Mathematik I

Vorlesungsskript

Falk Jonatan Strube

Vorlesung von Herrn Meinhold

16. November 2015

Inhaltsverzeichnis

I. Elementare Grundlagen 1

1. Aussagen und Grundzüge der Logik 1

2. Mengen 1

3. Zahlen 1
3.1. Gruppen, Ringe, Körper 1
3.2. Zahlentheorie 2
3.3. Reelle Zahlen 5
3.3.1. Algebraische Struktur 5
3.3.2. Zahlendarstellung im Computer 7
3.3.3. Ordnungsstruktur 10

Teil I. Elementare Grundlagen

1. Aussagen und Grundzüge der Logik

2. Mengen

3. Zahlen

3.1. Gruppen, Ringe, Körper

- Gegeben sei eine Menge M und eine zweistellige Operation $\circ$ (d.h. Abb. von $M \times M$ in M)
  Bezeichnung: $(M, \circ)$, analog $(M, \circ, *)$
- Die Operation $\circ$ heißt *kommutativ*, wenn $a \circ b = b \circ a$ für alle $a, b \in M$.
- Die Operation $\circ$ heißt *assoziativ*, wenn $(a \circ b) \circ c = a \circ (b \circ c)$ für alle $a, b, c \in M$.

Def. 1:
$(M, \circ)$ heißt *Gruppe*, wenn gilt:

1.) Die Operation $\circ$ ist assoziativ

2.) Es gibt genau ein *neutrales Element* $e \in M$ mit $a \circ e = e \circ a = a$ (für alle $a \in M$)

3.) Es gibt zu jedem $a \in M$ genau ein *inverses Element* a^{-1} mit $a \circ a^{-1} = a^{-1} \circ a = e$

4.) Eine Gruppe heißt *ABELsch*, wenn zusätzlich folgendes gilt:
$\circ$ ist kommutativ

Def. 2:
$(M, \oplus, *)$ heißt *Ring*, wenn gilt:

1.) $(M, \oplus)$ ist eine ABELsche Gruppe.

2.) Die Operation $*$ ist assoziativ.

3.) Es gelten für beliebige $a, b, c \in M$:
$a * (b \oplus c) = (a * b) \oplus (a * c)$
$(a \oplus b) * c = (a * c) \oplus (b * c)$ (Distributivgesetze)

4.) Ein Ring heiß *kommutativer Ring*, wenn gilt:
$*$ ist kommutativ

Def. 3:
$(M, \oplus, *)$ heißt *Körper*, wenn gilt:

1.) $(M, \oplus, *)$ ist ein Ring
(mit dem neutralen Element E_0 für die Operation $\oplus$)

2.) $(M \setminus \{E_0\}, *)$ ist eine ABELsche Gruppe
(mit dem neutralen Element E_1 für die Operation $*$)

3.2. Zahlentheorie

- Eine natürliche Zahl $p > 1$, die nurch durch 1 und sich selbst teilbar ist heißt *Primzahl*.
- Jede natürliche Zahl $n > 1$ ist entweder eine Primzahl, oder sie lässt sich als Produkt von Primzahlen schreiben.
  Diese sogenannte *Primfaktorzerlegung* ist bis auf die Reihenfolge der Faktoren eindeutig.

Def. 4:
Zwei natürliche zahlen aus $\mathbb{N}^*$ heißen *teilerfremd*, wenn sie außer 1 keine gemeinsamen teiler besitzen.

- Es sei $a \in \mathbb{Z}$ und $m \in \mathbb{N}^*$. Dann gibt es eine eindeutige Darstellung der Gestalt $\boxed{a = q \cdot m + r}$ mit $\boxed{0 \leq r < m}$ und $q \in \mathbb{Z}$.
  Bezeichnung: m ... *Modul* r ... (kleinste nichtnegative) *Rest modulo m* ($r \equiv mod(a, m)$)
- Zur Erinnerung: a und b seien ganze Zahlen, $m \in \ltimes^*$, dann $a \equiv b(mod\ \ m)$ [a kongruent $b\ modulo\ m$]
  $\Leftrightarrow$ a und b haben den gleicher Rest $modulo\ m$
  $\Leftrightarrow$ $a - b$ ist durch m teilbar (d.h. $\exists k \in \mathbb{Z} \quad a - b = k \cdot m$)

  Satz 1:
  Es sei $a \equiv b(mod\ m)$, $c \equiv d(mod\ m)$, dann gilt: $a + c \equiv b + d(mod\ m)$ und $a \cdot c \equiv b \cdot d(mod\ m)$ (d.h. in Summen und Produktenn darf jede Zahl durch einen beliebigen Vertreter der gleichen Restklasse ersetzt werden).

