Kapitel 2

Reelle Zahlen

2.1 Der Körper der reellen Zahlen

Definition 2.1 (Gruppe)**.** *Sei G eine Menge und $\circ$ eine Verknüpfung auf G (d. h. $\forall x, y \in G.\ x \circ y \in G$, $x \circ y$ ist eindeutig). Das Paar $(G, \circ)$ heißt eine* Gruppe*, wenn folgende Eigenschaften erfüllt sind:*

1. $\forall x, y, z \in G.\ (x \circ y) \circ z = x \circ (y \circ z)$ *(Assoziativität)*

2. Es gibt ein Element $n \in G$ mit der Eigenschaft $\forall x \in G.\ n \circ x = x \circ n = x$ (Existenz des neutralen Elements*)*

3. Zu jedem $x \in G$ gibt es genau ein $\bar{x} \in G$ mit der Eigenschaft $x \circ \bar{x} = \bar{x} \circ x = n$ (Existenz des inversen Elements*)*

Falls zusätzlich $\forall x, y \in G.\ x \circ y = y \circ x$ (Kommutativität) gilt, spricht man von einer kommutativen *(oder* abelschen*) Gruppe.*

Beispiel 2.1 (Gruppe)**.** *1. Sei $G = \mathbb{Z}$ und $\circ$ die übliche Addition, dann ist $(\mathbb{Z}, +)$ eine kommutative Gruppe. Dabei ist $0 \in \mathbb{Z}$ das neutrale Element und $-x$ für $x \in \mathbb{Z}$ das inverse Element.*

2. Sei G die Menge der Paare (x, y) mit $x, y \in \mathbb{Q}$ und $x^2 + y^2 \neq 0$. Sei $\circ$ definiert durch $(x, y) \circ (u, v) = (xu - yv, xv + yu)$. Dann ist $(G, \circ)$ eine Gruppe mit dem neutralen Element $(1, 0)$ und dem inversen Element $(\frac{x}{x^2+y^2}, \frac{-y}{x^2+y^2})$.

Die Gruppenstruktur dient als Grundlage zur Definition des Körpers, der eine Struktur mit zwei Verknüpfungen definiert und die Basis für die Definition der reellen Zahlen bildet.

Definition 2.2 (Körper)**.** *Auf einer Menge K sind zwei Verknüpfungen $+$ und $\cdot$ mit folgenden Eigenschaften gegeben:*

1. $(K, +)$ ist eine kommutative Gruppe mit neutralem Element 0,

2. $(K \setminus \{0\}, \cdot)$ ist eine kommutative Gruppe mit neutralem Element 1,

3. $\forall x, y, z \in K.\ x \cdot (y + z) = (x \cdot y) + (x \cdot z)$ und $(x + y) \cdot z = (x \cdot z) + (y \cdot z)$ (Distributivgesetze).

Dann wird $(K, +, \cdot)$ *ein* Körper *genannt.*

Beispiel 2.2 (Körper)**.**

- $(\mathbb{Q}, +, \cdot)$ *ist ein Körper*
- $(\mathbb{R}, +, \cdot)$ *ist ein Körper*[1]

Wir bezeichnen das inverse Element für x bezüglich der Addition als $-x$. Zu jedem $x \in K$ existiert also ein $-x \in K$ mit $x + (-x) = 0$. Die Subtraktion wird durch $x - y = x + (-y)$ erklärt.
Wir bezeichnen das inverse Element für x bezüglich der Multiplikation als x^{-1} beziehungsweise $\frac{1}{x}$. Zu jedem $x \in K$ existiert also ein x^{-1} mit $x \cdot x^{-1} = 1$. Das inverse Element x^{-1} wird auch Kehrwert genannt.
Die Zahl 0 hat eine Sonderrolle, da sie kein inverses Element bezüglich der Multiplikation besitzt, d. h. die Division durch 0 ist nicht definiert.

