Rentenversicherungsnummer, Euro-Banknoten, IdentCode Deutsche Post, EAN (=Europäische Artikel-Nummer, „Strichcode“), Matrikelnummern,...

Formal:

Berechnung der Prüfziffer: $d_{m+1} := f(d_1, ..., d_m)$

Prüfung: $P(d_1, ..., d_m, d_{m+1}) := d_{m+1} - f(d_1, ..., d_m) \begin{cases} \neq 0 \Rightarrow \text{ sicher Fehler} \\ = 0 \Rightarrow \text{vmtl. kein Fehler} \end{cases}$

Die Funktion f ist allgemein (mindestens Sender und Empfänger) bekannt.

Recht einfach und daher beliebt ist es, ein f folgender Struktur zu verwenden:

$$f(d_1, ..., d_m) :\equiv -\sum_{i=1}^{m} g_i \, d_i \equiv \sum_{i=1}^{m} (n - g_i) \, d_i \ (\text{mod } n)$$

Somit

$$P(d_1, ..., d_{m+1}) :\equiv d_{m+1} + \sum_{i=1}^{m} g_i\, d_i \overset{(g_{m+1}:=1)}{\equiv} \sum_{i=1}^{m+1} g_i\, d_i \pmod{n}$$

Die g_i, $i = 1, ..., m$ heißen *Gewichte*. Sie sind noch geeignet zu wählen. Wie?

Häufigste Fehler (und diese sollen sicher erkannt werden):

(i) Einzelfehler: Höchstens ein d_i ist falsch.

(ii) Nachbarvertauschungsfehler: Ein d_i und ein d_{i+1} wurden permutiert (sonst alles korrekt).

(iii) Vertauschungsfehler: Ein d_i und ein d_j $(i \neq j)$ wurden permutiert (sonst alles korrekt).

Zu (i): Wie müssen $g_1, ..., g_m, n$ gewählt sein, damit Einzelfehler sicher erkannt werden?
Statt der korrekten Nachricht $d_1, ..., d_{m+1}$ werde $d_1, ..., \tilde{d}_{i_0}, ..., d_{m+1}$ übertragen, $\tilde{d}_{i_0} \neq d_{i_0}$, wobei die Position $i_0 \in \{1, ..., m+1\}$ unbekannt sei.

Es ist also

$$P(d_1, ..., d_{i_0}, ..., d_{m+1}) = \sum_{i=1}^{m+1} g_i\, d_i \equiv 0 \pmod{n}, \quad \text{und}$$
$$P(d_1, ..., \tilde{d}_{i_0}, ..., d_{m+1}) = \sum_{i \in \{1,...,m+1\} \setminus \{i_0\}} g_i\, d_i + g_{i_0} \tilde{d}_{i_0} \equiv: c \overset{!}{\not\equiv} 0 \pmod{n}$$

wobei das "$c \overset{!}{\not\equiv} 0 \pmod{n}$" der Erkennung des Fehlers entspricht.

Die Differenz der beiden Gleichungen ergibt

$$g_{i_0} \tilde{d}_{i_0} - g_{i_0}\, d_{i_0} \equiv c \overset{!}{\not\equiv} 0 \pmod{n}$$

somit

$$[g_{i_0}]_n \cdot [\tilde{d}_{i_0} - d_{i_0}]_n = [c]_n \overset{!}{\neq} [0]_n$$

Für beliebige $[\tilde{d}_{i_0} - d_{i_0}]_n \neq [0]_n$ soll also immer $[g_{i_0}]_n \cdot [\tilde{d}_{i_0} - d_{i_0}]_n = [c]_n \neq [0]_n$ sein. Das ist nach Satz 1 genau dann der Fall, wenn $[g_{i_0}]_n$ invertierbar ist.

Satz (Erkennbarkeit von Einzelfehlern)
Einzelfehler werden genau dann sicher erkannt, wenn für alle $i = 1, ..., m$ gilt:

$$[g_i]_n \text{ ist invertierbar.}$$

Das ist genau dann der Fall, wenn $[g_i]_n \in \mathbb{Z}_n^*$, also wenn $ggT(g_i, n) = 1$ für alle i.

Beispiel: Im Fall $n := 10$ kommen als Gewichte g_i nur die Zahlen $1, 3, 7, 9$ in Frage.

Da (ii) ein Spezialfall von (iii) ist, machen wir gleich mit (iii) weiter:

Zu (iii): Wie müssen $g_1, ..., g_m, n$ gewählt sein, damit Vertauschungsfehler sicher erkannt werden?
Statt der korrekten Nachricht $d_1, ..., d_i, ..., d_j, ..., d_{m+1}$ werde $d_1, ..., d_j, ..., d_i, ..., d_{m+1}$ übertragen, wobei die Positionen $1 \leq i < j \leq m+1$ unbekannt seien und $d_i \neq d_j$.

Es ist also

$$P(d_1, ..., d_i, ..., d_j, ..., d_{m+1}) = g_1 d_1 + ... + g_i d_i + ... + g_j d_j + ... + g_{m+1} d_{m+1} \equiv 0 \pmod{n}, \quad \text{und}$$
$$P(d_1, ..., d_j, ..., d_i, ..., d_{m+1}) = g_1 d_1 + ... + g_i d_j + ... + g_j d_i + ... + g_{m+1} d_{m+1} \equiv: c \overset{!}{\not\equiv} 0 \pmod{n}.$$

Die Differenz der beiden Gleichungen ergibt, dass für beliebiges $[d_j - d_i]_n \neq [0]_n$ zu fordern ist:

$$\begin{aligned} & g_i (d_j - d_i) + g_j (d_i - d_j) \overset{!}{\not\equiv} 0 \pmod{n} \\ \Leftrightarrow \quad & (g_i - g_j)(d_j - d_i) \overset{!}{\not\equiv} 0 \pmod{n} \\ \Leftrightarrow \quad & [g_i - g_j]_n \cdot [d_j - d_i]_n \overset{!}{\neq} [0]_n \end{aligned}$$

Nach Satz 1 ist dies für beliebiges $[d_j - d_i]_n \neq [0]_n$ genau dann erfüllt, wenn $[g_i - g_j]_n$ invertierbar ist.

Ergebnis:

Satz (Erkennbarkeit von Vertauschungsfehlern)
Vertauschungsfehler werden genau dann sicher erkannt, wenn für alle $i, j = 1, ..., m+1$ mit $i \neq j$ gilt:

$[g_i - g_j]_n$ ist invertierbar.

Das ist genau dann der Fall, wenn $[g_i - g_j]_n \in \mathbb{Z}_n^*$, also wenn $ggT(|g_i - g_j|, n) = 1$ für alle $i \neq j$.

Es folgt für (ii):

Satz (Erkennbarkeit von Nachbarvertauschungsfehlern)
Nachbarvertauschungsfehler werden genau dann sicher erkannt, wenn für alle $i = 1, ..., m$ gilt:

$[g_i - g_{i+1}]_n$ ist invertierbar.

Das ist genau dann der Fall, wenn $[g_i - g_{i+1}]_n \in \mathbb{Z}_n^*$, also wenn $ggT(|g_i - g_{i+1}|, n) = 1$ für alle $i = 1, ..., m$.

Beispiele: s. Tafel und Übung.

Ein auf der Fehlererkennung aufbauender Gedanke:

Fehlerkorrektur: Da üblicherweise die Stelle i, an der ein Einzelfehler passiert ist, nicht bekannt ist, kann man aus der Prüfziffer die Originalnachricht i.a. *nicht* rekonstruieren.
Ist aber eines der d_i, genannt d_{i_0}, *unlesbar* (d.h. man kennt insbesondere die Position i_0), so kann man die Prüfziffer benutzen um d_{i_0} zu rekonstruieren, sofern $[g_{i_0}]_n$ invertierbar ist:
Gesucht ist d_{i_0}, so dass

$$\begin{aligned} & P(d_1, ..., d_{i_0}, ..., d_{m+1}) = \sum_{i=1}^{m+1} g_i\, d_i \overset{!}{\equiv} 0 \pmod{n} \\ \iff\ & [g_{i_0}]_n [d_{i_0}]_n + \sum_{i \in \{1,...,m+1\} \setminus \{i_0\}} [g_i]_n [d_i]_n \overset{!}{=} [0]_n \\ \iff\ & [d_{i_0}]_n = -\,[g_{i_0}]_n^{-1} \sum_{i \in \{1,...,m+1\} \setminus \{i_0\}} [g_i]_n [d_i]_n \end{aligned}$$

Weiterführende Bemerkung: Verbesserung der Fehlererkennung und -korrektur ist durch Verwendung von *mehreren* Prüfziffern möglich.
Weitere Ideen/Literatur zum Thema Fehlererkennung und -korrektur: z.B. *Reed-Solomon-Codes*, verwenden Polynome über endlichen Körpern.
Verwendung: Datenverkehr zu Voyager-Sonden (70er Jahre), Fehlerkorrektur in CD-Laufwerken, DVBT (digitales Video-Format) und DAB (digitales Radio).

(Die Idee in groben Zügen: Zum Wort $d_1, ..., d_m$ betrachtet man das Polynom $p(x) = d_1 + d_2 x + ... + d_m x^{m-1}$ und berechnet für fest vorgegebene Stellen $x_1, ..., x_{\tilde{m}}$ (mit $\tilde{m} > m$) die Funktionswerte $p(x_1), ..., p(x_{\tilde{m}})$; diese werden übertragen. Ist z.B. $\tilde{m} = m+2$, und ist nur 1 Einzelfehler passiert, so kann nicht nur der Fehler erkannt, sondern sogar die Stelle des Fehlers ermittelt werden (Prüfe: Welche $\tilde{m}-1$ der $\tilde{m}$ Datenpunkte liegen auf einem Polynom vom Grad $m-1$?), und es kann sogar die falsch übertragene/fehlende Information rekonstruiert werden.
Die Berechnung, von Funktionswerten auf die Koeffizienten eines Polynoms zu schließen, erfordert *'Division'*. Man rechnet daher in *(endlichen) Körpern* $\mathbb{F}_{p^n}$, mit p prim.)

3.2 Algebra und Anwendungen in der Kryptografie (RSA-Verschlüsselung)

Def. (Nullteilerfreiheit)
Ein Ring $(R,+,\cdot)$ heißt *nullteilerfrei*, wenn

$$\forall a,b\in R:\ (\ a\cdot b=0\quad\Rightarrow\quad a=0\ \vee\ b=0\ ).$$

Elemente $a,b\in R$ mit $a\cdot b=0$ nennt man *Nullteiler*.

Körper sind immer nullteilerfrei.

