Vorlesung

„Diskrete Mathematik und Numerische Mathematik für Informatiker“

Universität Frankfurt

Sommersemester 2016

Gegenstand der Diskreten Mathematik:
Studium von endlichen Mengen, die im allgemeinen mit einer zusätzlichen Struktur ausgestattet sind.

Gegenstand der Numerischen Mathematik:
Näherungsweise Berechnung mathematischer Größen, die nicht - oder nur mit unverhältnismäßig großem Aufwand - exakt bestimmt werden können.

Literatur:

Peter Hartmann:
Mathematik für Informatiker
Springer Vieweg
6. Auflage 2014

Martin Aigner:
Diskrete Mathematik
Vieweg 1993

Josef Stör:
Numerische Mathematik 1
Springer Verlag 2002

Inhaltsverzeichnis

§1 Elementare Zählprinzipien

Für eine endliche Menge M bezeichne

$$|M| = \#M = card(M)$$

stets die Kardinalität von M, das ist die Anzahl der Elemente von M.

Beispiele:

(i) $|\emptyset| = 0$

(ii) $|\{1, 2, 3\}| = 3$

(iii) $|\{2, 4, 6, 8, 10\}| = 5$

(iv) $|\{p \in \mathbb{N} | 2 \leq p \leq 100, \text{ p ist Primzahl}\}| = 25$

Häufig ist die Kardinalität einer endlichen Menge zu bestimmen, die mit einer zusätzlichen Struktur versehen ist oder aus anderen Mengen konstruiert worden ist.

Anzahl-Untersuchungen beruhen in der Regel auf folgenden Grundtatsachen:

(Z1) Zwischen zwei endlichen Mengen M , N gibt es genau dann eine Bijektion, wenn $|M| = |N|$ ist.

(Z2) Für zwei disjunkte endliche Mengen M, N ist

$$|M \cup N| = |M| + |N|.$$

Dabei heißen M und N disjunkt, wenn $M \cap N = \emptyset$ ist.

Satz 1.1. *Für endliche Mengen A,B gilt:*

$$|A \times B| = |A| \cdot |B|$$

Veranschaulichung für $|A| = 3, |B| = 4, A = \{a_1, a_2, a_3\}, B = \{b_1, b_2, b_3, b_4\}$.

Abbildung 1: Jeder Stern in obigem Rechteck, das genau $|A| \cdot |B|$ Felder hat, entspricht einem Paar $(a, b) \in A \times B$.

Satz 1.1 läßt sich - durch vollständige Induktion nach k - leicht verallgemeinern:

Satz 1.2. *Für endliche Mengen $M_1, \ldots, M_k$ gilt:*

$$|M_1 \times \cdots \times M_k| = |M_1| \cdots |M_k|.$$

Insbesondere gilt im Falle $M = M_1 = \cdots = M_k$:

$$|M^k| = \underbrace{|M \times \cdots \times M|}_{k-mal} = |M|^k.$$

Bemerkung 1.3. *Für endliche Mengen M, N setzen wir*

$$M^N := \{f : N \to M \mid f \text{ ist eine Abbildung}\}.$$

Dann ist $|M^N| = |M|^{|N|}$.

Beispiel:
Sei etwa $N = \{1, 2, 3\}$. Wir identifizieren jedes $(m_1, m_2, m_3) \in M^3$ mit der Abbildung $f : \{1, 2, 3\} \to M$, definiert durch

$$f(1) = m_1, \quad f(2) = m_2, \quad f(3) = m_3.$$

Wir schreiben dann

$$(m_1, m_2, m_3) \;\hat{=}\; f, \; M^3 \;\hat{=}\; M^{\{1,2,3\}}.$$

Konventionen 1.4. *Für eine Menge M setzen wir*

$$\mathcal{P}(M) \coloneqq \{A | A \subseteq M\}.$$

$\mathcal{P}(M)$ heißt die Potenzmenge von M.
Für $k \leq |M|$ setze

$$\mathcal{P}_k(M) \coloneqq \binom{M}{k} \coloneqq \{A \in \mathcal{P}(M) | \#A = k\}.$$

Satz 1.5. *Eine $n-$elementige Menge M besitzt genau 2^n Teilmengen.*

Beweis. Die Abbildung $\alpha : \{0,1\}^M \to \mathcal{P}(M)$, definiert durch

$$\alpha(f) \coloneqq \{m \in M | f(m) = 1\}$$

ist eine Bijektion. Damit liefern (Z1) und Satz 1.2:

$$|\mathcal{P}(M)| = |\{0,1\}^M| = 2^n.$$

□

Konventionen 1.6. *Es seien $M_1, \ldots, M_k$ beliebige Mengen, die paarweise disjunkt sind, es ist also $M_i \cap M_j = \emptyset$ für $i \neq j$.*
Für die Vereinigung $M_1 \cup \cdots \cup M_k$ dieser Mengen schreiben wir dann auch $M_1 \dot\cup \ldots \dot\cup M_k$.
$M_1 \dot\cup \ldots \dot\cup M_k$ heißt auch die disjunkte Vereinigung der Mengen $M_1, \ldots, M_k$.

Satz 1.7. *Es seien M, N endliche Mengen mit $1 \leq m \coloneqq |M| \leq n \coloneqq |N|$. Dann gibt es genau $n \cdot (n-1) \cdots (n-m+1)$ injektive Abbildungen von M in N.*

Beweis. Sei ohne Einschränkung $N = \{1, \ldots, n\}$. Wir führen Induktion nach m.
Im Falle $m = 1$ gibt es genau n injektive Abbildungen von M in N wie behauptet.
Sei nun $m \geq 2$, und sei $a_0 \in M$ beliebig. Für $1 \leq k \leq n$ setzen wir

$$I_k \coloneqq \{f : M \to N | f \text{ ist injektiv und } f(a_0) = k\}.$$

Dann ist

$$I \coloneqq I_1 \dot\cup \ldots \dot\cup I_n$$

die Menge der injektiven Abbildungen von M in N. Weiter gilt für $1 \leq k \leq n$:

$$|I_k| = |\{g : M \backslash \{a_0\} \to N \backslash \{k\} | g \text{ ist injektiv}\}|,$$

denn jede injektive Abbildung $g : M \backslash \{a_0\} \to N \backslash \{k\}$ korrespondiert -umkehrbar eindeutig- zu der Abbildung $f \in I_k$ mit

$$f(m) \coloneqq \begin{cases} g(m) & \text{für } m \neq a_0 \\ k & \text{für } m = a_0. \end{cases}$$

Nach Induktionsannahme gibt es $(n-1)\cdot(n-2)\cdots(n-m+1)$ injektive Abbildungen von $M\backslash\{a_0\}$ in $N\backslash\{k\}$. Daher folgt - mittels wiederholter Anwendung von (Z2):

$$|I| = \sum_{k=1}^{n} |I_k| = n\cdot(n-1)\cdot(n-2)\cdots(n-m+1)$$

wie gewünscht. □

Konventionen 1.8. *Es sei $n \in \mathbb{N}$ und $J_n := \{1,\ldots,n\}$. Dann ist S_n die Menge der Permutationen von J_n, das sind die Bijektionen von J_n in sich.*

Satz 1.9. *Für $n \in \mathbb{N}$ ist $|S_n| = n! = n\cdot(n-1)\cdots 2\cdot 1$.*

Beweis. Mit $M = N = J_n$ folgt die Behauptung direkt aus Satz 1.7, da jede injektive Abbildung $\alpha : J_n \to J_n$ automatisch auch bijektiv ist. □

Satz 1.10. *Sei N eine endliche Menge und $0 \le k \le n := |N|$. Dann gilt:*

$$|\mathcal{P}_k(N)| = \left|\binom{N}{k}\right| = \binom{n}{k} = \frac{n!}{k!\cdot(n-k)!}.$$

Beweis. Zu beweisen ist „nur“ die mittlere Gleichung.
Ist $k = 0$, so ist die Formel trivial, denn es ist

$$\left|\binom{N}{0}\right| = |\{\emptyset\}| = 1 = \binom{n}{0}.$$

Sei also $1 \le k \le n$ sowie $M := J_k = \{1,\ldots,k\}$.
Das Bild jeder injektiven Abbildung $f : J_k \to N$ ist eine $k-$elementige Teilmenge A von N, und für jede solche Menge $A \in \mathcal{P}_k(N)$ gibt es - nach Satz 1.9, angewandt auf k statt n- genau $k!$ verschiedene injektive Abbildungen f mit $f(J_k) = A$.
Zusammen mit Satz 1.7 folgt also

$$|\mathcal{P}_k(N)| = \frac{1}{k!}\cdot n\cdot(n-1)\cdots(n-k+1) = \frac{n!}{k!\cdot(n-k)!} = \binom{n}{k}.$$

□

Beispiel:
Beim Zahlenlotto „6 aus 49“ gibt es

$$\binom{49}{6} = \frac{49\cdot 48\cdot 47\cdot 46\cdot 45\cdot 44}{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6} = 13983816$$

Möglichkeiten.

Bemerkung 1.11. *Ist $f : M \to N$ eine Abbildung und $A \subseteq N$, so setzen wir*

$$f^{-1}(A) \coloneqq \{m \in M | f(m) \in A\}.$$

Für $a \in N$ ist also insbesondere

$$f^{-1}(\{a\}) = \{m \in M | f(m) = a\}.$$

Satz 1.12. ***„Das Schubfachprinzip“ :***

Es seien M, N endliche Mengen, setze $m \coloneqq |M|, n \coloneqq |N|$, und sei $k \in \mathbb{N}$ mit $m > n \cdot k$. Ist dann $f : M \to N$ eine Abbildung, so gibt es ein $a \in N$ mit $|f^{-1}(\{a\})| > k$.

Interpretation:
Ist $m > n \cdot k$, und verteilt man m Elemente auf n Fächer, so gibt es ein Fach, in das mindestens $k + 1$ Elemente gelegt werden.

Beweis von Satz 1.12: Wir nehmen an $\forall a \in N : \quad |f^{-1}(\{a\})| \leq k$.
Dann folgt:

$$m = |M| = |\underbrace{\dot{\bigcup}}_{a \in N} f^{-1}(\{a\})| = \sum_{a \in N} |f^{-1}(\{a\})| \leq n \cdot k.$$

Das widerspricht aber der Voraussetzung $m > n \cdot k$. □

Satz 1.13. *Es seien A und G endliche Mengen, und $\mathcal{R} \subseteq A \times G$ sei eine Relation. Für $a \in A$ und $g \in G$ setzen wir*

$$r(a) \coloneqq |\{g \in G | a \overset{\mathcal{R}}{\sim} g\}|, s(g) \coloneqq |\{a \in A | a \overset{\mathcal{R}}{\sim} g\}|.$$

Dann folgt:

i) Es ist

$$(\star) \qquad \sum_{a \in A} r(a) = \sum_{g \in G} s(g).$$

ii) Es gebe natürliche Zahlen k und l, sodass für alle $a \in A$ und alle $g \in G$ gilt:

$$r(a) = k \quad , \quad s(g) = l.$$

Dann folgt:

$$(\star\star) \qquad |A| \cdot k = |G| \cdot l.$$

Beweis. i) Beide Seiten in $(\star)$ stimmen überein mit

$$|\mathcal{R}| = |\{(a, g) \in A \times G | a \overset{\mathcal{R}}{\sim} g\}|.$$

ii) ist Spezialfall von i).

□

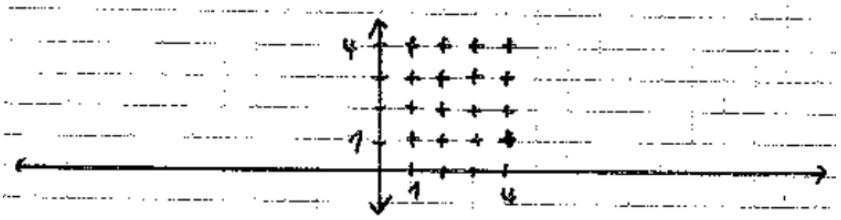
Beispiel 1.14:
Seien $q, n \in \mathbb{N}$, sei $A := \{1, 2, \ldots, q\}^n \subseteq \mathbb{R}^n$, und sei G die Menge aller achsenparallelen-affinen-Geraden in $\mathbb{R}^n$, die A schneiden.
Wir setzen $\mathcal{R} := \{(a, g) \in A \times G | a \in g\}$.
Mit den Bezeichnungen in Satz 1.13 gilt dann:

$$r(a) = n \text{ für alle } a \in A, s(g) = q \text{ für alle } g \in G.$$

Satz 1.13 ii) liefert daher wegen $|A| = q^n$:

$$|G| = \frac{|A| \cdot n}{q} = n \cdot q^{n-1}$$

Abbildung 2: Skizze für q=4 , n=2



Definition 1.15. *Sei E eine endliche Menge, und $w : \mathcal{P}(E) \to \mathbb{R}^+ \dot{\cup} \{0\}$ sei eine Funktion. Für $e \in E$ schreiben wir*

$$w(e) := w(\{e\}).$$

i) w heißt eine Gewichtsfunktion, wenn für alle $A \subseteq E$ gilt:

$$w(A) = \sum_{e \in A} w(e).$$

ii) Ist w eine Gewichtsfunktion mit $w(E) = 1$, so heißt das Paar (E, w) ein endlicher (oder diskreter) Wahrscheinlichkeitsraum.

Bemerkung:

i) Definition 1.15 i) besagt:
Für eine Gewichtsfunktion stimmt das Gewicht einer Menge überein mit der Summe der Gewichte der Elemente dieser Menge.

ii) Für jede Gewichtsfunktion $w : \mathcal{P}(E) \to \mathbb{R}^+ \dot{\cup} \{0\}$ ist $w(\emptyset) = 0$.

Beispiele:

i) $w : \mathcal{P}(E) \to \mathbb{R}^+ \dot{\bigcup} \{0\}$, definiert durch $w(A) := |A|$ ist eine Gewichtsfunktion.

ii) Definiere $w : \mathcal{P}(E) \to \mathbb{R}^+ \bigcup \{0\}$ durch $w(A) := \dfrac{|A|}{|E|}$. Dann ist (E, w) ein Wahrscheinlichkeitsraum. w wird auch Gleichverteilung auf E genannt.
$w(A) = \dfrac{|A|}{|E|}$ ist die Anzahl der für A „günstigen“ Fälle, geteilt durch die Anzahl der „möglichen“ Fälle.
Beim Werfen eines „idealen“ Würfels können wir speziell $E = \{1, 2, 3, 4, 5, 6\}$ und $w(A) = \dfrac{|A|}{6}$ für $A \subseteq E$ annehmen.

Satz 1.16. *Sei E eine endliche Menge, und $w : \mathcal{P}(E) \to \mathbb{R}^+ \bigcup \{0\}$ sei eine Gewichtsfunktion. Dann gilt:*

i) Sind $I_1, \ldots, I_n$ paarweise disjunkte Teilmengen von E, so folgt für $I := \dot{\bigcup}_{1 \le j \le n} I_j$:

$$w(I) = \sum_{j=1}^{n} w(I_j).$$

ii) Für alle $A, B \subseteq E$ gilt:

$$w(A) + w(B) = w(A \cup B) + w(A \cap B).$$

Beweis. i) Laut Definition der Gewichtsfunktion folgt:

$$\sum_{j=1}^{n} w(I_j) = \sum_{j=1}^{n} \sum_{e \in I_j} w(e) = \sum_{e \in I} w(e) = w(I).$$

ii) $A \backslash B, B \backslash A$ und $A \cap B$ sind paarweise disjunkte Teilmengen von $A \cup B$ mit

$$(A \backslash B) \,\dot\cup\, (B \backslash A) \,\dot\cup\, (A \cap B) = A \cup B.$$

i) liefert daher:

$$\begin{aligned} & w(A \cup B) + w(A \cap B) \\ = {} & w(A \backslash B) + w(B \backslash A) + w(A \cap B) + w(A \cap B) \\ = {} & (w(A \backslash B) + w(A \cap B)) + w((B \backslash A) + w(A \cap B)) \\ = {} & w(A) + w(B). \end{aligned}$$

□

§2 Elementare Zahlentheorie

Definition 2.1. *Es seien $a, b \in \mathbb{Z}$ mit $a \neq 0$. Die Zahl a heißt ein Teiler von b , falls $\frac{b}{a} \in \mathbb{Z}$ ist.*
In dem Fall sagen wir auch:

b ist ein Vielfaches von a	*oder*
b ist durch a teilbar	*oder*
a teilt b.	

Wir schreiben dann auch: $a \mid b$.
Ist a kein Teiler von b, so schreiben wir: $a \nmid b$.

Beispiele:

3|6 ; 7|49 ; -13|39 .

Bemerkung 2.2. *i) Für alle $a \in \mathbb{Z}\backslash\{0\}$ gilt:*

$$a \mid a \ , \ -a \mid a \ , \ 1 \mid a \ , \ -1 \mid a \ , \ a \mid 0.$$

ii) Sind $a, b \in \mathbb{Z}\backslash\{0\}$ und $c \in \mathbb{Z}$ mit $a \mid b$ und $b \mid c$, so folgt auch: $a \mid c$.

iii) Sind $a_1, a_2 \in \mathbb{Z}\backslash\{0\}$ und $b_1, b_2 \in \mathbb{Z}$ mit $a_1 \mid b_1$ und $a_2 \mid b_2$, so folgt auch: $(a_1 \cdot a_2) \mid (b_1 \cdot b_2)$.

iv) Sind $a, b \in \mathbb{Z}\backslash\{0\}$ mit $a \mid b$ und $b \mid a$, so ist $a = b$ oder $a = -b$.

v) Jede Zahl $b \in \mathbb{Z}\backslash\{0\}$ hat nur endlich viele Teiler; diese liegen alle im Intervall $[-b, b]$.

vi) Sind $b, c \in \mathbb{Z}$ und $a \in \mathbb{Z}\backslash\{0\}$ mit $a \mid b$ und $a \mid c$, so folgt für alle $m, n \in \mathbb{Z}$:

$$a \mid (m \cdot b + n \cdot c).$$

Definition 2.3. *Sei M eine nichtleere Menge und $\mathcal{R}$ eine Relation auf M. $\mathcal{R}$ heißt eine Äquivalenzrelation, falls gilt:*

(I) $\mathcal{R}$ ist reflexiv ; das heißt, für alle $x \in M$ gilt: $x \mathcal{R} x$.

(II) $\mathcal{R}$ ist symmetrisch ; das heißt:
Sind $x, y \in M$ mit $x \mathcal{R} y$, so gilt auch $y \mathcal{R} x$.

(III) $\mathcal{R}$ ist transitiv ; das heißt:
Sind $x, y, z \in M$ mit $x \mathcal{R} y$ und $y \mathcal{R} z$, so gilt auch: $x \mathcal{R} z$.

Ist $\mathcal{R}$ eine Äquivalenzrelation auf M, so heißen $x, y \in M$ äquivalent, falls $x \mathcal{R} y$ gilt. Auf diese Weise wird M in lauter Äquivalenzklassen eingeteilt.

Beispiele:

i) Die Gleichheitsrelation auf einer nichtleeren Menge M ist eine Äquivalenzrelation ; die ist gegeben durch
$$\mathcal{R} := \{(x, y) \in M^2 \mid x = y\}.$$
Jede Äquivalenzklasse enthält nur ein Element.

ii) Die volle Menge M^2 ist ebenfalls eine Äquivalenzrelation ; das heißt, für alle $x, y \in M$ gilt:
$$x \mathcal{R} y.$$
In diesem Beispiel gibt es nur eine Äquivalenzklasse.

iii) Sei $f : M \to N$ eine Abbildung, und setze
$$\mathcal{R} := \{(x, y) \in M^2 \mid f(x) = f(y)\}.$$
Dann ist $\mathcal{R}$ eine Äquivalenzrelation. Zwei Elemente $x, y \in M$ liegen genau dann in der gleichen Äquivalenzklasse, wenn ihre Bilder unter f übereinstimmen.

Definition 2.4. *Sei $m \in \mathbb{N}$. Zwei Zahlen $a, b \in \mathbb{Z}$ heißen kongruent modulo m , falls $b - a$ durch m teilbar ist, falls also gilt: $m \mid (b - a)$.*
Wir schreiben dann auch: $a \equiv b \mod m$.

Beispiel:

$$3 \equiv 7 \mod 2 \ , \ 2 \equiv 5 \mod 3 \ , \ -3 \equiv 7 \mod 10.$$

Satz 2.5. *Ist $m \in \mathbb{N}$ fest, so ist die Relation „$\equiv$“ eine Äquivalenzrelation auf $\mathbb{Z}$.*

Nachweis der Transitivität. Seien $a, b, c \in \mathbb{Z}$ mit $a \equiv b \mod m$ und $b \equiv c \mod m$. Dann ist sowohl $b - a$ als auch $c - b$ ein Vielfaches von m. Folglich ist auch $c - a = (c - b) + (b - a)$ ein Vielfaches von m; das heißt: $a \equiv c \mod m$. □

Satz 2.6. *Sei $m \in \mathbb{N}$, und seien $a_1, a_2, b_1, b_2 \in \mathbb{Z}$ mit*
$$a_1 \equiv b_1 \mod m \ , \ a_2 \equiv b_2 \mod m.$$
Dann gilt auch:

i) $a_1 + a_2 \equiv b_1 + b_2 \mod m$,

ii) $a_1 - a_2 \equiv b_1 - b_2 \mod m$,

iii) $a_1 \cdot a_2 \equiv b_1 \cdot b_2 \mod m$.

Kongruenzen können also addiert, subtrahiert und multipliziert werden.

Beweis von iii). Wir erhalten:

$$b_1 \cdot b_2 - a_1 \cdot a_2 = b_1 \cdot (b_2 - a_2) + (b_1 - a_1) \cdot a_2.$$

Die rechte Seite ist durch m teilbar, weil $b_2 - a_2$ und $b_1 - a_1$ durch m teilbar sind. □

Satz 2.7. ***Die Division mit Rest:***
Seien $m \in \mathbb{N}$ und $a \in \mathbb{Z}$. Dann gibt es genau eine Zahl $q \in \mathbb{Z}$ und genau eine Zahl $r \in \mathbb{Z}$ mit:

$$a = q \cdot m + r \quad , \quad 0 \leq r \leq m - 1.$$

Das bedeutet:
a kann auf eindeutige Weise durch m mit einem Rest r, $0 \leq r \leq m-1$, dividiert werden.

Beweis.
Nachweis der Eindeutigkeit:
Seien $q_1, q_2, r_1, r_2 \in \mathbb{Z}$ mit $0 \leq r_1, r_2 \leq m - 1$ und

$$a = q_1 \cdot m + r_1 = q_2 \cdot m + r_2.$$

Dann folgt einerseits $|r_1 - r_2| < m$ und andererseits $r_1 - r_2 = m \cdot (q_2 - q_1)$,
also $r_1 \equiv r_2 \mod m$. Das ist nur möglich, wenn $r_1 = r_2$ ist. Wegen $m \neq 0$ folgt dann auch: $q_1 = q_2$.

Nachweis der Existenz:
Sei $q \in \mathbb{Z}$ die größte Zahl mit $q \cdot m \leq a$, und setze $r := a - q \cdot m$.
Dann ist $r \geq 0$. Laut Wahl von q ist weiter $(q + 1) \cdot m > a$, also $r = a - q \cdot m < m$ und folglich $r \leq m - 1$. □

Definition 2.8. *Für $n, m \in \mathbb{Z}$ mit $(n, m) \neq (0, 0)$ bezeichnet $\mathrm{ggT}(n, m)$ den größten gemeinsamen Teiler von n und m, also die größte Zahl $t \in \mathbb{N}$ mit $t|n$ und $t|m$. Ist $\mathrm{ggT}(n, m) = 1$, so heißen n und m teilerfremd.*

Bemerkung 2.9. *Sind $m, n, q \in \mathbb{Z}$ mit $(n, m) \neq (0, 0)$, so ist eine Zahl $t \in \mathbb{N}$ genau dann ein gemeinsamer Teiler von n und m, wenn sie ein gemeinsamer Teiler von m und $q \cdot m + n$ ist.*
Man beachte dazu: $n = -q \cdot m + (q \cdot m + n)$.
Insbesondere gilt:

$$\mathrm{ggT}(n, m) = \mathrm{ggT}(q \cdot m + n, m).$$

Satz 2.10. ***Der Euklidische Algorithmus zur Berechnung des*** ggT
Es seien $a_0, a_1 \in \mathbb{N}$ mit $a_1 \leq a_0$. Dann kann $d \coloneqq \mathrm{ggT}(a_0, a_1)$ nach dem folgenden Euklidischen Algorithmus berechnet werden:

Schritt 0:
Falls $a_1 | a_0$, setze $d \coloneqq a_1$.
Stopp!

Schritt 1:
Bestimme $q_1 \in \mathbb{Z}$ und $a_2 \in \mathbb{N}$ mit
(EA1) $\quad a_0 = q_1 \cdot a_1 + a_2 \quad , \quad 0 < a_2 \leq a_1 - 1.$
Falls $a_2 | a_1$, setze $d \coloneqq a_2$.
Stopp!

Schritt i, $i > 1$:
Wähle $q_i \in \mathbb{Z}$ und $a_{i+1} \in \mathbb{N}$ mit
(EAi) $\quad a_{i-1} = q_i \cdot a_i + a_{i+1} \quad , \quad 0 < a_{i+1} \leq a_i - 1.$
Falls $a_{i+1} | a_i$, setze $d \coloneqq a_{i+1}$.
Stopp!

Beweis. Dass der Algorithmus abbricht, folgt sofort aus der Tatsache, dass die natürlichen Zahlen $a_0, a_1, a_2, \ldots$ immer echt kleiner werden. Weiter liefert Bemerkung 2.9 für alle möglichen i:

$$\mathrm{ggT}(a_{i-1}, a_i) = \mathrm{ggT}(q_i \cdot a_i + a_{i+1}, a_i) = \mathrm{ggT}(a_{i+1}, a_i).$$

Bricht der Algorithmus nach j Schritten ab, so gilt $a_{j+1} | a_j$, und für $d \coloneqq a_{j+1}$ folgt induktiv:

$$d = \mathrm{ggT}(a_j, a_{j+1}) = \mathrm{ggT}(a_{j-1}, a_j) = \cdots = \mathrm{ggT}(a_0, a_1).$$

□

Weiter gilt

Satz 2.11. *Sind $a_0, a_1 \in \mathbb{N}$, so folgt für $d \coloneqq \mathrm{ggT}(a_0, a_1)$:*
Es gibt $b, c \in \mathbb{Z}$ mit

$$d = b \cdot a_0 + c \cdot a_1.$$

Der Beweis folgt induktiv direkt aus den obigen Formeln (EAi).

Beispiel:
Zu berechnen ist $\mathrm{ggT}(182, 325)$.