  Bsp. 1:
  a) $307 + 598 \equiv 1 + (-2) \equiv -1 \equiv 5(mod\ 6)$
  b) $307 \cdot 598 \equiv 1 \cdot (-2) \equiv -2 \equiv 4(mod\ 6)$
  c) $598^6 \equiv (-2)^6 \equiv 64 \equiv 4(mod\ 6)$

- Man wählt aus jeder Restklasse den kleinsten nichtnegativen Vertreter
  $\curvearrowright$ Menge von Resten $modulo\ m$: $\mathbb{Z}_m := \{0, 1, ..., m - 1\}$
  $\curvearrowright$ „modulare Arithmetik“: Operation $\oplus$ und $\odot$ für Zahlen aus $\mathbb{Z}_m$ erklärbar, in dem für das Ergebnis jeweils der kleinste nichtnegative Rest $modulo\ m$ gewählt wird (vgl. Satz 1)
  z.B. $\mathbb{Z}_7 = \{0, 1, ..., 6\}$, $\quad 5 \oplus 4 = 2$, da $5 + 4 \equiv 9 \equiv 2(mod\ 7)$ $\quad 5 \odot 6 = 2$, da $5 \cdot 6 \equiv 30 \equiv 2(mod\ 7)$
  Falls keine Verwechselung zu befürchten ist, wird die übliche Schreibweise $+$ und $\cdot$ anstelle von $\oplus$ und $\odot$ verwendet.

Def. 5:
Wenn es zu $c \in \mathbb{Z}_m$ eine Zahl $d \in \mathbb{Z}_m$ gibt, mit $c \cdot d \equiv 1(mod\ \ m)$ (bzw. $c \odot d \equiv 1$), so heißt d die *(multiplikative) modulare Inverse* zu c in $\mathbb{Z}_m$.
Bezeichnung: $d = c^{-1}$

Bsp. 2:
$c = 3 \in \mathbb{Z}_7$, wegen $3 \cdot 5 \equiv 1(mod\ 7)$ ist (in $\mathbb{Z}_7$) $3^{-1} = 5$.

Satz 2: Zu $a \in \mathbb{Z}_m, a \neq 0$, gibt es genau dann eine modulare Inverse in $\mathbb{Z}_m$, wenn a und m teilerfremd sind ($ggT(a, m) = 1$).

Satz 3: Es sei p eine Primzahl. Dann ist $(\mathbb{Z}_m, \oplus, \odot)$ ein Körper.
Bemerkung: Falls m keine Primzahl ist, so ist $(\mathbb{Z}_m, \oplus, \odot)$ ein kommutativer Ring.

EUKLIDischer Algorithmus

- Verfahren zur Ermittlung des größten gemeinsamen Teilers t zweier positiver natürlicher Zahlen, $t = ggT(a,b)$.
- In erweiterter Form bietet der Algorithmus eine Möglichkeit zur Bestimmung der modularen Inversen von a zum Modul m (mit $a < m$ und a, m teilerfremd).

Satz 4: (EUKLIDischer Algorithmus)
Es seien $a, b \in \mathbb{N}^*, a > b$. Man bildet die endliche Folge
$r_0 := b,\ r_1 = mod\ (a,b),\ r_2 = mod\ (r_0, r_1), ...,\ r_n = mod\ (r_{n-2}, r_{n-1})$, Abbruch falls $r_n = 0$.
In diesem Fall gist $\boxed{ggT(a,b) = r_{n-1}}$ (letzter nicht verschwindender Rest).
Bezeichnung: j-te Division ... $\boxed{r_{j-2} : r_{j-1} = q_j \text{ Rest } r_j}$ $(j = 1, ..., n)$ (dabei $r_1 := a$).

Satz 5: (erweiterter EUKLIDischer Algorithmus)
Zusätzlich zur Folge (r_n) aus Satz 4 bilde man die Folgen
$x_0 = 0,\ x_1 = 1,\ x_2 = x_0 - q_2 x_1, ...,\ x_j = x_{j-2} - q_j x_{j-1} \quad (j \leq n-1)$ und
$y_0 = 1,\ y_1 = -q_1,\ y_2 = y_0 - q_2 y_1, ...,\ y_j = y_{j-2} - q_j y_{j-1} \quad (j \leq n-1)$
Dann gilt für alle $j = 0, ...,\ n-1$: $\boxed{r_j = x_j \cdot a + y_j \cdot b}$
Insbesondere gilt $\boxed{ggT(a,b) = x_{n-1} \cdot a + y_{n-1} \cdot b}$

Diskussion:

1.) Der Sinn der erweiterten EUKLIDischen Algorithmus besteht darin, in jedem Schrit den *Divisionsrest r als linearkombination von a und b mit ganzzahligen Koeffizienten x und y* darzustellen:
$r = x \cdot a + y \cdot b$
Der Mechanismus wird am besten im Rechenschema des nachfolgenden Bsp. 4 deutlich.