Denn: Sei $a\cdot b=0$. Falls $a=0$: Fertig. Andernfalls: a^{-1} existiert. Multipliziere mit a^{-1}. Es folgt $b=a^{-1}\cdot 0=0$. □

Bei Ringen, die keine Körper sind:

- $(\mathbb{Z},+,\cdot)$ ist nullteilerfrei
- Die Polynomringe $(\mathbb{Z}[X],+,\cdot)$, $(\mathbb{Q}[X],+,\cdot)$,$(\mathbb{R}[X],+,\cdot)$,$(\mathbb{C}[X],+,\cdot)$ sind nullteilerfrei
- $(\mathbb{Z}_n,+,\cdot)$, wobei n nicht prim ist, ist *nicht* nullteilerfrei
  (z.B. $[2]_4\cdot[2]_4=[0]_4$, aber $[2]_4\neq[0]_4$).
- Die Matrixringe $(\mathbb{Q}^{n\times n},+,\cdot),(\mathbb{R}^{n\times n},+,\cdot),(\mathbb{C}^{n\times n},+,\cdot)$ sind nicht nullteilerfrei:
  $\begin{pmatrix}1&0\\0&0\end{pmatrix}\cdot\begin{pmatrix}0&0\\0&1\end{pmatrix}=\begin{pmatrix}0&0\\0&0\end{pmatrix}$

Wozu Nullteilerfreiheit nützlich ist:

In nullteilerfreien Ringen ist folgende Schlussweise erlaubt:

$$a\cdot b=\tilde{a}\cdot b\ \wedge\ b\neq 0\quad\Rightarrow\quad a=\tilde{a}$$

Beweis:

$$\begin{aligned}&a\cdot b=\tilde{a}\cdot b\ \wedge\ b\neq 0\\ \Rightarrow\ &(a-\tilde{a})\cdot b=0\ \wedge\ b\neq 0\\ \overset{(*)}{\Rightarrow}\ &a-\tilde{a}=0\quad\Rightarrow\quad a=\tilde{a}\end{aligned}$$

Dabei wurde an der Stelle (*) die Nullteilerfreiheit verwendet. □

Beachte: Wir haben *nicht* benötigt, dass b^{-1} existiert! Die Argumentation ist also in nullteilerfreien Ringen, die keine Körper sind, wie z.B. $\mathbb{Z}$, erlaubt.

Primzahlen

Satz 5 (Satz von der Primfaktorzerlegung)
Jede Zahl $z \in \mathbb{Z}\setminus\{0\}$ besitzt eine *eindeutige* Darstellung (*Primfaktorzerlegung*)

$$z = \pm p_1 \cdot p_2 \cdot \ldots \cdot p_k \,,$$

wobei $k \in \mathbb{N}_0$ ist und $p_1, ..., p_k$ Primzahlen sind mit $p_1 \leq p_2 \leq \ldots \leq p_k$.
Anders formuliert: $z = \pm p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdot \ldots \cdot p_m^{\alpha_m}$, $m \in \mathbb{N}_0$, $\alpha_i \in \mathbb{N}$, $p_1 < p_2 < \ldots < p_m$ prim.

Beweis: OBdA sei $z > 0$.

(ii) **Existenz:** Beweis per Induktion nach z.
Induktionsanfang bei $z=1$ oder bei $z=2$:
$z=1$ hat Darstellung mit $k=0$; $z=2$ hat Darstellung mit $k=1$ und $p_1=2$.
Induktionsschritt ($z \to z+1$):

(α) Falls $z+1$ prim ist, hat $z+1$ die gesuchte Darstellung (mit $k=1$).

(β) Falls $z+1$ nicht prim ist, existieren (nach Def. Primzahl) $f_1, f_2 \in \{2, ..., n\}$ mit $z+1 = f_1 \cdot f_2$.
Nach Induktionsvoraussetzungen haben f_1, f_2 jeweils eine Primfaktorzerlegung. Aus diesen ergibt sich eine Primfaktorzerlegung von $z+1 = f_1 \cdot f_2$.

(i) **Eindeutigkeit:** Es gelte $z = p_1 \cdot \ldots \cdot p_k = \tilde{p}_1 \cdot \ldots \cdot \tilde{p}_{\tilde{k}}$.
Falls es darin $p_i, \tilde{p}_j$ gibt mit $p_i = \tilde{p}_j$, dann 'kürze' solange (d.h. Nullteilerfreiheit in $\mathbb{Z}$ ausnutzen), bis $p_i \neq \tilde{p}_j \; \forall \, i=1, ..., k, \; j=1, ..., \tilde{k}$.
Zu zeigen: $k = \tilde{k} = 0$.
Angenommen $k \geq 1$. Dann ist auch $\tilde{k} \geq 1$, denn andernfalls hat man Widerspruch dazu dass $\tilde{p}_1$ prim ist.
Somit $z = p_1 \cdot \ldots \cdot p_k \in \tilde{p}_1\mathbb{Z}$ mit $k \geq 1$.
Somit $[p_1]_{\tilde{p}_1} \cdot \ldots \cdot [p_k]_{\tilde{p}_1} = [\tilde{p}_1]_{\tilde{p}_1} = [0]_{\tilde{p}_1}$.
Wegen der Nullteilerfreiheit des Körpers(!) $\mathbb{F}_{\tilde{p}_1}$ folgt, dass einer der Faktoren $[p_j]_{\tilde{p}_1} = [0]_{\tilde{p}_1}$ ist.
Es folgt: $\exists \, m \in \mathbb{Z} : \; p_j = m \cdot \tilde{p}_1$.
Anwendung der Definition von Primzahlen auf p_j ergibt: $m = 1$ oder $\tilde{p}_1 = 1$.
Jeder der beiden Fälle führt auf einen Widerspruch. Also ist $k = 0$. Es folgt die Eindeutigkeit der Zerlegung. □

Weiterführende Bemerkung: Das Konzept "Primfaktorzerlegung" lässt sich auf die Polynomringe $\mathbb{Q}[X]$, $\mathbb{R}[X]$, $\mathbb{C}[X]$ übertragen: Das Analogon zu 'Primzahl' ist 'irreduzibles Polynom': Ein Polynom

heißt irreduzibel, wenn es sich nicht als Produkt von Polynomen mit echt niedrigerem Grad schreiben lässt.
(In $\mathbb{C}[X]$ sind nur lineare Polynome irreduzibel, in $\mathbb{R}[X]$ auch X^2+1, in $\mathbb{Q}[X]$ sogar X^2-2,...)
Abgesehen von konstanten Vorfaktoren hat jedes Polynom eine (bis auf Reihenfolge der Faktoren) eindeutige Darstellung als Produkt von irreduziblen Polynomen.

Schon seit der Antike ist bekannt:

Satz
Es gibt unendlich viele Primzahlen.

Beweis: Siehe erstes Semester, ggf. s. Übung.

Dagegen sind viele Fragen, die Primzahlen betreffen, offen, z.B.:

- Gibt es unendlich viele *Primzahlzwillinge*, d.h. Zahlen $p\in\mathbb{N}$, für die p und $p+2$ prim sind?
  ($[5,7],[11,13],[17,19],[29,31],...,[881,883],...,[1997,1999],...$,
  $3756801695685\cdot 2^{666669}\pm 1,...$)
- *Goldbachsche Vermutung*: Lässt sich jede gerade Zahl $n>2$ als Summe von zwei Primzahlen schreiben?
  $10=3+7$, $22=5+17$,...,$100=17+83$,...
  Bis $n=10^{18}$ stimmt's! (Preisgeld 10^6 Dollar)
- Gibt es unendlich viele Primzahlen, die die Form n^2+1 haben, $n\in\mathbb{N}$? (Landau-Vermutung)
- Liegt zwischen zwei aufeinanderfolgenden Quadratzahen immer mindestens eine Primzahl? (Legendre'sche Vermutung)
- Auch wenn in der folgenden Fragestellung Primzahlen nicht direkt auftauchen, sei hier erwähnt das folgende berühmte Problem aus der Zahlentheorie:
  Im Jahre 1995 bewiesen (von Andrew Wiles und Richard Taylor) wurde die aus dem 17. Jhdt(!) stammende **Fermat'sche Vermutung**, (heute: **Großer Fermat'scher Satz**):
  Die Gleichung $a^n+b^n=c^n$ besitzt keine ganzzahlige Lösungen (a,b,c,n) mit $n>2$.

Wie findet man alle Primzahlen $\leq n$, $n\in\mathbb{N}$ gegeben?

Klassische Methode: **Sieb des Eratosthenes**:

In der Liste $2,3,4,...,n$ streiche darin nacheinander alle Vielfachen von 2, dann alle Vielfachen von 3, dann alle Vielfachen von 5,...

Genauer:

Sieb des Eratosthenes

```
Stelle Liste 2,3,...,n auf.
k:=2
Wiederhole
  notiere k als Primzahl
  streiche alle Vielfachen von k aus der Liste
  setze k:=kleinste noch in der Liste vorhandene Zahl
Bis k^2>n
Notiere auch alle noch in der Liste vorhandenen Zahlen als Primzahlen.
```

Beispiel. Sei $n := 100$. Dann durchläuft k (nur) die Werte $2, 3, 5, 7$.
Beachte: Jede Zahl≤ 100, die nicht prim ist, muss einen Primfaktor ≤ 10,
also 2,3,5 oder 7, haben.

Bemerkungen. Man kann sich überlegen, dass es reicht, beim Streichen von Vielfachen von k mit $k^2, k^2+k, k^2+2k, ...$ zu beginnen statt bei $k, 2k, 3k, ...$
Es gibt Weiterentwicklungen (das sog. 'Sieb von Atkin'), die bei großem n ein wenig schneller sind (Aufwand offenbar $O(n)$ statt $O(n \log\log n)$).

Die Euler'sche Phi-Funktion

Def. (Euler'sche Phi-Funktion)
Die *Euler'sche ϕ-Funktion* $\phi : \mathbb{N} \to \mathbb{N}$ ist definiert als

$$\phi(n) := |\mathbb{Z}_n^*| = |\{k \in \{1, ..., n\} \mid ggT(k, n) = 1\}| .$$

Beispiel:
$\mathbb{Z}_{10}^* = \{1, 3, 7, 9\} \Rightarrow \phi(10) = 4$
$\mathbb{Z}_{17}^* = \{1, ..., 16\} \Rightarrow \phi(17) = 16$

Satz 6 (zur Berechnung von $\phi(n)$)

(a) Für Primzahlen p gilt $\phi(p) = p-1$.

(b) Für Primzahlpotenzen p^k gilt $\phi(p^k) = p^k - p^{k-1} = p^{k-1}(p-1)$.

(c) Für teilerfremde Zahlen $a, b \in \mathbb{N}$ gilt $\phi(a \cdot b) = \phi(a) \cdot \phi(b)$.

Folgerung: **Kennt man die Primfaktorzerlegung einer Zahl n, so kann man**

daraus $\phi(n)$ berechnen.