Wiederholte Division mit Rest liefert:

$$\begin{aligned} 325 &= 1 \cdot 182 + 143 \ , \\ 182 &= 1 \cdot 143 + 39 \ , \\ 143 &= 3 \cdot 39 + 26 \ , \\ 39 &= 1 \cdot 26 + 13 \ , \\ 26 &= 2 \cdot 13. \end{aligned}$$

Damit folgt: $13 = \mathrm{ggT}(182, 325)$.
Lesen wir die letzten Gleichungen von unten nach oben, so erhalten wir:

$$\begin{aligned} 13 = 39 - 1 \cdot 26 = 39 - (143 - 3 \cdot 39) &= 4 \cdot 39 - 143 \\ = 4 \cdot (182 - 143) - 143 &= 4 \cdot 182 - 5 \cdot 143 \\ = 4 \cdot 182 - 5 \cdot (325 - 182) & \\ = 9 \cdot 182 - 5 \cdot 325. & \end{aligned}$$

Definition 2.12. *Eine natürliche Zahl $p \in \mathbb{N}$ mit $p \geq 2$ heißt eine Primzahl, falls kein $k \in \mathbb{N}$ mit $1 < k < p$ und $k|p$ existiert. $\mathbb{P}$ bezeichne die Menge der Primzahlen.*

Bemerkung:
Die kleinsten Primzahlen sind:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29.$$

Satz 2.13. *Es seien $n, m \in \mathbb{N}$, und es sei p eine Primzahl mit $p|(n \cdot m)$. Dann gilt $p|n$ oder $p|m$.*

Beweis. Wir nehmen an, es gelte $p \nmid n$ und $p \nmid m$.
Wegen $p \in \mathbb{P}$ gilt dann: $1 = \mathrm{ggT}(p, n) = \mathrm{ggT}(p, m)$.
Nach zweimaliger Anwendung von Satz 2.11 existieren $a, b, c, d \in \mathbb{Z}$ mit:

$$a \cdot p + b \cdot n = 1 \quad , \quad c \cdot p + d \cdot m = 1.$$

Laut Voraussetzung ist weiter $k := \dfrac{n \cdot m}{p} \in \mathbb{N}$.
Damit folgt:

$$\begin{aligned} 1 & \\ = (a \cdot p + b \cdot n) \cdot (c \cdot p + d \cdot m) & \\ = p \cdot (a \cdot c \cdot p + a \cdot d \cdot m + b \cdot c \cdot n + b \cdot d \cdot k). & \end{aligned}$$

Das ist ein Widerspruch, denn 1 ist kein Vielfaches von p. □

Satz 2.14. ***Fundamentalsatz der Elementaren Zahlentheorie:***
Zu jeder natürlichen Zahl $n \geq 2$ gibt es Primzahlen $p_1, \ldots, p_s$ mit $p_1 \leq \cdots \leq p_s$ und

(Z) $$n = p_1 \cdots p_s = \textstyle\prod_{i=1}^{s} p_i.$$

Sind auch $q_1, \ldots, q_t$ Primzahlen mit $q_1 \leq \cdots \leq q_t$ und

(Z2) $$n = q_1 \cdots q_t = \textstyle\prod_{j=1}^{t} q_j,$$

so ist $s = t$ und $p_i = q_i$ für $1 \leq i \leq s$.

Beweis. Wir führen Induktion nach n.
Für $n = 2$ und $n = 3$ ist nichts zu zeigen, weil 2 und 3 selbst Primzahlen sind.
Sei nun $n \geq 4$, und sei p_1 der kleinste Primteiler von n; das ist die kleinste Primzahl mit $p_1 | n$.

Nachweis der Existenz:
Nach Induktionsannahme ist

$$\frac{n}{p_1} = p_2 \cdots p_s$$

für gewisse Primzahlen $p_2, \ldots p_s$ mit $p_2 \leq \cdots \leq p_s$.
Damit folgt auch (Z).

Nachweis der Eindeutigkeit:
Gelten (Z) und $(Z2)$, so liefert Satz 2.13 wegen $p_1 \in \mathbb{P}:\ \ p_1 | q_i$ für ein i mit $1 \leq i \leq t$. Wegen $q_i \in \mathbb{P}$ ist das nur möglich, wenn $p_1 = q_i$ ist. Weil p_1 der kleinste Primteiler von n ist, folgt: $p_1 = q_1$. (Z) und $(Z2)$ liefern also:

$$\frac{n}{p_1} = p_2 \cdots p_s = q_2 \cdots q_t.$$

Anwendung der Induktionsannahme auf die Zahl $\dfrac{n}{p_1}$ liefert die Behauptung. □

Bemerkung 2.15. *Satz 2.14 besagt auch:*
Jede natürliche Zahl $n \geq 2$ hat eine eindeutig bestimmte Primfaktorzerlegung der Gestalt

$$n = \prod_{i=1}^{k} p_i^{\alpha_i}$$

mit $p_1, \ldots, p_k \in \mathbb{P}\ ,\ p_1 < \cdots < p_k\ ,\ \alpha_1, \ldots, \alpha_k \in \mathbb{N}$.

Bemerkung 2.16. *Gegeben seien $n, m \in \mathbb{N}$ mit*

$$n = \prod_{i=1}^{k} p_i^{\alpha_i} \quad , \quad m = \prod_{i=1}^{k} p_i^{\beta_i},$$

wobei gelte:
$p_1, \ldots, p_k \in \mathbb{P}\ ,\ p_1 < \cdots < p_k$;
$\alpha_i, \beta_i \geq 0\ ,\ \max(\alpha_i, \beta_i) > 0$ *für* $1 \leq i \leq k$.

i) Eine Zahl $t \in \mathbb{N}$ ist genau dann ein gemeinsamer Teiler von n und m, wenn t die Gestalt

$$t = \prod_{i=1}^{k} p_i^{\gamma_i}$$

hat mit $0 \leq \gamma_i \leq \min(\alpha_i, \beta_i)$ für $1 \leq i \leq k$.
Insbesondere ist

$$\mathrm{ggT}(n, m) = \prod_{i=1}^{k} p_i^{\min(\alpha_i, \beta_i)} \quad ,$$

und jeder gemeinsame Teiler von n und m ist auch ein Teiler von $\mathrm{ggT}(n, m)$.

ii) Das kleinste gemeinsame Vielfache von n und m ist die Zahl

$$kgV(n,m) = \prod_{i=1}^{k} p_i^{\max(\alpha_i,\beta_i)}.$$

Jedes andere gemeinsame Vielfache von n und m ist nach Satz 2.14 auch ein Vielfaches von $kgV(n,m)$.

iii) Aus i) und ii) folgt:

$$\mathrm{ggT}(n,m) \cdot kgV(n,m) = \prod_{i=1}^{k} p_i^{\alpha_i+\beta_i} = n \cdot m.$$

Insbesondere gilt folgende Äquivalenz:

$$\mathrm{ggT}(n,m) = 1 \Leftrightarrow kgV(n,m) = n \cdot m.$$

Sind n und m teilerfremd, so sind die gemeinsamen Vielfachen von n und m also genau die Vielfachen von $n \cdot m$.

Satz 2.17. ***Der Chinesische Restsatz für 2 simultane Kongruenzen:***
Die natürlichen Zahlen m_1, m_2 seien teilerfremd, und setze $m \coloneqq m_1 \cdot m_2$. Weiterhin seien $a_1, a_2 \in \mathbb{Z}$ gegeben. Dann gibt es genau eine Zahl $a \in \mathbb{Z}$ mit $0 \leq a \leq m-1$, die die beiden folgenden Kongruenzen löst:

(CR0)

$$a \equiv a_i \mod m_i \quad \text{für } i \in \{1,2\}.$$

a kann wie folgt ermittelt werden:
Wähle $c_1, c_2 \in \mathbb{Z}$ mit

(CR1)

$$c_1 \cdot m_1 + c_2 \cdot m_2 = 1$$

und setze

(CR2)

$$q \coloneqq c_1 \cdot m_1 \cdot a_2 + c_2 \cdot m_2 \cdot a_1.$$

Dann ist a die eindeutig bestimmte ganze Zahl mit

(CR3)

$$a \equiv q \mod m \quad \text{und} \quad 0 \leq a \leq m-1.$$

Beweis. *Nachweis der Eindeutigkeit:*
Seien $a, a' \in \mathbb{Z}$ mit $0 \leq a, a' \leq m-1$ und

$$a \equiv a' \equiv a_i \mod m_i \quad \text{für } i \in \{1,2\}.$$

Dann ist $a - a'$ ein gemeinsames Vielfaches von m_1 und m_2 - und damit nach Bemerkung 2.16 *iii*) auch von $m = m_1 \cdot m_2$.
Aus $a \equiv a' \mod m$ und $0 \leq a, a' \leq m-1$ folgt nun $a = a'$.

Nachweis der Existenz:
Nach Satz 2.11 gibt es $c_1, c_2 \in \mathbb{Z}$, die (*CR1*) erfüllen. Für q, a wie in (*CR2*) bzw. (*CR3*) folgt dann

$$a \equiv q \equiv c_2 \cdot m_2 \cdot a_1 \equiv (1 - c_1 \cdot m_1) \cdot a_1 \equiv a_1 \mod m_1$$

und analog $a \equiv a_2 \mod m_2$. □

Bemerkung 2.18. *Ein ähnliches Ergebnis läßt sich für $v, v \geq 2$ beliebig, paarweise teilerfremde natürliche Zahlen $m_1, \ldots m_v$- etwa durch Induktion - beweisen:*
Ist $m \coloneqq m_1 \cdots m_v$, und sind $a_1, \ldots, a_v \in \mathbb{Z}$ gegeben, so gibt es genau eine Zahl $a \in \mathbb{Z}$ mit $0 \leq a \leq m-1$, die jede der folgenden Kongruenzen löst:

$$a \equiv a_i \mod m_i \quad \text{für } 1 \leq i \leq v.$$

Beispiel:
Bestimme die eindeutig bestimmte Zahl $a \in \mathbb{Z}$ mit $0 \leq a \leq 133 \cdot 92 - 1 = 12235$, die die folgenden Kongruenzen löst:

$$a \equiv 25 \mod 133 \quad , \quad a \equiv 17 \mod 92.$$

Zunächst gilt -etwa nach Anwendung des Euklidischen Algorithmus:

$$9 \cdot 133 - 13 \cdot 92 = 1.$$

Gemäß (*CR2*) setzen wir also

$$q \coloneqq 9 \cdot 133 \cdot 17 - 13 \cdot 92 \cdot 25 = 20349 - 29900 = -9551.$$

Dann ist $a \coloneqq -9551 + 12236 = 2685$ zu setzen.

Satz 2.19. *Sei $m \in \mathbb{N}$, und sei $a \in \mathbb{Z}$ mit $\mathrm{ggT}(a, m) = 1$. Dann gibt es ein $b \in \mathbb{Z}$ mit $a \cdot b \equiv 1 \mod m$.*

Beweis. Nach Satz 2.11 gibt es $b, c \in \mathbb{Z}$ mit $a \cdot b + m \cdot c = 1$. Damit folgt sofort:
$a \cdot b \equiv 1 \mod m$. □

Satz 2.20. ***Der kleine Satz von Fermat für Primzahlen:*** *Sei p eine Primzahl, und sei $a \in \mathbb{Z}$ mit $\mathrm{ggT}(a, p) = 1$. Dann gilt:*

$$a^{p-1} \equiv 1 \mod p.$$

Beweis. Nach Satz 2.19, angewandt auf $m = p$, gibt es ein $b \in \mathbb{Z}$ mit $a \cdot b \equiv 1 \mod p$. Nach Aufgabe 11*iii*) gilt weiter:

$$a^p \equiv a \mod p.$$

Multiplikation dieser Kongruenz mit b liefert:

$$a^{p-1} \equiv (b \cdot a) \cdot a^{p-1} \equiv b \cdot a^p \equiv b \cdot a \equiv 1 \mod p.$$

□

Definition 2.21. *Die Eulersche φ - Funktion $\varphi : \mathbb{N} \to \mathbb{N}$ ist definiert durch*

$$\varphi(n) := \#\{k \in \mathbb{N} \mid 1 \leq k \leq n \text{ , ggT}(k, n) = 1\}.$$

Wertetabelle:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\varphi(n)$	1	1	2	2	4	2	6	4	6	4	10	4	12	6	8	8

Satz 2.22. *Für jede Primzahl p und jedes $n \in \mathbb{N}$ gilt:*

$$\varphi(p^n) = p^n - p^{n-1} = p^{n-1} \cdot (p-1).$$

Insbesondere ist $\varphi(p) = p - 1$.

Beweis. Wir erhalten:

$$\begin{aligned} & \varphi(p^n) \\ = p^n - \#\{k \in \mathbb{N} \mid 1 \leq k \leq p^n \text{ , ggT}(p^n, k) > 1\} \\ = p^n - \#\{k \in \mathbb{N} \mid 1 \leq k \leq p^n \text{ , } p|k\} \\ = p^n - p^{n-1}. \end{aligned}$$

□

Satz 2.23. *Seien p und q zwei verschiedene Primzahlen. Dann gilt:*

$$\varphi(p \cdot q) = (p-1) \cdot (q-1) = \varphi(p) \cdot \varphi(q).$$

Beweis. Nach Satz 2.22 braucht nur die erste Gleichung bewiesen zu werden. - Dazu zählen wir alle Zahlen k mit $1 \leq k < p \cdot q$, die nicht teilerfremd zu $p \cdot q$ sind. Das sind einerseits die Vielfachen

$$p, 2p, \ldots, (q-1) \cdot p \qquad (q - 1 \text{ Stück})$$

von p sowie andererseits die Vielfachen

$$q, 2q, \ldots, (p-1) \cdot q \qquad (p - 1 \text{ Stück})$$

von q. Somit folgt:

$$\begin{aligned} \varphi(p \cdot q) &= (p \cdot q - 1) - ((q-1) + (p-1)) \\ &= p \cdot q - q - p + 1 = (p-1) \cdot (q-1). \end{aligned}$$

□

Allgemeiner kann man - etwa mit Hilfe des Chinesischen Restsatzes - beweisen:

Satz 2.24. *Für je zwei teilerfremde Zahlen $n, m \in \mathbb{N}$ gilt:*

$$\varphi(n \cdot m) = \varphi(n) \cdot \varphi(m).$$

Satz 2.25. *Seien p und q zwei verschiedene Primzahlen. Dann gilt für alle $m, k \in \mathbb{N}$:*

$$m^{k\cdot(p-1)\cdot(q-1)+1} \equiv m \mod p \cdot q.$$

Beweis. Weil p und q teilerfremd sind, reicht es aus Symmetriegründen, zu zeigen:

$$m^{k\cdot(p-1)\cdot(q-1)+1} \equiv m \mod p.$$

Dies ist trivial im Falle $p|m$.
Andernfalls ist $\mathrm{ggT}(m, p) = 1$, und dann liefert Satz 2.20:

$$\begin{aligned} m^{p-1} &\equiv 1 \mod p \\ \Rightarrow (m^{(p-1)})^{k\cdot(q-1)} &\equiv 1 \mod p \\ \Rightarrow m^{k\cdot(p-1)\cdot(q-1)+1} &\equiv m \mod p. \end{aligned}$$

□

Satz 2.25 wird im folgenden Abschnitt „Kryptologie“ wichtig; dort werden Verschlüsselungsvorschriften vorgestellt, die „leicht“ durchzuführen sind, wogegen die inverse Entschlüsselung - ohne zusätzliche Kenntnisse - „schwer“ ist.

§3 Kryptologie

Ziel der Kryptologie:
Konstruktion von Verschlüsselungsvorschriften, die „leicht“ durchzuführen sind, wobei aber die inverse Entschlüsselung - ohne weitere Kenntnisse, die nur der Empfänger hat- „schwer“ sein soll.

Beispiel einer Verschlüsselungsfunktion:
Für festes t mit $0 \leq t \leq 25$ wird jeder Buchstabe des gewöhnlichen Alphabets um t Stellen nach rechts (bzw. $26 - t$ Stellen nach links) verschoben.
Für $t = 4$ ergibt sich etwa:

$$\begin{array}{ccc} Klartext & & Geheimtext \\ & \rightarrow & \\ ZAUN & & DEYR \end{array}$$

Annahme:
Der Unbefugte weiß, dass nach diesem Algorithmus verschlüsselt wird, aber nur der Empfänger kennt von vornherein den Wert t.

Im allgemeinen ist es aber für den Unbefugten leicht, t zu ermitteln:
Es gibt in der Regel nur eine Möglichkeit, ein sinnvolles deutsches Wort (oder einen deutschen Text) zu erhalten.

Konventionen 3.1. *Gegeben seien zwei Mengen $\mathcal{A}$ und $\mathcal{B}$ mit mindestens 2, aber höchstens endlich vielen Elementen, genannt Alphabete.*
Zum Beispiel kann

$$\mathcal{A} = \{A, B, C, \ldots, Z\}$$

das gewöhnliche Alphabet in Block-Buchstaben sein und

$$\mathcal{B} = \mathcal{A} \quad oder \quad \mathcal{B} = \{k \in \mathbb{Z} \mid 0 \leq k \leq 9\}.$$

Die Elemente von $\mathcal{A}$ und $\mathcal{B}$ heißen Buchstaben.
Wir setzen

$$(\star) \qquad \mathcal{A}^\star := \{(A_1, \ldots A_n) | n \geq 1, A_i \in \mathcal{A} \text{ für } 1 \leq i \leq n\} = \dot{\bigcup_{n \geq 1}} \mathcal{A}^n$$

und definieren $\mathcal{B}^\star$ entsprechend.
$\mathcal{A}^\star$ ist also das System aller endlichen Folgen aus $\mathcal{A}$. Die Elemente aus $\mathcal{A}^\star$ bzw. $\mathcal{B}^\star$ heißen Texte über dem Alphabet $\mathcal{A}$ bzw. $\mathcal{B}$.

Definition 3.2. *i) Eine Chiffrierung oder Verschlüsselungsfunktion über $\mathcal{A}$ mit Werten in $\mathcal{B}^\star$ ist eine injektive Abbildung $f : \mathcal{A}^\star \to \mathcal{B}^\star$.*
Für einen Text $(A_1, \ldots A_n) \in \mathcal{A}^\star$ heißt $f(A_1, \ldots, A_n) \in \mathcal{B}^\star$ der verschlüsselte - oder chiffrierte - Text.

ii) Eine Chiffrierung $f : \mathcal{A}^\star \to \mathcal{B}^\star$ heißt eine Stromchiffrierung, wenn sie zeichenweise durchgeführt wird; das heißt, für alle $(A_1, \ldots, A_n) \in \mathcal{A}^\star$ gilt:

$$f(A_1, \ldots, A_n) = (f(A_1) \ldots f(A_n)).$$

Bemerkung 3.3. *Manchmal ist die injektive Abbildung f nur auf einer Teilmenge $\mathcal{M}$ von $\mathcal{A}^\star$ definiert, die eine Sprache über dem Alphabet $\mathcal{A}$ bezeichnet. Dabei heißt $\mathcal{M}$ der Klartextraum, und die Elemente von $\mathcal{M}$ heißen Klartexte oder sinnvolle Texte. Die Elemente der Bildmenge $f(\mathcal{M})$ werden Geheimtexte - oder Kryptogramme - genannt. Üblicherweise wird f aber auf ganz $\mathcal{A}^\star$ definiert, auch wenn nur Texte aus $\mathcal{M}$ chiffriert werden sollen.*

Beispiel 3.4:
Sei $\mathcal{A} = \{A, B, C, \ldots, Z\}$ das gewöhnliche Alphabet und

$$\mathcal{B} = \{k \in \mathbb{Z} \mid 0 \leq k \leq 9\}.$$

Auf $\mathcal{A} = \mathcal{A}^1$ wird f durch folgende Tabelle bestimmt:

A	B	C	D	E	F	G	H	I	J	K	L	M
00	01	02	03	04	05	06	07	08	09	10	11	12

N	O	P	Q	R	S	T	U	V	W	X	Y	Z
13	14	15	16	17	18	19	20	21	22	23	24	25

Für $0 \leq i \leq 25$ wird also dem $(i+1)$-ten Buchstaben die - zweistellige - Dezimaldarstellung der Zahl i zugeordnet.
Das heißt insbesondere:
Für $0 \leq i \leq 9$ beginnt die Dezimaldarstellung mit der Ziffer 0.
Auf $\mathcal{A}^\star$ wird f nun so definiert, dass f eine Stromchiffrierung ist.
Für $(A_1, \ldots, A_n) \in \mathcal{A}^n, n \geq 2$, setzen wir also:

$$f(A_1 \ldots A_n) = f(A_1)f(A_2) \ldots f(A_n) \in \mathcal{B}^{2n}.$$

Weil jeder Buchstabe aus $\mathcal{A}$ eineindeutig auf ein Element aus $\mathcal{B}^2$ abgebildet wird, ist f auf ganz $\mathcal{A}^\star$ injektiv.

Warnung:
Wird die Stromchiffrierung f im letzten Beispiel so modifiziert, dass bei den Bildern der Buchstaben $A - J$ jeweils die führende 0 ignoriert wird, so ist f nicht injektiv; denn es folgte:

$$f(BB) = f(B)f(B) = 11 = f(L).$$

Das Element 11 ließe sich also nicht eindeutig entschlüsseln.

Definition 3.5. *Sei $f : \mathcal{A}^\star \to \mathcal{B}^\star$ eine Chiffrierung und $\tilde{\mathcal{B}} \coloneqq f(\mathcal{A}^\star)$ der Bildraum von f. Die zugehörige -bijektive- Umkehrabbildung $f^{-1} : \tilde{\mathcal{B}} \to \mathcal{A}^\star$ heißt Dechiffrierung. Das zugehörige Kryptosystem besteht aus der Chiffrierung f und der inversen Dechiffrierung f^{-1}.*

Skizze:

$$\mathcal{A}^\star \xrightarrow{\quad f \quad} \tilde{\mathcal{B}} \xrightarrow{\quad f^{-1} \quad} \mathcal{A}^\star$$

Definition 3.6. ***Die Skytala - Verschlüsselung:***
Ein Text wird nach folgender Vorschrift chiffriert bzw. dechiffriert:

Schlüssel: $u \in \mathbb{N}$

Chiffrieren: *Schreibe den Klartext zeilenweise in ein Schema mit genau u Zeilen, in dem jede Spalte u oder $u-1$ Buchstaben aufweist. Man erhält den Geheimtext, indem der Text spaltenweise gelesen wird.*

Dechiffrieren: *Schreibe den Geheimtext spaltenweise in ein Schema mit genau u Zeilen. Man erhält den Klartext, indem man zeilenweise liest.*

Bemerkung 3.7. ***Geometrische Interpretation:***
Wickele ein schmales Band spiralförmig um einen Zylinder, und schreibe der Zylinderlänge nach eine Nachricht auf das Band. u ist - in Bezug auf die Anzahl der Buchstaben - der Umfang des Zylinders. Nach Abwickeln des Bandes erhält man den Geheimtext.

Beispiele:

i) Wir chiffrieren - für $u = 3$ - das Wort

STEUERSCHAETZER

und erhalten das Schema

STEUE
RSCHA
ETZER

Spaltenweises Lesen liefert den Geheimtext

SRETSTECZUHEEAR

ii) Gegeben sei der Geheimtext

NSAEUKCRSNK

Zum Entschlüsseln können verschiedene mögliche Umfänge u ausgetestet werden.

Ordnen wir den Text in $u = 4$ Zeilen an, so erhalten wir

$$NUS$$

$$SKN$$

$$ACK$$

$$ER$$

Das gesuchte Wort ist also:

$$NUSSKNACKER$$

Definition 3.8. *Eine bijektive Funktion $f : M_1 \to M_2$ heißt eine Einwegfunktion, wenn für $c \in \mathcal{M}_2$ die Ermittlung von $f^{-1}(c)$ - ohne zusätzliche Kenntnisse - „schwer“ ist.*

Bemerkung 3.9. *Für die heutigen Anwendungen der Kryptologie ist weniger gravierend, dass Definition 3.8 keine exakte mathematische Definition ist.*
Wesentlich ist: Die Berechnung der Bilder unter f ist erheblich einfacher als die Berechnung der Urbilder.
Häufig wird verlangt: Zum jetzigen Zeitpunkt ist kein polynomialer Algorithmus zur Berechnung der Urbilder bekannt.

Definition 3.10. *Sei $\mathcal{A}_0 \subseteq \mathcal{A}^\star$, und sei K eine Menge von Personen. Sei $(E_T)_{T\in K}$ eine Familie von Abbildungen von $\mathcal{A}_0$ nach $\mathcal{B}^\star$, genannt die öffentlichen Schlüsselfunktionen, und $(D_T)_{T\in K}$ sei eine Familie von Abbildungen von $\mathcal{B}^\star$ nach $\mathcal{A}^\star$, genannt die geheimen Schlüssel - Funktionen.*
Das System $(\mathcal{A}_0, \mathcal{B}^\star, (E_T)_{T\in K}, (D_T)_{T\in K})$ heißt ein Public - Key - Verschlüsselungssystem, falls gilt:

(I) Jede Funktion $E_T, T \in K$, ist eine Einwegfunktion.

(II) Für alle $T \in K$ und alle $m \in \mathcal{A}_0$ gilt:

$$D_T(E_T(m)) = m.$$

Bemerkung 3.11. *i) Die öffentlichen Schlüssel-Funktionen - oder kurz öffentlichen Schlüssel - dienen zum Chiffrieren und sind allgemein bekannt, während die geheimen - oder privaten - Schlüssel zum Dechiffrieren dienen und nur dem jeweiligen Besitzer bekannt sind.*

ii) Um einer festen Person $T \in K$ eine geheime Nachricht m zu übermitteln, kann ihr die verschlüsselte Nachricht $E_T(m)$ gesendet werden, die ja - nach (I) und (II) - nur von T selbst wieder „leicht" zu entschlüsseln ist.

iii) Die Rolle der - öffentlichen und geheimen - Schlüssel kann durch folgendes Bild eines Briefkastens illustriert werden:

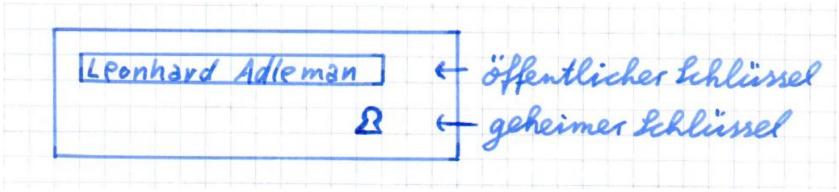


Abbildung 3: Der öffentliche Schlüssel ist das Namensschild mit der Öffnung. Nur der Besitzer kann das Fach mit seinem Schlüssel öffnen.

Der RSA - Algorithmus 3.12,
nach Rivest, Shamir, Adleman (1977):

Schritt I. Jede Person wählt zwei verschiedene - und große - Primzahlen p und q und berechnet $n = p \cdot q$. Ferner berechnet sie $\varphi(n) = (p-1) \cdot (q-1)$. Schließlich wählt sie eine Zahl e mit $1 \leq e < \varphi(n)$, die zu $\varphi(n)$ teilerfremd ist, und berechnet d mit $1 \leq d < \varphi(n)$ und

$$e \cdot d \equiv 1 \mod \varphi(n).$$

Schritt II. Als öffentlichen Schlüssel gibt sie das Paar (e, n) bekannt ; p und q bleiben geheim.
Der private Schlüssel ist d.

Schritt III. Wir setzen

$$\mathcal{C}_n := \{r \in \mathbb{Z} \mid 0 \leq r < n\}.$$

Ein Text T über einem zugrunde liegenden Alphabet wird mittels einer Stromchiffrierung als ein Element aus $\mathcal{C}_n^\star$ - also einer Folge aus $\mathcal{C}_n$ - dargestellt.

Schritt IV. Die Verschlüsselung von Elementen aus $\mathcal{C}_n$ erfolgt nach der Vorschrift $f_e : \mathcal{C}_n \to \mathcal{C}_n$, festgelegt durch

$$f_e(r) := r^e \mod n.$$

Bemerkung 3.13. *i) Die Entschlüsselung* $f_d : \mathcal{C}_n \to \mathcal{C}_n$ *ist festgelegt durch*

$$f_d(r) \equiv r^d \mod n.$$

ii) Wegen $e \cdot d \equiv 1 \mod \varphi(n)$ *gilt:*

$$k := \frac{e \cdot d - 1}{\varphi(n)} = \frac{e \cdot d - 1}{(p-1) \cdot (q-1)} \in \mathbb{N}.$$

Damit liefert Satz 2.25 für alle $m \in \mathcal{C}_n$ *:*

$$f_d(f_e(m)) \equiv (m^e)^d \equiv m^{e \cdot d} \equiv m^{k \cdot (p-1) \cdot (q-1)+1} \equiv m \mod n.$$

Wegen $m, f_d(f_e(m)) \in \mathcal{C}_n$ *bedeutet das:*

$$f_d(f_e(m)) = m.$$

Das heißt: Bedingung (II) *in Definition 3.10 ist erfüllt.*

iii) Für $m \in \mathcal{C}_n$ *ist die Berechnung von* $f_e(m)$ *einfach.*

iv) Die - prinzipiell analoge - Berechnung von $f_d(r)$ *für* $r \in \mathcal{C}_n$ *ist nur einfach, wenn der private Schlüssel* d *bekannt ist.*
d ließe sich - etwa mittels des Euklidischen Algorithmus - aus der Kongruenz $e \cdot d \equiv 1 \mod \varphi(n)$ *berechnen, wenn* $\varphi(n) = (p-1) \cdot (q-1)$ *bekannt wäre.*
Wäre neben n *auch* $\varphi(n)$ *öffentlich bekannt, so ließen sich* p *und* q *aus den Gleichungen*

$$n = p \cdot q \quad , \quad \varphi(n) = (p-1) \cdot (q-1)$$

berechnen.

v) Es ist kein schneller Algorithmus zur Faktorisierung großer Zahlen bekannt. Resümierend können wir sagen, dass auch Bedingung (I) *in Definition 3.10 erfüllt ist.*
Das bedeutet:
Nach unserem jetzigen Kenntnisstand ist das RSA-System ein sicheres Public - Key - Verschlüsselungssystem.