Einige Rechenregeln:

i) $x + z = y + z \Rightarrow x = y$

ii) $x \cdot z = y \cdot z \wedge z \neq 0 \Rightarrow x = y$ (Kürzungsregeln)

iii) $x \cdot 0 = 0 \cdot x = 0$ (Multiplikation mit 0)

iv) $-x = (-1) \cdot x$

v) $(-x) \cdot y = x \cdot (-y) = -(x \cdot y)$

vi) $(-x) \cdot (-y) = x \cdot y$

vii) $y, w \neq 0 \Rightarrow \frac{x}{y} + \frac{v}{w} = \frac{x \cdot w + v \cdot y}{y \cdot w}$

viii) $y, w \neq 0 \Rightarrow \frac{x}{y} \cdot \frac{v}{w} = \frac{x \cdot v}{y \cdot w}$

Beispiel 2.3 (Beweise einiger der obigen Regeln unter Nutzung der Körperaxiome)**.** *i)* $x + z = y + z \Rightarrow x = y$*:*
Addition von $-z$ *auf beiden Seiten der Gleichung liefert*

$$(x + z) + (-z) = (y + z) + (-z).$$

Mit dem Assoziativgesetz erhalten wir

$$x + (z + (-z)) = y + (z + (-z)).$$

Nach Definition von $-z$ *gilt* $z + (-z) = 0$*; daraus folgt*

$$x + 0 = y + 0$$

Da 0 *neutrales Element der Addition ist, ergibt sich schließlich*

$$x = y.$$

[1] Wir werden die reellen Zahlen als gegeben voraussetzen und ihre Eigenschaften in diesem Kapitel analysieren.

ii) $x \cdot z = y \cdot z \wedge z \neq 0 \Rightarrow x = y$*:*
Der Beweis ist analog zum Beweis von i) zu führen, wobei nun die Verknüpfung $\cdot$ *statt* $+$ *genutzt wird.*

iii)$x \cdot 0 = 0 \cdot x = 0$*:*
Es gilt auf Grund der Regel zur Addition des neutralen Elements und der Distributivgesetze:

$$x \cdot 0 = x \cdot (0 + 0) = (x \cdot 0) + (x \cdot 0)$$

sowie

$$x \cdot 0 = 0 + (x \cdot 0)$$

Damit gilt auch (unter Nutzung der Kürzungsregel ii)) $x \cdot 0 = 0$*. Unter Nutzung der Kommutativität gilt damit auch* $0 \cdot x = 0$*.*

vii)$y, w \neq 0 \Rightarrow \dfrac{x}{y} + \dfrac{v}{w} = \dfrac{x \cdot w + v \cdot y}{y \cdot w}$
Wir zeigen zuerst, dass $\dfrac{x}{y} = \dfrac{x \cdot v}{y \cdot v}$ *für* $v \neq 0$ *gilt. Nach Definition des inversen Elements bezüglich der Multiplikation gilt* $v \cdot v^{-1} = 1$*. Nach Definition des neutralen Elements bezüglich der Multiplikation gilt* $\frac{x}{y} \cdot 1 = \frac{x}{y}$ *und damit auch*

$$x \cdot y^{-1} = (x \cdot y^{-1}) \cdot 1 = (x \cdot y^{-1}) \cdot (v \cdot v^{-1}) = xv \cdot y^{-1}v^{-1} = \frac{xv}{yv} \, .$$

Daraus folgt, dass auch

$$\begin{array}{rcccl} \dfrac{x}{y} + \dfrac{v}{w} & = & \dfrac{x \cdot w}{y \cdot w} + \dfrac{y \cdot v}{y \cdot w} & = & (yw)^{-1}(xw) + (yw)^{-1}(yv) \\ & = & (yw)^{-1} \cdot ((xw) + (yv)) & = & \dfrac{xw + yv}{yw} \end{array}$$

gilt.

Konvention: Wir nutzen folgende Schreibweisen für $a \in \mathbb{R}$ und $n \in \mathbb{Z}$:

$$a^n := \begin{cases} 1 & \text{falls } n = 0 \\ a \cdot a^{n-1} & \text{falls } n > 0 \\ (a^{-n})^{-1} & \text{falls } n < 0 \text{ und } a \neq 0 \end{cases}$$

Der Sonderfall $a = 0$ ist nicht definiert für $n < 0$. Es gilt $0^0 = 1$ und $0^n = 0$ für $n > 0$. Zusätzlich gelten die sogenannten Potenzrechengesetze:

- $a^n \cdot a^m = a^{n+m}$
- $(ab)^n = a^n \cdot b^n$
- $(a^n)^m = a^{n \cdot m}$

a bezeichnet man als Basis und n bzw. m als Exponenten.