2.) Sind c und m teilerfremd, $1 \leq c < m$, d.h. $ggT(m,c) = 1$, so erhält man mit dem erweiterten EUKLIDischen Algorithmus ($a = m, b = c$) eine Darstellung in der Form $\boxed{1 = x \cdot m + y \cdot c}$.
$\curvearrowright y \cdot c \equiv 1 (mod\ m)$ und damit $c^{-1} \equiv y (mod\ m)$ (für die modulare Inverse muss eventuell noch der in $\mathbb{Z}_m$ liegende, zu y kongruente, Wert gebildet werden!).

Bsp. 3:
Man ermittle den größten gemeinsamen Teiler t sowie das kleinste gemeinsame Vielfache v der Zahlen 132 und 84.

- Es genügt der „einfache“ Algorithmus:

$$\begin{aligned} 132 : 84 &= 1 \text{ Rest } 48 \\ 84 : 48 &= 1 \text{ Rest } 36 \\ 48 : 36 &= 1 \text{ Rest } 12 \qquad \curvearrowright t = ggT(132, 84) = \underline{\underline{12}} \\ 36 : 12 &= 3 \text{ Rest } \boxed{0} \curvearrowright \text{ Ende.} \end{aligned}$$

- $v = \dfrac{a \cdot b}{t} = \dfrac{132 \cdot 84}{12} = \underline{\underline{924}} = kgV(132, 84)$

Bsp. 4:

Man ermittle die modulare Inverse von $\overbrace{11}^{b}$ zum Modul $\overbrace{25}^{a}$.

		$11 = 0 \cdot 25 + 1 \cdot 11$ (1)
$25 : 11 = 2$ Rest 3	$3 = 25 - 2 \cdot 11$	$3 = 1 \cdot 25 - 2 \cdot 11$ (2)
$11 : 3 = 3$ Rest 2	$2 = 11 - 3 \cdot 3$	$2 = -3 \cdot 25 + 7 \cdot 11$ (3)
$3 : 2 = 1$ Rest 1	$1 = 3 - 2$	$\boxed{1} = 4 \cdot 25 - 9 \cdot 11$
$2 : 1 = 2$ Rest 0		

$\curvearrowright (-9) \cdot 11 \equiv 1 (mod\ 25)$
$\curvearrowright 11^{-1} \equiv -9 \equiv 16 (mod\ 25)$, die Inverse von 11 in $\mathbb{Z}_{25}$ ist 16.
Zu den Schritten:

(1) $b = 0 \cdot a + 1 \cdot b$

(2) mittleres Feld als Linearkombination

(3) ab hier Rechnung links spaltenweise durchführen, dabei Faktoren a und b beibehalten.

EULERsche φ-Funktion, Satz von EULER

Def. 6:
Es sei $n \in \mathbb{N}^*$. Dann *EULERsche φ-Funktion*:
$\varphi(n) :=$ Anzahl der zu n teilerfremden Elemente aus $\{1, 2, ..., n\}$. Eigenschaften der φ-Funktion:

- Es sei p eine Primzahl, dann ist $\boxed{\varphi(p) = p - 1}$, $\boxed{\varphi(p^k) = p^{k-1}(p-1)}$ $(k \in \mathbb{N}^*$
- Falls $ggT(m, n) = 1$, so gilt $\varphi(m \cdot n) = \varphi(m) \cdot \varphi(n)$.
- Speziell: $n = p \cdot q$ (p, q Primzahlen), dann $\boxed{\varphi(n) = (p-1) \cdot (q-1)}$ (1).

Satz 6: (Satz von EULER)
Es sei $ggT(a, n) = 1$, dann gilt:
$\boxed{a^{\varphi(n)} \equiv 1 (mod\ n)}$ (2).

RSA-Verschlüsselung

- Die Formeln (1) und (2) [siehe oberhalb] bilden die Grundlage für die sogenannte RSA-Verschlüsselung (RIVES, SHAMIR, ADLEMAN - 1978)
- Schlüsselerzeugung:
  1.) Man wählt (in der Praxis sehr große) Primzahlen d und q.
  2.) $n := p \cdot q$, $m := \varphi(n) \overset{(1)}{=} (p-1)(q-1)$
  3.) e wird so gewählt, dass $ggT(e, m) = 1$
  4.) $d := e^{-1} (mod\ m)$ (modulare Inverse)
  5.) (n, e) ... öffentlicher Schlüssel
  (n, d) ... geheimer Schlüssel (geheim ist nur d)
  p, q und m werden nicht mehr benötigt, bleiben aber geheim!