Beispiel 3.14.:
Wir schreiben

$A \hat{=} 01, B \hat{=} 02, \ldots, Z \hat{=} 26, Leerzeichen \hat{=} 00,$
$\mathcal{A} = \{00, 01, 02, \ldots, 26\}.$

Weiter sei

$$\mathcal{B} = \{n_1 n_2 n_3 n_4 \mid 0 \leq n_i \leq 9 \text{ für } 1 \leq i \leq 4\}.$$

Dabei bedeutet, wie üblich, $n_1n_2n_3n_4$ die Zahl

$$1000 \cdot n_1 + 100 \cdot n_2 + 10 \cdot n_3 + n_4.$$

In dem Text

KOMME MORGEN ZURUECK

werden zunächst aufeinanderfolgende Bigramme (das sind Blöcke von zwei Buchstaben) verschlüsselt; damit erhalten wir folgende Folge von 4-Blöcken:

$$\begin{array}{c} 1115|1313|0500|1315|1807| \\ 0514|0026|2118|2105|0311 \end{array} \qquad (\star)$$

Wir wählen nun die Primzahlen $p = 47$, $q = 59$ und erhalten:

$$\begin{aligned} n &= 47 \cdot 59 = 2773, \\ \varphi(n) &= 46 \cdot 58 = 2668\ . \end{aligned}$$

Für $e = 17 = 2^4 + 1$ wird der Text in $(\star)$ nun vermöge $f_{17} \mod 2773$ verschlüsselt, indem jeder 4-Block mit $17 \mod 2773$ potenziert wird:

$$\begin{array}{c} 1379|2395|1655|0422|0482| \\ 1643|1445|0848|0747|2676 \end{array} \qquad (\star\star)$$

Beispielsweise erhalten wir modulo 2773:

$1115^2 \equiv 921$
$1115^4 \equiv 921^2 \equiv 2476,$
$1115^8 \equiv 2476^2 \equiv 2246,$
$1115^{16} \equiv 2246^2 \equiv 429,$
$1115^{17} \equiv 429 \cdot 1115 \equiv 1379.$

Die Verschlüsselung ist injektiv, weil alle möglichen 4-Blöcke -insbesondere die in $(\star)$- unterhalb von 2773 liegen.

Bemerkung 3.15,
siehe auch Beispiel 4.12. vom WS 2015/16:
Für $n \in \mathbb{N}$ mit $n \geq 2$ und $m \in \mathbb{Z}$ ist

$$m + n \cdot \mathbb{Z} \coloneqq \{m + n \cdot k \mid k \in \mathbb{Z}\} = \{a \in \mathbb{Z} \mid a \equiv m \mod n\}.$$

Ferner ist der Restklassenring modulo n die Menge

$$\mathbb{Z}/n \cdot \mathbb{Z} \coloneqq \{m + n \cdot \mathbb{Z} \mid 0 \leq m \leq n - 1\}.$$

Schreiben wir kurz $\overline{m} \coloneqq m + n \cdot \mathbb{Z}$ für $m \in \mathbb{Z}$, so sind die Addition und die Multiplikation in $\mathbb{Z}/n\mathbb{Z}$ gegeben durch:

$$\overline{m} + \overline{l} \coloneqq \overline{m + l} \quad , \quad \overline{m} \cdot \overline{l} \coloneqq \overline{m \cdot l} \quad \text{für } m, l \in \mathbb{Z}.$$

Beim Übergang von $\mathbb{Z}$ zu $\mathbb{Z}/n \cdot \mathbb{Z}$ werden zwei Zahlen, deren Differenz durch n teilbar ist, identifiziert.

Das bedeutet:
Das Rechnen in $\mathbb{Z}/n \cdot \mathbb{Z}$ ist gleichbedeutend mit dem Rechnen modulo n.

Definition 3.16. *Sei p eine ungerade Primzahl, und sei $2 \leq g \leq p - 1$.*

i) Die Funktion $E_g : \{1, \ldots, p-1\} \to \mathbb{Z}/p \cdot \mathbb{Z}$, definiert durch

$$E_g(x) \coloneqq \overline{g}^x \quad \text{mit } \overline{g} \coloneqq g + p \cdot \mathbb{Z}$$

heißt die Exponentialfunktion modulo p zur Basis g.

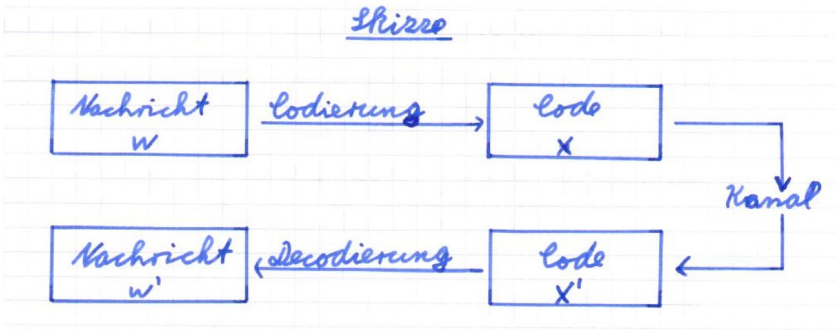
ii) Ist umgekehrt $\overline{y} = y + p \cdot \mathbb{Z} \in \mathbb{Z}/p \cdot \mathbb{Z}$ gegeben mit $p \nmid y$, so heißt jede Zahl $x \in \mathbb{Z}$ mit $\overline{g}^x = \overline{y}$ ein Diskreter Logarithmus von y modulo p zur Basis g.

Bemerkung 3.17. *i) Für p und g wie in Definition 3.16 gibt es nicht unbedingt zu jedem $y \in \mathbb{Z}$ mit $p \nmid y$ einen Diskreten Logarithmus von y modulo p zur Basis g. Allerdings kann - bei vorgegebener ungerader Primzahl p - die Zahl $g \in \{2, \ldots p-1\}$ so gewählt werden, dass zu dieser Basis g all diese Logarithmen existieren.*
Eine Basis g mit dieser Eigenschaft heißt auch Primitivwurzel modulo p.

ii) Nach obigen Ausführungen ist die Berechnung von Werten von Exponentialfunktionen modulo p einfach, nicht aber die Berechnung von Diskreten Logarithmen.

§4 Codierungstheorie

Bemerkung 4.1. ***Prinzip der Nachrichtenübertragung:***
Nachrichten werden vom Sender codiert ; die so codierte Mitteilung wird über einen Kanal an den Empfänger geschickt und schließlich von diesem decodiert.



Störungen im Kanal können bewirken, dass Nachrichten verfälscht werden. Die Konsequenz in obiger Skizze könnte sein:

$$x \neq x' \quad \textit{und damit auch} \quad w \neq w'.$$

Die Codierungstheorie beschäftigt sich damit , solche Übertragungsfehler zu erkennen und möglichst zu korrigieren. Dabei sollten Codierungen so gestaltet werden, dass sich verschiedene Codewörter „gut genug“ - zum Beispiel an mindestens drei Stellen - unterscheiden.
In leichter Abwandlung zur entsprechenden Notation in den Konventionen 3.1 fixieren wir für dieses Kapitel:

Definition 4.2. *Für eine endliche nichtleere Menge A sei*

$$A^* := \{\square\} \bigcup \bigcup_{n=1}^{\infty} A^n.$$

Die Elemente von A^* werden A-Wörter genannt. $\square$ heißt das leere Wort ; wir schreiben auch:

$$A^0 = \{\square\}.$$

Die Länge eines Wortes $\tilde{a} \in A^*$ ist gegeben durch

$$l(\tilde{a}) := l \ , \ \textit{falls } \tilde{a} \in A^l.$$

Definition 4.3. *Es seien A, B endliche nichtleere Mengen. Ein $A-B-Code$ ist eine injektive Abbildung $c: \ A \to B^* \backslash \{\square\}$. B heißt das zugrundeliegende Alphabet; die Menge $c(A)$ heißt die Menge der Codewörter von c.*
Ist $c: \ A \to B^* \backslash \{\square\}$ ein $A-B-Code$, so erweitern wir c auf folgende Weise zu einer Abbildung $c^*: \ A^* \to B^*$:

$$c^*(\square) := \square \quad ,$$
$$c^*(a_1 \dots a_n) := c(a_1) \dots c(a_n).$$

Beispiele 4.4:

i) Sei $A \subseteq \mathbb{N}$ endlich und nicht leer sowie

$$B := \{n \in \mathbb{Z} \mid 0 \leq n \leq 9\}.$$

Weiter sei $d: A \to B^* \backslash \{\square\}$ die übliche Dezimaldarstellung, wobei 0 nie Anfangsglied von einem $d(n), n \in A$, ist. Dann ist d ein $A - B - Code$.

ii) Die Abbildung, die jedem Objekt das zugehörige deutsche Wort über dem „gewöhnlichen“ Alphabet zuordnet, ist im Sinne von Definition 4.3 kein Code. Zum Beispiel ist „Tor“ ein Wort mehrfacher Bedeutung (sowohl für „Pforte“ als auch für „Narr“).

iii) Es sei $A = \{\alpha, \beta, \gamma\}$, $B = \{0, 1\}$ und

$$c(\alpha) := 0 \quad , \quad c(\beta) := 1 \quad , \quad c(\gamma) := 00.$$

Dann ist c ein $A - B - Code$, aber die Abbildung c^* ist nicht injektiv. Zum Beispiel ist

$$c^*(\alpha\alpha) = c^*(\gamma) = 00.$$

Definition 4.5. *Ein $A - B - Code$ heißt ein Präfix - Code (oder sofort entzifferbar), falls kein Codewort Präfix - das heißt Anfangsabschnitt - eines anderen ist.*

Satz 4.6. *Für jeden Präfix-Code $c: A \to B^* \backslash \{\square\}$ ist neben c auch c^* injektiv.*

Beweis. Wir zeigen durch Induktion nach m:
Zu $b_1 \dots b_m \in c^*(A^*)$ gibt es genau ein $\tilde{a} \in A^*$ mit $c^*(\tilde{a}) = b_1 \dots b_m$.
Für $m = 0$ ist nichts zu zeigen.
Sei nun $m \geq 1$. Weil c ein Präfix - Code ist, gibt es genau ein i mit $1 \leq i \leq m$ und $b_1 \dots b_i \in c(A)$.
Sei $a_1 \in A$ das Element mit $b_1 \dots b_i = c(a_1)$.
Dann ist $b_{i+1} \dots b_m \in c^*(A^*)$; nach Induktionsannahme gibt es also genau ein $\tilde{a}' \in A^*$ mit $b_{i+1} \dots b_m = c^*(\tilde{a}')$. Dann ist $\tilde{a} := a_1\tilde{a}'$ das eindeutig bestimmte Element in A^* mit $c^*(\tilde{a}) = b_1 \dots b_m$. □

Beispiele:

i) Sind alle Codewörter von c gleich lang, so ist c ein Präfix-Code.

ii) Es sei $A = \{\alpha, \beta, \gamma\}$, $B = \{0, 1\}$ und

$$c(\alpha) := 00 \quad , \quad c(\beta) := 01 \quad , \quad c(\gamma) := 1.$$

Dann ist c ein Präfix - Code.
Beispielsweise ist

$$00|01|01|00|1|01|00 \quad = \quad c^*(\alpha\beta\beta\alpha\gamma\beta\alpha).$$

Konvention 4.7. *Es sei* $c: A \to B^*\backslash\{\square\}$ *ein* $A-B-$*Code sowie*

$$n \coloneqq \max\{l(c(a)) \mid a \in A\}.$$

Für $1 \leq m \leq n$ *und* $w \in B^m$ *setzen wir*

$$F(w) = F_n(w) \coloneqq \{ww' \mid w' \in B^{n-m}\}.$$

$F_n(w)$ *besteht also aus denjenigen* $B-$*Wörtern der Länge* n*, die* w *als Präfix haben.*

Lemma 4.8. *Unter der gerade getroffenen Konvention sind folgende Aussagen äquivalent:*

(i) $c: A \to B^*\backslash\{\square\}$ *ist ein Präfix-Code.*

(ii) Für je zwei Elemente $a, a' \in A$ *mit* $a \neq a'$ *sind die Mengen* $F(c(a))$ *und* $F(c(a'))$ *disjunkt.*

Beweis. $(i) \Rightarrow (ii)$:
Wir nehmen an, es gebe ein Element $b_1 \dots b_n \in F(c(a)) \cap F(c(a'))$.
Dann gibt es Indizes i, j mit

$$b_1 \dots b_i = c(a) \quad , \quad b_1 \dots b_j = c(a').$$

Aus der Injektivität von c auf A und der Annahme $a \neq a'$ folgt: $i \neq j$.
Das widerspricht aber der Voraussetzung, dass c ein Präfix-Code ist.

$(ii) \Rightarrow (i)$:
Seien $a, a' \in A$ mit $a \neq a'$. Aus (ii) folgt dann sofort:
$c(a)$ ist kein Präfix von $c(a')-$ und umgekehrt. □

Satz 4.9. *Es seien* A, B *endliche nichtleere Mengen, es sei* $|A| = z$, $|B| = b$*, und für jedes* $a \in A$ *sei* $n(a) \in \mathbb{N}$ *fixiert. Dann sind die folgenden Aussagen äquivalent:*

(i) Es gibt einen Präfix-Code $c: A \to B^*\backslash\{\square\}$ *mit*

$$l(c(a)) = n(a) \quad \textit{für alle } a \in A.$$

(ii) Es gilt die folgende Kraft'sche Ungleichung:

$$\sum_{a \in A} \frac{1}{b^{n(a)}} \leq 1.$$

Beweis. Wir setzen $n \coloneqq \max\{n(a) | a \in A\}$.
$(i) \Rightarrow (ii)$:
Ist c wie in (i), so folgt aus Lemma 4.8, $(i) \Rightarrow (ii)$, dass die Mengen $F(c(a))$ für $a \in A$ paarweise disjunkt sind. Daher folgt:

$$b^n = |B|^n \geq \sum_{a \in A} |F(c(a))| = \sum_{a \in A} b^{n-n(a)}.$$

Division durch b^n liefert die Kraft'sche Ungleichung.

$(ii) \Rightarrow (i)$:
Wir konstruieren den gewünschten Präfix-Code wie folgt:
Wir schreiben $A = \{a_1, \dots a_z\}$ und nehmen ohne Einschränkung an:

$$n(a_1) \leq n(a_2) \leq \cdots \leq n(a_z).$$

Sodann wählen wir $c(a_1) \in B^{n(a_1)}$ beliebig. Ist $1 \leq m < z$, und sind $c(a_1) \in B^{n(a_1)}, \dots, c(a_m) \in B^{n(a_m)}$ bereits gewählt, so folgt wegen $m < z$ aus der Kraft'schen Ungleichung:

$$\sum_{j=1}^{m} |F(c(a_j))| = \sum_{j=1}^{m} b^{n-n(a_j))} < b^n \cdot \sum_{j=1}^{z} \frac{1}{b^{n(a_j)}} \leq b^n.$$

Es gibt also ein Element $b_1 \dots b_n \in B^n$, das in keiner der Mengen $F(c(a_j))$ für $1 \leq j \leq m$ liegt.
Setzen wir nun $i := n(a_{m+1})$ und $c(a_{m+1}) = b_1 \dots b_i$, so folgt zunächst:
$l(c(a_{m+1})) = i = n(a_{m+1})$.
Ferner gilt nach Konstruktion:

$$n(a_j) \leq i \ , \ b_1 \dots b_n \notin F(c(a_j)) \ \text{ für alle } \ j \text{ mit } 1 \leq j \leq m.$$

Das bedeutet, dass kein $c(a_j), 1 \leq j \leq m$, Präfix von $c(a_{m+1}) = b_1 \dots b_i$ ist - und umgekehrt.
Insgesamt folgt somit durch Induktion:
c ist ein Präfix-Code. □

Definition 4.10. *Ein $A - B - Code$ $c: \ A \to B^* \backslash \{\square\}$ heißt eindeutig entzifferbar, wenn die zugehörige Abbildung $c^*: \ A^* \to B^*$ injektiv ist.*

Bemerkung 4.11. *Nach Satz 4.6 ist jeder Präfix-Code eindeutig entzifferbar. Ebenso ist jeder $A - B - Code$, in dem kein Codewort Endabschnitt eines anderen ist, eindeutig entzifferbar.*

Definition 4.12. *Es sei E eine beliebige nichtleere Menge, und $d: \ E \times E \to \mathbb{R}_0^+$ sei eine Abbildung. Dann heißt das Paar $M := (E, d)$ ein Metrischer Raum mit der Metrik d , falls folgende Axiome erfüllt sind:*

(M1) $\qquad d(x, y) = 0 \quad \Leftrightarrow \quad x = y.$

(M2) $\qquad \forall \ x, y \in E: \quad d(x, y) = d(y, x) \qquad$ *(Symmetrie).*

(M3) $\qquad \forall \ x, y, z \in E: \quad d(x, z) \leq d(x, y) + d(y, z).$

Die Ungleichung in $(M3)$ heißt Dreiecksungleichung.

Definition 4.13. *Sei B eine endliche nichtleere Menge und $n \in \mathbb{N}$.*
Der Hamming - Abstand in B^n ist die Abbildung $h: B^n \times B^n \to \{0, 1, \ldots, n\}$, definiert durch

$$h((x_1, \ldots, x_n), (y_1, \ldots, y_n)) := \#\{i \in \{1, \ldots, n\} : \quad x_i \neq y_i\}.$$

Bemerkung 4.14. *Der Hamming - Abstand h ist eine Metrik - auf der Grundmenge B^n.*

Definition 4.15. *Sei B endlich und nicht leer, und sei $n \in \mathbb{N}$.*

i) Teilmengen C von B^n werden (auch) als Codes (genauer: Block-Codes) bezeichnet.

ii) Für $x \in B^n$ und $0 \leq e \leq n$ ist die Hamming - Kugel um x mit Radius e definiert durch

$$K_e(x) := \{z \in B^n \mid h(x, z) \leq e\}.$$

iii) Sind $x, y \in B^n$, und ist $e := h(x, y)$, so gehe y aus x durch e Abänderungen - oder e Fehler - hervor. Für $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ heißt die Menge $E := \{j \mid x_j \neq y_j\}$ das Fehler - Muster bei dieser Abänderung.

iv) Ein Block - Code $C \subseteq B^n$ korrigiere bis zu e Fehler, falls für alle $c, d \in C$ mit $c \neq d$ gilt:

$$K_e(c) \bigcap K_e(d) = \emptyset.$$

v) $C \subseteq B^n$ entdecke bis zu e Fehler, falls gilt:

$$c, d \in C \ , \ c \neq d \ \Rightarrow h(c, d) \geq 2 \cdot e.$$

Interpretation:
Für $0 \leq e \leq n$ und $x \in B^n$ besteht $K_e(x)$ aus denjenigen $z \in B^n$, die aus x durch höchstens e Fehler hervorgehen. Korrigiert $C \subseteq B^n$ bis zu e Fehler, so gibt es zu jedem $z \in B^n$ also höchstens ein $x \in C$ mit $h(x, z) \leq e$. Das bedeutet:
Wenn nur Wörter aus C gesendet werden können und dabei maximal e Fehler unterstellt werden, so kann das gesendete Wort $x \in C$ aus dem empfangenen Wort $z \in B^n$ rekonstruiert werden.

C korrigiert genau dann bis zu e Fehler, falls gilt:

$$c, d \in C \ , \ c \neq d \Rightarrow h(c, d) \geq 2 \cdot e + 1.$$

Ist dagegen $h(c, d) = 2 \cdot e = 2 \cdot h(c, z) = 2 \cdot h(z, d)$ für ein $z \in K_e(c) \bigcap K_e(d)$, so geht z aus c und aus d jeweils durch e Fehler hervor.

Beispiele 4.16:
Sei jeweils $B = \{0, 1\}$ und $k \geq 1$.

i) Sei $n := 3k$ und

$$C := \{(x_1, \ldots, x_n) \in B^n \mid x_i = x_{i+k} = x_{i+2k} \text{ für } 1 \leq i \leq k\}.$$

C heißt der 2- fache Wiederholungscode .
C korrigiert bis zu einem Fehler.

ii) Sei $n := k + 1$ und

$$C := \{(x_1, \ldots, x_n) \in B^n \mid x_n \equiv x_1 + \cdots + x_{n-1} \mod 2\}.$$

C heißt Paritätscode. C entdeckt bis zu einem Fehler, kann ihn aber nicht korrigieren.

iii) Sei $n = 4$, $c := (0, 0, 0, 0)$, $d := (1, 1, 1, 1)$, $C := \{c, d\}$.
Dann entdeckt C bis zu 2 Fehler, kann sie aber nicht korrigieren:
$z := (0, 0, 1, 1)$ kann sowohl aus c als auch aus d durch 2 Fehler hervorgegangen sein.

Satz 4.17. ***Die Hamming - Schranke:***
Sei $|B| = b$, $n \in \mathbb{N}$, und $C \subseteq B^n$ sei ein bis zu e Fehler korrigierender Code. Dann gilt:

$$|C| \leq b^n \cdot \left(\sum_{k=0}^{e} \binom{n}{k} \cdot (b-1)^k \right)^{-1}.$$

Beweis. Nach Voraussetzung ist $K_e(c) \bigcap K_e(d) = \emptyset$ für alle $c, d \in C$ mit $c \neq d$. Für $c \in C$ und $0 \leq k \leq e$ gibt es genau $\binom{n}{k} \cdot (b-1)^k$ Wörter in B^n mit Hamming - Abstand k zu c ; also folgt:

$$b^n = |B|^n \geq |\dot{\bigcup_{c \in C}} K_e(c)| = \sum_{c \in C} |K_e(c)|$$
$$= |C| \cdot \sum_{k=0}^{e} \binom{n}{k} \cdot (b-1)^k.$$

Damit folgt die Behauptung. □

Bemerkung 4.18. *Für $n \in \mathbb{N}$ mit $n \geq 2$ sind folgende Aussagen äquivalent:*

(i) Es gibt einen Körper mit genau n Elementen.

(ii) n ist eine Primzahlpotenz.

Ist $n = p$ selbst eine Primzahl, so ist $\mathbb{Z}/p \cdot \mathbb{Z}$ ein Körper ; siehe Aufgabe 23.
Ist $n = p^k$ für $p \in \mathbb{P}$ und $k \geq 2$, so ist $\mathbb{Z}/n \cdot \mathbb{Z}$ aber kein Körper ; der Körper mit p^k Elementen hat eine andere Struktur !

Beispiel 4.19
Ist $K = \{0, 1, a, b\}$ der Körper mit $4 = 2^2$ Elementen - mit neutralem Element 0 bzw. 1 bezüglich der Addition bzw. der Multiplikation , so ergeben sich die folgende Additions - und Multiplikationstabelle:

$+$	0	1	a	b
0	0	1	a	b
1	1	0	b	a
a	a	b	0	1
b	b	a	1	0

$\cdot$	0	1	a	b
0	0	0	0	0
1	0	1	a	b
a	0	a	b	1
b	0	b	1	a

Bemerkung 4.20. - *Siehe auch §5, WS 2015/16*
Es sei K ein endlicher Körper und $n \in \mathbb{N}$.

i) Für $x = (x_1, \ldots, x_n) \in K^n$ und $y = (y_1, \ldots, y_n) \in K^n$ setzen wir

$$x + y \coloneqq (x_1 + y_1, \ldots, x_n + y_n).$$

Für $\lambda \in K$ sei ferner

$$\lambda \cdot (x_1, \ldots, x_n) \coloneqq (\lambda \cdot x_1, \ldots, \lambda \cdot x_n).$$

ii) Eine Teilmenge V von K^n heißt ein Vektorraum , falls gilt:

(V1) Es ist $0 \in V$.

(V2) Für $x, y \in V$ ist auch $x + y \in V$.

(V3) Für $x \in V$ und $\lambda \in K$ ist auch $\lambda \cdot x \in V$.

iii) Ähnlich wie für den Grundkörper $\mathbb{R}$ läßt sich zeigen:
Jeder Vektorraum $V \subseteq K^n$ hat eine Basis B mit $|B| \leq n$; das ist eine Teilmenge $B = \{v_1, \ldots, v_k\}$ von V mit folgender Eigenschaft:
Zu jedem $v \in V$ existieren eindeutig bestimmte Elemente $\lambda_1, \ldots, \lambda_k \in K$ mit

$$v = \sum_{i=1}^{k} \lambda_i \cdot v_i.$$

iv) Je zwei verschiedene Basen B_1, B_2 eines Vektorraums $V \subseteq K^n$ haben gleich viele Elemente. Ihre gemeinsame Kardinalität heißt die Dimension von V und wird mit $\dim V$ bzw. $\dim_K V$ bezeichnet.

Definition 4.21. *Sei K ein endlicher Körper, und sei $n \in \mathbb{N}$.*

i) Jeder Vektorraum $C \subseteq K^n$ wird auch ein Linearer Code über K genannt.
Ist $k \coloneqq \dim_K C$, so heißt C auch ein (n, k) - Code.

ii) Die Distanz $d(C)$ eines linearen Codes $C \neq \{0\}$ ist gegeben durch

$$d(C) := \min_{\substack{a,b \in C \\ a \neq b}} h(a,b) = \min_{\substack{a,b \in C \\ a \neq b}} h(a-b,0) = \min_{c \in C \setminus \{0\}} w(c) \ ,$$

wobei $w(c) := h(c,0)$ ist.

Beispiel:
Es sei $K := \mathbb{Z}/2 \cdot \mathbb{Z}$, $n \in \mathbb{N}$ beliebig sowie

$$C := \{(0,\ldots,0),(1,\ldots,1)\} \subseteq K^n.$$

Dann ist C ein $(n,1)-$ Code mit $d(C) = n$.

Konventionen 4.22. *Für einen Vektorraum $C \subseteq K^n$ ist der zu C gehörige orthogonale Vektorraum $C^\perp$ gegeben durch*

$$C^\perp := \{(u_1,\ldots,u_n) \in K^n \mid \sum_{i=1}^{n} u_i \cdot v_i = 0 \text{ für alle } (v_1,\ldots,v_n) \in C\}.$$

Dann ist $\dim C + \dim C^\perp = n$.
Sei $k := \dim C$ und $m := n - k$.
Ist M eine Matrix über K mit m Zeilen und n Spalten, deren Zeilen eine Basis von $C^\perp$ bilden, so ist

$$C = \{c \in K^n \mid M \cdot c^T = 0\}.$$

Solch eine Matrix M wird auch Kontrollmatrix für C genannt.

Beispiele 4.23:
Wir betrachten $K := \mathbb{Z}/2 \cdot \mathbb{Z}$.

i) Für

$$M = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

erhalten wir den 2-fachen Wiederholungscode

$$C = \{(c_1,c_2,c_3) \in K^3 \mid c_1 = c_2 = c_3\} = \{(0,0,0),(1,1,1)\}.$$

ii) Für

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

erhalten wir den Fano - Code

$$C = \{c \in K^7 \mid M \cdot c^T = 0\}.$$

Satz 4.24. *Sei K ein endlicher Körper und $C \subseteq K^n$ ein $(n,k)-$Code mit Kontrollmatrix M. Für $d \leq n$ sind dann folgende Aussagen äquivalent:*

(i) $d(C) \geq d$.

(ii) *Je* $d-1$ *Spalten von* M *sind linear unabhängig. Insbesondere korrigiert* C *für* $e \leq \dfrac{n-1}{2}$ *genau dann bis zu* e *Fehler, wenn je* $2e$ *Spalten von* M *linear unabhängig sind.*

Beweis. Ist die Äquivalenz von (i) und (ii) bewiesen, so folgt auch die letzte Behauptung mit $d \coloneqq 2e+1$.

Wir schreiben $M = (v_1 \dots v_n)$ mit Spaltenvektoren $v_1, \dots, v_n \in K^m$.