2.2 Anordungsaxiome

Definition 2.3 (Ordnung)**.** *Sei M eine Menge und ν eine Relation auf M (d. h. eine Teilmenge von $M \times M$). Für $(x, y) \in \nu$ schreiben wir $x \nu y$. Die Relation ν heißt eine* Ordnung *und (M, ν) ist eine* geordnete Menge*, falls folgende Bedingungen erfüllt sind:*

1. $\forall x \in M.\ x \nu x$ *(Reflexivität)*

2. $\forall x, y \in M.\ x \nu y \wedge y \nu x \ \Rightarrow\ x = y$ *(Antisymmetrie)*

3. $\forall x, y, z \in M.\ x \nu y \ \wedge\ y \nu z \ \Rightarrow\ x \nu z$ *(Transititivität)*

Gilt darüber hinaus

4. $\forall x, y \in M.\ x \nu y \ \vee\ y \nu x,$

so heißt ν eine lineare *(oder* totale*)* Ordnung *und (M, ν) eine* linear *(oder* total*)* geordnete Menge.

Beispiel 2.4 (Ordnung)**.**

- *Sei $\mathcal{P}(M)$ die Potenzmengen von M. Dann stellt für $x, y \in \mathcal{P}(M)$*

$$x \nu y \Leftrightarrow x \subseteq y$$

eine Ordnung, aber keine lineare Ordnung dar (nachrechnen).

- *Die Menge $\mathbb{N}$ mit der üblichen Relation $\leq$ ist eine linear geordnete Menge.*

- *Ein Programm besteht aus 4 Modulen, die wir der Einfachheit halber mit 1 bis 4 nummerieren. Zwischen diesen Modulen gibt es Abhängigkeiten, so benötigt Modul 3 die Eingaben von 1 und Modul 4 benötigt die Ausgaben der anderen 3 Module. Damit wird eine Ordnung definiert. Wenn das Programm ausgeführt werden soll, so gibt es verschiedene Möglichkeiten die einzelnen Module sequentiell oder auch teilweise parallel auszuführen, so dass die Ordnung eingehalten wird. So sind $(1, 2, 3, 4)$, $(2, 1, 3, 4)$, $(1, 3, 2, 4)$ mögliche Ausführungsreihenfolgen. Wir können aber 1 und 2 parallel ausführen und danach 3 und 4 sequentiell.*

Definition 2.4 ($\mathbb{R}$ als linear geordneter Körper)**.** *Wir definieren eine lineare Ordnung $\leq$ ("kleiner oder gleich") auf $\mathbb{R}$, sodass $(\mathbb{R}, \leq)$ eine linear geordnete Menge mit folgenden Eigenschaften ist:*

1. $\forall x, y, z \in \mathbb{R}$. *falls* $x \leq y \ \Rightarrow\ x + z \leq y + z$ *(Verträglichkeit mit der Addition)*

2. $\forall x, y, z \in \mathbb{R}$. *falls* $x \leq y \ \wedge\ 0 \leq z \ \Rightarrow\ xz \leq yz$ *(Verträglichkeit mit der Multiplikation)*

Die Beziehung $x \leq y$ heißt Ungleichung.

Definition 2.5. *Seien $x, y \in \mathbb{R}$:*

1. $y \geq x$ *("größer oder gleich") bedeutet* $x \leq y$

2. $x < y$ *("kleiner") bedeutet* $x \leq y \ \wedge\ x \neq y$

3. $x > y$ *("größer") bedeutet* $x \geq y \ \wedge \ x \neq y$

4. y *heißt* nichtnegativ *(bzw.* positiv*) wenn* $0 \leq y$ *(bzw.* $0 < y$*)*

5. x *heißt* nichtpositiv *(bzw.* negativ*) wenn* $x \leq 0$ *(bzw.* $x < 0$*).*

Das Rechnen mit Ungleichungen ist von großer praktischer Relevanz. Deshalb werden in den nachfolgenden Sätzen, die wir auch beweisen werden, die notwendigen Rechenregeln eingeführt.

Satz 2.6. *Für je zwei Elemente* $x, y \in \mathbb{R}$ *gilt genau eine der drei Beziehungen:* $x < y$, $x = y$, $x > y$.