- Verschlüsselung:
  Klartext a teilerfremd zu n verschlüsseln mit e, d.h. $b :\equiv a^e (mod\ n)$ bilden ($b\ldots$ Geheimtext)
- Entschlüsselung:
  Der Empfänger und Besitzer des geheimen Schlüssels bildet $b^d(mod\ \ m)$ und erhält $b^d \equiv a(mod\ n)$ denn $b^d \equiv (a^e)^d \equiv a^{ed} \equiv a^{1+k\cdot m} \equiv a \cdot \underbrace{\left(a^{\varphi(n)}\right)}_{\equiv 1 \text{ wegen Satz 5}}{}^{k} \equiv a(mod\ n)$.
- Praktische Durchführung vgl Übungsaufgabe 2.4

3.3. Reelle Zahlen

$\mathbb{R}\ldots$ Menge der reellen Zahlen.
Auf $\mathbb{R}$ existiert eine algebraische Struktur und eine Ordnungsstruktur.

3.3.1. Algebraische Struktur

$(\mathbb{R}, +, \cdot)$ mit den arithmetischen Operationen $+$ (Addition) und $\cdot$ (Multiplikation) ist ein Körper.

Def. 7:

a.) $\boxed{0! := 1,\ n! = n\cdot(n-1)!}$ mit $n \in \mathbb{N}^*$ *Fakultät* (rekursive Funktion)

b.) Sei $\alpha \in \mathbb{R}, k \in \mathbb{N}^*$, dann sei $\boxed{\binom{\alpha}{0} := 1, \binom{\alpha}{k} := \frac{\alpha}{k}\binom{\alpha-1}{k-1}}$ Binominalkoeffizient α über k.

d.h. $\boxed{\binom{\alpha}{k} = \frac{\alpha(\alpha-1)(\alpha-2)...(\alpha-k-1)}{k!}}$

Diskussion:

1.) Für $k, n \in \mathbb{N},\ 0 \le k \le n$ gilt $\boxed{\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k!(n-k)!}}$

2.) Für $k \in \mathbb{N}, \alpha \in \mathbb{R}$ gilt $\boxed{\binom{\alpha}{k} + \binom{\alpha}{k+1} = \binom{\alpha+1}{k+1}}$

3.) *Binomischer Lehrsatz*: $(a+b)^k = \sum_{k=0}^{n}\binom{n}{k} a^{n-k}\cdot b^k = a^n + n\cdot a^{n-1}\cdot b + \binom{n}{2} a^{n-2}\cdot b^2 + ... + b^n$

Pascalsches Dreieck:

$$
\begin{array}{ccccccccccc|l}
 & & & & & 1 & & & & & & n=0 \\
 & & & & 1 & & 1 & & & & & 1 \\
 & & & 1 & & 2 & & 1 & & & & 2 \\
 & & 1 & & 3 & & 3 & & 1 & & & 3 \\
 & 1 & & 4 & & 6 & & 4 & & 1 & & 4 \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 & 5
\end{array}
$$

(Spalte $\hat{=} k$ in $\binom{n}{k}$)

Stellenwertsysteme:

- Es sei $k > 1$ eine natürliche Zahl (die sogenannte Basis)
- $x = (\underbrace{x_p x_{p-1} ... x_1 x_0}_{\text{Vorkomma}}, \underbrace{x_{-1} x_{-2} ... x_{-q}}_{\text{Nachkomma}})_b$
  $:= \underbrace{x_p \cdot b^p + x_{p-1} \cdot b^{p-1} + ... + x_1 \cdot b^1 + x_0 \cdot b^0}_{\text{Vorkomma}} + \underbrace{x_{-1} \cdot b^{-1} + x_{-2} \cdot b^{-2} + ... + x_{-q} \cdot b^{-q}}_{\text{Nachkomma}}$
  heißt Darstellung zur Basis b (*).

Bsp. 5:

- $b = 2 \ldots$ *Dual- oder Binärsystem* (Ziffern $\{0, 1\}$)
- $b = 3 \ldots$ Trialsystem
- $b = 10 \ldots$ *Dezimalsystem*
- $b = 16 \ldots$ *Hexadezimalsystem* (Ziffern $\{0, 1, 2, ..., 9, \underset{10}{A}, \underset{11}{B}, \underset{12}{C}, \underset{13}{D}, \underset{14}{E}, \underset{15}{F}\}$

z.B. $(47)_{10} = (101111)_2 = (1202)_3 = (57)_8 = (2F)_{16}$

Übergang von einem Ziffernsystem zu einem anderen
z.B. $p = 3,\ q = 2$

$$\begin{aligned} x &= x_3 b^3 + x_2 b^2 + x_1 b^1 + x_0 + x_{-1} b^{-1} + x_{-2} b^{-2} \\ &= (x_3 b^2 + x_2 x^1 + x_1) b + x_0 + (x_{-1} + x_{-2} b^{-1}) b^{-1} \\ &= ((x_3 b^1 + x_2) b + x_1) b + x_0 + (x_{-1} + x_{-2} b^{-1}) b^{-1} \end{aligned}$$