$(i) \Rightarrow (ii)$:
Es sei $1 \leq i_1 < i_2 < \dots < i_{d-1} \leq n$, und es seien $\lambda_{i_1}, \dots, \lambda_{i_{d-1}} \in K$ mit

$$\sum_{j=1}^{d-1} \lambda_{i_j} \cdots v_{i_j} = 0.$$

Definiere $c = (c_1, \dots, c_n) \in C$ durch

$$c_\nu \coloneqq \begin{cases} \lambda_{i_j}, & \text{falls } \nu = i_j \text{ für passendes } j \\ 0 & \text{sonst.} \end{cases}$$

Dann ist $w(c) \leq d-1$, also $c = 0$ nach (i).
Somit ist $\lambda_{i_1} = \dots = \lambda_{i_{d-1}} = 0$; also gilt (ii).

$(ii) \Rightarrow (i)$:
Es gelte $d(C) < d$; sei also $c = (c_1, \dots, c_n) \in C$ mit $0 < w(c) < d$. Ist dann $I \coloneqq \{i | 1 \leq i \leq n \ , \ c_i \neq 0\}$, so sind die Spalten $v_i, i \in I$, wegen der Beziehung $M \cdot c^T = 0$ linear abhängig. □

Definition 4.25. *Es sei* B *eine nichtleere und endliche Menge, und es seien* $n, e \in \mathbb{N}$. *Ein Code* $C \subseteq B^n$ *heißt* e- perfekt *, wenn gilt:*

$$B^n = \dot{\bigcup_{c \in C}} K_e(c) \quad .$$

Das bedeutet: Zu jedem $x \in B^n$ *gibt es genau ein* $c \in C$ *mit* $x \in K_e(c)$.

Beispiel:
Es sei $B = \{0, 1\}$, $n = 2e+1 \geq 3$ und

$$C = \{(0, \dots, 0), (1, \dots, 1)\} \subseteq K^n.$$

Satz 4.26. ***(siehe auch Satz 4.17.)***
Es sei B *endlich und nicht leer, und es seien* $n, e \in \mathbb{N}$. *Für einen Code* $C \subseteq B^n$ *sind dann äquivalent:*

(i) C ist e- perfekt.

(ii) C korrigiert bis zu e Fehler, und es gilt:

$$|C| = |B|^n \cdot \left(\sum_{k=0}^{e} \binom{n}{k} \cdot (|B|-1)^k \right)^{-1} \quad .$$

Beweis. Siehe Übung Nr. 26.

Definition 4.27. *Sei K ein Körper mit q Elementen, sei $m \in \mathbb{N}$ und*

$$n := \frac{q^m - 1}{q-1} = q^{m-1} + q^{m-2} + \cdots + q + 1 \ .$$

Ein Hamming-Code ist ein $(n, n-m)$- Code über K, der bis zu einem Fehler korrigiert.

Satz 4.28. *Jeder Hamming - Code ist 1-perfekt.*

Beweis. Es seien K, q, m, n wie in Definition 4.27. Ist nun $C \subseteq K^n$ ein Hamming-Code mit $\dim C = n - m$, so folgt:

$$\begin{aligned} |C| \cdot \textstyle\sum_{k=0}^{1} \binom{n}{k} \cdot (q-1)^k &= q^{n-m} \cdot (1 + n \cdot (q-1)) \\ &= q^{n-m} \cdot q^m = q^n = |K|^n \ . \end{aligned}$$

Damit liefert Satz 4.26 , $(ii) \Rightarrow (i)$, die Behauptung.

□

Bemerkung 4.29. *Es sei wieder K ein Körper mit q Elementen, sei nun $m \geq 2$ sowie*

$$n := \frac{q^m - 1}{q-1} \ .$$

Auf $K^m \backslash \{0\}$ betrachten wir folgende Äquivalenzrelation $\mathcal{R}$:
Wir schreiben $v \mathcal{R} w$, falls ein $\lambda \in K \backslash \{0\}$ existiert mit $\lambda \cdot v = w$.
Jede Äquivalenzklasse besitzt $|K \backslash \{0\}| = q - 1$ Elemente; es gibt also genau

$$\frac{|K^m \backslash \{0\}|}{q-1} = \frac{q^m - 1}{q-1} = n$$

Äquivalenzklassen.
Zwei Vektoren $v, w \in K^m \backslash \{0\}$ sind nun genau dann linear unabhängig, wenn sie nicht äquivalent sind. Es gibt also eine Matrix M über K mit n Spalten der Länge m, die paarweise linear unabhängig sind.
Für den $(n, n-m)-$Code

$$C := \{c \in K^n \mid M \cdot c^T = 0\}$$

mit Kontrollmatrix M folgt daher aus Satz 4.24.:
C ist ein Hamming-Code.

Beispiele 4.30:

i) Für $K = \mathbb{Z}/2 \cdot \mathbb{Z}$ ist sowohl der 2-fache Wiederholungscode als auch der Fano-Code - aus den Beispielen 4.23 - ein Hamming - Code.

ii) Sei $K = \mathbb{Z}/3 \cdot \mathbb{Z} = \{\overline{0}, \overline{1}, \overline{2}\}$.
Die Kontrollmatrix M sei - für $m = 2$ und $n = 4$ - gegeben durch:

$$\begin{pmatrix} \overline{1} & \overline{0} & \overline{1} & \overline{1} \\ \overline{0} & \overline{1} & \overline{1} & \overline{2} \end{pmatrix} \quad .$$

Der zu M gehörige Hamming-Code ist dann

$$\begin{aligned} C &= \{c \in K^4 \mid M \cdot c^T = 0\} \\ &= \{(\overline{0},\overline{0},\overline{0},\overline{0}), (\overline{1},\overline{0},\overline{1},\overline{1}), (\overline{2},\overline{0},\overline{2},\overline{2}), (\overline{0},\overline{1},\overline{1},\overline{2}), (\overline{0},\overline{2},\overline{2},\overline{1}), \\ &\quad (\overline{1},\overline{1},\overline{2},\overline{0}), (\overline{2},\overline{2},\overline{1},\overline{0}), (\overline{1},\overline{2},\overline{0},\overline{2}), (\overline{2},\overline{1},\overline{0},\overline{1})\}. \end{aligned}$$

Die ist ein $(4, 2)$- Code.

Definition 4.31. *Es sei K ein endlicher Körper und $C \subseteq K^n$ ein (n, k) - Code über K. Ist $\{g_1, \ldots, g_k\}$ eine Basis des Zeilenvektorraumes C, so heißt die $k \times n$ - Matrix*

$$G \coloneqq \begin{pmatrix} g_1 \\ \vdots \\ g_k \end{pmatrix}$$

eine Generatormatrix für C.
Hat G die Form

$$G = (I_k \;\; G_1)$$

mit der $k \times k$ - Einheitsmatrix I_k und einer Matrix $G_1 \in \boldsymbol{Mat}_{k\times(n-k)}(K)$, so heißt die Generatormatrix G systematisch.

Bemerkung 4.32. *i) Mit obigen Bezeichnungen gilt:*

$$C = \{\sum_{i=1}^{k} w_i \cdot g_i \mid w_i \in K\}.$$

Identifizieren wir die Nachrichten aus C mit den $|K|^k$ Vektoren aus K^k, so codieren wir mittels des Isomorphismus $\Phi : \; K^k \to C$, gegeben durch

$$\Phi(w_1, \ldots, w_k) \coloneqq \sum_{i=1}^{k} w_i \cdot g_i \quad .$$

Die Decodierung erfolgt durch Lösung eines Linearen Gleichungssystems.
Liegt eine systematische Generatormatrix zugrunde, so besteht die Nachricht gerade aus den ersten k Symbolen des Codewortes.

ii) Die Generatormatrizen G für einen Code C sind genau die Kontrollmatrizen für $C^\perp$. Ist $G = (I_k \;\; G_1)$ eine systematische Generatormatrix für C, so ist $M \coloneqq (-G_1^T \;\; I_{n-k})$ eine Kontrollmatrix für C.

Definition 4.33. *Die Informationsrate r eines (n,k) - Codes ist gegeben durch*

$$r \coloneqq \frac{k}{n}.$$

Beispiel 4.34.:
Sei $n \geq 2$, und sei C der $(n, n-1)$ - Code über $\mathbb{Z}/2 \cdot \mathbb{Z}$ mit der systematischen Generatormatrix

$$G \coloneqq \left.\begin{pmatrix} 1 & & 0 & 1 \\ & \diagdown & & \vdots \\ 0 & & 1 & 1 \end{pmatrix}\right\} n-1 \text{ Zeilen}$$

Dann ist

$$C = \{(w_1, \ldots, w_n) \in (\mathbb{Z}/2 \cdot \mathbb{Z})^n \mid w_n = w_1 + \cdots + w_{n-1}\}$$

der Paritätscode.
$M \coloneqq (1 \ldots 1)$ ist die Kontrollmatrix für C, und die Informationsrate ist $r = \dfrac{n-1}{n}$.
Der zugehörige Duale Code $C^\perp = \{(0, \ldots, 0), (1, \ldots, 1)\}$ hat M als Generatormatrix, G als Kontrollmatrix und Informationsrate $\dfrac{1}{n}$.

Beispiel 4.35, Die Reed-Solomon - Codes:
Sei $K = \{0, 1, a_1, \ldots, a_{q-2}\}$ „der“ Körper mit q Elementen - für eine vorgegebene Primzahlpotenz q.
Für festes $d \in \mathbb{N}$ mit $2 \leq d \leq q+1$ sei

$$M \coloneqq \begin{pmatrix} 0 & 1 & 1 & 1 & \ldots & 1 \\ \vdots & 0 & 1 & a_1 & \ldots & a_{q-2} \\ \vdots & \vdots & \vdots & a_1^2 & \ldots & a_{q-2}^2 \\ 0 & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & a_1^{d-2} & \ldots & a_{q-2}^{d-2} \end{pmatrix}$$

Die $(d-1) \times (q+1)$ - Matrix M hat den Rang $d-1$; genauer sind je $d-1$ Spalten von M linear unabhängig.
Der $(q+1, q+2-d)$- Code C mit Kontrollmatrix M heißt Reed-Solomon-Code.
Weil trivialerweise je d Spalten von M linear abhängig sind, liefert Satz 4.24:

$$d(C) = d.$$

Die Informationsrate $r = r(q,d)$ ist gegeben durch

$$r = \frac{q+2-d}{q+1}.$$

Für festes d ist also $\lim\limits_{q \to \infty} r(q,d) = 1$.

§5 Zahlendarstellung auf Digitalrechnern und Fehleranalyse

Bemerkung 5.1. ***Zahlendarstellung:***
Bei den Digitalrechnern werden reelle Zahlen durch endlich viele physikalische Zustände dargestellt. Weil nur endlich viele reelle Zahlen dargestellt werden können, müssen die anderen approximiert werden.
Sei $g \in \mathbb{N}$ fest mit $g \geq 2$. Zu $r \in \mathbb{R}$ gibt es ein eindeutig bestimmtes $j \in \{0,1\}$, ein eindeutig festgelegtes $l \in \mathbb{Z}$ und zu $k \in \mathbb{Z}$ mit $k \leq l$ eindeutig bestimmte $a_k \in \mathbb{Z}$ mit :

(5.1) $r = (-1)^j \cdot \sum_{k=-\infty}^{l} a_k \cdot g^k$,
(5.1a) $a_l \neq 0$ *für* $r \neq 0$, $j = l = 0$ *für* $r = 0$,
(5.1b) $a_k \in S_g := \{0, 1, \ldots, g-1\}$ *für alle* $k \leq l$,
$a_k < g - 1$ *für unendlich viele* $k < l$.

g hängt dabei von der Maschine ab;
in der Regel ist $g = 10$ (Dezimaldarstellung) oder $g = 2$ (Dualdarstellung).

Bemerkung 5.2. ***Die Festpunktdarstellung:***
Bei der Festpunktdarstellung sind die Zahlen s und t der Stellen vor bzw. nach dem Komma durch die Maschine fixiert. Die darstellbaren Zahlen r haben die Gestalt

(5.2) $r = \pm \sum_{k=-t}^{s-1} b_k \cdot g^k$ *mit* $b_k \in S_g$ *für* $-t \leq k < s$.

Darstellung auf der Maschine:

$\pm b_{s-1} \ldots b_0$	$b_{-1} \ldots b_{-t}$

Die darstellbaren Zahlen sind in dem offenen Intervall $(-g^s, g^s)$ äquidistant verteilt mit dem Abstand g^{-t}.

Die Festpunktdarstellung ist beispielsweise angemessen für kaufmännische Rechnungen.

Bemerkung 5.3. ***Die Gleitpunktdarstellung:***
Zu $r \in \mathbb{R}\backslash 0$ gibt es eindeutig bestimmte $a \in \mathbb{R}$ mit $\frac{1}{g} \leq |a| < 1$ und $b \in \mathbb{Z}$ mit

(5.3) $r = a \cdot g^b$.

a heißt die Mantisse und b der Exponent von r.

Für gewisse, durch die Maschine fixierte, $s, t \in \mathbb{N}$ wird a bzw. b durch t bzw. s Stellen und das Vorzeichen dargestellt.
Ist r exakt darstellbar, so ist

$$-(g^s - 1) \leq b \leq g^s - 1$$

und folglich

$$-g^{(g^s-1)} < r < g^{(g^s-1)}.$$

Beispiele:

Für $g = 10, s = 2$ und $t = 8$ erhält man folgende Darstellungen:

$$\begin{array}{rcl} 5138 & \mapsto & \boxed{0,51380000 \mid 04} \\ -18,23197 & \mapsto & \boxed{-0,18231970 \mid 02} \\ 0,03164 & \mapsto & \boxed{0,31640000 \mid -01} \\ -0,03164 & \mapsto & \boxed{-0,31640000 \mid -01} \end{array}$$

Bemerkung 5.4. ***Rundung:***
Es sei $\mathcal{M}$ die - endliche - Menge der in der Maschine darstellbaren Zahlen - auch Maschinenzahlen- genannt.
Eine Zahl $r \in \mathbb{R}$ zu runden heißt:
Suche ein $\tilde{r} \in \mathcal{M}$ mit :

$$|r - \tilde{r}| \leq |r - \rho| \quad \textit{für alle } \rho \in \mathcal{M}.$$

Im Falle der Festpunktdarstellung setzen wir für r wie in (5.1) und $l \leq s - 1$:

$$rd(r) := \begin{cases} (-1)^j \cdot \sum_{k=-t}^{l} a_k \cdot g^k & \textit{falls } a_{-t-1} < \dfrac{1}{2} \cdot g \\ (-1)^j \cdot \left(\sum_{k=-t}^{l} a_k \cdot g^k + g^{-t} \right) & \textit{falls } a_{-t-1} \geq \dfrac{1}{2} \cdot g. \end{cases} \tag{5.4}$$

Wir erhalten dann die - von r unabhängige - Abschätzung

$$|r - rd(r)| \leq \frac{1}{2} \cdot g^{-t} \ . \tag{5.4 a}$$

Im Fall der Gleitpunktdarstellung setzen wir

$$gl(0) := 0.$$

$$gl(a \cdot g^b) := rd(a) \cdot g^b \quad \textit{für } \ \frac{1}{g} \leq |a| < 1, \ b \in \mathbb{Z}. \tag{$\widetilde{5.4}$}$$

Für den absoluten Fehler $|r - gl(r)|$ erhalten wir die - von $r = a \cdot g^b$ abhängige - Abschätzung

$$|r - gl(r)| = |a - rd(a)| \cdot g^b \leq \frac{1}{2} \cdot g^{-t} \cdot g^b. \tag{$\widetilde{5.4a}$}$$

Für den relativen Fehler $\dfrac{|r - gl(r)|}{|r|}$, der den absoluten Fehler in Relation zur betrachteten Zahl $r \neq 0$ setzt, erhalten wir die - von $r = a \cdot g^b$ unabhängige - Abschätzung

$$\frac{|r - gl(r)|}{|r|} \leq \frac{1}{2} \cdot g^{-t+b} \cdot \frac{g}{g^b} = \frac{1}{2} \cdot g^{-t+1} =: eps. \tag{$\widetilde{5.4b}$}$$

eps heißt Maschinengenauigkeit.
Zu jedem $r \in \mathbb{R}\backslash\{0\}$ gibt es also ein $\varepsilon \in [-eps, eps]$ mit:

$(\widetilde{5.4c})$ $$gl(r) = r \cdot (1 + \varepsilon).$$

Beispiele für $g = 10,\ t = 8,\ s = 2$

$$gl(246813795) = 246813800 = 0,2468138 \cdot 10^9 \ ,$$
$$gl(-\frac{1}{15}) = gl(-0,0\overline{6}) = -0,66666667 \cdot 10^{-1} \ ,$$
$$gl(e) = gl(2,7182818) = 0,27182818 \cdot 10^1 \ .$$

Bemerkung 5.5. ***Rechenoperationen in der Gleitpunktdarstellung***
Für $r_1, r_2 \in \mathcal{M}$ setzen wir

(5.5a) $r_1 \oplus r_2 \coloneqq gl(r_1 + r_2)$,

(5.5b) $r_1 \ominus r_2 \coloneqq gl(r_1 - r_2)$,

(5.5c) $r_1 \odot r_2 \coloneqq gl(r_1 \cdot r_2)$,

(5.5d) $r_1 \oslash r_2 \coloneqq gl(r_1 \ \div \ r_2)$ *für* $r_2 \neq 0$.

Das seien jeweils die Ergebnisse, die von der Maschine berechnet werden.

Bemerkung 5.6. *Die Gleitpunktoperationen genügen nicht den üblichen Gesetzen der arithmetischen Operationen.*
Zum Beispiel gilt:

$$eps = \frac{1}{2} \cdot g^{-t+1} = \min\{x \in \mathcal{M} | 1 \oplus x > 1\}.$$

Für $0 \leq x < eps,\ x \in \mathcal{M}$, ist also $1 \oplus x = 1$.

Gegenbeispiel zum Assoziativgesetz:

Sei $g = 10,\ t = 4,\ s = 2$ sowie

$a = 0,1349 \cdot 10^{-1} = 0,01349\ ;\ b = 0,3368 \cdot 10^2 = 33,68\ ;\ c = -0,3367 \cdot 10^2 = -33,67.$

Dann folgt:

$$\begin{aligned} a + b + c &= 0,02349 \ ; \\ (a \oplus b) \oplus c &= 33,69 \oplus (-33,67) = 0,02000 \ ; \\ a \oplus (b \oplus c) &= 0,01349 \oplus 0,01 = 0,02349 \ . \end{aligned}$$

Konvention. *Für $r \in \mathbb{R}\backslash\{0\}$ mit gerundetem $r' \in \mathcal{M}$ bezeichne*

(5.6) $$\varepsilon_r \coloneqq \frac{|r - r'|}{|r|}$$

den relativen Fehler.

Satz 5.7. ***Fehlerfortpflanzung***
- exakte Rechnung mit falschen Daten:

Es seien $r_1, r_2 \in \mathbb{R}\backslash\{0\}$ *; sind* $r_1', r_2' \in \mathcal{M}$ *die gerundeten Werte, so gilt also:*

$$\varepsilon_{r_1} \cdot |r_1| = |r_1 - r_1'| \ , \ \varepsilon_{r_2} \cdot |r_2| = |r_2 - r_2'| \ . \tag{5.7}$$

Ferner nehmen wir an, dass $r_1' + r_2'$, $r_1' - r_2'$, $r_1' \cdot r_2'$ *und* $\dfrac{r_1'}{r_2'}$ *exakt berechnet werden. Dann gelten folgende Abschätzungen für die relativen Fehler* $\varepsilon_{r_1 \circ r_2}$ *, wobei* $\circ$ *für die 4 Grundrechenarten steht:*

(5.7a) $\varepsilon_{r_1+r_2} \leq \dfrac{1}{|r_1 + r_2|} \cdot (\varepsilon_{r_1} \cdot |r_1| + \varepsilon_{r_2} \cdot |r_2|)$ *, falls* $r_1 + r_2 \neq 0$;

(5.7b) $\varepsilon_{r_1-r_2} \leq \dfrac{1}{|r_1 - r_2|} \cdot (\varepsilon_{r_1} \cdot |r_1| + \varepsilon_{r_2} \cdot |r_2|)$ *, falls* $r_1 \neq r_2$;

(5.7c) $\varepsilon_{r_1 \cdot r_2} \leq \varepsilon_{r_1} + \varepsilon_{r_2} + \varepsilon_{r_1} \cdot \varepsilon_{r_2} \approx \varepsilon_{r_1} + \varepsilon_{r_2}$;

(5.7d) $\varepsilon_{r_1 \div r_2} \leq \varepsilon_{r_1} + (\varepsilon_{r_2} + \varepsilon_{r_1} \cdot \varepsilon_{r_2}) \cdot \dfrac{|r_2|}{|r_2'|} \approx \varepsilon_{r_1} + \varepsilon_{r_2}$.

Beweis. Verifikation von (5.7a):
Wir erhalten im Falle $r_1 + r_2 \neq 0$:

$$\begin{aligned} \varepsilon_{r_1+r_2} &= \frac{|(r_1 + r_2) - (r_1' + r_2')|}{|r_1 + r_2|} \leq \frac{|r_1 - r_1'| + |r_2 - r_2'|}{|r_1 + r_2|} \\ &= \frac{1}{|r_1 + r_2|} \cdot (\varepsilon_{r_1} \cdot |r_1| + \varepsilon_{r_2} \cdot |r_2|) \quad . \end{aligned}$$

Verifikation von (5.7b):
Ist $r_1 \neq r_2$, so folgt:

$$\begin{aligned} \varepsilon_{r_1-r_2} &= \frac{|(r_1 - r_2) - (r_1' - r_2')|}{|r_1 - r_2|} \leq \frac{|r_1 - r_1'| + |r_2' - r_2|}{|r_1 - r_2|} \\ &= \frac{1}{|r_1 - r_2|} \cdot (\varepsilon_{r_1} \cdot |r_1| + \varepsilon_{r_2} \cdot |r_2|) \quad . \end{aligned}$$

Verifikation von (5.7c):
Wir erhalten:

$$\begin{aligned}\varepsilon_{r_1 \cdot r_2} &= \frac{|r_1 \cdot r_2 - r_1' \cdot r_2'|}{|r_1 \cdot r_2|} = \frac{|r_1 \cdot (r_2 - r_2') + (r_1 - r_1') \cdot r_2'|}{|r_1 \cdot r_2|} \\ &= \frac{|r_1 \cdot (r_2 - r_2') + (r_1 - r_1') \cdot r_2 - (r_1 - r_1') \cdot (r_2 - r_2')|}{|r_1 \cdot r_2|} \\ &\leq \frac{|r_1| \cdot |r_2 - r_2'|}{|r_1| \cdot |r_2|} + \frac{|r_1 - r_1'| \cdot |r_2|}{|r_1| \cdot |r_2|} + \frac{|r_1 - r_1'| \cdot |r_2 - r_2'|}{|r_1| \cdot |r_2|} \\ &= \varepsilon_{r_2} + \varepsilon_{r_1} + \varepsilon_{r_1} \cdot \varepsilon_{r_2} \quad .\end{aligned}$$

Verifikation von (5.7d):
Es folgt:

$$\begin{aligned}\varepsilon_{r_1 \div r_2} \cdot \left|\frac{r_1}{r_2}\right| &= \left|\frac{r_1}{r_2} - \frac{r_1'}{r_2'}\right| = \frac{|r_2' \cdot (r_1 - r_1') - r_1' \cdot (r_2 - r_2')|}{|r_2 \cdot r_2'|} \\ &= \frac{|r_2' \cdot (r_1 - r_1') - r_1 \cdot (r_2 - r_2') + (r_1 - r_1') \cdot (r_2 - r_2')|}{|r_2 \cdot r_2'|} \\ &\leq \frac{|r_2'| \cdot |r_1 - r_1'|}{|r_2| \cdot |r_2'|} + \frac{|r_1| \cdot |r_2 - r_2'|}{|r_2| \cdot |r_2'|} + \frac{|r_1 - r_1'| \cdot |r_2 - r_2'|}{|r_2| \cdot |r_2'|}\end{aligned}$$

Multiplikation mit $\frac{|r_2|}{|r_1|}$ liefert:

$$\begin{aligned}\varepsilon_{r_1 \div r_2} &\leq \frac{|r_1 - r_1'|}{|r_1|} + \frac{|r_2 - r_2'|}{|r_2'|} + \frac{|r_1 - r_1'|}{|r_1|} \cdot \frac{|r_2 - r_2'|}{|r_2'|} \\ &= \varepsilon_{r_1} + (\varepsilon_{r_2} + \varepsilon_{r_1} \cdot \varepsilon_{r_2}) \cdot \frac{|r_2|}{|r_2'|} \quad .\end{aligned}$$

□

Bemerkung 5.8. ***Analyse der Fehlerfortpflanzung - Auslöschung:***
Bei der Multiplikation und der Division verhält sich der relative Fehler unproblematisch. Bei Vernachlässigung sehr kleiner Produkte wie $\varepsilon_{r_1} \cdot \varepsilon_{r_2}$ addieren sich die relativen Fehler im ungünstigsten Fall.
Die Addition ist ebenfalls unproblematisch, wenn r_1 und r_2 gleiches Vorzeichen haben; gegebenenfalls ist $\varepsilon_{r_1+r_2} \leq \varepsilon_{r_1} + \varepsilon_{r_2}$.

Auch die Subtraktion ist unproblematisch, wenn $r_1 - r_2$ nicht ungefähr 0 ist. Gilt jedoch $r_1 \approx r_2$, so kann $\varepsilon_{r_1-r_2}$ gemäß (5.7b) groß werden ! Dieses Verhalten heißt Fehlerverstärkung durch Auslöschung.

Beispiel 5.9:

Wir betrachten das folgende lineare Gleichungssystem:

$$3x = 10 \quad \wedge \quad 5x - \frac{1}{1000} \cdot y = 16,66 \ .$$

Die exakte Lösung berechnet sich wie folgt:

$$x = \frac{10}{3} = 3,\overline{3} \ ;$$
$$y = 1000 \cdot (16,\overline{6} - 16,66) = 1000 \cdot 0,00\overline{6} = 6,\overline{6} = 6\frac{2}{3}.$$

Ein Rechner mit vierstelliger Gleitpunktdarstellung liefert jedoch folgende gerundete Werte:

$$x' = 3,333 \ ;$$
$$y' = 1000 \odot (5 \odot 3,333 \ominus 16,66) = 1000 \cdot (16,67 - 16,66) = 10 \ .$$

Während der Wert x' noch möglichst genau ist (bei Berücksichtigung von nur 4 Stellen), ergibt sich aber für y' ein mehr als nur ungenauer Wert!