Beweis: In zwei Schritten:

1. Mindestens einer der drei Beziehungen trifft zu

2. Höchstens eine der drei Beziehungen trifft zu.

Zu 1.: Die Linearität der Ordnung besagt, dass

$$x \leq y \text{ oder } x \geq y$$

gilt. Im ersten Fall gilt dann ($x \leq y$ und $x \neq y$) oder ($x \leq y$ und $x = y$), d. h. es gilt $x < y$ oder $x = y$. Im zweiten Fall gilt entsprechend ($x \geq y$ und $x \neq y$) oder ($x \geq y$ und $x = y$), also $x > y$ oder $x = y$. Damit ist 1. bewiesen.

Zu 2.: Da $x < y$ gleichzeitig $x \neq y$ bedeutet, können $x < y$ und $x = y$ nicht gleichzeitig erfüllt sein. Entsprechend können $x > y$ und $x = y$ nicht gleichzeitig erfüllt sein. Es bleibt zu zeigen, dass $x < y$ und $x > y$ nicht gleichzeitig gelten können. Wäre dies der Fall, so würde auch $x \leq y$ und $x \geq y$ gelten, woraus $x = y$ folgt, sodass weder $x < y$ noch $x > y$ gelten kann. Dies stellt einen Widerspruch dar. □

Satz 2.7. *Für alle* $a, b, c, d \in \mathbb{R}$ *gilt:*

1. $a < b \ \Rightarrow \ a + c < b + c$ *(Verträglichkeit mit der Addition)*

2. $a \leq b \wedge c \leq d \ \Rightarrow \ a + c \leq b + d$
   $a < b \wedge c \leq d \ \Rightarrow \ a + c < b + d$

3. $a < b \wedge 0 < c \ \Rightarrow \ ac < bc$ *(Verträglichkeit mit der Multiplikation)*

4. $0 \leq a \leq b \wedge 0 \leq c \leq d \ \Rightarrow \ ac \leq bd$
   $0 \leq a < b \wedge 0 < c \leq d \ \Rightarrow \ ac < bd$

5. $a \leq b \wedge c < 0 \ \Rightarrow \ ac \geq bc$
   $a < b \wedge c < 0 \ \Rightarrow \ ac > bc$

6. $0 < a \ \Rightarrow \ 0 < \dfrac{1}{a}$
   $0 < a < b \ \Rightarrow \ \dfrac{1}{b} < \dfrac{1}{a}$

7. $0 < 1$

Beweis: Hier nur teilweise:

1. Nach Definition von $<$ gilt $a < b \Rightarrow a \leq b$. Also folgt aufgrund der Verträglichkeit mit der Addition $a + c \leq b + c$. Gleichheit kann nicht gelten, da aus $a + c = b + c$ auch $a = b$ folgt, was aber $a < b$ widerspricht.

3. Da $a < b \Rightarrow a \leq b$ und $c > 0 \Rightarrow c \geq 0$ gilt:

$$ac \leq bc$$

Wäre $ac = bc$ so würde, da $c \neq 0$, $a = b$ folgen, was im Widerspruch zu $a < b$ steht.

5. Es gilt $-c > 0$ und damit gilt nach 3. $-ac \leq -bc$ (bzw. $ac < bc$). Nach 1. ergibt sich durch Addition von $ac + bc$ auf beiden Seiten $bc \leq ac$ (bzw. $bc < ac$).

6. Wegen $a \neq 0$ existiert $\frac{1}{a} \neq 0$. Wäre $\frac{1}{a} < 0$, so würde aus 5. die folgende Ungleichung folgen

$$0 < a \text{ und } 0 \cdot (\frac{1}{a}) > a \cdot \frac{1}{a} \Rightarrow 1 < 0$$

Dies kann nicht sein, da dann mit 5.

$$1 < 0 \Rightarrow 1 \cdot 1 > 0 \cdot 0$$

gilt. Dies stellt einen Widerspruch dar. Also kann weder $\frac{1}{a} < 0$ noch $\frac{1}{a} = 0$ gelten. Es bleibt nur $\frac{1}{a} > 0$.

7. Wir haben in 6. gezeigt, dass $1 < 0$ nicht gelten kann. Falls $1 = 0$, so würde

$$0 = 0 \cdot (x + y) = 1 \cdot (x + y) = 1 \cdot x + 1 \cdot y = x + y$$

für alle $x, y \in \mathbb{R}$ gelten.