Beispiel: $4200 = 2^3 \cdot 3 \cdot 5^2 \cdot 7 \;\Rightarrow\; \phi(4200) = \phi(2^3) \cdot \phi(3) \cdot \phi(5^2) \cdot \phi(7) = (8-4) \cdot 2 \cdot (25-5) \cdot 6 = 960$

Beweis des Satzes:

(a) trivial

(b) Von 1 bis p^k gibt es insgesamt p^k viele Zahlen.
Wie viele von denen haben gemeinsamen Teiler mit p^k? Nur genau diejenigen Zahlen, die durch p teilbar sind. Also $1 \cdot p$, $2 \cdot p$,...,$p^{k-1} \cdot p$; das sind p^{k-1} viele.
Somit bleiben $p^k - p^{k-1}$ viele teilerfremde Zahlen übrig.

(c) Später! (Folgt aus dem sog. *Homomorphiesatz* (=Satz 9)).

Wo kommt die Euler'sche Phi-Funktion vor?

1. Prüfsummen: Gute Fehlererkennung $\rightarrow$ großes $\phi(n)$ wünschenswert

2. Kleiner Fermat'scher Satz (später)
Folgerung daraus: RSA-Verschlüsselung

3. **Konstruktionen mit Zirkel und Lineal:**
Ausgehend von $\{0,1\} \subseteq \mathbb{C}$, welche Punkte$\in \mathbb{C}$ sind mit Zirkel und Lineal konstruierbar?
Sei $\{0,1\} \subseteq M \subseteq \mathbb{C}$ diese Menge.
Man kann zeigen:

   (i) $\mathbb{Q} \subseteq M \subseteq \mathbb{C}$, $\mathbb{Q} + i\,\mathbb{Q} \subseteq M \subseteq \mathbb{C}$ (Skizzen siehe Tafel)

   (ii) M ist ein Körper

   (iii) $x \in M \cap \mathbb{R}^+ \;\Rightarrow\; \sqrt{x} \in M$ (Skizze siehe Tafel)

   (iv) Gauß zeigte 1796 (im zarten Alter von 18/19 Jahren), dass die 17-te Einheitswurzel $x_1 := \exp(2\pi i/17) \in M$.
   Daraus folgt: Das regelmäßige 17-Eck ist mit Zirkel und Lineal konstruierbar (dies war das erste mal seit der Antike, dass man etwas Neues m.Z.u.L. konstruieren konnte).
   Idee: Obiges x_1 ist, so wie auch alle anderen 17-ten Einheitswurzeln $1 = x_0, x_1, ..., x_{16}$, Nullstelle des Polynoms $p(X) := X^{17} - 1$. Faktorisiere dieses Polynom (schwierig!): Da $x_0 = 1$ eine Nullstelle ist, ist $X - 1$ ein Linearfaktor. Abspaltung dieses Faktors liefert Polynom $\tilde{p}(X) = 1 + X + X^2 + ... + X^{16}$; dessen Nullstellen sind $x_1, ..., x_{16}$. Gauß' geniale Idee bestand darin, einen Zusammenhang zur Gruppentheorie zu sehen: Die Nullstellenmenge $x_0, ..., x_{16}$ kann mit $(\mathbb{Z}_{17}, +)$ identifiziert werden; $\mathbb{Z}_{17}^*$ entspricht der Nullstellenmenge $x_1, ..., x_{16}$ von $\tilde{p}$. In $(\mathbb{Z}_{17}^*, \cdot)$ kann man eine Untergruppe $\langle 2 \rangle$ der Ordnung 8 finden; die Nullstellen $x_1, ..., x_{16}$ werden aufgeteilt, je nachdem, ob ihr Analogon in $\mathbb{Z}_{17}^*$ zu $\langle 2 \rangle$ gehört oder nicht. Dies ermöglichte Gauß, $\tilde{p}$ als Produkt zweier Polynome vom Grad 8 zu schreiben; auf die Details gehen wir hier nicht ein; Überlegungen basieren auf Stoff von S.18-23. Jedes der beiden Polynome 8-ten Grades wird mit analoger Technik auf Polynome vom Grad 4 reduziert, usw... Somit konnte er explizit herleiten, dass
   $\mathrm{Re}(x_1) = \frac{1}{16}(-1 + \sqrt{17} + \sqrt{2(17 - \sqrt{17})} + 2\sqrt{17 + 3\sqrt{17} - \sqrt{2(17 - \sqrt{17})} - 2\sqrt{2(17 + \sqrt{17})}})$,

$\mathrm{Im}(x_1) = \sqrt{1-\mathrm{Re}(x_1)^2}$, und die hierin vorkommenden Rechenoperationen, Grundrechenarten und Quadratwurzelziehen, sind m.Z.u.L. durchführbar, also ist das regelmäßige 17-Eck m.Z.u.L. konstruierbar!

(v) Diese Technik lässt sich exakt so anwenden, wenn n prim ist und die Form $2^{2^k}+1$ hat. (Es ist dann übrigens $\phi(n) = 2^{2^k}$.)
Wenige Jahre später konnte Gauß zeigen: Das Polynom $p(X) := X^n - 1$ hat genau dann alle n Nullstellen in M ("zerfällt über M in Linearfaktoren"), wenn $\phi(n)$ eine Potenz von 2 ist.
Das bedeutet: **Das regelmäßige n-Eck ist genau dann m.Z.u.L. konstruierbar, wenn $\phi(n)$ eine Potenz von 2 ist.**

Tabelle der Konstruierbarkeit von regelmäßigen n-Ecken:

n	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	...	$2^{16}+1$	...
$\phi(n)$	2	2	4	2	6	4	6	4	10	4	12	6	8	8	16	6	18	8	12	10	22	...	2^{16}	...
konstruierbar	J	J	J	J	N	J	N	J	N	J	N	N	J	J	J	N	N	J	N	N	N		J	

Eine Folgerung:
Kann man jeden **Winkel** m.Z.u.L. **dreiteilen**?
Antwort: Nein. Denn andernfalls würde aus der Konstruierbarkeit des reg. Dreiecks die Konstruierbarkeit des reg. Neunecks folgen. Widerspruch.

(vi) Ferner: $\pi \notin M$, d.h. die "**Quadratur des Kreises**" ist nicht möglich.

(vii) Ferner: $\sqrt[3]{2} \notin M$

Mit *algebraischen* Mitteln konnte man viele geometrische Fragen, die bereits in der Antike aufgeworfen wurden, rund 2000 Jahre später beantworten.

Gruppen und Untergruppen

Def. (Untergruppe)
Sei $(G, *)$ eine Gruppe. Eine Teilmenge $\emptyset \neq U \subseteq G$ heißt *Untergruppe* von G, falls

(i) $\forall a, b \in U : \; a * b \in U$

(ii) $\forall a \in U : \; a^{-1} \in U$

Jede Untergruppe ist mit der ererbten Verknüpfung eine Gruppe. Assoziativität ist klar, und auch das Enthaltensein des Neutralen Elements muss nicht extra gefordert werden:
$a \in U \overset{(ii)}{\Rightarrow} a^{-1} \in U \overset{(i)}{\Rightarrow} e_G = a * a^{-1} \in U$.

Bsp.: $(G, *) := (\mathbb{Z}, +)$ hat $U := \{a \in \mathbb{Z} \mid a \text{ ist gerade}\} = 2\mathbb{Z}$ als Untergruppe; jedoch ist $\{a \in \mathbb{Z} \mid a \text{ ist ungerade}\}$ keine Untergruppe, da (i) verletzt ist (auch: da das neutrale Element von G nicht enthalten ist).

Wie findet man Untergruppen einer Gruppe?

Eine Methode (mit der man *einige* Untergruppen aufspüren kann): Für eine beliebige endliche Gruppe $(G,*)$ und beliebiges $a\in G$ ist immer

$$\langle a\rangle := \{e, a, a*a, a*a*a, ...\} = \{\underbrace{a*...*a}_{k \text{ mal}} \mid k\in\mathbb{N}_0\}$$

eine Untergruppe von $(G,\cdot)$. Wenn man die Gruppe 'multiplikativ' schreibt, also $G,\cdot)$, dann schreibt man kurz

$$\langle a\rangle := \{a^k \mid k\in\mathbb{N}_0\}$$

Man nennt sie *die von a erzeugte Untergruppe*. Eine von einem Element erzeugte Gruppe nennt man auch *zyklisch*. Falls es ein $k\in\mathbb{N}$ gibt mit $a^k = 1$, so ist das kleinste $k\in\mathbb{N}$ mit dieser Eigenschaft die *Ordnung* von a, $\mathrm{ord}(a)$; es ist dann offensichtlich $\langle a\rangle = \{1, a, a^2, ..., a^{\mathrm{ord}(a)-1}\}$; falls es kein solches k gibt (was nur in unendlichen Gruppen G möglich ist), setzt man $\mathrm{ord}(a)=\infty$.

Bemerkungen: Oben haben wir die Gruppe G 'multiplikativ' geschrieben. In einer 'additiv' geschriebenen Gruppe $(G,+)$ würde man $\langle a\rangle = \{0, a, 2a, 3a, ...\}$ bzw., im Fall einer *endlichen* Ordnung, $\langle a\rangle = \{0, a, 2a, ..., (\mathrm{ord}(a)–1)a\}$ schreiben; dabei sei $2a := a+a$, $3a := a+a+a$,...
Und: In einer *unendlichen* Gruppe G muss man in obigen Ausagen $k\in\mathbb{Z}$ statt $k\in\mathbb{N}_0$ nehmen.

Beispiele:

- $(\mathbb{Z}_n,+)$ ist, für $n\in\mathbb{N}$ beliebig, eine zyklische Gruppe, denn $\mathbb{Z}_n = \langle 1\rangle$
- $\mathbb{Z}_5^* = \{1,2,3,4\}$ hat folgende zyklische Untergruppen: $\langle 1\rangle = \{1\} \subsetneq \mathbb{Z}_5^*$
  $\langle 2\rangle = \{1,2,4,3\} = \mathbb{Z}_5^*$ (denn $2^2=4$, $2^3=8\equiv 3$, usw.)
  $\langle 3\rangle = \{1,3,4,2\} = \mathbb{Z}_5^*$
  $\langle 4\rangle = \{1,4\} \subsetneq \mathbb{Z}_5^*$ (denn $4^2=16\equiv 1$)
  $\mathbb{Z}_5^*$ ist also zyklisch, denn sie wird von 2 (und von 3) erzeugt.
- Beispiel für eine nicht-zyklische Gruppe: $(\mathbb{Z}_2\times\mathbb{Z}_2,+)$, denn:
  $$\left\langle \begin{pmatrix}0\\0\end{pmatrix}\right\rangle = \left\{\begin{pmatrix}0\\0\end{pmatrix}\right\} \subsetneq \mathbb{Z}_2\times\mathbb{Z}_2$$
  $$\left\langle \begin{pmatrix}1\\0\end{pmatrix}\right\rangle = \left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}\right\} \subsetneq \mathbb{Z}_2\times\mathbb{Z}_2$$
  $$\left\langle \begin{pmatrix}0\\1\end{pmatrix}\right\rangle = \left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\} \subsetneq \mathbb{Z}_2\times\mathbb{Z}_2$$
  $$\left\langle \begin{pmatrix}1\\1\end{pmatrix}\right\rangle = \left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}\right\} \subsetneq \mathbb{Z}_2\times\mathbb{Z}_2$$

Eine weitere Möglichkeit, Untergruppen zu finden, bietet der gleich folgende Satz 7.