Beispiel 5.10. Der Horner - Algorithmus

Sei $m \in \mathbb{N}$, seien $a_0, \ldots, a_m \in \mathbb{R}$, und definiere die Polynom- Funktion $f : \ \mathbb{R} \to \mathbb{R}$ durch

$$f(x) \coloneqq \sum_{j=0}^{m} a_j \cdot x^j = a_m \cdot x^m + \cdots + a_1 \cdot x + a_0$$
$$= (\ldots (a_m \cdot x + a_{m-1}) \cdot x + \cdots + a_1) \cdot x + a_0.$$

Für $x \in \mathbb{R}$ ist der Horner - Algorithmus zur Berechnung von $f(x)$ gegeben durch:

$s_0 \coloneqq a_m$
$for \ j = 1 \ \ until \ \ m \ \ do$
$s_j \coloneqq s_{j-1} \cdot x + a_{m-j} \ \ .$

Für $m = 2$ werden - bei exakten Anfangsdaten - folgende Werte geliefert:

Exakte Werte	Werte im Rechner
$s_0 \coloneqq a_2$	$s_0 \coloneqq a_2$
$s_1 \coloneqq s_0 \cdot x + a_1$	$\tilde{s_1} \coloneqq s_0 \odot x \oplus a_1$
$s_2 \coloneqq s_1 \cdot x + a_0$	$\tilde{s_2} \coloneqq \tilde{s_1} \odot x \oplus a_0$

Aus $(\widetilde{5.4c})$ folgt - für passende $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 \in [-eps, eps]$:

$$
\begin{aligned}
s_0 &= a_2 \ , \\
\tilde{s_1} &= ((s_0 \cdot x) \cdot (1+\varepsilon_1) + a_1) \cdot (1+\varepsilon_2) \\
&= a_2 \cdot (1+\varepsilon_1) \cdot (1+\varepsilon_2) \cdot x + a_1 \cdot (1+\varepsilon_2) \quad , \\
\tilde{s_2} &= ((\tilde{s_1} \cdot x) \cdot (1+\varepsilon_3) + a_0) \cdot (1+\varepsilon_4) \\
&= a_2 \cdot (1+\varepsilon_1) \cdot (1+\varepsilon_2) \cdot (1+\varepsilon_3) \cdot (1+\varepsilon_4) \cdot x^2 \\
&\quad + a_1 \cdot (1+\varepsilon_2) \cdot (1+\varepsilon_3) \cdot (1+\varepsilon_4) \cdot x + a_0 \cdot (1+\varepsilon_4).
\end{aligned}
$$

§6 Interpolation durch Polynome

Bemerkung 6.1. *i) Für $n \in \mathbb{Z}$ mit $n \geq 0$ hat ein Polynom vom Grad n über $\mathbb{R}$ die Gestalt*

$$(*) \qquad P(x) = \sum_{k=0}^{n} a_k \cdot x^k \quad \text{mit } a_0, \ldots, a_n \in \mathbb{R} \quad \text{und } a_n \neq 0.$$

Dem Null-Polynom ordnen wir keinen Grad zu. Ist P wie in $(*)$, so heißt a_n der Leitkoeffizient von P.

ii) Für $n \geq 0$ setzen wir ferner

$$\prod_n := \left\{ \sum_{k=0}^{n} a_k \cdot x^k | a_0, \ldots, a_n \in \mathbb{R} \right\}.$$

Die Menge $\prod_n$ enthält also das Null-Polynom sowie diejenigen Polynome, die höchstens den Grad n haben.

iii) Es sei P ein Polynom vom Grad n mit Leitkoeffizient a. Weiter nehmen wir an: P habe n paarweise verschiedene Nullstellen $x_1, \ldots, x_n \in \mathbb{R}$. Dann gilt für alle $x \in \mathbb{R}$:

$$(**) \qquad P(x) = a \cdot (x - x_1) \cdots (x - x_n) = a \cdot \prod_{k=1}^{n} (x - x_k).$$

Das bedeutet:
P läßt sich vollständig faktorisieren und kann keine weitere Nullstelle besitzen. Insbesondere besitzt jedes Polynom vom Grad n höchstens n Nullstellen.

Definition 6.2. *Für $n \geq 0$ seien $n+1$ paarweise verschiedene reelle Zahlen $x_0, x_1, \ldots, x_n$ gegeben. Dann sind die zugehörigen Lagrange'schen Interpolationspolynome $L_0, L_1, \ldots, L_n$ vom Grad n definiert durch*

$$(6.2) \qquad L_i(x) := \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}$$

$$= \frac{(x - x_0) \cdots (x - x_{i-1}) \cdot (x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0) \cdots (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdots (x_i - x_n)} \quad \text{für } 0 \leq i \leq n.$$

Bemerkung 6.3. *Für $x_0, \ldots, x_n, L_0, \ldots, L_n$ wie in Definition 6.2 gilt:*

$$(6.3) \qquad L_i(x_j) = \delta_{ij} = \begin{cases} 0 & \text{für } i \neq j \\ 1 & \text{für } i = j. \end{cases}$$

Bemerkung 6.4, Die Lagrange'sche Interpolationsaufgabe:
Gegeben seien $n+1$ Stützpunkte

$$(x_i, f_i) \in \mathbb{R}^2 \ , \ 0 \leq i \leq n \ , \ \text{mit } x_i \neq x_j \text{ für } i \neq j.$$

Bestimme ein Polynom $P \in \prod_n$ mit

$$(6.4) \qquad P(x_i) = f_i \ \text{ für } 0 \leq i \leq n.$$

Satz 6.5. *Die Lagrange'sche Interpolationsaufgabe* (6.4) *ist stets eindeutig lösbar. Insbesondere ist das gesuchte Interpolationspolynom P gegeben durch*

$$P(x) = \sum_{i=0}^{n} f_i \cdot L_i(x) = f_0 \cdot L_0(x) + \cdots + f_n \cdot L_n(x). \tag{6.5}$$

Beweis. Nachweis der Existenz.
Sei P wie in (6.5). Zunächst hat P höchstens den Grad n , weil $L_0, \ldots, L_n$ - genau - den Grad n haben.

Weiter liefert (6.3) für $0 \leq j \leq n$:

$$P(x_j) = \sum_{i=0}^{n} f_i \cdot L_i(x_j) = f_j$$

wie gewünscht.

Nachweis der Eindeutigkeit:
Es sei auch $Q \in \prod_n$ mit $Q(x_i) = f_i$ für $0 \leq i \leq n$. Für $R := P - Q \in \prod_n$ folgt dann:

$$R(x_i) = f_i - f_i = 0 \quad \text{für } 0 \leq i \leq n.$$

Das Polynom R hat also mindestens $n+1$ Nullstellen, nämlich $x_0, \ldots, x_n$.
Wegen $R \in \prod_n$ ist das nach Bemerkung 6.1*iii*) nur möglich, wenn R das Null-Polynom ist. Damit folgt: $P = Q$. □

Im folgenden werden Algorithmen zur Berechnung von Interpolationspolynomen vorgestellt, die besser geeignet sind als die Formel (6.5).

Konvention 6.6. *Im folgenden seien $n+1$ Stützpunkte*

$$(x_i, f_i) \in \mathbb{R}^2 \text{ für } 0 \leq i \leq n \text{ mit } x_i \neq x_j \text{ für } i \neq j$$

fixiert.
Für paarweise verschiedene Indizes $i_0, \ldots, i_k \in \{0, \ldots, n\}$ bezeichne $P_{i_0 \ldots i_k}$ das - nach Satz 6.5 eindeutig bestimmte - Polynom in $\prod_k$ mit

$$P_{i_0 \ldots i_k}(x_{i_l}) = f_{i_l} \quad \text{für } 0 \leq l \leq k. \tag{6.6}$$

Satz 6.7. *i) Für $0 \leq i \leq n$ gilt:*

$$P_i(x) = f_i \quad \text{für alle } x \in \mathbb{R}. \tag{6.7a}$$

ii) Für $0 < k \leq n$ und $i_0 \ldots, i_k \in \{0, \ldots, n\}$ mit $|\{i_0, \ldots, i_k\}| = k+1$ gilt:

$$P_{i_0 \ldots i_k}(x) = \frac{(x - x_{i_0}) \cdot P_{i_1 \ldots i_k}(x) - (x - x_{i_k}) \cdot P_{i_0 \ldots i_{k-1}}(x)}{x_{i_k} - x_{i_0}}. \tag{6.7b}$$

Beweis. i) gilt nach (6.6) - mit $k = 0$ und $i_0 = i$, weil P_i als Element von $\prod_0$ konstant sein muss.

ii) Wir führen Induktion nach k.
Es bezeichne $q(x)$ die „rechte“ Seite in (6.7b).
Klar ist: $q(x) \in \prod_k$.
Zu zeigen ist noch:
$q(x_{i_l}) = f_{i_l}$ für $0 \le l \le k$.
Wir erhalten - laut Induktionsannahme:

$$q(x_{i_0}) = \frac{-(x_{i_0} - x_{i_k}) \cdot P_{i_0 \ldots i_{k-1}}(x_{i_0})}{x_{i_k} - x_{i_0}} = f_{i_0} \ ,$$
$$q(x_{i_k}) = \frac{(x_{i_k} - x_{i_0}) \cdot P_{i_1 \ldots i_k}(x_{i_k})}{x_{i_k} - x_{i_0}} = f_{i_k}$$

sowie für $0 < l < k$:

$$q(x_{i_l}) = \frac{(x_{i_l} - x_{i_0}) \cdot P_{i_1 \ldots i_k}(x_{i_l}) - (x_{i_l} - x_{i_k}) \cdot P_{i_0 \ldots i_{k-1}}(x_{i_l})}{x_{i_k} - x_{i_0}}$$
$$= \frac{(x_{i_l} - x_{i_0}) - (x_{i_l} - x_{i_k})}{x_{i_k} - x_{i_0}} \cdot f_{i_l} = f_{i_l}.$$

□

Konvention 6.8. *Zur Abkürzung setzen wir im folgenden*

(6.8a) $P_{i+j,j} \coloneqq P_{i\ i+1 \ldots i+j}$ *für* $0 \le j \le i + j \le n$.

Nach (6.7b) *gilt dann also für* $j > 0$*:*

(6.8b) $$P_{i+j,j}(x) = \frac{(x - x_i) \cdot P_{i+j,j-1}(x) - (x - x_{i+j}) \cdot P_{i+j-1,j-1}(x)}{x_{i+j} - x_i}.$$

Bemerkung 6.9, Das Neville-Schema- für n=3 :

$P_{0123} = P_{3,3}$ *wird mittels* (6.8b) *nach folgendem Schema berechnet:*

	$k=0$	$k=1$	$k=2$	$k=3$
x_0	$f_0 = P_{0,0}$			
		$P_{1,1}$		
x_1	$f_1 = P_{1,0}$		$P_{2,2}$	
		$P_{2,1}$		$P_{3,3}$
x_2	$f_2 = P_{2,0}$		$P_{3,2}$	
		$P_{3,1}$		
x_3	$f_3 = P_{3,0}$			

Beispiel: Gegeben seien die Stützpunkte

$(x_0, f_0) = (1, 2)$, $(x_1, f_1) = (2, 1)$, $(x_2, f_2) = (3, 2)$.
Gesucht ist $P \in \prod_2$ mit $P(x_i) = f_i$ für $0 \le i \le 2$. Wir erhalten - mittels (6.8*b*) :

$$\begin{aligned} P_{1,1}(x) &= \frac{(x - x_0) \cdot f_1 - (x - x_1) \cdot f_0}{x_1 - x_0} \\ &= (x-1) \cdot 1 - (x-2) \cdot 2 = -x + 3 \quad ; \end{aligned}$$

$$\begin{aligned} P_{2,1}(x) &= \frac{(x - x_1) \cdot f_2 - (x - x_2) \cdot f_1}{x_2 - x_1} \\ &= (x-2) \cdot 2 - (x-3) \cdot 1 = x - 1 \quad ; \end{aligned}$$

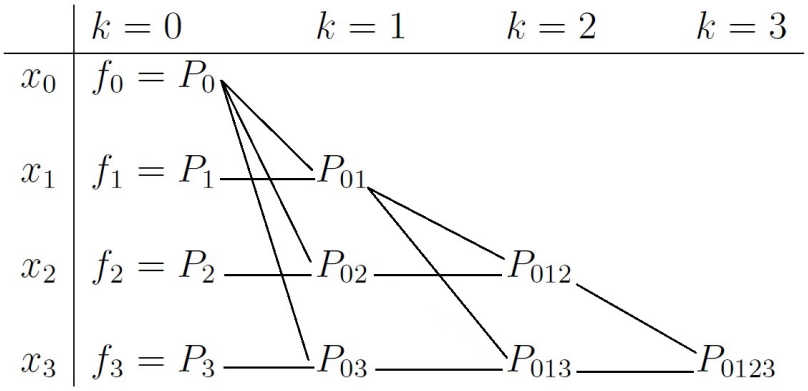
$$\begin{aligned} P_{2,2}(x) &= \frac{(x - x_0) \cdot P_{2,1}(x) - (x - x_2) \cdot P_{1,1}(x)}{x_2 - x_0} \\ &= \frac{1}{2}\left((x-1) \cdot (x-1) - (x-3) \cdot (-x+3)\right) \\ &= \frac{1}{2} \cdot (x^2 - 2x + 1 + x^2 - 6x + 9) \\ &= x^2 - 4x + 5. \end{aligned}$$

Vor der Formulierung des nahe verwandten Aitken - Schemas bemerken wir zunächst:

Bemerkung 6.10. *Sind $i_0, \ldots, i_k \in \{0, \ldots, n\}$ paarweise verschieden, so gilt für jede Bijektion σ : $\{i_0, \ldots, i_k\} \to \{i_0, \ldots, i_k\}$:*

(6.10) $$P_{i_0 \ldots i_k} = P_{\sigma(i_0) \ldots \sigma(i_k)}.$$

Bemerkung 6.11, Das Aitken-Schema - für n=3 :
P_{0123} wird mittels (6.7*b*) *- und* (6.10) *- nach folgendem Schema berechnet:*



Definition 6.12. *Für $0 \le i \le i + k \le n$ ist die k-te dividierte Differenz $f[x_i, x_{i+1}, \ldots, x_{i+k}]$ rekursiv definiert durch:*

(6.12a) $$f[x_i] := f_i,$$

(6.12b) $$f[x_i, \ldots, x_{i+k}] := \frac{f[x_{i+1}, \ldots, x_{i+k}] - f[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}$$
für $0 \le i < i+k \le n$.

Beispiel:

Es ist $f[x_i, x_{i+1}] = \dfrac{f_{i+1} - f_i}{x_{i+1} - x_i}$ für $0 \le i < n$.

Satz 6.13, Die Newton - Darstellung:
Für alle $x \in \mathbb{R}$ *gilt*:

(6.13) $$P_{0\ldots n}(x) = f[x_0] + \sum_{j=1}^{n} f[x_0, \ldots, x_j] \cdot \left(\prod_{k=0}^{j-1} (x - x_k) \right).$$

Beweisskizze - durch Induktion nach n:

· Für $n = 0$ ist (6.13) trivial.

· Für $n \ge 0$ und $0 \le j \le n$ gilt:

$$P_{0\ldots n}(x_j) = f_j = P_{0\ldots n\ n+1}(x_j).$$

Das bedeutet:

$$\exists a \in \mathbb{R}\ \ \forall x \in \mathbb{R}: \quad P_{0\ldots nn+1}(x) = P_{0\ldots n}(x) + a \cdot (x - x_0) \cdot (x - x_1) \cdots (x - x_n).$$

· a ist der Koeffizient von x^{n+1} des Polynoms $P_{0\ldots nn+1}$.

· Aus (6.7a),(6.7b),(6.12a) und (6.12b) folgt
- induktiv - durch Koeffizientenvergleich:

$$a = f[x_0, x_1, \ldots, x_{n+1}].$$

□

Bemerkung 6.14. *Setzen wir* $a_j := f[x_0, \ldots, x_j]$ *für* $0 \le j \le n$, *so gilt gemäß* (6.13) *für alle* $x \in \mathbb{R}$:

(6.14*a*) $$P(x) = a_0 + a_1(x - x_0) + a_2 \cdot (x - x_0) \cdot (x - x_1) + \cdots + a_n \cdot (x - x_0) \cdots (x - x_{n-1}).$$

Ähnlich wie beim Horner - Algorithmus folgt weiter:

(6.14*b*) $$P(x) = (\ldots(a_n \cdot (x - x_{n-1}) + a_{n-1}) \cdot (x - x_{n-2}) + \cdots + a_1) \cdot (x - x_0) + a_0.$$

Konvention 6.15. *Für* $a_0, \ldots, a_m \in \mathbb{R}$ *bezeichne* $I[a_0, \ldots, a_m] \subseteq \mathbb{R}$ *das kleinste Intervall, das* $a_0, \ldots, a_m$ *enthält.*

Satz 6.16. ***Das Restglied bei der Polynom - Interpolation:***
Es seien $x_0, \ldots, x_n, \bar{x} \in \mathbb{R}$ *mit* $x_i \ne x_j$ *für* $0 \le i < j \le n$, *sei* $I := I[x_0, \ldots, x_n, \bar{x}]$, *und* $f: \ I \to \mathbb{R}$ *sei* $n+1$ *mal differenzierbar. Sei* $P = P_{0\ldots n} \in \prod_n$ *die Lösung der Lagrange'schen Interpolationsaufgabe*

(6.16a) $P(x_i) = f(x_i)$ *für* $0 \le i \le n$,

und definiere $\omega \in \prod_{n+1}$ *durch*

(6.16b) $\omega(x) \coloneqq \prod_{j=0}^{n} (x - x_j)$.

Dann gibt es ein $\xi \in I$ *mit:*

(6.16c) $$f(\bar{x}) - P(\bar{x}) = \omega(\bar{x}) \cdot \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Beweis. (6.16*c*) ist klar im Falle $\bar{x} \in \{x_0, \ldots, x_n\}$; gelte also:

$$\bar{x} \neq x_j \text{ für alle } j \text{ mit } 0 \le j \le n.$$

Definiere $F: \ I \to \mathbb{R}$ durch

$(\star)$ $$F(x) \coloneqq f(x) - P(x) - \frac{f(\bar{x}) - P(\bar{x})}{\omega(\bar{x})} \cdot \omega(x).$$

F besitzt in I mindestens die - paarweise verschiedenen $n+2$ Nullstellen $x_0, \ldots, x_n, \bar{x}$.
$(n+1)$ - fache Anwendung des Satzes von Rolle (siehe Satz 10.14 im WS 2015/16) liefert:
Für $1 \le j \le n+1$ besitzt $F^{(j)}$ in I mindestens $n+2-j$ Nullstellen. Insbesondere besitzt $F^{(n+1)}$ eine Nullstelle $\xi \in I$.
Wegen $P^{(n+1)} \equiv 0$ und $\omega^{(n+1)} \equiv (n+1)!$ folgt damit aus $(\star)$:

$$0 = f^{(n+1)}(\xi) - \frac{f(\bar{x}) - P(\bar{x})}{\omega(\bar{x})} \cdot (n+1)!$$

Äquivalenzumformung liefert (6.16*c*). □

Bemerkungen 6.17. *i) Außerhalb von* $I[x_0, \ldots, x_n]$ *wächst* $|\omega(x)|$ *stark an. Bei der Verwendung von* $P_{0\ldots n}$ *zur Approximation von* f *außerhalb von* $I[x_0, \ldots, x_n]$ *ist daher ein großer Fehler zu erwarten; man spricht dann von Extrapolation.*

ii) Für $m \in \mathbb{N}$ *sei*

$$\triangle_m = \{a = x_0^{(m)} < x_1^{(m)} < \cdots < x_m^{(m)} = b\}$$

eine Einteilung eines gegebenen Intervalls $[a, b] \subsetneq \mathbb{R}$ *in* m *Teilintervalle.*
Für eine gegebene Funktion $f : [a, b] \to \mathbb{R}$*, die überall stetig ist, bezeichne* $P_{\triangle_m} \in \prod_m$ *die Lösung der Interpolationsaufgabe*

$$P_{\triangle_m}(x_i^{(m)}) = f(x_i^{(m)}) \quad \textit{für } 0 \le i \le m.$$

Es gilt dann der

Satz von Faber:
Zu jeder Folge von Intervalleinteilungen $(\triangle_m)_{m\in\mathbb{N}}$ *von* $[a, b]$ *in jeweils* m *Teilintervalle gibt es eine stetige Funktion* $f: \ [a, b] \to \mathbb{R}$*, sodass die Folge der Polynome* $(P_{\triangle_m})_{m\in\mathbb{N}}$ *nicht gleichmäßig gegen* f *konvergiert.*

§7 Spline Interpolation

Konventionen 7.1. *Im folgenden seien $a, b \in \mathbb{R}$ mit $a < b$, und für ein $n \in \mathbb{N}$ seien $n+1$ Knoten $x_0, x_1, \ldots, x_n$ gegeben mit*

(7.1a) $$a = x_0 < x_1 \cdots < x_n = b.$$

Setze

(7.1b) $$T \coloneqq \{x_0, x_1, \ldots, x_n\}.$$

Schließlich sei

(7.1c) $$I_\nu \coloneqq [x_{\nu-1}, x_\nu] \text{ für } 1 \leq \nu \leq n.$$

Definition 7.2. *Sei $m \in \mathbb{N}$. Eine Funktion $s : [a, b] \to \mathbb{R}$ heißt Spline-Funktion (oder Spline) vom Grad m zur Knotenmenge T, wenn gilt:*

(S1) *s ist $m-1$ mal stetig differenzierbar.*

(S2) *Auf jedem Teilintervall $I_\nu, 1 \leq \nu \leq n$, stimmt s mit einem Polynom $P_\nu \in \prod_m$ überein.*

$S_{m,n}(T)$ bezeichne die Menge der Splines vom Grad m zur Knotenmenge T.

Beispiele 7.3:

i) Die Splinefunktionen vom Grad 1 sind genau die Streckenzüge mit Anfangspunkt $(x_0, s(x_0))$, Endpunkt $(x_n, s(x_n))$ und Eckpunkten $(x_1, s(x_1)), \ldots, (x_{n-1}, s(x_{n-1}))$. s ist durch Vorgabe der Werte $s(x_0), s(x_1), \ldots, s(x_n)$ eindeutig bestimmt.

ii) Die Splines vom Grad 3 heißen kubische Splines.

Konvention 7.4. *Für $t \in \mathbb{R}$ und $k \in \mathbb{N}$ setzen wir*

(7.4) $$(t)_+^k \coloneqq \begin{cases} t^k & \text{für } t > 0 \\ 0 & \text{für } t \leq 0. \end{cases}$$

Satz 7.5. *Seien m, s und die Polynome $P_\nu, 1 \leq \nu \leq n$, wie in Definition 7.2. Dann gilt:*

i) Zu jedem ν mit $1 \leq \nu \leq n-1$ gibt es ein $a_\nu \in \mathbb{R}$, so dass für alle $x \in \mathbb{R}$ gilt:

(7.5a) $$P_{\nu+1}(x) = P_\nu(x) + a_\nu \cdot (x - x_\nu)^m.$$

ii) Für alle $x \in [a, b]$ gilt:

(7.5b) $$s(x) = P_1(x) + \sum_{\nu=1}^{n-1} a_\nu \cdot (x - x_\nu)_+^m.$$

Beweis. i) s ist insbesondere an der Stelle x_ν (mindestens) $m-1$ mal stetig differenzierbar. Das bedeutet:

$$P_{\nu+1}^{(j)}(x_\nu) = P_\nu^{(j)}(x_\nu) \quad \text{für} \quad 0 \leq j \leq m-1.$$

Es gibt daher ein Polynom Q_ν, das die folgende Identität erfüllt:

$$P_{\nu+1}(x) - P_\nu(x) = Q_\nu(x) \cdot (x - x_\nu)^m.$$

Wegen $P_{\nu+1} - P_\nu \in \prod_m$ folgt:

$$Q_\nu \equiv a_\nu \quad \text{für ein } a_\nu \in \mathbb{R}.$$

ii) Sei etwa $x \in [x_{k-1}, x_k]$ für passendes k mit $1 \le k \le n$.
Im Falle $k = 1$ ist $s(x) = P_1(x)$ wie gewünscht.
Für $2 \le k \le n$ liefert wiederholte Anwendung von i):

$$\begin{aligned} s(x) = P_k(x) &= P_{k-1}(x) + a_{k-1} \cdot (x - x_{k-1})^m \\ &= \dots \\ &= P_1(x) + \sum_{\nu=1}^{k-1} a_\nu \cdot (x - x_\nu)^m \\ &= P_1(x) + \sum_{\nu=1}^{n-1} a_\nu \cdot (x - x_\nu)_+^m. \end{aligned}$$

□

Im folgenden werden nur noch kubische Splines betrachtet; es ist also $m = 3$.

Satz und Definition 7.6. *Zu vorgegebenen $f_0, f_1, \dots, f_n \in \mathbb{R}$ gibt es mindestens einen Spline $s \in S_{3,n}(T)$ mit*

(7.6) $$s(x_j) = f_j \quad \textit{für } 0 \le j \le n.$$

s ist durch jeweils eine der folgenden Randbedingungen eindeutig bestimmt:

(I) s ist ein natürlicher Spline; das heißt:

(7.6a) $$s''(a) = s''(b) = 0.$$

(II) s ist ein periodischer Spline; das heißt:

Es ist $f_0 = f_n$ sowie
(7.6b) $$s^{(j)}(a) = s^{(j)}(b) \textit{ für } 0 \le j \le 2.$$

(III) Für vorgegebene $\gamma_1, \gamma_2 \in \mathbb{R}$ gilt:

(7.6c) $$s'(a) = \gamma_1, \quad s'(b) = \gamma_2.$$

Beweis. Gemäß Satz 7.5 machen wir für den gesuchten Spline $s \in S_{3,n}(T)$ den Ansatz

$$s(x) = \sum_{j=0}^{3} \alpha_j \cdot x^j + \sum_{\nu=1}^{n-1} a_\nu \cdot (x - x_\nu)_+^3 \quad \text{für } x \in [a, b]$$

mit $\alpha_0, \ldots, \alpha_3;\ a_1, \ldots, a_{n-1} \in \mathbb{R}$.
Die Bedingungen (7.6) liefern $n+1$ lineare Gleichungen für diese Koeffizienten.
Jede der Bedingungen (I),(II),(III) liefert genau 2 weitere lineare Gleichungen - für diese Koeffizienten.
Im Falle von (II) beachte man dabei , dass die Bedingung $s(a) = s(b)$ auf die - geforderte - Bedingung $f_0 = f_n$ hinausläuft - und nicht auf eine neue unabhängige Bedingung für s. Wir erhalten also ein lineares Gleichungssystem mit n+3 Gleichungen und n+3 Variablen.
Zum Nachweis der Existenz und Eindeutigkeit der Lösung s können wir annehmen, dass ein homogenes lineares Gleichungssystem vorliegt. Das bedeutet insbesondere:

$$f_0 = \cdots = f_n = 0;$$
$$\gamma_1 = \gamma_2 = 0 \ \text{ in Fall (III).}$$

Zu zeigen ist: $\quad s \equiv 0$.
In allen drei Fällen gilt:

$$(*) \qquad s''(a) \cdot s'(a) = s''(b) \cdot s'(b).$$

Wir setzen nun:

$$I := \int_a^b (s''(x))^2 \, \mathrm{d}x.$$

Auf jedem offenen Intervall $(x_{\nu-1}, x_\nu)$, $1 \le \nu \le n$, stimmt s''' mit einer Konstanten c_ν überein.
Damit folgt durch partielle Integration:

$$\begin{aligned} & I \\ =& \sum_{\nu=1}^{n} \int_{x_{\nu-1}}^{x_\nu} (s''(x))^2 \, \mathrm{d}x \\ =& \sum_{\nu=1}^{n} \left(s'' \cdot s' \Big|_{x_{\nu-1}}^{x_\nu} - \int_{x_{\nu-1}}^{x_\nu} c_\nu \cdot s'(x) \, \mathrm{d}x \right) \\ =& s''(b) \cdot s'(b) - s''(a) \cdot s'(a) - \sum_{\nu=1}^{n} c_\nu \cdot (s(x_\nu) - s(x_{\nu-1})) \\ =& 0. \end{aligned}$$

Die letzte Gleichung folgt dabei aus $(*)$ und der Forderung $s(x_j) = f_j = 0$ für $0 \le j \le n$.
Wegen $(s'')^2 \ge 0$ und der Stetigkeit von s'' auf $[a, b]$ folgt weiter: $\quad s'' \equiv 0$.
Das bedeutet: $\quad s \in \prod_1$.
Schließlich folgt nun wegen $s(a) = s(b) = 0$: $\quad s \equiv 0$. □

Beispiel:

Sei $T := \{0, 1, 2\}$. Wir suchen den natürlichen Spline $s \in S_{3,2}(T)$ mit:

$$s(0) = 1 \ \ , \ \ s(1) = 0 \ \ , \ \ s(2) = 11 \ \ ; \ \ s''(0) = s''(2) = 0.$$

Mit dem Ansatz

$$s(x) = \alpha \cdot x^3 + \beta \cdot x^2 + \gamma \cdot x + \delta + a \cdot (x-1)_+^3$$

folgt weiter für alle $x \in [0,2]$:

$$\begin{aligned} s'(x) =& 3\alpha \cdot x^2 + 2\beta \cdot x + \gamma + 3a \cdot (x-1)_+^2 \ , \\ s''(x) =& 6\alpha \cdot x + 2\beta + 6a \cdot (x-1)_+ \ . \end{aligned}$$

Damit erhalten wir folgendes lineares Gleichungssystem:

$$\begin{aligned} & \delta = 1 \ \wedge \ \alpha + \beta + \gamma + \delta = 0 \ \wedge \ 8\alpha + 4\beta + 2\gamma + \delta + a = 11 \\ \wedge \ & \beta = 0 \ \wedge \ 12\alpha + 2\beta + 6a = 0. \end{aligned}$$

Äquivalenzumformung liefert:

$$\begin{aligned} & \beta = 0 \ \wedge \ \delta = 1 \\ \wedge \ & \alpha + \gamma = -1 \ \wedge \ 8\alpha + 2\gamma + 1 + a = 11 \ \wedge \ a = -2\alpha \\ \Leftrightarrow \ & \beta = 0 \ \wedge \ \delta = 1 \ \wedge \ a = -2\alpha \ \wedge \ 6\alpha + 2\gamma = 10 \ \wedge \ \gamma = -1 - \alpha \\ \Leftrightarrow \ & \beta = 0 \ \wedge \ \delta = 1 \ \wedge \ \alpha = 3 \ \wedge \ \gamma = -4 \wedge a = -6 \end{aligned}$$

Die gesuchte Spline - Funktion $s \in S_{3,2}(T)$ ist also gegeben durch:

$$\begin{aligned} s(x) &= 3 \cdot x^3 - 4 \cdot x + 1 - 6 \cdot (x-1)_+^3 \\ &= \begin{cases} 3 \cdot x^3 - 4 \cdot x + 1 & \text{für } \ 0 \le x \le 1 \\ -3 \cdot x^3 + 18x^2 - 22x + 7 & \text{für } \ 1 < x \le 2. \end{cases} \end{aligned}$$

Konvention 7.7. *Für $a, b \in \mathbb{R}$ mit $a < b$ und eine stetige Funktion $g\colon \quad [a,b] \to \mathbb{R}$ setzen wir:*

$$\|g\|_\infty \coloneqq \max\{|g(x)| \ : \quad x \in [a,b]\}. \tag{7.7}$$

Satz 7.8. *Die Funktion $f\colon \ [a,b] \to \mathbb{R}$ sei 4 mal stetig differenzierbar. Weiter sei $f \in S_{3,n}(T)$ der - nach Satz und Definition 7.6 (III) - eindeutig bestimmte kubische Spline mit*

$$\begin{aligned} s(x_\nu) = f(x_\nu) \quad & \textit{für } 0 \le \nu \le n \ , \\ s'(a) = f'(a) \ \ , \ \ & s'(b) = f'(b). \end{aligned}$$

Weiter setze

$$h \coloneqq \max_{1 \le \nu \le n} (x_\nu - x_{\nu-1}).$$

Dann gilt:

$$\|f - s\|_\infty \le \frac{5}{384} \cdot h^4 \cdot \|f^{(4)}\|_\infty. \tag{7.8}$$

Beweisidee: Für $1 \leq \nu \leq n$ sei $u_\nu \in \prod_3$ die eindeutig bestimmte Lösung der Interpolationsaufgabe

$$\begin{aligned} u_\nu(x_{\nu-1}) &= f(x_{\nu-1}) \quad , \quad u_\nu(x_\nu) = f(x_\nu) \; ; \\ u'_\nu(x_{\nu-1}) &= f'(x_{\nu-1}) \quad , \quad u'_\nu(x_\nu) = f'(x_\nu) \; . \end{aligned}$$

Dann ist die Funktion $u\colon \; [a,b] \to \mathbb{R}$, definiert durch

$$u(x) \coloneqq u_\nu(x) \quad , \text{ falls } x \in [x_{\nu-1}, x_\nu]$$

wohldefiniert und stetig differenzierbar.