Damit gilt, dass weder $1 < 0$ noch $1 = 0$ gelten kann, sodass $1 > 0$ gelten muss.

Die restlichen Beweisteile können zur Übung durch den Leser durchgeführt werden. □

Wir schreiben $\max(x, y) = x$, falls $x \geq y$ und y sonst und $\min(x, y) = x$, falls $x \leq y$ und y sonst, Die Schreibweise kann auf Mengen $X \subset \mathbb{R}$ ausgedehnt werden. D. h. $x = \max(X) \Leftrightarrow x \in M$ und $\forall y \in M.y \leq x$ und $x = \min(X) \Leftrightarrow x \in M$ und $\forall y \in M.y \geq x$. Es sollte beachtet werden, dass nicht für alle Mengen ein Maximum bzw. Minimum existiert.

2.3 Betrag und Dreiecksungleichungen

Definition 2.8 (Betrag)**.** *Für $x \in \mathbb{R}$ heißt*

$$|x| := \begin{cases} x & \text{falls } x \geq 0 \\ -x & \text{falls } x < 0 \end{cases}$$

der (Absolut-)Betrag *von x.*

Satz 2.9. *Der Absolutbetrag hat folgende Eigenschaften*

1. $\forall x \in \mathbb{R}.\ |x| \geq 0 \ \wedge\ (|x| = 0 \ \Rightarrow\ x = 0)$*)*

2. $\forall x, y \in \mathbb{R}.\ |x \cdot y| = |x| \cdot |y|$

3. $\forall x, y \in \mathbb{R}.\ |x + y| \leq |x| + |y|$ *(Dreiecksungleichung)*

Beweis:

1. Folgt unmittelbar aus der Definition

2. Trivial für $x, y \geq 0$. Ansonsten sei $x = \pm x_0, y = \pm y_0$ mit $x_0, y_0 \geq 0$, daher gilt:

$$|x \cdot y| = |\pm x_0 \cdot y_0| = |x_0 \cdot y_0| = |x_0| \cdot |y_0| = |x| \cdot |y|$$

3. Da $x \leq |x|$ und $y \leq |y|$ folgt aus Satz 2.7.2 $x + y \leq |x| + |y|$ ebenso wie $-(x+y) = -x - y \leq |x| + |y|$, da ebenfalls $-x \leq |x|$ und $-y \leq |y|$. Damit gilt auch $|x + y| \leq |x| + |y|$.

□

Satz 2.10. *Für* $a, x, \varepsilon \in \mathbb{R}$ *mit* $\varepsilon > 0$ *gilt:*

1. $|x| < \varepsilon \ \Leftrightarrow\ x < \varepsilon$ *und* $-\varepsilon < x \ \Leftrightarrow\ -\varepsilon < x < \varepsilon$

2. $|x - a| < \varepsilon \ \Leftrightarrow\ a - \varepsilon < x < a + \varepsilon$

3. Die Aussagen 1. und 2. gelten auch, wenn $<$ *durch* $\leq$ *ersetzt wird.*

Beweis:

1. Sei $|x| < \varepsilon$, da $-x \leq |x|$ und $x \leq |x|$ folgt aus $|x| < \varepsilon$: $x < \varepsilon$ und $-x < \varepsilon$. Durch Multiplikation mit -1 folgt $x > -\varepsilon$.

2. Ersetze in 1. x durch $x - a$, sodass $|x - a| < \varepsilon \Leftrightarrow -\varepsilon < x - a < \varepsilon$. Addiere a, sodass $a - \varepsilon < x < a + \varepsilon$.

3. Beweis analog zu 1. und 2. unter Verwendung von $\leq$ statt $<$.

□

Satz 2.11 (Bernoullische Ungleichung)**.** *Für alle* $x \in \mathbb{R}$ *mit* $x \geq -1$ *und alle* $n \in \mathbb{N}_0$ *gilt*

$$(1 + x)^n \geq 1 + n \cdot x$$

Beweis: Per Induktion über n.