Gruppenhomomorphismen

Def. (Homomorphismus, Isomorphismus von Gruppen)
Seien $(G,*),(H,\circ)$ Gruppen. eine Abbildung $f : G \to H$ heißt *(Gruppen-)Homomorphismus*, falls

$$f(a*b) = f(a) \circ f(b) \quad \forall a,b \in G.$$

Ist f zusätzlich bijektiv, so heißt f *(Gruppen-)Isomorphismus*, und die Gruppen G und H heißen dann *isomorph*: $G \cong H$

Isomorphie bedeutet anschaulich: Die Gruppen/Verknüpfungstabellen haben die "gleiche Struktur"; die Gruppen sind bis auf Umbenennung/Umordnung der Elemente "identisch".

Beispiele für Gruppenisomorphismen:

- $(\mathbb{Z}_3^*,\cdot) \cong (\mathbb{Z}_2,+)$; betrachte dazu die beiden Verknüpfungstabellen; ein Isomorphismus ist $1 \mapsto 0$, $2 \mapsto 1$
- **Eine endliche zyklische Gruppe $(\langle a\rangle,*)$ ist immer isomorph zu einer Gruppe $(\mathbb{Z}_n,+)$, $n = \mathbf{ord}(a)$.** Isomorphismus: $\mathbb{Z}_n \to \langle a\rangle$, $[k]_n \mapsto a^k$
- Sei $K \subset \mathbb{C}$ der Einheitskreis.
  $([0,2\pi),+) \cong (K,\cdot)$ (wobei die Addition modulo 2π auszuführen ist).
  Ein Isomorphismus ist $f(\varphi) = e^{i\varphi}$, denn diese Abbildung ist bijektiv und es gilt $f(\varphi_1+\varphi_2) = e^{i(\varphi_1+\varphi_2)} = e^{i\varphi_1} \cdot e^{i\varphi_2} = f(\varphi_1)\cdot f(\varphi_2)$.
- Sei $K_n \subset K \subset \mathbb{C}$ die Menge aller n-ten komplexen Einheitswurzeln. Dann ist $(K_n,\cdot) \cong (\mathbb{Z}_n,+)$.

Gruppenhomomorphismen führen auf Untergruppen:

Satz 7
Für einen Gruppenhomomorphismus $f : (G,*) \to (H,\circ)$ gilt:

(i) $f(1_G) = 1_H$

(ii) $f(a^{-1}) = f(a)^{-1}\ \forall\, a \in G$

(iii) Ist f Isomorphismus, dann auch f^{-1}.

(iv) Bild(f) ist eine Untergruppe von H

(v) Kern$(f) := \{a \in G \mid f(a) = 1_H\}$ ist eine Untergruppe von G

(vi) f injektiv $\iff$ Kern$(f) = \{1_G\}$

Beachte, dass hier die Gruppen "multiplikativ geschrieben" wurden; man hätte sie genau so gut "additiv" schreiben können, d.h. mit $0_G, 0_H, -x$ an Stelle von $1_G, 1_H, x^{-1}$.

Beachte die Analogie zu linearen Abbildungen $f : V \to W$ zwischen Vektorräumen; siehe 1. Semester:

$f(0_V) = 0_W$, $f(-a) = -f(a)$, Bild(f) ist Unterraum von W, Kern(f) ist Unterraum von V, f ist genau dann injektiv, wenn Kern$(f) = \{0\}$;

Die Beweise für obigen Satz können analog zu den entsprechenden Beweisen aus dem 1. Semester geführt werden!

Einige der Beweise (i)-(vi): s. Tafel

Skizze mit $G, H, \text{Kern}(f), \text{Bild}(f)$: s. Tafel.

Beispiele für Untergruppen, als Kern/Bild eines Homomorphismus 'gefunden':

1.) $f : (\mathbb{Z}, +) \to (\mathbb{Z}, +)$, $f(z) := 3z$
Bild$(f) = 3\mathbb{Z}$ ist Untergruppe von $(\mathbb{Z}, +)$
Kern$(f) = \{0\}$ ist Untergruppe von $(\mathbb{Z}, +)$; f ist injektiv

2.) $f : (\mathbb{Z}_{15}, +) \to (\mathbb{Z}_{10}, +)$, $f([z]_{15}) := [2z]_{10}$
ist wohldef. u. ein Homomorphismus (ggf. s. Ü.)
Bild$(f) = \{[0]_{10}, [2]_{10}, [4]_{10}, [6]_{10}, [8]_{10}\} = \langle [2]_{10} \rangle$ ist Untergruppe von $\mathbb{Z}_{10}$
Kern$(f) = \{[0]_{15}, [5]_{15}, [10]_{15}\} = \langle [5]_{15} \rangle$ ist Untergruppe von $\mathbb{Z}_{15}$

Bemerkungen: Möglicherweise fällt hier folgendes auf:
(a) Die Größe der Untergruppen Bild(f), Kern(f) ist hier jeweils ein Teiler der Größe der jeweiligen übergeordneten Gruppe. Ist das Zufall? Dazu gleich mehr (→Satz 8)
(b) Es besteht hier ein einfacher Zusammenhang zwischen der Elementezahl von $\mathbb{Z}_{15}$ und der von Bild(f), Kern(f): $15 = 8 \cdot 3$. Ist das Zufall? Dazu später mehr!

Um den Effekt (a) genauer zu untersuchen, wollen wir nochmal an das Konzept, mit dem wir $\mathbb{Z}_n$ aus $\mathbb{Z}$ konstruiert haben (über Äquivalenzklassen), erinnern, und dieses verallgemeinern:

Wir hatten auf $\mathbb{Z}$ die Äquivalenzrelation $a \sim b :\Leftrightarrow b - a \in n\,\mathbb{Z}\, \forall a, b \in \mathbb{Z}$ verwendet, und $\mathbb{Z}_n$ als die Menge der entstehenden Äquivalenzklassen, $\mathbb{Z}_n := \{[0], ..., [n-1]\}$ definiert. Dazu stellen wir fest:

- "$b - a = -a + b$" ist die Verknüpfung eines Elements und des (Additiv-)Inversen eines anderen Elements; in multiplikativer Gruppenschreibweise würde dies "$a^{-1} \cdot b$" lauten.
- $(n\mathbb{Z}, +)$ ist eine Untergruppe der Gruppe $(\mathbb{Z}, +)$.

Wir verallgemeinern die Konstruktion von $\mathbb{Z}_n$ aus $\mathbb{Z}$ wie folgt:

1. An Stelle von $(\mathbb{Z}, +)$ nehmen wir eine beliebige Gruppe $(G, *)$.
2. An Stelle der Untergruppe $n\mathbb{Z}$ von $\mathbb{Z}$ nehmen wir eine beliebige Untergruppe U von $(G, *)$.

Also:
Sei $(G, *)$ eine Gruppe und $U \subseteq G$ eine Untergruppe. Wir definieren auf G die Relation $\sim_U$ oder kurz $\sim$:

$$a \sim b \ :\Longleftrightarrow\ a^{-1} * b \in U$$

Man kann leicht nachrechnen (ggf. siehe Tafel): Dies ist eine *Äquivalenzrelation*!

Die Menge der Äquivalenzklassen bezeichnen wir mit G/U:

$$G/U := \{[a]_U \,|\, a \in G\}$$

(G/U ist also das, was im Fall $G = \mathbb{Z}$, $U = n\mathbb{Z}$ das $\mathbb{Z}_n$ war; in neuer Notation also $\mathbb{Z}_n := \mathbb{Z}/(n\mathbb{Z})$.)

Wir wollen die Partitionierung von G, die die Relation $\sim$ erzeugt, genauer untersuchen: Es ist

$$\begin{aligned} [a]_U &= \{b \in G \,|\, a \sim b\} = \{b \in G \,|\, a^{-1} * b \in U\} = \{b \in G \,|\, \exists\, u \in U :\ a^{-1} * b = u\} \\ &= \{b \in G \,|\, \exists\, u \in U :\ b = a * u\} = \{a * u \,|\, u \in U\} = a * U\,. \end{aligned}$$

Insbesondere ist

$$[1_G]_U = 1_G * U = U\,,$$

d.h. eine der erzeugten Klassen ist die Untergruppe U selbst.

Wir wollen uns nun noch überlegen: Alle Klassen sind "gleich groß", d.h. haben gleiche Kardinalität, d.h. es gibt eine Bijektion von einer zur anderen Klasse.

Das ist ganz einfach: Sei $a \in G$ beliebig. Zwischen der Klasse U und der Klasse $[a]_U = a * U$ lautet eine Bijektion wir folgt:

$$U \to a * U, \quad u \mapsto a * u$$

Dass diese Abbildung injektiv und surjektiv ist, ist trivial.

Wichtiges Ergebnis dieser Überlegung (im Fall, dass G *endliche* Gruppe ist):

Alle Klassen sind gleich groß; G wird in gleich große Klassen unterteilt. Insbesondere ist also die Anzahl der Elemente der Gruppe G das Produkt aus der Anzahl der Elemente der Untergruppe U und der Anzahl der Klassen:

$$|G| = |G/U| \cdot |U|$$

(Daher auch die Schreibweise G/U, die an einen Quotienten erinnern soll: "G wird mittels U in gleich große Klassen unterteilt/geteilt.")

Das zentrale Ergebnis dieser Überlegung:

Satz 8 (Lagrange, 1736-1813)
Sei G eine endliche Gruppe und U eine Untergruppe von G. Dann ist die Elementezahl von U ein Teiler der Elementezahl von G.

Eine Folgerung:
Die Ordnung eines beliebigen Elements $a \in G$, also das kleinste $k \in \mathbb{N}$, für das $a^k = 1_G$ ist, ist immer ein Teiler der Elementezahl von G.
Grund: Die Menge $\{a^0, a^1, ..., a^{k-1}\} \subseteq G$ ist eine Untergruppe von G.

Beispiele für die Anwendung von Satz 8:

1. In Gruppe $(\mathbb{Z}_{15}, +)$ muss jede Untergruppe aus 1,3,5 oder 15 Elementen bestehen, und jedes Element hat die Ordnung 1,3,5 oder 15.
   (Z.B. $[6]_{15}$ hat Ordnung 5: $\langle [6]_{15} \rangle = \{[6]_{15}, [12]_{15}, [3]_{15}, [9]_{15}, [0]_{15}\}$.