Relativ leicht zu zeigen ist - bei Übertragung der Idee des Beweises von Satz 6.16:

(7.8a) $$\|f - u\|_\infty \leq \frac{1}{384} \cdot h^4 \cdot \|f^{(4)}\|_\infty .$$

Schwer zu zeigen ist:

(7.8b) $$\|u - s\|_\infty \leq \frac{1}{96} \cdot h^4 \cdot \|f^{(4)}\|_\infty .$$

(7.8) folgt aus (7.8a) und (7.8b).

Beispiel:
Definiere $f\colon \; [0, 2\pi] \to \mathbb{R}$ durch $f(x) \coloneqq \sin x$.
Setze $n = 4, x_0 \coloneqq 0, x_1 \coloneqq \frac{\pi}{2}, x_2 \coloneqq \pi, x_3 \coloneqq \frac{3}{2} \cdot \pi, x_4 \coloneqq 2\pi$.
Für den zugehörigen Spline $s \in S_{3,n}(T)$ mit

$$s(x_\nu) = \sin x_\nu \;\; \text{für } 0 \leq \nu \leq 4 \; , \quad s'(0) = s'(2\pi) = 1$$

erhalten wir mit $h = \frac{\pi}{2}$ mittels (7.8) die folgende Abschätzung:

$$\|f - s\|_\infty \leq \frac{5}{384} \cdot \left(\frac{\pi}{2}\right)^4 \cdot 1 < 0,08 .$$

§8 Nullstellen - Bestimmung durch Iterationsverfahren

Problemstellung 8.1:
Gegeben sei eine stetige Funktion $f: I \to \mathbb{R}$ auf einem Intervall I. Bestimme ein (oder mehrere oder alle) $\xi \in I$ mit $f(\xi) = 0$.

Ansatz 8.2:
Ausgehend von einem Startwert $x_0 \in I$ berechne weitere Näherungswerte $x_n, n \in \mathbb{N}$, für ξ mit Hilfe einer Iterationsfunktion $\Phi: I \to I$; das heißt es soll gelten:

$$x_n := \Phi(x_{n-1}) \text{ für } n \in \mathbb{N}. \tag{8.2}$$

An Φ stellen wir die folgenden Minimalforderungen:

(I) Für alle $x \in I$ mit $\Phi(x) = x$ ist $f(x) = 0$; das heißt: Jeder Fixpunkt von Φ ist Nullstelle von f.

(II) Φ ist stetig.

Lemma 8.3. *Es seien (I) und (II) erfüllt, und für festes $x_0 \in I$ sei die Folge $(x_n)_{n\geq 0}$ rekursiv durch (8.2) gegeben.*
Besitzt diese Folge einen Grenzwert $\xi_0 \in I$, so gilt:

$$f(\xi_0) = 0.$$

Beweis. Aus (II) und (8.2) folgt:

$$\xi_0 = \lim_{n\to\infty} x_n = \lim_{n\to\infty} \Phi(x_{n-1}) = \Phi(\lim_{n\to\infty} x_{n-1}) = \Phi(\xi_0).$$

Aus (I) folgt daher: $f(\xi_0) = 0$. □

Beispiel 8.4:
Zur Bestimmung der kleinsten positiven Nullstelle der Cosinus-Funktion (also $\frac{\pi}{2}$) machen wir folgenden Ansatz:
Definiere $f, \Phi: \mathbb{R} \to \mathbb{R}$ durch

$$f(x) := \cos x \quad , \quad \Phi(x) := x + \cos x.$$

Die Minimalanforderungen (I) und (II) sind erfüllt; insbesondere gilt für $x \in \mathbb{R}$ folgende Äquivalenz:

$$\Phi(x) = x \quad \Leftrightarrow \quad f(x) = 0.$$

Mit dem Startwert $x_0 = 1$ erhalten wir mittels (8.2) folgende Näherungswerte für $\frac{\pi}{2}$:

$$\begin{aligned} x_1 &= 1 + \cos 1 \approx 1,540302306 \ , \\ x_2 &= x_1 + \cos x_1 \approx 1,570791601 \ , \\ x_3 &= x_2 + \cos x_2 \approx 1,570796327 \ , \end{aligned}$$

$$x_n = x_{n-1} + \cos x_{n-1} \approx x_3 \text{ für } n \geq 4.$$

Dabei ist $x_3 = \frac{\pi}{2} + \theta$ mit $|\theta| < 10^{-9}$.

Definition 8.5. *Eine Folge $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R}$ heißt Cauchy-Folge, falls gilt:*

Zu jedem $\epsilon > 0$ gibt es ein $N = N(\epsilon) \in \mathbb{N}$, so dass für alle $n, m \in \mathbb{N}$ mit $n, m \geq N$ gilt: $|x_n - x_m| < \epsilon$.

Bemerkung 8.6. *Eine Folge $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R}$ ist genau dann eine Cauchy-Folge, wenn sie konvergiert.*
Ist nämlich $x = \lim_{n\to\infty} x_n$, so gibt es zu vorgegebenem $\epsilon > 0$ ein $N = N(\epsilon) \in \mathbb{N}$, so dass für alle $n \in \mathbb{N}$ mit $n \geq N$ gilt: $|x_n - x| < \frac{\epsilon}{2}$.
Dann folgt für alle $n, m \geq N$:

$$|x_n - x_m| \leq |x_n - x| + |x - x_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Somit ist jede konvergente Folge eine Cauchy-Folge.
Zum Beweis der Umkehrung siehe Übung Nr.41.

Warnung:
Eine Cauchy-Folge mit lauter rationalen Gliedern muss ihren Grenzwert nicht in $\mathbb{Q}$ haben. Beispielsweise ist die irrationale Zahl $\sqrt{2}$ Grenzwert einer Folge von rationalen Zahlen.

Definition 8.7. *Es seien $a, b \in \mathbb{R}$ mit $a < b$. Eine Abbildung $g : [a, b] \to [a, b]$ heißt kontrahierend, falls ein $L \in \mathbb{R}$ mit $0 \leq L < 1$ existiert, so dass für alle $x, y \in [a, b]$ gilt:*

$$|g(x) - g(y)| \leq L \cdot |x - y|. \tag{8.7}$$

Bemerkung 8.8. *Jede kontrahierende Abbildung $g : [a, b] \to [a, b]$ ist stetig.*
Umgekehrt ist jedoch etwa die Abbildung $f : [0, 1] \to [0, 1]$, gegeben durch $f(x) := x^2$ zwar stetig, aber nicht kontrahierend.

Satz 8.9. ***Der Banachsche Fixpunktsatz:***
Seien $a, b \in \mathbb{R}$ mit $a < b$, und die Abbildung $g : [a, b] \to [a, b]$ sei kontrahierend; sei also $L \in [0, 1)$, so dass (8.7) für alle $x, y \in [a, b]$ gilt. Dann besitzt g genau einen Fixpunkt ξ_0. Genauer gilt:
Sei $x_0 \in [a, b]$ beliebig, und definiere die Folge $(x_n)_{n\geq 0}$ rekursiv durch $x_{n+1} := g(x_n)$ für $n \geq 0$. Dann gilt für alle $n \geq 1$:

$$|x_n - \xi_0| \leq \tfrac{L^n}{1-L} \cdot |x_1 - x_0|. \tag{8.9a}$$

Insbesondere folgt:

$$\lim_{n\to\infty} x_n = \xi_0. \tag{8.9b}$$

Beweis. Wir zeigen zunächst, dass g höchstens einen Fixpunkt besitzt.
Seien also $\xi_1, \xi_2 \in [a, b]$ mit $g(\xi_1) = \xi_1$ und $g(\xi_2) = \xi_2$. Zu zeigen ist: $\xi_1 = \xi_2$.
(8.7) liefert:

$$|\xi_1 - \xi_2| = |g(\xi_1) - g(\xi_2)| \leq L \cdot |\xi_1 - \xi_2|.$$

Wegen $L < 1$ ist das nur möglich, wenn $|\xi_1 - \xi_2| = 0$ - und damit $\xi_1 = \xi_2$ ist.

Zum Nachweis der Existenz sei nun die Folge $(x_n)_{n\geq 0}$ wie im Satz. Wir zeigen nun:

(8.9c) $$|x_n - x_{n-1}| \leq L^{n-1} \cdot |x_1 - x_0| \quad \text{für alle } n \in \mathbb{N} \ ,$$

(8.9d) $$|x_{n+t} - x_n| \leq \frac{L^n}{1-L} \cdot |x_1 - x_0| \quad \text{für alle } n, t \in \mathbb{N}.$$

(8.9c) folgt sofort durch Induktion:
(8.9c) ist trivial für $n = 1$.
Im Induktionsschritt folgt mittels (8.7):

$$|x_{n+1} - x_n| = |g(x_n) - g(x_{n-1})| \leq L \cdot |x_n - x_{n-1}| \leq L^n \cdot |x_1 - x_0|.$$

Nun folgt weiter für alle $n, t \in \mathbb{N}$ - mittels der Formel für die endliche geometrische Reihe:

$$\begin{aligned}
|x_{n+t} - x_n| &= \left| \sum_{j=0}^{t-1} (x_{n+j+1} - x_{n+j}) \right| \\
&\leq \sum_{j=0}^{t-1} |x_{n+j+1} - x_{n+j}| \leq \sum_{j=0}^{t-1} L^{n+j} \cdot |x_1 - x_0| \\
&= (L^n + L^{n+1} + \cdots + L^{n+t-1}) \cdot |x_1 - x_0| \\
&= \frac{L^n - L^{n+t}}{1-L} \cdot |x_1 - x_0| \\
&\leq \frac{L^n}{1-L} \cdot |x_1 - x_0|.
\end{aligned}$$

Damit ist auch (8.9d) bewiesen.
Sei nun $\epsilon > 0$. Dann gibt es ein $N = N(\epsilon) \in \mathbb{N}$ mit:

$$\frac{L^N}{1-L} \cdot |x_1 - x_0| < \epsilon.$$

(8.9d) liefert nun auch, dass für alle $n, m \in \mathbb{N}$ mit $m > n \geq N$ (und $t := m - n$) gilt:

$$|x_m - x_n| \leq \frac{L^n}{1-L} \cdot |x_1 - x_0| < \epsilon.$$

Das bedeutet: Die Folge $(x_n)_{n\in\mathbb{N}}$ ist eine Cauchy-Folge in $[a, b]$- und besitzt daher nach Bemerkung 8.6 einen Grenzwert $\xi \in [a, b]$. (8.9d) liefert weiter durch Grenzübergang - für $t \to \infty$:

$$|\xi - x_n| \leq \frac{L^n}{1-L} \cdot |x_1 - x_0| \quad \text{für alle } n \in \mathbb{N}.$$

Damit folgen (8.9a) und (8.9b) für $\xi_0 := \xi$. Schließlich folgt aus der Stetigkeit von g:

$$g(\xi_0) = g(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} g(x_n) = \lim_{n\to\infty} x_{n+1} = \xi_0.$$

ξ_0 ist also Fixpunkt von g - und aufgrund der eingangs gezeigten Eindeutigkeit unabhängig von dem speziell gewählten Anfangswert $x_0 \in [a,b]$. □

Definition 8.10. ***Das Newton - Verfahren:***
Sei $I \subseteq \mathbb{R}$ ein Intervall, und $f: I \to \mathbb{R}$ sei stetig differenzierbar mit $f'(x) \neq 0$ für alle $x \in I$.
Definiere $\Phi: I \to \mathbb{R}$ durch

$$\Phi(x) := x - \frac{f(x)}{f'(x)}. \tag{8.10}$$

Ist $\Phi(I) \subseteq I$, so heißt das durch die Iterationsfunktion Φ gegebene Iterationsverfahren das Newton-Verfahren zur Bestimmung einer Nullstelle ξ von f.

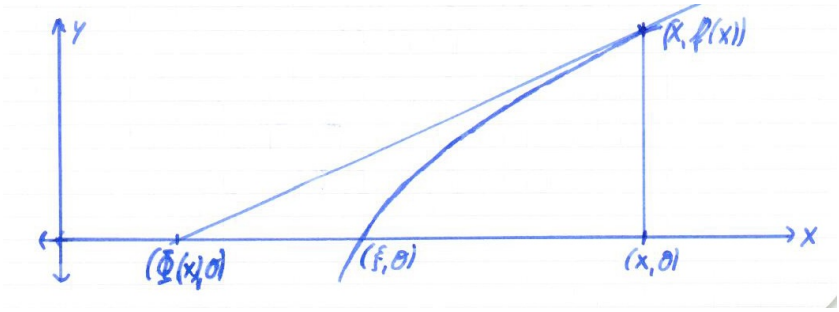
Bemerkung 8.11. *i) Seien f, Φ wie in Definition 8.10 - mit $\Phi(I) \subseteq I$.*
Dann erfüllt Φ auch die geforderten Minimalforderungen (I) und (II).

ii) (8.10) besagt auch:

$$\frac{f(x)}{x - \Phi(x)} = f'(x). \tag{8.11}$$

Das bedeutet:
Der Punkt $(\Phi(x), 0)$ ist der Schnittpunkt der x-Achse mit der Tangente an f durch den Punkt $(x, f(x))$.

Skizze:



Satz 8.12. *Sei $f: I \to \mathbb{R}$ zweimal stetig differenzierbar auf einem Intervall $I \subseteq \mathbb{R}$, sei $f'(x) \neq 0$ für alle $x \in I$, und sei $\xi \in I$ eine Nullstelle von f, die kein Randpunkt von I ist. Dann gibt es ein $\delta > 0$ mit $[\xi - \delta, \xi + \delta] \subseteq I$, so dass für Φ wie in (8.10) folgt:*

i) Es gibt ein $L \in [0,1)$, so dass für alle $x \in [\xi - \delta, \xi + \delta]$ gilt:

$$|\Phi'(x)| \leq L. \tag{8.12a}$$

ii) Es ist $\Phi([\xi - \delta, \xi + \delta]) \subseteq [\xi - \delta, \xi + \delta]$.

iii) Für jeden Startwert $x_0 \in [\xi - \delta, \xi + \delta]$ folgt für die durch (8.2) gegebene Folge $(x_n)_{n\geq 0}$:

(8.12b) $\lim_{n\to\infty} x_n = \xi.$

Beweis. i) Weil f auf I zweimal stetig differenzierbar ist, ist Φ auf I (mindestens) einmal stetig differenzierbar, und für alle $x \in I$ folgt:

$$\Phi'(x) = 1 - \frac{f'(x)^2 - f(x)\cdot f''(x)}{f'(x)^2} = \frac{f(x)\cdot f''(x)}{f'(x)^2}.$$

Insbesondere ist $\Phi'(\xi) = 0$. Weil Φ' stetig ist, gibt es daher sogar zu vorgegebenem $L \in (0,1)$ ein $\delta > 0$ mit $[\xi - \delta, \xi + \delta] \subseteq I$, so dass für alle $x \in [\xi - \delta, \xi + \delta]$ gilt:

$$|\Phi'(x)| \leq L.$$

ii) Es seien $x, y \in [\xi - \delta, \xi + \delta]$. Dann gibt es nach dem Mittelwertsatz ein $\eta \in I\,[x, y]$ mit:

$$\Phi(x) - \Phi(y) = \Phi'(\eta)\cdot(x - y).$$

i) liefert damit weiter für alle $x, y \in [\xi - \delta, \xi + \delta]$:

(8.12c) $|\Phi(x) - \Phi(y)| \leq L \cdot |x - y|.$

Insbesondere folgt wegen $\Phi(\xi) = \xi$:

$$|\Phi(x) - \xi| = |\Phi(x) - \Phi(\xi)| \leq |x - \xi| \text{ für alle } x \in [\xi - \delta, \xi + \delta]\,.$$

Das bedeutet: $\Phi([\xi - \delta, \xi + \delta]) \subseteq [\xi - \delta, \xi + \delta]\,.$

iii) Nach (8.12c) ist Φ auf dem abgeschlossenen Intervall $[\xi - \delta, \xi + \delta]$ kontrahierend. Somit folgt aus dem Banachschen Fixpunktsatz, angewandt auf die Restriktion g von Φ auf $[\xi - \delta, \xi + \delta]$:

ξ ist der einzige Fixpunkt von g, und (8.12b) gilt.

□

Beispiel:
Wir berechnen numerisch die - eindeutig bestimmte - Nullstelle der - streng monoton steigenden - Funktion $f : \mathbb{R} \to \mathbb{R}$, gegeben durch

$$f(x) := x^3 + 3x + 1.$$

Für alle $x \in \mathbb{R}$ ist dann

$$f'(x) = 3x^2 + 3 > 0,$$

und die durch (8.10) gegebene Iterationsfunktion $\Phi : \mathbb{R} \to \mathbb{R}$ ist gegeben durch:

$$\Phi(x) = x - \frac{x^3 + 3x + 1}{3x^2 + 3}.$$

Mit dem Startwert $x_0 = 0$ erhalten wir weiter:

$$x_1 = -\frac{1}{3} \quad , \quad x_2 = -\frac{29}{90} = -0,3\bar{2} \quad ,$$
$$x_3 \approx -0,322185355 \ ,$$
$$x_n \approx x_3 \quad \text{für } n \geq 4.$$

Definition 8.13. ***Die Sekantenmethode:***
Für eine stetige und injektive Funktion $f : I \to \mathbb{R}$ auf einem Intervall $I \subseteq \mathbb{R}$ betrachte folgendes Iterationsverfahren zur Bestimmung einer Nullstelle ξ von f:
Wähle zwei verschiedene Startwerte $x_0, x_1 \in I$.
Sei $n \geq 1$, und seien $x_0, x_1, \ldots, x_n$ bereits bestimmt.
Stopp, falls $x_n \notin I$.
Setze $x_m := x_n$ für alle $m > n$, falls $x_{n-1} = x_n$.
Andernfalls setze

$$x_{n+1} := x_n - f(x_n) \cdot \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} = \frac{x_{n-1} \cdot f(x_n) - x_n \cdot f(x_{n-1})}{f(x_n) - f(x_{n-1})}. \tag{8.13}$$

Dieses Rekursionsverfahren heißt die Sekantenmethode.

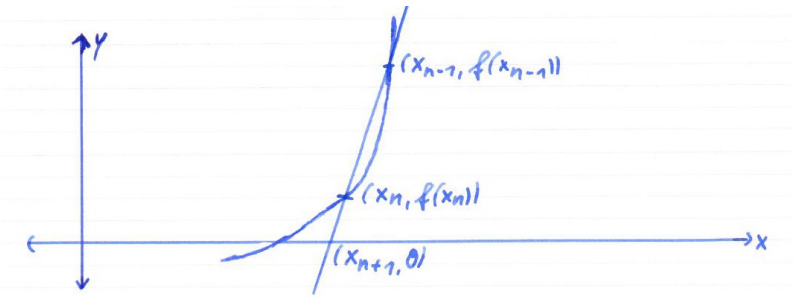
Bemerkungen 8.14. *i) Ist $x_{n+1} = x_n$ für ein $n \geq 1$ und ist n minimal mit dieser Eigenschaft, so ist $\xi := x_n$ nach (8.13) eine Nullstelle von f.*

ii) (8.13) besagt auch:

$$\frac{f(x_n)}{x_n - x_{n+1}} = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}. \tag{8.14}$$

Das bedeutet:
Der Punkt $(x_{n+1}, 0)$ ist der Schnittpunkt der x-Achse mit der Sekante zu f durch die Punkte $(x_{n-1}, f(x_{n-1}))$ und $(x_n, f(x_n))$.

Skizze:



iii) Bei der Regula falsi wird die Sekantenmethode dahingehend modifiziert, dass $f(x_{n-1})$ und $f(x_n)$ stets unterschiedliche Vorzeichen aufweisen - solange $f(x_n) \neq 0$ ist.

Das hat - unter anderem - den Vorteil, dass x_{n+1} stets zwischen x_{n-1} und x_n und damit auch in dem gegebenen Intervall I liegt.
Insbesondere sind die Startwerte $x_0, x_1 \in I$ so zu wählen, dass $f(x_0) \cdot f(x_1) \leq 0$ ist.

In Analogie zu Satz 8.12 gilt folgendes- allerdings schwieriger zu beweisendes - Ergebnis:

Satz 8.15. *Sei $f: I \to \mathbb{R}$ zweimal stetig differenzierbar auf einem Intervall $I \subseteq \mathbb{R}$, sei $f'(x) \neq 0$ für alle $x \in I$, und sei $\xi \in I$ eine Nullstelle von f, die kein Randpunkt von I ist. Dann gibt es ein $\delta > 0$, so dass für alle $x_0, x_1 \in [\xi - \delta, \xi + \delta]$ mit $x_0 \neq x_1$ die Sekantenmethode eine Folge $(x_n)_{n \geq 0}$ liefert mit $\lim_{n \to \infty} x_n = \xi$.*

§9 Fehleranalyse Linearer Gleichungssysteme

Aufgabenstellung 9.1:
(siehe auch Bemerkung 6.37 im WS 2015/16)
Gegeben sei folgendes lineares Gleichungssystem über $\mathbb{R}$ - mit n Variablen $x_1, \ldots, x_n$ und n Gleichungen:

$$\begin{aligned}
&(9.1.1) \qquad a_{11} \cdot x_1 + \cdots + a_{1n} \cdot x_n = b_1 \quad , \\
&\quad \vdots \\
&(9.1.i) \qquad a_{i1} \cdot x_1 + \cdots + a_{in} \cdot x_n = b_i \quad , \\
&\quad \vdots \\
&(9.1.n) \qquad a_{n1} \cdot x_1 + \cdots + a_{nn} \cdot x_n = b_n \quad .
\end{aligned}$$

Mit $A \coloneqq (a_{ij})_{1 \le i,j \le n} \in \mathbf{Mat}_{n \times n}(\mathbb{R})$ und
$b \coloneqq (b_1, \ldots, b_n)^T \in \mathbb{R}^n$, $x \coloneqq (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ können wir auch schreiben:

(9.1) $\qquad A \cdot x = b.$

Wir nehmen im folgenden an: $\det A \neq 0$.
Dann gibt es genau eine Lösung von (9.1), nämlich

(9.1') $\qquad x = A^{-1} \cdot b.$

Zur praktischen Lösung von (9.1) liegt folgender Ansatz nahe:

Eliminiere x_1 mittels (9.1.1) aus (9.1.2)-(9.1.n),
eliminiere x_2 mittels (9.1.2) aus (9.1.3) - (9.1.n),

$\vdots$

eliminiere x_{n-1} mittels (9.1.n-1) aus (9.1.n),
berechne x_n aus (9.1.n),
berechne x_{n-1} aus (9.1.n-1),

$\vdots$

berechne x_1 aus (9.1.1).

Dieses Gaußsche Eliminationsverfahren funktioniert immer unter der Voraussetzung $\det A \neq 0$, wobei zusätzlich zwischen einzelnen Eliminationsschritten eventuell noch gewisse Gleichungen zu vertauschen sind.

Im folgenden soll das Gaußsche Eliminationsverfahren dadurch analysiert werden , dass A als Produkt einfacherer Matrizen geschrieben wird.