Induktionsanfang $n = 0$**:**
$(1 + x)^0 = 1 = 1 + 0 \cdot x$ und damit auch $(1 + x)^0 \geq 1 + 0 \cdot x$

Induktionsschritt $n \to n + 1$**:**
Wir nehmen an, dass $(1 + x)^n \geq 1 + nx$ gilt. Da $1 + x \geq 0$ gilt dann auch

$$\begin{aligned}
(1 + x)^{n+1} &= (1 + x) \cdot (1 + x)^n \\
&\geq (1 + x) \cdot (1 + n \cdot x) \\
&= 1 + n \cdot x + x + n \cdot x^2 \\
&= 1 + (n + 1) \cdot x + n \cdot x^2 \\
&\geq 1 + (n + 1) \cdot x \qquad (\text{da } n > 0 \text{ und } x^2 \geq 0)
\end{aligned}$$

□

Archimedisches Axiom

Das Archimedische Axiom stellt den Zusammenhang zwischen natürlichen und reellen Zahlen her und lautet: Zu jedem $x \in \mathbb{R}$ gibt es ein $n \in \mathbb{N}$ mit $x < n$.
Für $x, y \in \mathbb{R}$ mit $0 < x < y$ existiert ein $n \in \mathbb{N}$, sodass $n \cdot x > y$ ist. Dies wird ohne Beweis anmerkt. Diese Aussage kann man sehr einfach geometrisch interpretieren. Wenn man zwei Stecken auf einer Geraden betrachtet, so muss man die kürzere nur oft genug abtragen, um die Länge der längeren Strecke zu übertreffen.

Abbildung 2.1: Geometrische Interpretation des Archimedischen Axioms.

Ein Körper, in dem das archimedische Axiom gilt, heißt archimedisch geordnet.
Die Körper $\mathbb{R}$ und $\mathbb{Q}$ sind archimedisch geordnet. Es sollte angemerkt werden, dass wir $\mathbb{R}$ bisher noch nicht formal vollständig definiert haben. Dies werden wir erst im Laufe des Kapitels tun.

2.4 Darstellung von Zahlen im Rechner

Die Darstellung von Zahlen im Rechner gehört üblicherweise nicht zu den Inhalten einer Analysis-Vorlesung in der Mathematik und wird darüber hinaus in anderen Informatikvorlesungen detailliert behandelt. Wir wollen deshalb nur einen kurzen Überblick über die Thematik erlangen, um die damit verbundenen Auswirkungen beim numerischen Rechnen mit dem Computer einordnen zu können.

2.4.1 Basis der Zahlendarstellung

Natürliche und ganze Zahlen

Natürliche bzw. ganze Zahlen können durch Zeichenketten fester Länge dargestellt werden.

Definition 2.12 (Stellenwertsystem)**.** *Seien* $a_0, a_1, \ldots, a_n \in \{0, 1, \ldots, 9\}$ *Ziffern, dann ist* $p = \sum_{i=0}^{n} a_i \cdot 10^i \in \mathbb{N}_0$ *und es existiert eine Darstellung* $b_0, b_1, \ldots, b_m \in \{0, 1\}$*, sodass* $p = \sum_{i=0}^{m} b_i \cdot 2^i \in \mathbb{N}_0$

- Ein zusätzliches Vorzeichenbit erlaubt die Darstellung von ganzen Zahlen.
- Die vorgegebene Länge definiert den darstellbaren Zahlenbereich.

Beispiel 2.5.
$2 \cdot 10^2 + 3 \cdot 10^1 + 7 \cdot 10^0 = 1 \cdot 2^7 + 1 \cdot 2^6 + 1 \cdot 2^5 + 0 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$

In einem Rechner hat der Zahlenbereich zur Darstellung ganzer Zahlen eine feste Länge. Unter Umständen gibt es verschiedene Bereiche wie zum Beispiel `short`, `int` und `long` in Java, die eine Länge von 15, 31 und 63 Bit ohne Vorzeichenbit haben. Damit lassen sich die entsprechenden Zahlenbereiche darstellen. Die Festlegung einer festen Stellenzahl bedingt zwangsläufig, dass mathematische Operationen dazu führen können, dass der darstellbare Zahlenbereich zur Darstellung des Ergebnisses nicht ausreicht und es zu einem Fehler (d. h. Überlauf kommt).
Bei der Darstellung von rationalen oder reellen Zahlen im Rechner unterscheidet man zwischen Festkomma- und Gleitkommadarstellungen, die wir im Folgenden kurz betrachten wollen.