2. $(\mathbb{Z}_{15}^*, \cdot)$ hat $\phi(15) = \phi(3) \cdot \phi(5) = 2 \cdot 4 = 8$ Elemente.
   (Kontrolle: $\mathbb{Z}_{15}^* = \{[1], [2], [4], [7], [8], [11], [13], [14]\}$)
   Somit hat jede Untergruppe von $(\mathbb{Z}_{15}^*, \cdot)$ die Elementezahl 1,2,4 oder 8, und jedes Element der Gruppe hat die Ordnung 1,2,4 oder 8.
   Insbesondere muss somit für *jedes* Element $[a]_{15}$ der Gruppe $[a]_{15}^8 = [1]_{15}$ sein, also $a^8 \equiv 1 (\text{mod } 15)$ für alle $[a]_{15} \in \mathbb{Z}_{15}^*$!
   (In der Tat ist z.B. $2^8 = 256 = 17 \cdot 15 + 1 \equiv 1 \,(\text{mod } 15)$,
   und $7^8 = 5764801 = 384320 \cdot 15 + 1 \equiv 1 \,(\text{mod } 15)$.)

   Bemerkung: In diesem Zahlbeispiel gilt $x^{\phi(n)} \equiv 1 \,(\text{mod } n) \, \forall x \in \mathbb{Z}_n^*$.
   Ob dies immer so ist, wird später noch in Satz 12 untersucht.

Rechnen auf G/U:

Im Fall $(G, *) = (\mathbb{Z}, +)$, $U = n\,\mathbb{Z}$ hatten wir auf G/U eine Verknüpfung definiert per $[a] + [b] := [a + b]$ und so eine *Gruppenstruktur* bekommen auf $\mathbb{Z}_n$.
Ist das auch im Fall beliebiger Gruppen $(G, *)$, $U \subseteq G$, möglich?
Wir probieren also, auf der Menge der Klassen G/U eine Operation

$$[a]_U \circ [b]_U := [a * b]_U \quad \forall a, b \in G$$

zu definieren. Wir müssen die *Wohldefiniertheit* prüfen.
Dazu seien $\tilde{a} \in [a]_U$, $\tilde{b} \in [b]_U$, und wir müssen $[a * b]_U = [\tilde{a} * \tilde{b}]_U$ zeigen.

Dazu: Wir wissen $a \sim \tilde{a}$, $b \sim \tilde{b}$, also, nach Def. der Relation, $a^{-1} * \tilde{a} \in U$, $b^{-1} * \tilde{b} \in U$ (und zu zeigen: $(a*b)^{-1} * (\tilde{a}*\tilde{b}) \in U$).

$\Rightarrow\ (a*b)^{-1} * (\tilde{a}*\tilde{b}) = b^{-1} * \underbrace{a^{-1} * \tilde{a}}_{\in U} * \tilde{b} \overset{?}{\in} U$?

Um hier weiterzukommen, brauchen wir eine zusätzliche Forderung: Am einfachsten ist es, wenn wir uns auf *kommutative* Gruppen $(G, *)$ beschränken. Dann können wir weiterrechnen: $b^{-1} * a^{-1} * \tilde{a} * \tilde{b} = \underbrace{a^{-1} * \tilde{a}}_{\in U} * \underbrace{b^{-1} * \tilde{b}}_{\in U} \in U$. $\Rightarrow$ Wohldefiniertheit $\square$

So wie im Fall $(G, *) = (\mathbb{Z}, +)$, $U = n\,\mathbb{Z}$, $G/U = \mathbb{Z}_n$ kann man leicht unter Verwendung der Gruppeneigenschaften von $(G, *)$ nachrechnen, dass $(G/U, \circ)$ alle Gruppeneigenschaften erfüllt; man bezeichnet die Gruppe $(G/U, \circ)$ als *Quotienten-* oder auch als *Faktorgruppe*.

Ergebnis:

Satz u. Def. (Quotientengruppe)
Sei $(G, *)$ eine kommutative(!) Gruppe und $U \subseteq G$ eine Untergruppe.
Dann ist auf G/U die Verknüpfung

$$[a]_U \circ [b]_U := [a * b]_U \quad \forall a, b \in G$$

wohldefiniert, und $(G/U, \circ)$ ist eine (kommutative) Gruppe, die sogenannte *Quotienten-* oder *Faktorgruppe*.

Weiterführende Bemerkung: Die Forderung nach Kommutativität von G lässt sich abschwächen: Man kann durchaus nicht-kommutative Gruppen G betrachten, wenn man sich bei der Wahl der Untergruppe U einschränkt, und zwar auf solche, für die $g * U = U * g \,\forall\, g \in G$ gilt (d.h. $\forall\, g \in G, u \in U\ \exists\, \tilde{u} \in U:\ g*u = \tilde{u}*g$). Eine solche Untergruppe von G nennt man auch *Normalteiler* von G. (In einer *kommutativen* Gruppe ist somit jede Untergruppe Normalteiler; in einer nicht-kommutativen Gruppe G sind zumindest die trivialen Untergruppen $\{1_G\}$ und G Normalteiler.) Die Bedeutung des Normalteiler-Konzepts erkannte zuerst Galois (um 1830).

Wichtigstes Beispiel für Quotientengruppen (wie bereits mehrfach erwähnt):
$(G, *) = (\mathbb{Z}, +)$, $U = n\,\mathbb{Z}$, $G/U = \mathbb{Z}/(n\mathbb{Z}) =: \mathbb{Z}_n$.

Eine nette **Anwendung des Rechnens auf G/U: Struktursatz für endliche abelsche Gruppen**:

Ziel: Wollen "Ordnung" in die Gesamtheit aller endlichen abelschen Gruppen bringen:

Eine naheliegende Frage: Sei $n \in \mathbb{N}$ gegeben; wie viele verschiedene Gruppen mit genau n Elementen gibt es? Die Beantwortung dieser anscheinend einfachen Frage ist überaschend schwierig! Fragt man stattdessen "Wie viele verschiedene *abelsche*

Gruppen mit genau n Elementen gibt es?", so kann man dies deutlich leichter beantworten:

Idee: Sei $(G,*)$ eine beliebige(!) endliche abelsche Gruppe mit $G \neq \{1_G\}$.
Wähle ein $a \in G$ mit $a \neq 1_G$, d.h. mit $\mathrm{ord}(a) > 1$.
Betrachte die Untergruppe $U := \langle a \rangle$; $k := \mathrm{ord}(a) = |U|$.
Wir betrachten die zu U gehörenden Nebenklassen $[1_G]_U = U$, $[x_2]_U = x_2 * U, ..., [x_m]_U = x_m * U$ (dabei $m = |G|/k$). Wir wissen, dass diese Nebenklassen eine Partitionierung von G bilden; man kann somit alle Elemente von G in Form der folgenden Tabelle anordnen:

$$\begin{array}{ccccccc}
1_G & a & a^2 & ... & a^{k-1} & & \text{(dies sind die Elemente von } [1_G]_U = U\text{)} \\
x_2 & x_2 * a & x_2 * a^2 & ... & x_2 * a^{k-1} & & \text{(dies sind die Elemente von } [x_2]_U = x_2 * U\text{)} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots \\
x_m & x_m * a & x_m * a^2 & ... & x_m * a^{k-1} & & \text{(dies sind die Elemente von } [x_m]_U = x_m * U\text{)}
\end{array}$$

In einer Tabelle hat man immer eine Bijektion "Menge der Zeilen" $\times$ "Menge der Spalten" $\to$ "Menge der Tabelleneinträge". Hier also: Es gibt eine Bijektion $(G/\langle a \rangle) \times \langle a \rangle \to G$, nämlich $([x_i]_U, a^j) \mapsto x_i * a^j$. Man zeigt unter Verwendung der Kommutativität(!) von G, dass diese Bijektion ein Homomorphismus, somit ein Isomorphismus ist. Weil außerdem $\langle a \rangle$, als zyklische k-elementige Gruppe, zu $\mathbb{Z}_k$ isomorph ist, folgt insgesamt:

$$G \cong (G/\langle a \rangle) \times \langle a \rangle \cong (G/\langle a \rangle) \times \mathbb{Z}_k$$

Anschließend wiederholt man den Vorgang für $G/\langle a \rangle$ anstelle von G, so oft, bis man bei der Gruppe $\{1\}$ angekommen ist. Man hat dann eine Darstellung

$$\begin{array}{l}
G \cong \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times ... \times \mathbb{Z}_{n_k}, \quad k \geq 1,\ n_i \geq 2, \text{ insbes.} \\
|G| = n_1 \cdot n_2 \cdot ... \cdot n_k
\end{array}$$

Jede endliche abelsche Gruppe hat diese Struktur; man bezeichnet daher diese Darstellbarkeit als *Struktursatz endlicher abelscher Gruppen.*

Beispiel: Wie viele abelsche Gruppen mit genau 75 Elementen gibt es?
Antwort: Es ist $75 = 5 \cdot 5 \cdot 3$.
Somit gibt es die abelschen Gruppen $\mathbb{Z}_5 \times \mathbb{Z}_5 \times \mathbb{Z}_3$, $\mathbb{Z}_{25} \times \mathbb{Z}_3$, $\mathbb{Z}_5 \times \mathbb{Z}_{15}$, $\mathbb{Z}_{75}$. Mehr gibt es nicht; d.h. jede andere kommutative Gruppe mit 75 Elementen muss zu einer dieser Gruppen isomorph sein.

Weiterführende Bemerkung: Das obige Zerlegungsprinzip kann man auch auf *nicht-kommutative* Gruppen G anwenden, sofern man zumindest eine "echte" Untergruppe findet, die *Normalteiler* ist. Nicht jede nicht-kommutative Gruppe hat so einen "echten" Normalteiler. Die Klassifizierung sämtlicher *nicht-kommutativer* endlicher Gruppen ist somit *erheblich* schwieriger und war bis vor einigen Jahren noch Gegenstand von Forschung.

Obiger "Struktursatz" für endliche abelsche Gruppen lässt sich noch verschärfen, denn:

Eine Gruppe der Form $\mathbb{Z}_p \times \mathbb{Z}_q$, wobei p und q teilerfremd sind, kann man immer als $\langle(1,1)\rangle$ schreiben, d.h. als zyklische Gruppe, die vom Element $(1,1)$ erzeugt wird, und dieses Element hat die Ordnung $p \cdot q$, somit $\mathbb{Z}_p \times \mathbb{Z}_q \cong \mathbb{Z}_{pq}$.
Haben p und q einen echten gemeinsamen Teiler, so sind $\mathbb{Z}_p \times \mathbb{Z}_q$ und $\mathbb{Z}_{pq}$ *nicht* isomorph.

Dies wird als Satz 11 später noch bewiesen. Momentan soll reichen:

Veranschaulichung/Beispiel: $(\mathbb{Z}_2 \times \mathbb{Z}_3, +)$ wird erzeugt vom Element $(1,1)$, denn
$(1,1) + (1,1) = (0,2)$,
$(1,1) + (1,1) + (1,1) = (1,0)$, ...,
$(1,1) + (1,1) + (1,1) + (1,1) + (1,1) + (1,1) = (0,0)$,
d.h. $(\mathbb{Z}_2 \times \mathbb{Z}_3, +) \cong (\mathbb{Z}_6, +)$.