Grundlage: fortgesetztes Klammern:

$$\begin{aligned} x = ((....((x_p b + x_{p-1}) b + x_{p-2}) b + ... + x_2) b + x_1) b + x_0 + \\ ((...(x_{-q} b^{-1} + x_{-(q-1)}) b^{-1} + ... + x_{-2}) b^{-1} + x_{-1}) b^{-1} \end{aligned}$$

(**)

Bsp. 6: Übergang Dezimalsystem $\rightarrow$ anderes System

- ganzer Teil: fortgesetzte Division durch b und Restabspaltung liefert b-Ziffern in der Reihenfolge $x_0, x_1, x_2, ...$
- gebrochener Teil: fortgesetzte Multiplikation mit b und Abspaltung des ganzzahligen Anteils liefert b-Ziffern in der Reihenfolge $x_{-1}, x_{-2}, ...$

z.B. Dezimalsystem $\rightarrow$ Hexadezimalsystem ($b = 16$)
$x = 435, 9$
ganzer Teil:

$435 : 16 =$	27	Rest 3	$\rightarrow x_0$
$27 : 16 =$	1	Rest 11	$\rightarrow x_1$
$1 : 16 =$	$\boxed{0}$	Rest 1	$\rightarrow x_2$

gebrochener Teil:

$0,9 \cdot 16 =$	0.4	$+14$	$\rightarrow x_{-1}$
$0,4 \cdot 16 =$	$\boxed{0.4}$	$+6$	$\rightarrow x_{-2}$ (Periode, da gleicher „Nachkommarest“)

$\curvearrowright x = (1B3, E\overline{6})_{16}$

Bsp. 7: Übergang anderer Systeme $\rightarrow$ Dezimalsystem
Entweder direkte Auswertung von (*) (v.a. beim Dualsystem $\rightarrow$ Addition von 2er-Potenzen) *oder* Klammern in (**) von innen nach außen berechnen (zweckmäßig HORNER-Schema).

- ganzer Teil: beginnend mit x_p
  ABB1
- gebrochener Teil: beginnend mit x_{-q}
  ABB 2

$x = (1E2, B8)_{16}$
ganzer Teil:

$$\begin{array}{c|ccc} & 1 & 14 & 2 \\ 16 & & 16 & 480 \\ \hline & 1 & 30 & \boxed{482} \end{array}$$

gebrochener Teil:

$$\begin{array}{c|ccc} & 8 & 11 & * \\ \frac{1}{16} = 0,0625 & & 0,5 & \boxed{0,71875} \\ \hline & 8 & 11,5 & \end{array}$$

$\curvearrowright x = \underline{\underline{482,71875}}$

Bsp. 8: Hexadezimalsystem $\leftrightarrow$ Dualsystem
4 Dualziffern entsprechen einer Hexadezimalziffer ($2^4 = 16$) $\curvearrowright$ 4er Gruppen von Dualziffern ab Komma bilden.

a.) $(A8C, B2)_{16} = (1010\ 1000\ 1100,\ 1011\ 1011\ 001(0))_2$

b.) $((0)110\ 1110,\ 101(0))_2 = (6E, A)_{16}$

3.3.2. Zahlendarstellung im Computer

1.) Ganze Binärzahlen in Zweierkomplementdarstellung (n Bit, meist $n = 8, 16, 32, 64$)

- Bsp.: $n = 8$ $\quad (100)_{10} = (64)_{16}$

0	1	1	0	0	1	0	0
2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0

  MSB: most siginficant bit
  (LSB: least significant bit)
- Um auch negative Zahlen darstellen zu können, wird das MSB als Vorzeichen reserviert. Negative Zahlen $-a$ ($1 \leq a \leq 2^{n-1}$) werden im sogenannten Zweierkomplement $\overline{a} := 2^n - a$ dargestellt ($\curvearrowright \overline{a} \geq 2^{n-1} \curvearrowright MSB = 1$)
- Nightnegtavie Zahlen $0 \leq a \leq 2^{n-1} - 1$ werden unverändert dargestellt ($MSB = 0$)
- Damit Darstellung ganzer Zahlen von -2^{n-1} bis $2^{n-1} - 1$ (Anzahl 2^n) möglich, $n = 8$: $-128 bis 127$.
- Umwandlung negativer Zahlen $\rightarrow$ Zweierkomplement

Bsp. 9: $n = 8$, umzuwandeln sei -100 (dezimal)
zwei Möglichkeiten:

1.) (für die Handrechnung): $\overline{100} = \underbrace{2^8}_{=256} - 100 = \underbrace{156}_{(9C)_{16}} = \boxed{1 \mid 0 \; 0 \; 1 \mid 1 \; 1 \; 0 \; 0}$

Bemerkung: Das Zweierkomplement der positiven Zahl 100 ist die positive Zahl $156 = \overline{100}$, diese wird wegen $MSB = 1$ als negative Zahl -100 interpretiert.