Festkommadarstellung

Bei der Festkommadarstellung wird eine feste Zahl von Vor- und Nachkommastellen vorgegeben. Wenn wir von m Vorkommastellen und n Nachkommastellen ausgehen, so wird jede Zahl durch $n+m$ Bits dargestellt, d. h.

$$a_0, a_1, \ldots, a_n, a_{n+1}, \ldots, a_{n+m} \in \{0,1\}, \text{ sodass } x = \sum_{i=-n}^{m} a_{i+n} \cdot 2^i \in \mathbb{R}.$$

Damit wird natürlich nur eine Teilmenge von $\mathbb{Q}$ dargestellt, da für jedes darstellbare x gilt

$$x = \frac{1}{2^n} \sum_{i=0}^{m+n} a_i \cdot 2^i \quad \in \mathbb{Q}.$$

Wie bei den ganzen Zahlen werden negative Zahlen durch das Vorzeichenbit dargestellt und die vorgegebene Länge definiert darstellbaren Zahlenbereich.

Gleitkommadarstellung

Besser geeignet als Festkommazahlen, sind die so genannten Gleitkommazahlen, die auch als Fließkommazahlen bezeichnet werden. Eine Gleitkommazahl zur Basis b besteht aus eine Mantisse M und dem Exponenten E. Es gilt $x = M \cdot b^E$ wobei $E \in \mathbb{Z}$ und die Mantisse so gewählt wird, dass $b^{-1} \leq |M| < 1$, um zu einer eindeutigen Darstellung zu gelangen. Es gilt

- $M = \pm 0.m_1 m_2 \ldots m_t = \pm \sum_{j=1}^{t} m_j \cdot b^{-j}$ und
- $E = e_{s-1} \ldots e_1 e_0 = \pm \sum_{j=0}^{s-1} e_j \cdot b^j$.

Jedes als Gleitkommazahl darstellbare Zahl x gehört zu $\mathbb{Q}$, aber nicht jedes $x \in \mathbb{Q}$ ist darstellbar, auch wenn der Darstellungsbereich besser genutzt wird als bei der Festkommadarstellung. Es ist sogar so, dass eine Darstellung für eine Basis b nicht garantiert, dass eine endliche Darstellung zu einer anderen Basis b' existiert. Ein Beispiel ist die Darstellung von $\frac{1}{10}$, die zur Basis 10 problemlos ist aber zur Basis $b = 2$ keine endliche Darstellung besitzt, sondern nur approximiert werden kann.

2.4.2 Grenzen der Darstellung

Im Rechner gibt es üblicherweise standardisierte Darstellungen, die vom Institute of Electrical and Electronics Engineers (IEEE) festgelegt wurden:

- Single precision $t = 23$ und $E \in [-126, 127]$
  Die größte darstellbare Zahl $x_{max} \approx 3.40 \cdot 10^{38}$
  Die kleinste positive darstellbare Zahl $x_{min} \approx 1.18 \cdot 10^{-38}$
- Double precision $t = 52$ und $E \in [-1022, 1023]$
  Die größte darstellbare Zahl $x_{max} \approx 1.80 \cdot 10^{308}$
  Die kleinste positive darstellbare Zahl $x_{min} \approx 2.23 \cdot 10^{-308}$
- Zusätzlich spezielle Symbole ±INF oder NaN (**N**ot **a** **N**umber)
- Spezielle Software erlaubt die Nutzung frei definierbarer Darstellungen bei denen die Länge der Mantisse und des Exponenten definiert werden kann.

Rundungsfehler

Jede Zahl muss durch eine darstellbare Zahl dargestellt (bzw. approximiert) werden. Dabei tritt der im Folgenden kurz beschriebene Effekt auf:

- Sei $x = a \cdot 2^e$ mit $0.5 \leq a < 1$ und $x_{min} \leq x \leq x_{max}$.
- Seien u, v zwei benachbarte darstellbare Zahlen mit $u \leq x \leq v$.
- Sei $u = 2^e \cdot \sum_{i=1}^{t} b_i \cdot 2^{-i}$ dann ist $v = 2^e \cdot (\sum_{i=1}^{t} b_i \cdot 2^{-i} + 2^{-t})$ (wir nehmen zur Vereinfachung an, dass $\sum_{i=1}^{t} b_i \cdot 2^{-i} + 2^{-t}$ keinen Überlauf erzeugt), sodass $v - u = 2^{e-t}$ und $|rd(x) - x| \leq \frac{1}{2}(v - u) = 2^{e-t-1}$ mit $rd(x)$ als optimaler Darstellung von x.
- Der relative Fehler ist $\frac{|rd(x)-x|}{x} \leq \frac{2^{e-t-1}}{a \cdot 2^e} \leq 2^{-t}$. Der Wert 2^{-t} wird als relative Maschinengenauigkeit bezeichnet.