Definition 9.2. *i) Eine Matrix $L = (l_{ij})_{1 \le i,j \le n} \in \mathbf{Mat}_{n \times n}(\mathbb{R})$ heißt eine untere Dreiecksmatrix, wenn für alle i, j mit $1 \le i < j \le n$ gilt: $l_{ij} = 0$.*

L hat dann also folgende Gestalt:

$$L = \begin{pmatrix} * & & & \\ & * & & 0 \\ & & \ddots & \\ * & & & * \end{pmatrix}.$$

ii) Eine Matrix $R = (r_{ij})_{1 \le i,j \le n} \in \mathbf{Mat}_{n\times n}(\mathbb{R})$ *heißt eine <u>obere Dreiecksmatrix</u>, wenn für alle* i, j *mit* $1 \le j < i \le n$ *gilt:* $r_{ij} = 0$.
R hat dann folgende Gestalt:

$$R = \begin{pmatrix} * & & & \\ & * & & * \\ & & \ddots & \\ 0 & & & * \end{pmatrix}.$$

iii) Eine Matrix $P \in \mathbf{Mat}_{n\times n}(\mathbb{R})$ *heißt eine <u>Permutationsmatrix</u>, wenn P aus der Einheitsmatrix entsteht, indem gewisse Zeilen - oder Spalten - permutiert werden.*

Beispiele für Permutationsmatrizen:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Bemerkungen 9.3. *i) Das Produkt zweier unterer (bzw. oberer) Dreiecksmatrizen ist wieder eine untere (bzw. obere) Dreiecksmatrix.*
Für gegebene Koeffizienten $l_{ij} \in \mathbb{R}$ *mit* $1 \le j < i \le n$ *setzen wir*

$$L_j := \begin{pmatrix} 1 & & & & \\ & \ddots & & & 0 \\ & & 1 & & \\ & & l_{j+1j} & \ddots & \\ & 0 & \vdots & & 1 \\ & & l_{nj} & & \end{pmatrix} \quad \text{für } 1 \le j \le n-1. \tag{9.3a}$$

Dann folgt:

$$L_1 \cdots L_{n-1} = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & 0 \\ \vdots & & \ddots & & \\ & & & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{nn-1} & 1 \end{pmatrix}. \tag{9.3b}$$

Auf der linken Seite in (9.3b) darf dabei die Reihenfolge der Faktoren nicht vertauscht werden !

ii) Eine (untere oder obere) Dreiecksmatrix ist genau dann invertierbar, wenn alle Koeffizienten der Hauptdiagonalen von 0 verschieden sind; denn das Produkt dieser Koeffizienten ist die Determinante.
Gegebenenfalls ist die zugehörige inverse Matrix wieder eine untere bzw. obere Dreiecksmatrix.
Speziell folgt für L_j wie in (9.3a):

$$L_j^{-1} = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{j+1\,j} & \ddots & \\ & 0 & \vdots & & \\ & & -l_{nj} & & 1 \end{pmatrix} \quad \text{für } 1 \le j \le n-1. \tag{9.3c}$$

Bemerkung 9.4. ***Durchführung des Gaußschen Eliminationsverfahrens mit Hilfe von Dreiecksmatrizen:***
Wir nehmen hier der Einfachheit halber an:
Bei der Durchführung des Gaußschen Eliminationsverfahrens brauchen in (9.1) keine Zeilen (mehr) vertauscht zu werden. Für $1 \le k \le n-1$ erhalten wir nach $k-1$ Eliminationsschritten eine Matrix der Gestalt

$$(A^{(k-1)}, b^{(k-1)}) = \left(\begin{array}{ccc|c} a_{11}^{(k-1)} & * & & \\ & \ddots & * & \\ 0 & & a_{k-1\,k-1}^{(k-1)} & b^{(k-1)} \\ & 0 & B^{(k-1)} & \end{array}\right)$$

ausgehend von $A^{(0)} = A, b^{(0)} = b$.
Dabei gilt nun:

$$\begin{aligned} B^{(k-1)} &= (a_{ij}^{(k-1)})_{k \le i,j \le n} \in \mathbf{Mat}_{(n-k+1)\times(n-k+1)}(\mathbb{R}), \\ b^{(k-1)} &\in \mathbb{R}^n. \end{aligned}$$

Sodann setzen wir:

$$l_{ik} := a_{ik}^{(k-1)} / a_{kk}^{(k-1)} \quad \text{für } k < i \le n. \tag{9.4a}$$

Mit L_k wie in (9.3a) folgt dann - zusammen mit (9.3c):

$$(A^{(k)}, b^{(k)}) := L_k^{-1} \cdot (A^{(k-1)}, b^{(k-1)})$$
$$= \left(\begin{array}{cccc|c} a_{11}^{(k)} & * & & * & \\ & \ddots & & & \\ & 0 & a_{kk}^{(k)} & & b^{(k)} \\ & & & B^{(k)} & \\ & & 0 & & \end{array}\right).$$

für passende Koeffizienten $a_{ij}^{(k)} \in \mathbb{R}$, $B^{(k)} \in \mathbf{Mat}_{(n-k)\times(n-k)}(\mathbb{R})$, $b^{(k)} \in \mathbb{R}^n$.

Am Ende erhalten wir eine obere Dreiecksmatrix

$$R := A^{(n-1)} = \begin{pmatrix} a_{11}^{(n)} & \cdots & a_{1n}^{(n)} \\ & \ddots & \vdots \\ 0 & & a_{nn}^{(n)} \end{pmatrix}$$

Dabei gilt - laut Durchführung der Eliminationsschritte:

$$L_{n-1}^{-1} \cdot L_{n-2}^{-1} \cdots L_2^{-1} \cdot L_1^{-1} \cdot A = R.$$

Nach Bemerkung 9.3 ist mit den Matrizen $L_1, \ldots, L_{n-1}$ auch

$$L := (L_{n-1}^{-1} \cdots L_1^{-1})^{-1} = L_1 \cdots L_{n-1}$$

eine untere Dreiecksmatrix, und es folgt:

$$A = L \cdot R.$$

Berücksichtigen wir auch den Fall, dass vor bzw. während des Eliminationsverfahrens eventuell noch Zeilen zu vertauschen sind, so erhalten wir allgemein:

Satz und Definition 9.5. *Zu der gegebenen $n \times n$- Matrix A über $\mathbb{R}$ mit $\det A \neq 0$ gibt es eine Permutationsmatrix P, eine untere Dreiecksmatrix L und eine obere Dreiecksmatrix R, so dass gilt:*

$$P \cdot A = L \cdot R. \tag{9.5}$$

Die Produktdarstellung durch die Matrix $L \cdot R$ heißt eine Dreieckszerlegung der Matrix $P \cdot A$

Man beachte dabei:
$P \cdot A$ entsteht aus A, indem die Zeilen der Matrix A permutiert werden.

Bemerkung 9.6. *Ist speziell eine Dreieckszerlegung der Gestalt*

$$A = L \cdot R$$

gegeben, so kann die Lösung eines linearen Gleichungssystems $A \cdot x = b$ wie in (9.1) zurückgeführt werden auf die beiden - einfacher zu lösenden - Dreieckssysteme

(9.6a) $$L \cdot y = b,$$

(9.6b) $$R \cdot x = y.$$

Sind (9.6a) und (9.6b) gelöst, so folgt nämlich:

(9.6c) $$A \cdot x = L \cdot (R \cdot x) = L \cdot y = b.$$

Beispiel 9.7:
Wir lösen folgendes lineare Gleichungssystem:

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 3 \\ 3 & 1 & 6 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 7 \\ 2 \end{pmatrix}.$$

Wir setzen

$$A := \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 3 \\ 3 & 1 & 6 \end{pmatrix} \quad , \quad L_1 := \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

und erhalten:

$$L_1^{-1} \cdot A = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 3 \\ 3 & 1 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & -2 & 3 \end{pmatrix} =: A^{(1)}.$$

Mit $L_2 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}$ folgt dann weiter:

$$L_2^{-1} \cdot A^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & -2 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix} =: R.$$

Mit $L := L_1 \cdot L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix}$ folgt nun:

$$A = L_1 \cdot A^{(1)} = L_1 \cdot L_2 \cdot R = L \cdot R.$$

Wir erhalten also die folgenden beiden Dreieckssysteme:

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 7 \\ 2 \end{pmatrix} \quad , \quad \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}.$$

Die Lösung des ersten linearen Gleichungssystems ist gegeben durch:

$$\begin{aligned} y_1 &= 4, \\ y_2 &= 7 - 2y_1 = 7 - 8 = -1, \\ y_3 &= 2 - 3y_1 - 2y_2 = 2 - 12 + 2 = -8. \end{aligned}$$

Das zweite lineare Gleichungssystem lautet nun also:

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ -1 \\ -8 \end{pmatrix}.$$

Die Lösung ist gegeben durch:

$$\begin{aligned} x_3 &= -8, \\ x_2 &= -1 \cdot (-1 - x_3) = 1 + x_3 = -7, \\ x_1 &= 4 - x_2 - x_3 = 19. \end{aligned}$$

Probe:

$$\begin{aligned} x_1 + x_2 + x_3 &= 4, \\ 2x_1 + x_2 + 3x_3 &= 38 - 7 - 24 = 7, \\ 3x_1 + x_2 + 6x_3 &= 57 - 7 - 48 = 2. \end{aligned}$$

Fragestellung 9.8
Sei $A \in \mathbf{Mat}_{n\times n}(\mathbb{R})$ invertierbar und $b \in \mathbb{R}^n \backslash \{0\}$. Wie stark kann die Lösung $\tilde{x} \in \mathbb{R}^n$ eines gestörten linearen Gleichungssystems

(9.8) $$(A + \Delta A) \cdot \tilde{x} = b + \Delta b$$

die Lösung $x \in \mathbb{R}^n \backslash \{0\}$ von

(9.1) $$A \cdot x = b$$

verfälschen ?

Um diese Fragestellung zu behandeln, müssen wir eine Möglichkeit haben, die „Größe“ eines Vektors und einer Matrix durch eine reelle Zahl zu messen. - Dies geschieht nun mittels einer Norm.

Definition 9.9, siehe auch Aufgabe Nr. 22:
Eine Abbildung $\| \ \| \colon \mathbb{R}^n \to \mathbb{R}$ *heißt eine Norm auf* $\mathbb{R}^n$*, wenn gilt:*

(N1) Es ist $\|x\| \geq 0$ *für alle* $x \in \mathbb{R}^n$ *; ferner gilt folgende Äquivalenz:*

$$\|x\| = 0 \Leftrightarrow x = 0.$$

(N2) Für alle $x \in \mathbb{R}^n$ *und alle* $\lambda \in \mathbb{R}$ *gilt:*

$$\|\lambda \cdot x\| = |\lambda| \cdot \|x\|.$$

(N3) Für alle $x, y \in \mathbb{R}^n$ gilt die Dreiecksungleichung:

$$\|x + y\| \leq \|x\| + \|y\|.$$

Beispiele 9.10:
Es sei jeweils $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$.

i) $\|x\|_1 := \sum_{i=1}^n |x_i|$ (Summennorm)

ii) $\|x\|_2 := \sqrt{\sum_{i=1}^n x_i^2}$ (Euklidische Norm)

iii) $\|x\|_p := \sqrt[p]{\sum_{i=1}^n |x_i|^p}$ für $p \in \mathbb{N}$ (l_p-Norm)

iv) $\|x\|_\infty := \|x\|_{\sup} := \max_{1 \leq i \leq n} |x_i|$ (Maximumnorm)

Bemerkung 9.11. *i) Je zwei Normen $\| \ \|$ und $\| \ \|'$ auf $\mathbb{R}^n$ sind in folgendem Sinne äquivalent:*
Es gibt - von $\| \ \|$, $\| \ \|'$ und n abhängige - positive reelle Zahlen c_1, c_2, so dass gilt:

$$\forall x \in \mathbb{R}^n : \quad c_1 \cdot \|x\| \leq \|x\|' \leq c_2 \cdot \|x\|. \tag{9.11}$$

ii) Definition 9.9 und die obigen Beispiele können für $m, n \in \mathbb{N}$ direkt auf $\mathbf{Mat}_{m \times n}(\mathbb{R})$ übertragen werden, weil $\mathbf{Mat}_{m \times n}(\mathbb{R})$ mit $\mathbb{R}^{m \cdot n}$ identifiziert werden kann.

Definition 9.12. *Sei $\| \ \| : \mathbb{R}^n \to \mathbb{R}$ eine Norm. Dann ist die zugehörige Matrixnorm $\| \ \| : \ \mathbf{Mat}_{n \times n}(\mathbb{R}) \to \mathbb{R}$ definiert durch*

$$\|A\| := \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|A \cdot x\|}{\|x\|} = \max_{\substack{x \in \mathbb{R}^n \\ \|x\| = 1}} \|A \cdot x\|. \tag{9.12}$$

Bemerkungen 9.13. *i) In der rechten Seite in (9.12) wird das Maximum aus Stetigkeitsgründen angenommen. Die rechte Gleichung in (9.12) gilt wegen*

$$\|A \cdot (\lambda \cdot x)\| = \|\lambda \cdot (A \cdot x)\| = |\lambda| \cdot \|A \cdot x\|$$

für alle $x \in \mathbb{R}^n$ (mit $\|x\| = 1$) und alle $\lambda \in \mathbb{R}$.

ii) Die Normeigenschaften der Matrixnorm folgen direkt aus den Normeigenschaften der gegebenen Norm.

iii) Die zugehörige Matrixnorm wird manchmal auch mit „lub“ bezeichnet (das bedeutet: “least-upper-bound-Norm“);
Schreibweise: $\|A\| = lub(A)$.

iv) Für eine gegebene Norm $\| \ \| : \ \mathbb{R}^n \to \mathbb{R}$ ist die zugehörige Matrixnorm $\| \ \|$ die kleinste Norm auf $\mathbf{Mat}_{n \times n}(\mathbb{R})$, so dass für alle $A \in \mathbf{Mat}_{n \times n}(\mathbb{R})$ und alle $x \in \mathbb{R}^n$ gilt:

$$\|A \cdot x\| \leq \|A\| \cdot \|x\|. \tag{9.13}$$

Dies folgt direkt aus (9.12).

Satz und Definition 9.14. *Die zur Maximumnorm* $\| \ \|_\infty : \mathbb{R}^n \to \mathbb{R}$ *gehörige Matrixnorm* $\| \ \|_\infty : \mathbf{Mat}_{n\times n}(\mathbb{R}) \to \mathbb{R}$ *ist die Zeilensummennorm, gegeben durch*

$$\|A\|_\infty = \max_{1\le i\le n} \left(\sum_{j=1}^n |a_{ij}|\right) \tag{9.14}$$

für $A = (a_{ij})_{1\le i,j\le n}$.

Beweis. Wir erhalten mittels (9.12):

$$\begin{aligned}
\|A\|_\infty &= \max_{\substack{x\in\mathbb{R}^n \\ \|x\|_\infty=1}} \|A\cdot x\|_\infty \\
&= \max_{\substack{x\in\mathbb{R}^n \\ \|x\|_\infty=1}} \left(\max_{1\le i\le n} \left|\sum_{j=1}^n a_{ij}\cdot x_j\right|\right) \qquad \left(\text{mit } x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\right) \\
&= \max_{1\le i\le n} \left(\sum_{j=1}^n a_{ij}\cdot sign(a_{ij})\right) \qquad \text{(für optimales } i \text{ setze hier } x_j = sign(a_{ij}) \text{ für } 1\le j\le n) \\
&= \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}|.
\end{aligned}$$

□

Im folgenden sei $\| \ \| : \mathbb{R}^n \to \mathbb{R}$ eine Norm; die zugehörige Matrixnorm werde ebenfalls mit $\| \ \|$ bezeichnet.

Definition 9.15. *Die Kondition einer invertierbaren Matrix* $A \in \mathbf{Mat}_{n\times n}(\mathbb{R})$ *ist gegeben durch*

$$cond(A) \coloneqq \|A\| \cdot \|A^{-1}\|. \tag{9.15}$$

Lemma 9.16. *i) Für die n- reihige Einheitsmatrix* I_n *gilt:*

$$\|I_n\| = 1. \tag{9.16a}$$

ii) Für alle $B, C \in \mathbf{Mat}_{n\times n}(\mathbb{R})$ *gilt:*

$$\|B\cdot C\| \le \|B\| \cdot \|C\|. \tag{9.16b}$$

iii) Ist $A \in \mathbf{Mat}_{n\times n}(\mathbb{R})$ *invertierbar, so gilt:*

$$cond(A) \ge 1. \tag{9.16c}$$

Beweis. i) (9.12) impliziert direkt:

$$\|I_n| = \max_{\substack{x\in\mathbb{R}^n \\ \|x\|=1}} \|I_n\cdot x\| = \max_{\substack{x\in\mathbb{R}^n \\ \|x\|=1}} \|x\| = 1.$$

ii) ist trivial für $C = 0$. Sei nun $C \neq 0$. Dann ist

$$T \coloneqq \{x \in \mathbb{R}^n \ : \ \|C \cdot x\| > 0\} \neq \emptyset.$$

(9.12) liefert weiter:

$$\begin{aligned}\|B \cdot C\| &= \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|(B \cdot C) \cdot x\|}{\|x\|} \\ &= \max_{x \in T} \left(\frac{\|B \cdot (C \cdot x)\|}{\|C \cdot x\|} \cdot \frac{\|C \cdot x\|}{\|x\|} \right) \\ &\leq \max_{y \in \mathbb{R}^n \setminus \{0\}} \frac{\|B \cdot y\|}{\|y\|} \cdot \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|C \cdot x\|}{\|x\|} \\ &= \|B\| \cdot \|C\|.\end{aligned}$$

iii) Aus i) und ii) folgt:

$$cond(A) = \|A\| \cdot \|A^{-1}\| \geq \|A \cdot A^{-1}\| = \|I_n\| = 1.$$

□

Lemma 9.17. *Sei $F \in \mathbf{Mat}_{n \times n}(\mathbb{R})$ mit $\|F\| < 1$. Dann ist die Matrix $I_n + F$ invertierbar, und es gilt:*

$$\|(I_n + F)^{-1}\| \leq \frac{1}{1 - \|F\|}. \tag{9.17}$$

Beweis. Aus (9.13) folgt für alle $x \in \mathbb{R}^n \setminus \{0\}$:

$$\begin{aligned}\|(I_n + F) \cdot x\| = \|x + F \cdot x\| &\geq \|x\| - \|F \cdot x\| \\ \geq \|x\| - \|F\| \cdot \|x\| &= (1 - \|F\|) \cdot \|x\| > 0;\end{aligned}$$

also ist $I_n + F$ invertierbar.
Lemma 9.16 ii) liefert nun für $B \coloneqq F$ und $C \coloneqq (I_n + F)^{-1}$:

$$\begin{aligned}1 &= \|(I_n + F) \cdot C\| \geq \|C\| - \|F \cdot C\| \geq \|C\| - \|F\| \cdot \|C\| \\ &= \|C\| \cdot (1 - \|F\|) > 0.\end{aligned}$$

Damit folgt auch:

$$\|(I_n + F)^{-1}\| = \|C\| \leq \frac{1}{1 - \|F\|}.$$

□

Warnung:
Die Lemmata 9.16 und 9.17 gelten nicht für beliebige Matrix - Normen.

Nun können wir zeigen:

Satz 9.18. *Neben den in der Fragestellung 9.8 getroffenen Konventionen gelte zusätzlich:*

(9.18a) $$\|\Delta A\| \cdot \|A^{-1}\| < 1.$$

Dann ist die gestörte Gleichung (9.8) eindeutig lösbar, und für den relativen Fehler $\frac{\|x-\tilde{x}\|}{\|x\|}$ *gilt:*

(9.18) $$\frac{\|x-\tilde{x}\|}{\|x\|} \leq \frac{cond(A)}{1-\|\Delta A\| \cdot \|A^{-1}\|} \cdot \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|}\right).$$

Beweis. Nach Voraussetzung ist

$$\|A^{-1} \cdot \Delta A\| \leq \|A^{-1}\| \cdot \|\Delta A\| < 1.$$

Nach Lemma 9.17 ist somit die Matrix $I_n + A^{-1} \cdot \Delta A$ invertierbar, und es gilt:

(9.18b) $$\|(I_n + A^{-1} \cdot \Delta A)^{-1}\| \leq \frac{1}{1-\|A^{-1} \cdot \Delta A\|} \leq \frac{1}{1-\|A^{-1}\| \cdot \|\Delta A\|}.$$

Weiter ist auch $A + \Delta A = A \cdot (I_n + A^{-1} \cdot \Delta A)$ invertierbar; also gibt es genau ein $\tilde{x} \in \mathbb{R}^n$, das (9.8) löst.
Wir setzen nun $\Delta x \coloneqq \tilde{x} - x$ und erhalten:

$$\begin{aligned}
&(A + \Delta A) \cdot (x + \Delta x) = b + \Delta b \\
\Rightarrow &\Delta A \cdot x + (A + \Delta A) \cdot \Delta x = \Delta b && \text{(nach (9.1))} \\
\Rightarrow &(A + \Delta A) \cdot \Delta x = \Delta b - \Delta A \cdot x \\
\Rightarrow &(I_n + A^{-1} \cdot \Delta A) \cdot \Delta x = A^{-1} \cdot (\Delta b - \Delta A \cdot x) \\
\Rightarrow &\Delta x = (I_n + A^{-1} \cdot \Delta A)^{-1} \cdot A^{-1} \cdot (\Delta b - \Delta A \cdot x).
\end{aligned}$$

Also folgt mittels (9.18b) und wiederholter Anwendung von (9.13):

(9.18c) $$\|\Delta x\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\| \cdot \|\Delta A\|} \cdot (\|\Delta b\| + \|\Delta A\| \cdot \|x\|).$$

Weiter gilt wegen $\|b\| = \|A \cdot x\| \leq \|A\| \cdot \|x\|$:

$$\begin{aligned}
\|\Delta b\| + \|\Delta A\| \cdot \|x\| &= \|b\| \cdot \frac{\|\Delta b\|}{\|b\|} + \|A\| \cdot \|x\| \cdot \frac{\|\Delta A\|}{\|A\|} \\
&\leq \|A\| \cdot \|x\| \cdot \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|}\right).
\end{aligned}$$

Zusammen mit (9.18c) folgt

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \cdot \|A\|}{1-\|A^{-1}\| \cdot \|\Delta A\|} \cdot \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|}\right)$$

wie gewünscht. □

Bemerkung 9.19. *(9.18) ist eine gute Abschätzung, falls cond(A) „nicht wesentlich" größer als 1 ist.*

Im Gegensatz dazu studieren wir folgendes

Beispiel 9.20:
Sei $A := \begin{pmatrix} 1 & 0 \\ 1000 & 1 \end{pmatrix}$. Dann ist $A^{-1} = \begin{pmatrix} 1 & 0 \\ -1000 & 1 \end{pmatrix}$.
Wir betrachten ein zugehöriges lineares Gleichungssystem

$$\begin{aligned} & x_1 && = b_1 \\ \wedge \quad & 1000x_1 + x_2 && = b_2. \end{aligned}$$

Eine kleine Änderung von b_1 bewirkt - bei gleich bleibender Matrix A - natürlich die gleiche Änderung von x_1. Diese kann aber eine große Änderung von x_2 bewirken.

Mit der durch $\| \ \|_\infty$ induzierten Kondition $cond = cond_\infty$ gilt nach Satz 9.14:

$$\|A\| = \|A^{-1}\| = 1000 + 1 = 1001$$

und damit

$$cond_\infty(A) = \|A\| \cdot \|A^{-1}\| = 1001^2 > 10^6.$$

Bemerkung 9.21. ***Rundungsfehleranalyse beim Gaußschen Eliminationsverfahren***
Sobald die Gleitpunktoperationen nicht mehr exakt durchgeführt werden können, sollte während der Durchführung der Eliminationsschritte folgendes beachtet werden:

i) Die Matrizen $A^{(k-1)}$ sollten - für $1 \leq k < n$ - <u>equilibriert</u> sein; das heißt:

$$\sum_{j=1}^{n} |a_{ij}^{(k-1)}| \approx \sum_{j=1}^{n} |a_{\nu_j i}^{(k-1)}| \quad \text{für alle } i, \nu$$

Das kann durch geeignete Zeilen- Multiplikationen erreicht werden.

ii) Sodann sollte das <u>Pivotelement</u> $a_{kk}^{(k-1)}$, durch das im k-ten Eliminationsschritt dividiert wird, dem Betrage nach möglichst groß sein; das heißt:

$$|a_{ik}^{(k-1)}| \leq |a_{kk}^{(k-1)}| \quad \text{für } k < i \leq n$$

und folglich

$$l_{ik} := a_{ik}^{(k-1)} / a_{kk}^{(k-1)} \in [-1, 1] \quad \text{für } k < i \leq n.$$

Das kann durch eine Zeilen - Vertauschung erreicht werden.

Bemerkung 9.22. ***Das lineare Ausgleichsproblem***
In der Praxis sind oft die Werte gewisser Größen $x_1, \ldots, x_n \in \mathbb{R}$ zu bestimmen, die nicht direkt gemessen oder berechnet werden können. Statt dessen kann man aber andere Größen $y_1, \ldots, y_m \in \mathbb{R}$ ermitteln, die - im einfachsten Fall - linear von $x_1, \ldots, x_n$ abhängen; das heißt, es gibt eine Matrix $A \in \mathbf{Mat}_{n\times n}(\mathbb{R})$ mit:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = A \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}. \tag{9.22}$$

Die Matrix A sei dabei ebenfalls bekannt. Damit $x_1, \ldots, x_n$ eindeutig bestimmt sind, muss $m \geq n$ sein. Für $m > n$ ist (9.22) im allgemeinen ein überbestimmtes lineares Gleichungssystem, das gewöhnlich keine Lösung besitzt, weil $y_1, \ldots, y_m$ mit unvermeidbaren Messfehlern behaftet sind. - Weil aber die Informationen mit der Zahl der durchgeführten Experimente wachsen, ist es eher unzweckmäßig, $m = n$ statt $m \geq n$ zu fordern. Statt dessen wird versucht, eine geeignet erscheinende Ersatzaufgabe zu lösen.

Die Behandlung einer der beiden folgenden Aufgaben liegt nahe, wenn $a_1, \ldots, a_m$ die Zeilenvektoren der Matrix A beschreiben:

i) Suche $x_1, \ldots x_n \in \mathbb{R}$ mit

$$\| \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} - A \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \|_1 = \sum_{k=1}^{m} |y_k - a_k \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}| \stackrel{!}{=} Min. \tag{9.22a}$$

ii) Suche $x_1, \ldots, x_n \in \mathbb{R}$ mit

$$\| \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} - A \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \|_2^2 = \sum_{k=1}^{m} \left(y_k - a_k \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right)^2 \stackrel{!}{=} Min. \tag{9.22b}$$

(9.22b) heißt ein lineares Ausgleichsproblem. Es mag eher auf der Hand liegen, (9.22a) statt (9.22b) zu betrachten. Die Lösung von (9.22b) ist aber nicht nur einfacher, sondern auch wahrscheinlichkeitstheoretisch eher gerechtfertigt.

Satz 9.23. *Es seien $m, n \in \mathbb{N}$ mit $m \geq n$, und es seien $A \in \mathbf{Mat}_{m\times n}(\mathbb{R})$ und $y = (y_1, \ldots, y_m)^T \in \mathbb{R}^m$ gegeben. Dann besitzt das lineare Ausgleichsproblem*

$$\|y - A \cdot x\|_2^2 \stackrel{!}{=} Min$$

mindestens eine Lösung $x_0 \in \mathbb{R}^n$.
Ist $\mathbb{L}$ die Menge aller Lösungen dieses linearen Ausgleichsproblems, so gilt:

$$\mathbb{L} = \{x \in \mathbb{R}^n | A \cdot x = A \cdot x_0\} = \{x \in \mathbb{R}^n | A^T \cdot A \cdot x = A^T \cdot y\}. \tag{9.23}$$

Beweis. Es sei

$$U_0 := \{A \cdot x | x \in \mathbb{R}^n\};$$

das ist der von den Spalten von A aufgespannte Unterraum von $\mathbb{R}^m$. Zu dem vorgegebenen $y \in \mathbb{R}^m$ gibt es genau ein $u \in U_0$ und genau ein $w \in U_0^\perp$ mit

(9.23a) $$y = u + w.$$

Laut Definition von U_0 gibt es - mindestens - ein $x_0 \in \mathbb{R}^n$ mit $A \cdot x_0 = u$.
Für dieses x_0 setzen wir nun

$$\mathbb{L}_1 := \{x \in \mathbb{R}^n | A \cdot x = A \cdot x_0\}, \mathbb{L}_2 := \{x \in \mathbb{R}^n | A^T \cdot A \cdot x = A^T \cdot y\}.$$

Zu zeigen ist: $\mathbb{L} = \mathbb{L}_1 = \mathbb{L}_2$.
Dazu zeigen wir zunächst: $A^T \cdot w = 0$.
Dazu genügt es, zu zeigen:

$$\forall v \in \mathbb{R}^n : \quad v^T \cdot A^T \cdot w = 0.$$

Wegen $v^T \cdot A^T \cdot w \in \mathbb{R}$ erhalten wir:

$$v^T \cdot A^T \cdot w = (v^T \cdot A^T \cdot w)^T = w^T \cdot A \cdot v = \langle w, A \cdot v \rangle = 0.$$

Dabei gilt die letzte Gleichung wegen $A \cdot v \in U_0$ und wegen $w \in U_0^\perp$.
Aus der Beziehung $A^T \cdot w = 0$ folgt nun mittels (9.23a):

$$A^T \cdot A \cdot x_0 = A^T \cdot u = A^T \cdot (y - w) = A^T \cdot y.$$

Es ist also $x_0 \in \mathbb{L}_2-$ und damit auch $\mathbb{L}_1 \subseteq \mathbb{L}_2$.
Nun sei $x \in \mathbb{L}_2$. Ist dann $v \in \mathbb{R}^n$ beliebig, so folgt mit $z := A \cdot v - A \cdot x$:

$$\begin{aligned} z^T \cdot (y - A \cdot x) &= (v^T \cdot A^T - x^T \cdot A^T) \cdot (y - A \cdot x) \\ =&(v^T - x^T) \cdot A^T \cdot (y - A \cdot x) = (v^T - x^T) \cdot (A^T \cdot y - A^T \cdot A \cdot x) = 0; \end{aligned}$$

denn es ist $x \in \mathbb{L}_2$. Damit folgt weiter:

$$\begin{aligned} \|y - A \cdot v\|_2^2 &= \|(y - A \cdot x) + (A \cdot x - A \cdot v)\|_2^2 = \|(y - A \cdot x) - z\|_2^2 \\ =&\|y - A \cdot x\|_2^2 + \|z\|_2^2 \geq \|y - A \cdot x\|_2^2. \end{aligned}$$

Dabei gilt Gleichheit genau im Falle $z = 0$.
Das bedeutet: v löst das Minimierungsproblem (9.22) genau dann, wenn $A \cdot v = A \cdot x$ ist.
Da $x \in \mathbb{L}_2$ beliebig gewählt wurde, folgt insbesondere: $\mathbb{L}_2 \subseteq \mathbb{L}$.
Wenden wir die letzte Überlegung speziell an für $x = x_0$, so folgt auch: $\mathbb{L} = \mathbb{L}_1$.
Damit ist alles bewiesen. □

Bemerkung 9.24. *Die n Gleichungen des linearen Gleichungssystems*

(9.24) $$A^T \cdot A \cdot x = A^T \cdot y$$

heißen auch die Normalgleichungen für x.