2.) (am schnellsten): Rechts (beim LSB) beginnend alle Ziffern bis einschließlich der ersten 1 übernehmen (unverändert lassen), für alle höherwertigen Ziffern Z das *Einerkomplement* $1 - z$ bilden:

$(100)_{10} = \boxed{0 \mid 1 \; 1 \; 0 \mid 0 \; 1 \; 0 \; 0}$

$\boxed{1 \mid 0 \; 0 \; 1 \mid 1 \; 1 \; 0 \; 0} = (\overline{100})_{10} = (156)_2$

Rückumwandlung (Zahl mit $MSB = 1 \to$ neg. Zahl) analog:
$\overline{156} = 256 - 156 = 100 \to \underline{\underline{-100}}$
Die Subtraktion wird damit auf die Addition des Zweierkomplements zurückgeführt.

Bsp. 10: $a = 64 - 100 = 64 + (-100)$

$$
\begin{aligned}
64 = 2^6 &= 0100\,0000 \\
-100 &= 1001\,1100 \quad + \\
Summe &= 1101\,1100 \quad \text{Ergebins negativ} \\
36 &= 0010\,0100 \quad \text{dargestellst ist aber } \overline{z} = 2^n - z
\end{aligned}
$$

$\curvearrowright$ Ergebnis: $a = -36 = -z$
Ein *Überlauf* (Ergebnis $\geq 2^{n-1}$ oder $< -2^{n-1}$) ensteht in folgenden Fällen ($\to$ ERROR!)

	a	b	$a+b$	
MSB	0	0	1	($MSB = 0$ erwartet!)
MSB	1	1	0	($MSB = 1$ erwartet!)

Bemerkung: Für die Handrechnung (z.B. $2 - 5 =: a$) kleinere zahl von größerer Subtrahieren $a = -(5-2)$, dabei genügt es für n die Binärstellenzahl des Minuenden $(5)_{10} = (101)_2$ also $n = 3$ zu verwenden. Es wird dabei ausschließlich mit nicht-negativen Zahlen gerechnet $(0, 1, ..., 2^n - 1)$:
$(5-2)_{10} = ((5 + 2^n - 2) - 2^n)_{10} = (5 + \overline{2} - 2^n)_{10}$
$(2)_{10} = (010)_2 \curvearrowright \overline{2} = (110)_2$
$(5)_{10} = (101)_2$

$$
\begin{array}{ll}
5: & 101 \\
\overline{2}: & 110 \quad + \\
1 & 011
\end{array}
$$

vordere Stelle 2^n ignorieren
$\curvearrowright 5 - 3 = 3 = (011)_2 \curvearrowright \underline{\underline{a = -3}}$

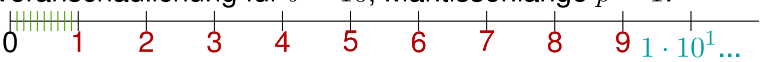
2.) Gleitkommasysteme
$\boxed{x = v \cdot m \cdot b^e}$ dabei

- $v = (-1)^V \ldots$ *Vorzeichen* $\begin{cases} V = 0 & \text{positive Zahl} \\ V = 1 & \text{negative Zahl} \end{cases}$
- $m \ldots$ *Mantisse*, Stellenzahl p, die Mantisse heißt *normalisiert* falls sie folgende Gestalt besitzt:

$m_1, m_2, ..., m_p$ oder $0, m_1, m_2, ..., m_p$ mit $m \neq 0$. Dabei sind $m_1, m_2, ..., m_p$ die Ziffern zur Basis b.

- e ... *Exponent*, ganzzahlig $e_{min} \leq e \leq e_{max}$.

In jedem Gleitkommasystem sind nur endlich viele Zahlen darstellbar, die Menge der reellen zahlen ist aber überabzählbar (unendlich).
Gleitkommazahlen liegen auf der Zahlengeraden diskret verteilt (fester Exponent $\curvearrowright$ gleiche Abstände, wächst Exponent um k, so wachsen die Abstände auf b^k-fache!)
Veranschaulichung für $b = 10$, Mantissenlänge $p = 1$:



Exponent 0: $1 \cdot 10^0, 2 \cdot 10^0, ..., 9 \cdot 10^0$
Exponent -1: $1 \cdot 10^{-1}, 2 \cdot 10^{-1}, ..., 9 \cdot 10^{-1}$
Exponent 1: $1 \cdot 10^1, 2 \cdot 10^1, ..., 9 \cdot 10^1$
Rundung: Zahlen, die nicht in diesem „Raster“ enthalten sind, werden auf dei nächstgelegene darstellbare Gleitkommazahl gerundet. Falls die Zahl genau in der Mitte zwischen zwei darstellbaren Zahlen liegt, wird auf die gerade Zahl gerundet (Bsp. $3,75 \to 3,8$ oder $4,65 \to 4,6$ bei Rundung auf eine Stelle nach dem Komma).