Damit kann man im obigen Beispiel der abelschen Gruppen mit 75 Elementen einige eliminieren: $\mathbb{Z}_5 \times \mathbb{Z}_5 \times \mathbb{Z}_3 \cong \mathbb{Z}_5 \times \mathbb{Z}_{15}$, $\mathbb{Z}_{25} \times \mathbb{Z}_3 \cong \mathbb{Z}_{75}$.
Es gibt also, bis auf Isomorphie, letztendlich nur *zwei* verschiedene abelsche Gruppen mit 75 Elementen, nämlich
$\mathbb{Z}_{75}$ $[\cong \mathbb{Z}_{25} \times \mathbb{Z}_3]$ und
$\mathbb{Z}_5 \times \mathbb{Z}_{15}$ $[\cong \mathbb{Z}_5 \times \mathbb{Z}_5 \times \mathbb{Z}_3]$.

Unter Ausnutzung u.a. dieses Effekts kann man den obigen Struktursatz verschärfen zu:

Satz (Struktursatz endlicher abelscher Gruppen)
Jede endliche abelsche Gruppe G ist isomorph zu einer Gruppe der Form

$$\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times ... \times \mathbb{Z}_{n_k}, \quad k \geq 1,$$

wobei $n_1 \geq 2$ und n_1 Teiler von n_2, n_2 Teiler von n_3, ..., n_{k-1} Teiler von n_k ist. Es ist weiterhin $|G| = n_1 \cdot n_2 \cdot ... \cdot n_k$.

Der Homomorphiesatz:

Im folgenden geht es darum, mehr über die Struktur der Quotientengruppe $(G/U, \circ)$ herauszufinden, und zwar in dem Fall, dass sich U als Kern eines Gruppenhomomorphismus auffassen lässt (s. Satz 7 Teil (v)):

Satz 9 (Homomorphiesatz)
Seien G und H Gruppen und $f : G \to H$ ein Gruppenhomomorphismus.
Dann sind die Gruppen $G/\mathrm{Kern}(f)$ und $\mathrm{Bild}(f)$ isomorph, d.h.

$$G/\mathrm{Kern}(f) \cong \mathrm{Bild}(f)\,;$$

ein Isomorphismus ist

$$\tilde{f} : G/\mathrm{Kern}(f) \longrightarrow \mathrm{Bild}(f)\,,$$
$$[a]_{\mathrm{Kern}(f)} \longmapsto \tilde{f}([a]_{\mathrm{Kern}(f)}) := f(a)\,.$$

Das heißt:

- Die Anzahl der Äquivalenzklassen, in die G bzgl. der Relation $a \sim b \Leftrightarrow a^{-1} \circ b \in \mathrm{Kern}(f)$ zerlegt wird, ist gleich der Anzahl der Bildwerte von f,
- die Struktur (Größe und Verknüpfungstabelle) der 'abstrakten' Quotientengruppe $(G/\mathrm{Kern}(f), \cdot)$ kann man an der Gruppe $\mathrm{Bild}(f)$ ablesen.

Folgerung aus dem Satz: Da für die Elementezahl $|G/\mathrm{Kern}(f)| = \frac{|G|}{|\mathrm{Kern}(f)|}$ gilt (s. Formel vor Satz 8), und nach Satz 9 $|G/\mathrm{Kern}(f)| = |\mathrm{Bild}(f)|$ gilt, folgt

$$|G| = |\mathrm{Kern}(f)| \cdot |G/\mathrm{Kern}(f)| = |\mathrm{Kern}(f)| \cdot |\mathrm{Bild}(f)|$$

Skizze: s. Tafel

Dies ist der Effekt, den wir bereits im Beispiel 2 nach Satz 7 beobachtet haben:

Wiederholung des Beispiels 2 nach Satz 7, und Erklärung mittels Homomorphiesatz:
Bsp. $f : (\mathbb{Z}_{15}, +) \to (\mathbb{Z}_{10}, +),\ f([z]_{15}) := [2z]_{10}$

$\mathrm{Bild}(f) = \{[0]_{10}, [2]_{10}, [4]_{10}, [6]_{10}, [8]_{10}\} = \langle [2]_{10} \rangle$ ist Untergruppe von $\mathbb{Z}_{10}$
$\mathrm{Kern}(f) = \{[0]_{15}, [5]_{15}, [10]_{15}\} = \langle [5]_{15} \rangle$ ist Untergruppe von $\mathbb{Z}_{15}$

Für die Elementezahl der 3 beteiligten Gruppen gilt, wie Satz 9 voraussagt, in der Tat $15 = 3 \cdot 5$. Der Satz sagt weiterhin, dass die abstrakte Gruppe $\mathbb{Z}_{15}/\{[0]_{15}, [5]_{15}, [10]_{15}\}$ die Struktur von Gruppe $(\{[0]_{10}, [2]_{10}, [4]_{10}, [6]_{10}, [8]_{10}\}, +)$ hat (somit von $(\mathbb{Z}_5, +)$).

Bemerkung zum Homomorphiesatz: Quotientengruppen hatten wir eigentlich nur in *kommutativen* Gruppen gebildet. Im Homomorphiesatz jedoch fehlt die Anforderung 'kommutativ', und es wird dennoch die Quotientengruppe $G/\mathrm{Kern}(f)$ gebildet.
Erklärung: Man könnte zwar das Wort 'Kommutativ' in den Homomorphiesatz einfügen, das ist jedoch nicht nötig; der Homomorphiesatz gilt jedoch in der Tat auch für nicht-kommutative Gruppen

G; Hintergrund: Eine Untergruppe der Form $U = \text{Kern}(f)$ ist *immer* Normalteiler, wie man zeigen kann.

Beweis des Homomorphiesatzes: Entfällt aus Zeitgründen;
man zeigt:
1. $\tilde{f}$ ist wohldefiniert
2. $\tilde{f}$ ist Homomorphismus
3. $\tilde{f}$ ist injektiv
(Die Surjektivität von $\tilde{f}$ ist trivial.)

Im Zusammenhang mit dem Struktursatz endlicher abelscher Gruppen hatten wir bereits erwähnt bzw uns anschaulich überlegt, dass $\mathbb{Z}_m \times \mathbb{Z}_n \cong \mathbb{Z}_{mn}$, sofern n und m teilerfremd sind. Diese Aussage sowie einige andere werden im folgenden Satz gebündelt:

Satz 11
Seien $m, n \in \mathbb{N}$.

(a) Falls m und n teilerfremd sind, dann sind die Gruppen $(\mathbb{Z}_{nm}, +)$ und $(\mathbb{Z}_m \times \mathbb{Z}_n, +)$ isomorph; und zwar ist die Abbildung

$$\begin{aligned} \tilde{f} : \mathbb{Z}_{mn} &\longrightarrow \mathbb{Z}_m \times \mathbb{Z}_n, \\ [a]_{mn} &\longmapsto [a]_m \times [a]_n \end{aligned}$$

ein Isomorphismus *("Chinesischer Restsatz"; China, 3. Jhdrt. n.Chr.(!)).*

(b) Falls m und n einen echten gemeinsamen Teiler haben, dann sind $(\mathbb{Z}_{nm}, +)$ und $(\mathbb{Z}_m \times \mathbb{Z}_n, +)$ *nicht* isomorph.

(c) Die Einschränkung von obigem $\tilde{f}$ auf $\mathbb{Z}^*_{mn}$,

$$\tilde{f} : \mathbb{Z}^*_{mn} \longrightarrow \mathbb{Z}^*_m \times \mathbb{Z}^*_n \,,$$

ist, falls m und n teilerfremd, wohldefiniert und ebenfalls bijektiv.

Die Aussage (c) ist interessant, denn sie besagt:
Folgerung aus dem Satz, Teil (c):
Für teilerfremde $m, n \in \mathbb{N}$ ist $\underbrace{|\mathbb{Z}^*_{mn}|}_{=\phi(mn)} = \underbrace{|\mathbb{Z}^*_m|}_{=\phi(m)} \cdot \underbrace{|\mathbb{Z}^*_n|}_{\phi(n)}$.
Wir haben somit die Rechenregel zur Berechnung der Euler'schen Phi-Funktion, Satz 6-(c), bewiesen!

Zur Veranschaulichung von Aussagen (a) und (b):
Zu (a): Die Aussage (a) besagt: Alle Zahlen$a$, die bzgl. mn den gleichen Rest haben, haben auch

bzgl. m und bzgl. n den gleichen Rest, und umgekehrt (sofern m, n teilerfremd). Der Rest bzgl. mn determiniert die Reste bzgl. m und n, sowie auch umgekehrt.
Nehmen wir z.B. $m=3$ und $m=7$ (sind teilerfremd!); $mn=21$. Sei bekannt, dass eine Zahl a bzgl. 21 den Rest 8 habe. Also $a \in \{8, 21+8, 2\cdot 21+8, 3\cdot 21+8, ...\}$. Dann sind dadurch die Reste von a bzgl. 3 und 7 festgelegt; diese sind 2 und 1.
Auch umgekehrt: Eine Zahl habe (i) bzgl. 3 den Rest 2 und (ii) bzgl. 7 den Rest 1. Aus (i) folgt, dass der Rest der Zahl bzgl. 21 in $\{2, 5, 8, 11, 14, 17, 20\}$ liegt, aus (ii) folgt dass dieser Rest in $\{1, 8, 15\}$ liegt, also muss diese Zahl bzgl. 21 den Rest 8 haben.

Zu (b): Die Teilerfremdheit von m, n ist wichtig. Haben m, n gemeinsame Teiler, dann funktioniert Obiges nicht: Seien z.B. $m=10$ und $n=15$ (also nicht teilerfremd). Dann haben die beiden Zahlen 30 und 60 gleiche Reste bezugl. 10 und 15. Aber bzgl. $mn=150$ haben beide Zahlen unterschiedliche Reste.