Beispiel 9.25
Wir lösen die Normalgleichungen für das überbestimmte lineare Gleichungssystem

$$x_1 - x_2 = 0 \quad , \quad x_1 + x_2 = 0 \quad , \quad 5x_1 + x_2 = 12.$$

Hier ist

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 5 & 1 \end{pmatrix} \quad , \quad A^T = \begin{pmatrix} 1 & 1 & 5 \\ -1 & 1 & 1 \end{pmatrix} \quad , \quad y = \begin{pmatrix} 0 \\ 0 \\ 12 \end{pmatrix}.$$

Damit folgt weiter:

$$A^T \cdot A = \begin{pmatrix} 27 & 5 \\ 5 & 3 \end{pmatrix} \quad , \quad A^T \cdot y = \begin{pmatrix} 60 \\ 12 \end{pmatrix}.$$

Die Normalgleichungen lauten nun also:

$$27x_1 + 5x_2 = 60 \quad \wedge \quad 5x_1 + 3x_2 = 12.$$

Als Lösung erhalten wir - etwa mittels der Cramerschen Regel:

$$x_1 = 2\frac{1}{7} \quad , \quad x_2 = \frac{3}{7}.$$

In diesem Zusammenhang ist schließlich noch folgender Satz von Interesse, der etwa mit Hilfe des Orthogonalisierungsverfahrens von Gram - Schmidt (siehe Satz 7.10 im WS 2015/16) bewiesen werden kann:

Satz 9.26. ***Die QR- Zerlegung:***
Seien $m, n \in \mathbb{N}$ mit $m \geq n$, und sei $A \in \mathbf{Mat}_{m\times n}(\mathbb{R})$. Dann gibt es eine Matrix $Q \in \mathbf{Mat}_{m\times n}(\mathbb{R})$, deren Spalten paarweise orthogonal zueinander sind, und eine obere Dreiecksmatrix $R \in \mathbf{Mat}_{n\times n}(\mathbb{R})$, so dass gilt:

$$A = Q \cdot R. \tag{9.26}$$

§10 Numerische Berechnung von Integralen

Aufgabenstellung 10.1:
Es seien $a, b \in \mathbb{R}$ mit $a < b$, und $f\colon\ [a,b] \to \mathbb{R}$ sei eine stetige Funktion. Das Integral

$$I(f) \coloneqq \int_a^b f(x)\,\mathrm{d}x \tag{10.1}$$

ist näherungsweise zu berechnen.

Bemerkung 10.2. ***Approximation von f durch Interpolationspolynome:***
Für fixiertes $n \in \mathbb{N}$ setzen wir $h \coloneqq \frac{1}{n} \cdot (b-a)$, und die äquidistant verteilten Stützstellen $x_0, x_1, \ldots, x_n \in [a,b]$ seien gegeben durch

$$x_i \coloneqq a + i \cdot h \quad \text{für } 0 \le i \le n. \tag{10.2a}$$

Weiter sei $P_n \in \Pi_n$ das eindeutig bestimmte Interpolationspolynom mit

$$P_n(x_i) = f_i \coloneqq f(x_i) \quad \text{für } 0 \le i \le n. \tag{10.2b}$$

Mit

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \quad \text{für } 0 \le i \le n \tag{10.2c}$$

folgt dann also:

$$P_n(x) = \sum_{i=0}^{n} f_i \cdot L_i(x). \tag{10.2d}$$

Schließlich definieren wir für $0 \le i \le n$ die Funktion $\varphi_i\colon\ [0,n] \to \mathbb{R}$ durch

$$\varphi_i(s) \coloneqq \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{s - j}{i - j} \tag{10.2e}$$

und setzen

$$\alpha_i \coloneqq \int_0^n \varphi_i(s)\,\mathrm{d}s. \tag{10.2f}$$

Man beachte dabei: $\alpha_0, \alpha_1, \ldots, \alpha_n$ hängen zwar von n, aber nicht von a, b und f ab.

Satz 10.3. ***Die Integrationsformeln von Newton - Cotes:***
Unter den in Bemerkung 10.2 getroffenen Konventionen gilt:

$$\int_a^b P_n(x)\,\mathrm{d}x = h \cdot \sum_{i=0}^{n} f_i \cdot \alpha_i. \tag{10.3}$$

Beweis. Definiere die Bijektion $\Psi\colon\ [0,n] \to [a,b]$ durch

(10.3a) $$\Psi(s) := a + h \cdot s.$$

Für $0 \le i \le n$ ist $\Psi(i) = x_i$, und für $0 \le s \le n$ und $x = \Psi(s)$ erhalten wir:

$$L_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{\Psi(s) - \Psi(j)}{\Psi(i) - \Psi(j)} = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{s-j}{i-j} = \varphi_i(s).$$

Also folgt mittels (10.2d) und Variablensubstitution:

$$\int_a^b P_n(x)\,\mathrm{d}x = \sum_{i=0}^{n} f_i \cdot \int_a^b L_i(x)\,\mathrm{d}x = \sum_{i=0}^{n} f_i \cdot \int_0^n L_i(\Psi(s)) \cdot \Psi'(s)\,\mathrm{d}s$$
$$= \sum_{i=0}^{n} f_i \cdot h \cdot \int_0^n \varphi_i(s)\,\mathrm{d}s = h \cdot \sum_{i=0}^{n} f_i \cdot \alpha_i.$$

□

Korollar 10.4. *Es gilt:*

(10.4) $$\sum_{i=0}^{n} \alpha_i = n.$$

Beweis. Weil die Koeffizienten $\alpha_0, \alpha_1, \ldots, \alpha_n$ nur von n und nicht von f abhängen, können wir die obigen Überlegungen auf $f \equiv 1$ anwenden. Dann ist auch $P_n \equiv 1$, und (10.3) liefert:

$$b - a = h \cdot \sum_{i=0}^{n} \alpha_i.$$

Wegen $h = \dfrac{1}{n} \cdot (b-a)$ folgt nun die Behauptung. □

Korollar 10.5. *Für $0 \le i \le n$ gilt:*

(10.5) $$\alpha_i = \alpha_{n-i}.$$

Beweis. Für $0 \le s \le n$ ist

$$\varphi_{n-i}(n-s) = \prod_{\substack{j=0 \\ j\neq n-i}}^{n} \frac{(n-s)-j}{(n-i)-j} = \prod_{\substack{j=0 \\ j\neq n-i}}^{n} \frac{s-(n-j)}{i-(n-j)}$$
$$= \prod_{\substack{k=0 \\ k\neq i}}^{n} \frac{s-k}{i-k} = \varphi_i(s).$$

Damit folgt weiter:

$$\alpha_{n-i} = \int_0^n \varphi_{n-i}(s)\,\mathrm{d}s = \int_0^n \varphi_{n-i}(n-s)\,\mathrm{d}s = \int_0^n \varphi_i(s)\,\mathrm{d}s = \alpha_i.$$

□

Satz 10.6. ***Die Trapezregel***
Sei speziell $n = 1$. Dann gilt:

(10.6a) $$\alpha_0 = \alpha_1 = \frac{1}{2}$$

und damit

(10.6b) $$\int_a^b P_1(x)\,\mathrm{d}x = (b-a)\cdot\frac{f(a)+f(b)}{2}.$$

Beweis. Aus Korollar 10.4 und Korollar 10.5 folgt:

$$\alpha_0 + \alpha_1 = 1 \qquad \wedge \qquad \alpha_0 = \alpha_1.$$

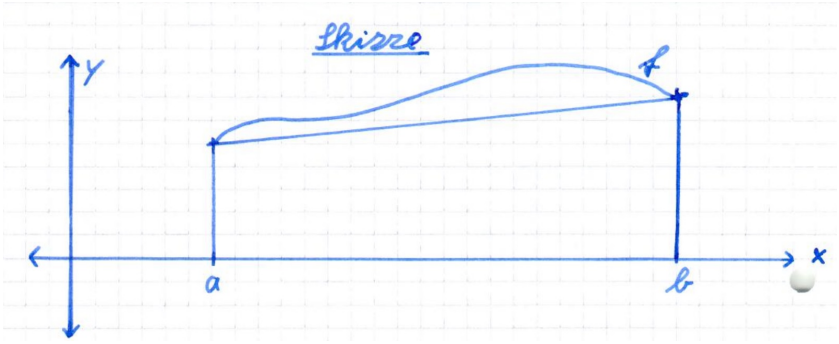
Damit erhalten wir (10.6a).
Somit folgt weiter (10.6b) aus Satz 10.3 und den Beziehungen

$$h = b - a \quad , \quad f_0 = f(a) \quad , \quad f_1 = f(b).$$

□

Bemerkung 10.7. ***Geometrische Interpretation:***
Ist $f(a) > 0$ und $f(b) > 0$, so folgt aus (10.6b):

$\int_a^b P_1(x)\,\mathrm{d}x$ ist der Flächeninhalt des Trapezes, das von den Eckpunkten $(a,0),(b,0),(b,f(b)),(a,f(a))$ aufgespannt wird.



Bemerkung 10.8. ***Die zusammengesetzte Trapezregel:***
Es sei $k \in \mathbb{N}$ und

(10.8a) $$t_i := a + \frac{i}{k}\cdot(b-a) \quad \textit{für } 0 \le i \le k.$$

$t_0, t_1, \ldots, t_k$ sind also äquidistant verteilte Stützstellen in $[a,b]$.
Wenden wir die Trapezregel an auf jedes Teilintervall $[t_{i-1}, t_i]$, $1 \le i \le k$, so ergibt sich für $I(f)$ der Näherungswert

(10.8b)
$$\begin{aligned} T_k(f) &:= \frac{b-a}{2k}\cdot\sum_{i=1}^{k}(f(t_{i-1})+f(t_i)) \\ &= \frac{b-a}{k}\cdot\left(\frac{f(a)}{2}+\sum_{i=1}^{k-1}f(t_i)+\frac{f(b)}{2}\right). \end{aligned}$$

Satz 10.9. ***Das Restglied bei der zusammengesetzten Trapezregel:***
Ist $f\colon [a,b] \to \mathbb{R}$ zweimal stetig differenzierbar, so gibt es ein $\xi = \xi_k \in [a,b]$ mit:

$$I(f) - T_k(f) = -\frac{(b-a)^3}{12k^2} \cdot f''(\xi_k). \tag{10.9a}$$

Weil f'' auf $[a,b]$ beschränkt ist, folgt insbesondere:

$$\lim_{k\to\infty} T_k(f) = I(f). \tag{10.9b}$$

Beispiel:
Definiere $f\colon [0,1] \to \mathbb{R}$ durch

$$f(x) := \sqrt{1+x^4}.$$

Für alle $x \in [0,1]$ ist

$$\begin{aligned} f'(x) &= \frac{2x^3}{\sqrt{1+x^4}}, \\ 0 \leq f''(x) &= \frac{6x^2+2x^6}{(1+x^4)^{3/2}} \leq f''(1) = 2^{\frac{2}{3}} = \sqrt{8}. \end{aligned}$$

Damit liefert (10.9a) für alle $k \in \mathbb{N}$:

$$-\frac{\sqrt{2}}{6k^2} \leq I(f) - T_k(f) \leq 0.$$

Speziell ist $\frac{\sqrt{2}}{6\cdot 5^2} < \frac{1}{100}$ und folglich

$$I(f) = T_5(f) - \Theta' \cdot 10^{-2} = 1{,}095 - \Theta \cdot 10^{-2}$$

mit passenden $\Theta', \Theta \in [0,1]$.

Satz 10.10. ***Die Simpson-Regel:***
Für $n = 2$ liefert (10.2f):

$$\alpha_0 = \frac{1}{3} \quad , \quad \alpha_1 = \frac{4}{3} \quad , \quad \alpha_2 = \frac{1}{3} \tag{10.10a}$$

und damit

$$\int_a^b P_2(x)\,\mathrm{d}x = \frac{b-a}{6} \cdot \left(f(a) + 4 \cdot f\left(\frac{a+b}{2}\right) + f(b) \right). \tag{10.10b}$$

Beweis. (10.2e) und (10.2f) liefern:

$$\begin{aligned} \alpha_2 &= \int_0^2 \frac{s-0}{2-0} \cdot \frac{s-1}{2-1}\,\mathrm{d}s = \frac{1}{2} \cdot \int_0^2 (s^2 - s)\,\mathrm{d}s \\ &= \frac{1}{2} \cdot \left(\frac{2^3}{3} - \frac{2^2}{2} \right) = \frac{4}{3} - 1 = \frac{1}{3}. \end{aligned}$$

Nach Korollar 10.5 ist dann auch $\alpha_0 = \frac{1}{3}$.
Schließlich liefert dann Korollar 10.4:

$$\alpha_1 = 2 - \alpha_0 - \alpha_2 = \frac{4}{3}.$$

Somit folgt weiter (10.10b) aus Satz 10.3 und den Beziehungen

$$h = \frac{1}{2} \cdot (b-a) \quad , \quad f_0 = f(a) \quad , \quad f_1 = f\left(\frac{a+b}{2}\right) \quad , \quad f_2 = f(b).$$

□

Bemerkung 10.11. ***Die zusammengesetzte Simpson-Regel:***
Nun sei $k \in \mathbb{N}$ gerade - und wieder

$$t_i \coloneqq a + \frac{i}{k} \cdot (b-a) \quad \textit{für } 0 \le i \le k. \tag{10.8a}$$

Ferner sei $k' \coloneqq \frac{k}{2}$.

Wenden wir sie Simpson-Regel an auf jedes Teilintervall $[t_{2i-2}, t_{2i}]$ für $1 \le i \le k'$, so ergibt sich für $I(f)$ der Näherungswert

$$\begin{aligned} S_k(f) &\coloneqq \frac{b-a}{6k'} \cdot \sum_{i=1}^{k'} (f(t_{2i-2}) + 4 \cdot f(t_{2i-1}) + f(t_{2i})) \\ &= \frac{b-a}{3k} \cdot (f(a) + 4 \cdot f(t_1) + 2 \cdot f(t_2) \\ &\qquad + 4 \cdot f(t_3) + 2 \cdot f(t_4) \\ &\qquad \cdots\cdots\cdots \\ &\qquad + 4 \cdot f(t_{k-1}) + f(b)). \end{aligned}$$

Satz 10.12. ***Das Restglied bei der zusammengesetzten Simpson - Regel:***
Ist $f\colon [a,b] \to \mathbb{R}$ viermal stetig differenzierbar, so gibt es zu jedem geraden $k \in \mathbb{N}$ ein $\xi = \xi_k \in [a,b]$ mit:

$$I(f) - S_k(f) = -\frac{(b-a)^5}{180 \cdot k^4} \cdot f^{(4)}(\xi_k). \tag{10.12a}$$

Insbesondere ist

$$\lim_{k \to \infty} S_k(f) = I(f). \tag{10.12b}$$

Beispiel:
Wir berechnen das Integral

$$\int_0^1 \frac{\cos x}{\sqrt{x}} \, \mathrm{d}x = \int_0^1 2 \cdot \cos(t^2) \, \mathrm{d}t$$

nach der zusammengesetzten Simpson - Regel. Für $f(t) \coloneqq 2 \cdot \cos(t^2)$ liefert (10.11) für $k = 2^j,\ 1 \leq j \leq 8$, die Formel

$$S_k(f) = \frac{2}{3} \cdot 2^{-j} \cdot \Big(\cos 0 + 4 \cdot \sum_{i=1}^{2^{j-1}} \cos(((2i-1) \cdot 2^{-j})^2) + 2 \cdot \sum_{i=1}^{2^{j-1}-1} \cos((2i \cdot 2^{-j})^2) + \cos 1 \Big).$$

Weiter folgt für alle $t \in [0,1]$:

$$\begin{aligned} f'(t) &= -4t \cdot \sin(t^2), \\ f''(t) &= -4 \cdot \sin(t^2) - 8t^2 \cdot \cos(t^2), \\ f'''(t) &= -24t \cdot \cos(t^2) + 16t^3 \cdot \sin(t^2), \\ f^{(4)}(t) &= -24 \cdot \cos(t^2) + 96t^2 \cdot \sin(t^2) + 32t^4 \cdot \cos(t^2). \end{aligned}$$

Damit erhalten wir weiter für alle $t \in [0,1]$:

$$|f^{(4)}(t)| \leq 96 + |32 \cdot t^4 - 24| \cdot |\cos(t^2)| \leq 96 + 24 = 120.$$

Satz 10.12 liefert also für $k = 2^j,\ 1 \leq j \leq 8$:

$$\left| \int_0^1 f(t)\,\mathrm{d}t - S_k(f) \right| \leq \frac{120}{180} \cdot k^{-4} = \frac{1}{3} \cdot 2^{-4j+1}.$$

Tabelle:

j	$S_k(f)$	$\frac{1}{3} \cdot 2^{-4j+1}$
1	$1,805317331$	$4,1666667 \cdot 10^{-2}$
2	$1,809002530$	$2,6041667 \cdot 10^{-3}$
3	$1,809048319$	$1,6276042 \cdot 10^{-4}$
4	$1,809048505$	$1,0172526 \cdot 10^{-5}$
5	$1,809048478$	$6,3578288 \cdot 10^{-7}$
6	$1,809048476$	$3,9736430 \cdot 10^{-8}$
7	$1,809048475$	$2,4835269 \cdot 10^{-9}$
8	$1,809048476$	$1,5522043 \cdot 10^{-10}$

Als guten Näherungswert für das gesuchte Integral erhalten wir also 1,809048476.

Warnung 10.13 Für größere n treten in den Integrationsformeln von Newton - Cotes auch negative Werte α_i auf. Diese Formeln werden daher numerisch unbrauchbar.

§11 Geometrische Aspekte der Graphentheorie

Definition 11.1. *Ein Graph $G = (V, E)$ besteht aus einer endlichen nichtleeren Menge V von Vertizes (oder Knoten) und einer Teilmenge E von $P_2(V) = \binom{V}{2}$; die Elemente von E heißen Kanten von G.*

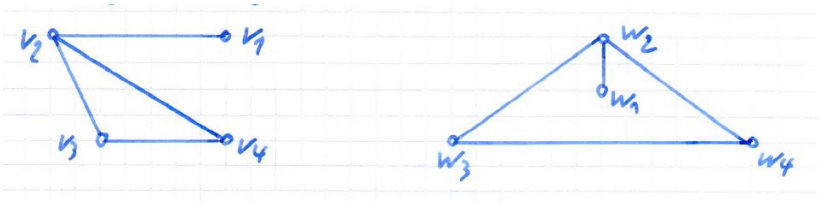
Definition 11.2. *Zwei Graphen $G = (V, E)$ und $G' = (V', E')$ heißen isomorph, falls es eine Bijektion $f : V \to V'$ gibt, so dass für $v, w \in V$ folgende Äquivalenz erfüllt ist:*

$$\{v, w\} \in E \Leftrightarrow \{f(v), f(w)\} \in E'. \tag{11.2}$$

In dem Fall heißt f ein Isomorphismus von G auf G'.

Bemerkung 11.3. *Graphen $G = (V, E)$ werden so in die Ebene $\mathbb{R}^2$ gezeichnet, dass die Vertizes durch Punkte dargestellt werden. Für $v, w \in V$ mit $\{v, w\} \in E$ werden die Punkte v und w durch eine Linie - und zwar möglichst durch eine Strecke - miteinander verbunden ; diese wird dann ebenfalls Kante genannt.*

Beispiel:
Die folgenden Graphen sind isomorph:



Ein Isomorphismus ist gegeben durch

$$f(v_i) := w_i \quad \text{für } 1 \leq i \leq 4.$$

Definition 11.4. *Es sei $G = (V, E)$ ein Graph.*

i) Ein Weg $W = (v_0, v_1, \ldots, v_n)$ in G ist eine endliche Folge von Vertizes mit

$$\{v_{i-1}, v_i\} \in E \quad \text{für alle } i \text{ mit } 1 \leq i \leq n.$$

v_0 heißt Anfangspunkt von W, und v_n heißt Endpunkt von W.

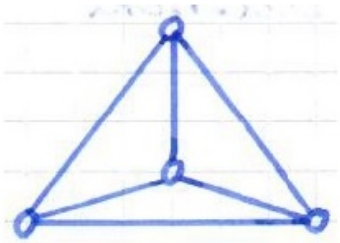
ii) Der Graph G heißt zusammenhängend, wenn gilt:
Zu je zwei verschiedenen Vertizes $v, w \in V$ gibt es einen Weg in G, der v als Anfangspunkt und w als Endpunkt hat.

iii) Ein Vertex $v \in V$ heißt isoliert, falls es gar keine Kante in E gibt, die v enthält.

iv) Zwei Vertizes $v, w \in V$ heißen benachbart, falls $\{v, w\} \in E$ ist. v und w heißen dann auch Nachbarn in G.

Definition 11.5. *Ein Graph $G = (V, E)$ heißt ein vollständiger Graph, wenn $E = \binom{V}{2}$ ist.*
Ist dabei $n := |V|$, so schreiben wir auch: $G = K_n$.

Darstellung von K_4



Definition 11.6. *Ein Graph $G = (V, E)$ heißt ein bipartiter Graph, wenn Teilmengen $A, B \subseteq V$ mit $A \cap B = \emptyset$ und $A \cup B = V$ existieren, so dass gilt:*

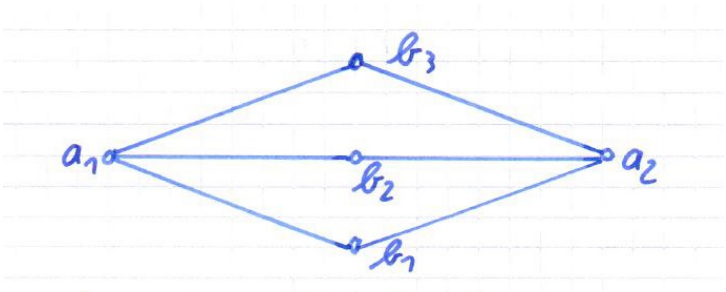
(11.6a) $$E \subseteq \{\{v, w\} \mid v \in A, w \in B\}.$$

Gilt sogar

(11.6b) $$E = \{\{v, w\} \mid v \in A, w \in B\},$$

so heißt G ein vollständig bipartiter Graph. Ist dabei $|A| = n$ und $|B| = m$, so schreiben wir auch: $G = K_{n,m}$.

Darstellung von $K_{2,3}$



Hier ist $A = \{a_1, a_2\}, B = \{b_1, b_2, b_3\}$.

Definition 11.7. *Ein Graph $G = (V, E)$ heißt planar (oder plättbar), wenn er so in die Ebene gezeichnet werden kann, dass sich verschiedene Linien nur in den Knoten schneiden.*

Satz 11.8. ***Die Eulersche Formel:***
Es sei $G = (V, E)$ ein planarer Graph, der gemäß Definition 11.7 in die Ebene gezeichnet ist. Dann unterteilt der Graph die Ebene in f Flächenstücke - für ein $f \in \mathbb{N}$, von denen eines unbeschränkt und die übrigen beschränkt sind. Ferner sei G zusammenhängend.
Setzen wir noch $v := |V|, e := |E|$, so gilt:

$$v - e + f = 2. \tag{11.8}$$

Bemerkung 11.9. *Gegeben sei ein Graph $G = (V, E)$. Wir betrachten folgende drei Möglichkeiten, eine Unterstruktur von G zu bilden:*

i) Es sei $e_0 = \{v_0, w_0\} \in E$. Entferne die Kante e_0 von E, betrachte also den neuen Graphen

$$G_1 := (V, E \backslash \{e_0\}).$$

ii) Es sei $|V| \geq 2$, und v_0 sei ein isolierter Vertex in G. Entferne diesen, betrachte also den neuen Graphen

$$G_2 := (V \backslash \{v_0\}, E).$$

iii) Es seien v_0, w_0 Nachbarn in G, die keinen weiteren gemeinsamen Nachbarn in $V \backslash \{v_0, w_0\}$ haben; setze $e_0 := \{v_0, w_0\}$.
Ziehe v_0, w_0 zu einem neuen Vertex u_0 - mit $u_0 \notin V$ - zusammen, setze also

$$V' := (V \backslash \{v_0, w_0\}) \cup \{u_0\}.$$

Betrachte ferner die - injektive - Abbildung $\pi\colon \quad E \backslash \{e_0\} \to \binom{V'}{2}$, definiert durch

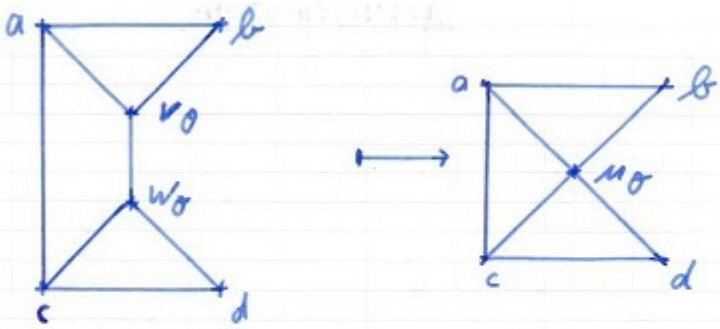
$$\begin{aligned}
&\pi(\{u, v\}) := \{u, v\} \ \textit{falls} \ \{u, v\} \cap \{v_0, w_0\} = \emptyset, \\
&\pi_0(\{u, v_0\}) := \{u, u_0\} \ \textit{falls} \ \{u, v_0\} \in E \backslash \{e_0\}, \\
&\pi_0(\{u, w_0\}) := \{u, u_0\} \ \textit{falls} \ \{u, w_0\} \in E \backslash \{e_0\}.
\end{aligned}$$

Betrachte nun den neuen Graphen

$$G_3 := (V', \pi(E \backslash \{e_0\})).$$

Die Konstruktionen in i) und ii) heißen elementare Einschränkungen von G; eine Konstruktion wie in iii) heißt eine elementare Kontraktion von G.

Skizze zu iii)



Definition 11.10. *Ein Minor eines gegebenen Graphen G ist jeder Graph, der, ausgehend von G, durch eine Folge von elementaren Einschränkungen und elementaren Kontraktionen entsteht.*

Theorem 11.11. ***Der Satz von Kuratowski:***
Ein endlicher Graph ist genau dann planar, wenn weder der vollständige Graph K_5, noch der vollständig bipartite Graph $K_{3,3}$ ein Minor von G ist.

Bemerkung 11.12. *Ein Platonischer Körper P ist eine konvexe Teilmenge von $\mathbb{R}^3$, dessen Oberfläche - für ein $n \in \mathbb{N}$ - aus lauter regelmäßigen n-Ecken besteht und in dem von jeder Ecke gleich viele - etwa k - Kanten ausgehen.*
In Platonischen Körpern gilt auch die Eulersche Formel; hier ist nun f die Anzahl der n-Ecke.

Klassifikation

Name	n	k	v	e	f
Tetraeder	3	3	4	6	4
Würfel	4	3	8	12	6
Oktaeder	3	4	6	12	8
Ikosaeder	3	5	12	30	20
Dodekaeder	5	3	20	30	12