Numerische Probleme beim Rechnen mit Gleitkommazahlen

- Kommutativ-, Assoziativ- und Distributivgesetze gelten im allgemeinen nicht mehr. Ursachen sind bspw. Ziffernauslöschung bei der Subtraktion fast gleicher Zahlen, Addition oder Subtraktion von Zahlen unterschiedlicher Größenordnung.

Bsp. 11:

1.) Man berechne $(a + b) + c$ und $a + (b + c)$ in einem System mit 3-Stelliger Mantisse: $a = 3,73 \cdot 10^6$, $b = -3,71 \cdot 10^6$ und $c = 6,42 \cdot 10^3$

   - $a + b = 0,02 \cdot 10^6 = 2,00 \cdot 10^4$ (Normalisierung)
     $c = 0,642 \cdot 10^4 = 0,64 \cdot 10^4$ (Exponentenangleichung und Rundung)
     $(a + b) + c = 2,64 \cdot 10^4 = \underline{\underline{26400}}$
   - $c = 0,00642 \cdot 10^6 = 0,01 \cdot 10^6$ (Exponentenangleichung und Rundung)
     $b + c = -3,,70 \cdot 10^6$
     $a + (b + c) = 0,03 \cdot 10^6 = 3,00 \cdot 10^4 = \underline{\underline{30000}}$
   - exakter Wert: $a + b + c = \underline{\underline{26420}}$

2.) Aufgabe der numerischen Mathematik ist es, die unvermeidlichen Genauigkeitsverluste beim Rechnen mit Maschinenzahlen durch optimale Organisation (Reihenfolge) der Rechnung und Fehleranalyse in Grenzen zu halten.

3.) Gleitkommaformat IEEE 754 (single precision, 32 Bit)
$x = v \cdot m \cdot b^e = (-1)^V \cdot 1, m_2\, m_3 ... m_2 4 \cdot 2^{E-B}$ ($b = 2$, Binärsystenm)

- Vorzeichen $V = 0 \curvearrowright$ positiv, $V = 1 \curvearrowright$ negativ (1 Bit)
- Mantisse m_1 im Binärsystem stets gleich 1.
  $\curvearrowright$ nur Abspeicherung von $\boxed{M = m_2\, m_3 ... m_2 4}$ (23 Bit)
- Exponent: Abgespeichert wird $\boxed{E := e + B}$
  mit dem sogenannten *Biaswert* $B = 127$ (Bias = Verzerrung) als nichtnegative 8-stellige Binärzahl, $e_{min} = -126$ ($E = 1$), $e_{max} = 127$ ($E = 254 = (1111\ 1110)_2$), die Grenzfälle

$E = (0000\ 0000)_2$ und $E = (1111\ 1111)_2$ sind für Sonderfälle (0, ∞, nichtdefinierte Werte) vorgesehen.

Abspeicherung in der Reihenfolge $\boxed{VEM}$ (32 Bit)

Bsp. 12: Umwandlung einer Dezimalzahl in das IEEE 754-Format (32-Bit), $x = 435,9$ (vgl. Bsp. 6)

1.) Konvertierung in Dualzahl (unter Verwendung von Bsp. 6/Hexadez.)
$x = 1B3,E\overline{6})_{16} = (1\ 1011\ 0011, 1100\ 0110\ 0110...)_2$

2.) Normalisierte Gleitkommadarstellung, Mantisse mit 23 Stellen nach dem Komma, Kommaverschiebung um 8 Stellen.
$\curvearrowright x = 1,\underbrace{1011\ 0011\ 1110\ 0110\ 0110\ 011}_{M}(0\ 0110...))_2 \cdot 2^8$ (Abrundung!)

3.) Exponent $e = 8 \curvearrowright E = e + B = 8 + 127 = 135 = \underbrace{(1000\ 0111)_2}_{E}$

4.) Vorzeichenbit $V = 0$ (da x positiv)

$\curvearrowright x:$ $\boxed{0\ 1000\ 0111\ 1011\ 0011\ 1110\ 0110\ 0110\ 011}$

Bsp. 13: IEEE 754$\rightarrow$ Dezimalzahl

$\boxed{1\ 1000\ 0011\ 0111\ 1100\ 0000\ 0000\ 0000\ 000}$

1.) $E = (1000\ 00111)_2 = 131 \curvearrowright e = E - B = 131 - 127 = \underline{\underline{4}}$

2.) $V = 1 \curvearrowright$ negativ, normalisierte Mantisse $1, M$
$\curvearrowright x = -(1,011111)_2 \cdot 2^4 = -(10111,11)_2$
$\curvearrowright x = -23,75$

Bemerkung:

1.) Neben dem single-Format gibt es in IEEE 754 das double-Format (54 Bit, V=1Bit, E=11Bit, M=52 Bit, B=1023) sowie das erweiterbare Format