Zum Beweis von Satz 11:
Die Aussage (a) ist uns zumindest anschaulich klar: $(1,1)$ ist ein erzeugendes Element von $\mathbb{Z}_m \times \mathbb{Z}_n$, somit ist $\mathbb{Z}_m \times \mathbb{Z}_n$ zyklisch, hat mn viele Elemente, ist also isomorph zu $\mathbb{Z}_{mn}$.
Ein formaler Beweis von (a) kann unter Verwendung des *Homomorphiesatzes* erfolgen:
zu (a): Man betrachte die Abbildung $f : \mathbb{Z} \to \mathbb{Z}_m \times \mathbb{Z}_n$, $a \mapsto [a]_m \times [a]_n$. Man zeigt, dass $\mathrm{Kern}(f) = mn\,\mathbb{Z}$ ist (ist elementar). Nach dem Homomorphiesatz ist somit die Quotientengruppe $\mathbb{Z}/\mathrm{Kern}(f)$ isomorph zu $\mathrm{Bild}(f)$, d.h. $\mathbb{Z}_{mn} = \mathbb{Z}/(mn\,\mathbb{Z}) = \mathbb{Z}/\mathrm{Kern}(f)$ ist isomorph zu $\mathrm{Bild}(f)$. Dann überlegt man sich noch, dass f surjektiv ist.

zu (b): Sei $t>1$ ein gemeinsamer Teiler von m und n; es sei also $n=\tilde{n}t$, $m=\tilde{m}t$. Wir wollen zeigen, dass es in $\mathbb{Z}_m \times \mathbb{Z}_n$ dann kein Element der Ordnung $mn = \tilde{m}\tilde{n}t^2$ gibt, denn da es in $\mathbb{Z}_{mn}$ ein Element dieser Ordnung gibt (nämlich $[1]_{mn}$), folgt, dass die Gruppen nicht isomorph sind.
Also: Zeige, dass alle Elemente aus $\mathbb{Z}_m \times \mathbb{Z}_n$ eine Ordnung $\leq \tilde{n}\tilde{m}t < \tilde{n}\tilde{m}t^2 = mn$ haben:
Für beliebiges $([a]_m, [b]_n) \in \mathbb{Z}_m \times \mathbb{Z}_n$ ist $\tilde{m}\tilde{n}t([a]_m, [b]_n) = ([\tilde{n}\tilde{m}ta]_{\tilde{m}t}, [\tilde{n}\tilde{m}tb]_{\tilde{n}t}) = ([0]_n, [0]_m)$. Fertig.
zu (c): Um von (a) auf (c) zu schließen, müssen wir nur zeigen, dass $[a]_m \times [a]_n$ genau dann in $\mathbb{Z}_m^* \times \mathbb{Z}_n^*$ liegt, wenn $[a]_{mn} \in \mathbb{Z}_{mn}^*$, für beliebiges $a \in \mathbb{Z}$.
Das wiederum ist gleichbedeutend dazu, dass man die Äquivalenz

$$\mathrm{ggT}(a,m)=1 \ \wedge \ \mathrm{ggT}(a,n)=1 \quad \Longleftrightarrow \quad \mathrm{ggT}(a,mn)=1\,.$$

zeigt. Dies wiederum ist, indem wir auf beiden Seiten zur Negation übergehen, äquivalent zu

$$\mathrm{ggT}(a,m)>1 \ \vee \ \mathrm{ggT}(a,n)>1 \quad \Longleftrightarrow \quad \mathrm{ggT}(a,mn)>1\,.$$

Dass dies wahr ist, ist klar: Gibt es einen gemeinsamen Teiler von a und m oder von a und n, dann ist dieser auch Teiler von mn. Und haben a und mn einen gemeinsamen Teiler, dann haben sie einen gemeinsamen Primfaktor. Dieser ist dann in a und in m oder n enthalten, somit haben a und m oder a und n diesen gemeinsamen Primfaktor. □

Bemerkung: Das $\tilde{f}$ aus (c) wird, anders als das $\tilde{f}$ aus (a), nicht als Homomorphismus bzw. Isomorphismus bezeichnet. Grund: Die in (a) vorkommenden Mengen bilden Gruppen bzgl. der *Addition*; das $\tilde{f}$, und somit auch dessen in (c) betrachtete Einschränkung, erfüllen die Homomophieeigenschaft bzgl. '+'. Die in (c) vorkommenden Mengen sind aber bzgl. '+' gar keine Gruppen (fehlende Abgeschlossenheit), sondern nur bzgl. '$\cdot$' sind sie Gruppen (bezüglich der Verknüpfung '$\cdot$' jedoch wird $\tilde{f}$ wohl kaum die Homomorphieeigenschaft besitzen.)

Der folgende Satz ist die wesentliche Grundlage für die RSA-Verschlüsselung; er lässt sich als Folgerung aus dem Satz 8 (Lagrange) auffassen:

Satz 12 (Kleiner Fermat'scher Satz)

(a) Für jede Primzahl p und jedes $x \in \mathbb{Z}$ ist

$$x^p \equiv x \pmod{p}$$

(b) Speziell für $x \in \mathbb{Z}$, das nicht durch die Primzahl p teilbar ist, ist

$$x^{p-1} \equiv 1 \pmod{p}$$

(c) Für $n \in \mathbb{N}$ und $x \in \mathbb{Z}$ mit $ggT(x,n) = 1$ (d.h. für $[x]_n \in \mathbb{Z}_n^*$)

$$x^{\phi(n)} \equiv 1 \pmod{n}$$

Bemerkungen:

- Während (a) und (b) auf Fermat zurückgehen, ist (c) eine Verallgemeinerung von (b), die auf Euler zurückgeht.
- Das "Beispiel 2 für die Anwendung von Satz 8" ist ein Zahlenbeispiel für die Gültigkeit des obigen Satzes.

Beweis:

- **Wir beginnen mit Aussage (c):**
  Nach Voraussetzung ist $ggT(x,n) = 1$, also liegt (nach Satz 4) $[x]_n$ in der Gruppe $(\mathbb{Z}_n^*, \cdot)$:

  $$[x]_n \in \mathbb{Z}_n^*$$

  Somit (siehe Folg. aus Satz 8 (Lagrange)) ist die Ordnung von $[x]_n$ ein Teiler der Gruppengröße $|\mathbb{Z}_n^*| = \phi(n)$:

  $$\phi(n) = \alpha \cdot \mathrm{ord}([x]_n) \quad \text{für ein } \alpha \in \mathbb{N}\,. \qquad (*)$$

  Somit

  $$[x^{\phi(n)}]_n \overset{(*)}{=} [x^{\alpha \cdot \mathrm{ord}([x]_n)}]_n \;=\; [x^{\mathrm{ord}([x]_n)}]_n^{\alpha} \overset{\text{Def.'ord'}}{=} [1]_n^{\alpha} \;=\; [1]_n\,.$$

  $$[x^{\phi(n)}]_n = [1]_n\,.$$

- **Wir zeigen: Aus (c) folgt (b):**
  Wähle das n in (c) als *Primzahl*. Dann ist $\phi(n) = n-1$ (Satz 6), und Aussage (b) steht da.

– **Wir zeigen Aussage (a):**
Falls x nicht durch p teilbar ist, kann (b) verwendet werden: (b) mit x multiplizieren ergibt (a).
Falls x durch p teilbar ist, kann (b) nicht verwendet werden, jedoch es ist $x \equiv 0 \pmod{p}$ sowie $x^p \equiv 0 \pmod{p}$, somit $x^p \equiv x \pmod{p}$. □

RSA-Verschlüsselung (Rivest, Shamir, Adleman, MIT 1977)

Allgemein zum Konzept 'Verschlüsselung':

Die Nachricht wird vorab in "Pakete" (Wörter)$\in \mathbb{Z}_n$ (n fest); d.h. die Nachricht besteht aus einer Aneinanderreihung (Tupel) von Elementen aus $\mathbb{Z}_n$.
Jedes Wort wird für sich verschlüsselt:

Verschlüsseln: Abbildung $E_n : \mathbb{Z}_n \longrightarrow \mathbb{Z}_n$, *injektiv(!), somit bijektiv*
Entschlüsseln: Umk'abb. $E_n^{-1} : \mathbb{Z}_n \longrightarrow \mathbb{Z}_n$

Hinsichtlich Geheimhaltung sind 2 Konzepte denkbar:

- **Klassische Variante: Symmetrische Verfahren:**
  Jeder, der *Ver*schlüsseln kann, kann auch *ent*schlüsseln.
  Beispiel: "**Ver-Odern**": n sei 2erpotenz, und im Binärsystem wird bitweise gerechnet (siehe Tafel):
  E_n, E_n^{-1} : Nachricht $\mapsto$ Nachricht *xor* Schluessel; hier ist $E_n^{-1} = E_n$
  Nachteil: Der Herausgeber des Schlüssels muss seinem Kommunikationspartner den *Ver*schlüsselungs-Schlüssel zusenden. Wird diese Zusendung von einem Unbefugten abgehört, so kann dieser auch *ent*schlüsseln.
  Bemerkung: Von der Antike bis in die 1970er Jahre gab es nur symmetrische Verfahren.

- **Asymmetrische Variante/Private-Public-Key-Variante:**
  Der Schlüsselerzeuger veröffentlicht nur seinen *Ver*schlüsselungsschlüssel E_n. Es ist nicht möglich bzw sehr schwer, daraus die Entschlüsselungsvorschrift E_n^{-1} zu gewinnen; diese ist geheim ('private'), während der Verschlüsselungsschlüssel also 'public' ist.
  (Nur) Der Schlüsselerzeuger kann also mit seinem eigenen, nur ihm bekannten Entschlüsselungsschlüssel Nachrichten *ent*schlüsseln. Dagegen wird die *Ver*schlüsselungsmethode nicht geheim gehalten.
  Dieses Konzept setzt also voraus, **dass man aus der Kenntnis der Abbildung E_n nicht leicht auf die Abbildung E_n^{-1} schließen kann**. Das bedeutet insbesondere, dass n sehr groß sein muss, denn andernfalls kann man sich bei Kenntnis von E_n eine Wertetabelle für E_n anlegen; dies liefert eine Wertetabelle für E_n^{-1}.
  Aus E_n soll man also nicht auf E_n^{-1} schließen können. Andererseits muss der

Schlüsselgenerierer ja irgendwie in der Lage sein, zueinander inverse Abbildungen E_n und E_n^{-1} generieren zu können. Wie geht das?

Wie konstruiert man $E_n, E_n^{-1} : \mathbb{Z}_n \to \mathbb{Z}_n$?

Idee: Fermat's Satz besagt $x^{\phi(n)} \equiv 1 \,(\text{mod } n)\, \forall x \in \mathbb{Z}_n^*$.
Es folgt: $x^{\phi(n)+1} \equiv x \,(\text{mod } n)\, \forall x \in \mathbb{Z}_n^*$
Das bedeutet: Die Abbildung $x \mapsto x^{\phi(n)+1}$ ist die Identität auf $\mathbb{Z}_n^*$!

Falls man also $\phi(n)+1$ als Produkt $\phi(n)+1 = e \cdot d$ schreiben kann, dann kann man

$$E_n x := E_{n,e} x :\equiv x^e \,(\text{mod } n)\,,$$
$$E_n^{-1} x := E_{n,d} x :\equiv x^d \,(\text{mod } n)$$

setzen, und es ist $E_{n,d}(E_{n,e}(x)) \equiv x^{de} \equiv x^{\phi(n)+1} \equiv x$, also $E_{n,d} \circ E_{n,e}(x) = x$, für alle $[x]_n \in \mathbb{Z}_n^*$ (um die Wörter $[x]_n \in \mathbb{Z}_n \setminus \mathbb{Z}_n^*$ müssen wir uns später noch kümmern).
(Dabei 'e' für encode, 'd' für decode)
Die Umkehrfunktion zu „ hoch e nehmen modulo n", also das Ziehen der e-ten Wurzel modulo n, lässt sich also als *Potenzieren* mit $d \in \mathbb{N}$ modulo n realisieren.