2.) Zahlbereiche single: $1,401 \cdot 10^{-45}...3,403 \cdot 10^{38}$, double: $4,941 \cdot 10^{-324}...1,798 \cdot 10^{308}$

3.3.3. Ordnungsstruktur

- Durch $\leqq$ (auch $\leq$) ist auf $\mathbb{R}$ eine vollständige Ordnungsrelation erklärt.
- Verträglichkeit mit der algebraischen Struktur (für alle $x, y, z \in \mathbb{R}$):

(1) $x \leq y \quad \Rightarrow \quad x + z \leqq y + z$

(2) $(x \leq y) \wedge (z \geq 0) \quad \Rightarrow \quad x \cdot z \leq y \cdot z$
$(x \leq y) \wedge (z \leq 0) \quad \Rightarrow \quad x \cdot z \geq y \cdot z$

Für die strikte Ordnung $<$ gilt:

$\boxed{(x < y) \wedge (z < 0) \quad \Rightarrow \quad x \cdot > y \cdot z}$

Def. 8:

Sei x eine reele Zahl. Dann heißt $|x| := \begin{cases} x & \text{für } x \geq 0 \\ -x & x < 0 \end{cases}$ der (absolute) Betrag von x.

ABB 5

Vorzeichenfunktion $sgn(x) := \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$

Diskussion: Es gilt:

1.) $|a - b|$ = „Abstand der Zahlen a und b auf der Zahlengeraden“
ABB 6
(speziell: $|a|$ = „Abstand von a zum Ursprung“)

2.) $a = |a| \cdot sgn(a)$

3.) $|a| = |-a|, |ab| = |a| \cdot |b|, \left|\frac{a}{b}\right| = \frac{|a|}{|b|}$ (falls $b \neq 0$)

4.) $|a \pm b| \leq |a| + |b|$ (*Dreiecksungleichung*) für alle $a, b \in \mathbb{R}$

Lösung von Ungleichung

Bsp. 14: (Ungleichung mit Beträgen)
Gesucht sei die Lösungmenge L der reellen Zahlen, die die Ungleichung $|x - 1| < 3 + \frac{1}{2}x$ (*) erfüllen.

- kritische Stelle(n): Nullstellen des Terms innerhalb der Betragsstriche d.h. $x = 1 \curvearrowright$ Fallunterscheidung
ABB 7

- 1. Fall: $x - 1 < 0$ d.h. $x < 1$
in (*):
$-(x - 1) < 3 + \frac{1}{2}x \Leftrightarrow$
$-\frac{3}{2}x < 2 \Leftrightarrow$
$\underline{\underline{x > -\frac{4}{3}}}$
$\curvearrowright L_1 = \left\{x|(x < 1) \wedge x > \frac{4}{3}\right\} = \left(-\frac{4}{3}, 1\right)$

- 2. Fall $x - 1 \geq 0$ d.h. $x \geq 1$
in (*):
$x - 1 < 3 + \frac{1}{2}x \Leftrightarrow$
$\frac{1}{2}x < 4 \Leftrightarrow$
$\underline{\underline{x < 8}}$
$\curvearrowright L_2 = \{x|(x \geq 1) \wedge (x < 8)\} = [1, 8)$

- $\Rightarrow L = L_1 \cup L_2 = \underline{\underline{\left(-\frac{4}{3}, 8\right)}}$

Bsp. 15: (Ungleichung mit gebrochen rationalen Termen)
$\frac{x}{x+1} < 1$ (*)

- kritische Stelle(n): Nenner-Nullstellen, hier: $x = -1$.
  ABB 7 (äquiv. mit -1)

- 1. Fall: $x < -1$
  in (*):
  $\Leftrightarrow x > x + 1 \Leftrightarrow 0 > 1$ (Widerspruch)
  $L_1 = \emptyset$.

- 2. Fall: $x > -1$
  in (*):
  $\Leftrightarrow x < x + 1 \Leftrightarrow 0 < 1$ (wahre Aussage)
  $L_2 = \{x | x > -1\} = (-1, \infty)$

- $\Rightarrow L = L_1 \cup L_2 = \underline{\underline{(-1, \infty)}}$

Bsp. 16: (quadratische Ungleichungen)
$x^2 + 3x < 10 \Leftrightarrow$
$\left(x + \frac{3}{3}\right)^2 - \frac{9}{4} < 10 \Leftrightarrow$ (vereinfacht durch quadratische Ergännzung)
$\left(x + \frac{3}{2}\right)^2 < \frac{49}{4} \Leftrightarrow$
$\left|x + \frac{3}{2}\right| < \frac{7}{2} \Leftrightarrow$ (Äquivalenz siehe Übung)
$\frac{-7}{2} < x + \frac{3}{2} < \frac{7}{2} \Leftrightarrow$
$-5 < x < 2$
$L = \underline{\underline{(-5, 2)}}$