Kennt man e und n, so kann man verschlüsseln; kennt man d und n, so kann man entschlüsseln.

Ob sich die obige Forderung an e und d, dass

$$e\,d \stackrel{!}{=} \phi(n)+1\,, \qquad (1)$$

für ein vorgegebenes $n \in \mathbb{N}$ überhaupt erfüllen lasst, mit $e, d > 1$, bzw. mit $e, d \gg 1$, ist gar nicht so klar. Glücklicherweise kann man die Forderung (1) abschwächen zu:

$$\exists\, \alpha \in \mathbb{N} : \quad e\,d \stackrel{!}{=} \alpha\,\phi(n)+1\,, \qquad (\tilde{1})$$

denn dann ist weiterhin

$$E_{n,d}(E_{n,e}(x)) \equiv x^{de} \equiv x^{\alpha\phi(n)+1} \equiv (\underbrace{x^{\phi(n)}}_{\equiv 1})^\alpha \cdot x \equiv 1^\alpha \cdot x \equiv x \,(\text{mod } n) \qquad (*)$$

für alle $[x]_n \in \mathbb{Z}_n^*$, also $E_{n,e} \circ E_{n,d} = id$.

Wir wollen nun e und d und n konstruieren, so dass $(\tilde{1})$ gilt. Dazu wende "mod $\phi(n)$" auf $(\tilde{1})$ an; wir bekommen, dass $(\tilde{1})$ äquivalent ist zu

$$[e]_{\phi(n)} \cdot [d]_{\phi(n)} \stackrel{!}{=} [1]_{\phi(n)}\,. \qquad (\tilde{\tilde{1}})$$

Dies besagt: **Zur Konstruktion von** e **und** d (bei vorgegebenem n) ist zu beachten, dass $[e]_{\phi(n)}$ und $[d]_{\phi(n)}$ invers zueinander sein müssen in $(\mathbb{Z}_{\phi(n)}^*, \cdot)$, insbesondere muss $e \in \mathbb{Z}_{\phi(n)}^*$ gewählt werden (d.h. $ggT(e, \phi(n)) \stackrel{!}{=} 1$), und setze dann $[d]_{\phi(n)} := [e]_{\phi(n)}^{-1}$

(dies kann auch bei großen Zahlen mit Euklidischem Divisionsalgorithmus berechnet werden!)

Wie sollte man n wählen? Dazu sind 2 Dinge zu beachten:

(I) Die Gleichung (*) soll auch für alle $[x]_n \in \mathbb{Z}_n \setminus \mathbb{Z}_n^*$ gelten. ($\rightsquigarrow$nächste Seite)

(II) Aus der Kenntnis des öffentlichen Schlüssels n, e soll es für Fremde schwer/unmöglich sein, d zu berechnen; $[d]_{\phi(n)} = [e]_{\phi(n)}^{-1}$. Das können wir dadurch erreichen, indem wir dafür sorgen, dass es für Fremde schwer ist, aus der Kenntnis von n den Wert $\phi(n)$ zu berechnen.

Idee: Wähle zwei **sehr große Primzahlen** p, q und setze

$$n := p \cdot q \,. \tag{2a}$$

mit $p \neq q$. Es ist dann nach Satz 6-(a+c)

$$\phi(n) = \phi(p) \cdot \phi(q) = (p-1) \cdot (q-1) \tag{2b}$$

n ist zwar öffentlich, nicht aber $p, q, \phi(n)$. Das Knacken des Codes erfordert Berechnung von $[d]_{\phi(n)} = [e]_{\phi(n)}^{-1}$, somit von $\phi(n)$. Das $\phi(n)$ aber ist aus n (soweit man weiß) nur berechenbar, indem man die Primfaktorzerlegung (2a) kennt/findet. Und das ist (soweit man weiß) sehr schwer! Somit wird (II) erfüllt.

Wir sind fast fertig. Es bleibt nur noch zu zeigen, dass (I) die Gleichung (*) auch für alle $[x]_n \in \mathbb{Z}_n \setminus \mathbb{Z}_n^*$ erfüllt wird.

Sei dazu $0 \leq x < n$ mit $[x]_n \in \mathbb{Z}_n \setminus \mathbb{Z}_n^*$, also mit $ggT(x, n) > 1$.
Da $n = pq$ als echte Teiler nur p und q hat (da p, q prim), muss x ein Vielfaches von p oder q sein (nicht jedoch von beiden, da $0 \leq x < n = pq$).
OBdA sei x ein Vielfaches von p und nicht Vielfaches von q.

Es folgt erstens
$\Rightarrow ggT(x, q) = 1 \overset{\text{(Fermat-(b))}}{\Longrightarrow} x^{q-1} \equiv 1 \,(\text{mod } q)$
$\Rightarrow x^{de} \overset{(\tilde{1})}{=} x^{\alpha\phi(n)+1} \overset{(2b)}{=} x^{\alpha(q-1)(p-1)+1} = (\underbrace{x^{q-1}}_{\equiv 1})^{\alpha(p-1)} \cdot x \equiv 1^{\alpha(p-1)} \cdot x = x \,(\text{mod } q)$
also $x^{de} \equiv x \,(\text{mod } q)$.

Und zweitens
$x \equiv 0 \,(\text{mod } p)$, somit auch $x^{de} \equiv 0 \,(\text{mod } p)$,
also $x^{de} \equiv x \,(\text{mod } p)$.

Erstens und zweitens zusammengenommen bedeutet

$$[x^{de}]_p \times [x^{de}]_q = [x]_p \times [x]_q \overset{\text{Chin.Rest-satz (a)}}{\Longrightarrow} [x^{de}]_n = [x]_n \qquad \square$$

Zusammenfassend: RSA-Verschlüsselung

Wähle große Primzahlen p, q, setze $n := p \cdot q$; wähle $e \in \{2, ..., \phi(n)-1\}$ mit $ggT(e, \phi(n)) = 1$; berechne $[d]_{\phi(n)} := [e]^{-1}_{\phi(n)}$ mittels Euklidischem Divisionsalgorithmus.

Veröffentliche nur n, e, nicht aber $d, p, q, \phi(n)$.
→ Der Verschlüsselungsschlüssel $E_{n,e}$, $E_{n,e}x := x^e \pmod n$ ist öffentlich.
Die Werte $p, q, \phi(n)$ können gelöscht werden, nachdem d, e berechnet wurden.

Zum Entschlüsseln $E_{n,d}x := x^d \pmod n$ ist d erforderlich. Nur der Erzeuger des Schlüssels kennt d. Aus der Kenntnis der öffentlichen Daten n, e das d zu berechnen ($[d]_{\phi(n)} = [e]^{-1}_{\phi(n)}$), ist schwer; man benötigt dazu $\phi(n)$. Obwohl n öffentlich ist, ist die Berechnung von $\phi(n) = (p-1)(q-1)$ anscheinend nur machbar, wenn man n in seine Primfaktoren p und q zerlegt hat, und das ist hart.

Letztendlich haben wir (unter Verwendung des kleinen Fermat'schen Satzes und der Rechenregel aus Satz 6-c, die auf Satz 11-c (Gauß) beruht) eine bijektive Funktion $\mathbb{Z}_n \to \mathbb{Z}_n$ konstruiert, für die, obwohl öffentlich bekannt, es kaum möglich ist, die Umkehrfunktion zu berechnen, solange man nicht über gewisse Zusatz-Information verfügt, die nur der Schlüsselerzeuger kennt.

Die Sicherheit des Verfahrens beruht also auf:

- Es ist (nach heutigem Kenntnisstand) hart, für eine große Zahl n, die Produkt zweier Primzahlen ist, das $\phi(n)$ zu berechnen, denn dazu benötigt man (nach heutigem Kenntnisstand) die Zerlegung von n in eben diese beiden Primzahlen, und diese zu finden ist (nach heutigem Kenntnisstand) hart.
- Es gibt übrigens keinen Beweis dafür, dass das Finden der Primfaktoren bzw. die Berechnung von $\phi(n)$ für große n grundsätzlich "schwierig" ist; es ist nur so, dass niemand einen schnellen Algorithmus dafür kennt. (Das Finden/Beweisen von "unteren Schranken" für den Mindest-Rechenaufwand, der zum Lösen eines Problems nötig ist, ist übrigens eine Fragestellung, die in der theoretischen Informatik Aufmerksamkeit findet.)
- Verschlüsseln ist Potenzieren mit e in $\mathbb{Z}_n$; *Ent*schlüsseln kann somit als e-tes Wurzelziehen in $\mathbb{Z}_n$ aufgefasst werden. Es ist (nach heutigem Kenntnisstand) hart, in $\mathbb{Z}_n$, für n groß, e-te Wurzeln zu ziehen (in $\mathbb{R}$ und somit in $\mathbb{Z}$ ist das Wurzelziehen viel leichter als in $\mathbb{Z}_n$: mit Newton-Verfahren).

Bemerkungen zur Umsetzung in die Praxis:

- Es werden werden Zahlen n der Größenordnung (mindestens) $n \approx 2^{2048} \approx 10^{617}$ verwendet (Empfehlung der Bundesnetzagentur, bis zum Jahr 2020 gültig), d.h. $2048/8 = 256$ Ascii-Zeichen pro Wort. Ab 2023 wird sogar $n \approx 2^{3000}$ empfohlen.

In 2005 wurden einige Zahlen $n = pq \approx 10^{200}$ faktorisiert.
In 2007 wurden einige Zahlen $n = pq \approx 2^{1024} = 10^{309}$ faktorisiert.
n sollte so groß gewählt werden, dass das Faktorisieren von n per Superrechner länger dauert als der Zeitrahmen, in dem die Nachricht geheim bleiben soll. Dabei auch künftige Leistungszuwächse von Rechnern berücksichtigen!

- Auch e sollte nicht zu klein gewählt werden (nicht kleiner als in etwa 2^{16}), andernfalls gibt es erfolgversprechende Angriffsverfahren.
  Auch für zu kleine d gibt es Angriffsmöglichkeiten.
- RSA ist auch zum **Signieren** geeignet ("**Authentifizierung**"):
  Jedermann kann ein x an den Code-Besitzer senden und diesen Auffordern, $y := E_{n,d}(x)$ zurückzuschicken. Der Sender des x kann $E_{n,e}(y)$ berechnen und vergleichen, ob $E_{n,e}(y) = x$. Da nur der Besitzer des Codes das Wissen $E_{n,d}$ zur Entschlüsselung hat, hat sich dieser authentifiziert.
- Da die RSA-Verschlüsselung einen gewissen Rechenaufwand beinhaltet, werden für längere Nachrichten heute meist andere, schnellere Verschlüsselungsverfahren eingesetzt. RSA wird dann aber häufig verwendet, um die besonders heikle Versendung von Schlüsseln besonders sicher zu gestalten.