Beispiel 2.6. *Wir schauen uns ein einfaches Beispiel an und betrachten einen Rechner, der mit einem Dezimalsystem mit vierstelliger Mantisse arbeitet. Damit ist* $t = 4$ *und die Maschinengenauigkeit lautet* $eps = 0.5 \cdot 10^{1-4} = 0.0005 = 0.05\%$. *Dies bedeutet, dass der Wert einer einzelnen Berechnung um bis zu* 0.05% *vom exakten Wert abweichen kann. Wenn wir als Beispiel die Berechnung* $1.492 \cdot 1.066 = 1.590472$ *untersuchen, so wird dieser Wert auf* 1.590 *gerundet und der relative Fehler beträgt*

$$\frac{1.590 - 1.590472}{1.590472} \approx 0.0003 = 0.03\%.$$

Rundungsfehler treten bei der Darstellung von Zahlen auf und auch bei (fast) jeder Berechnung. In extremen Fällen können Rundungsfehler über mehrere Rechenschritte so akkumuliert werden, dass die berechneten Ergebnisse völlig unbrauchbar werden. Die endliche Zahlendarstellung bedingt, dass Ergebnisse von Berechnungen eine vorgegebene Genauigkeit nicht unterschreiten können und darüber hinaus, die Genauigkeit unter Umständen auch von der Reihenfolge von Operationen abhängt. Wir werden diesen Aspekt, der zum Beispiel in der Numerik behandelt wird, in der Vorlesung nicht vertiefen.

2.5 Intervalle

Um die reellen Zahlen genauer zu charakterisieren benötigen wir einige weitergehende Konzepte, die im Folgenden definiert werden. Zuerst wird dazu die Menge $\mathbb{R}$ so erweitert, dass der Begriff des Unendlichen mit einbezogen wird.

Definition 2.13 (Erweiterung von $\mathbb{R}$)**.**
Die Menge $\hat{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ *heißt* erweiterte reelle Zahlengerade*. Es gilt* $-\infty < \infty$ *und* $-\infty < x < \infty$ *für alle* $x \in \mathbb{R}$*.*

Zentral für die folgenden Ergebnisse sind Intervalle auf $\mathbb{R}$ (bzw. $\hat{\mathbb{R}}$), für die folgende Schreibweisen verwendet werden:

a) Abgeschlossene Intervalle
Seien $a, b \in \mathbb{R}, a \leq b$, dann ist $[a, b] := \{x \in \mathbb{R} \mid a \leq x \leq b\}$.

b) Offene Intervalle
Seien $a, b \in \mathbb{R}, a < b$, dann ist $(a, b) := \{x \in \mathbb{R} \mid a < x < b\}$
(Man schreibt manchmal auch $]a, b[$ statt (a, b))

c) Halboffene Intervalle
Seien $a, b \in \mathbb{R}, a < b$:

$[a, b) := \{x \in \mathbb{R} \mid a \leq x < b\}$
$(a, b] := \{x \in \mathbb{R} \mid a < x \leq b\}$

d) Uneigentliche Intervalle
Sei $a \in \mathbb{R}$:

$[a, +\infty) := \{x \in \mathbb{R} \mid x \geq a\}$
$(a, +\infty) := \{x \in \mathbb{R} \mid x > a\}$
$(-\infty, a] := \{x \in \mathbb{R} \mid x \leq a\}$
$(-\infty, a) := \{x \in \mathbb{R} \mid x < a\}$

Definition 2.14 (Länge eines Intervalls)**.** *Für ein abgeschlossenes Intervall* $[a, b]$ *bezeichnet* $|[a, b]| = b - a$ *die* Länge des Intervalls*.*

Wir benutzen den Begriff der Länge für abgeschlossene Intervalle, wobei das Intervall $[a, a]$, das nur aus dem Punkt a besteht, die Länge 0 hat. Man kann die Längendefinition auf halboffene, offene und uneigentliche Intervalle erweitern. Uneigentliche Intervalle haben dann die Länge ∞.

Einige weitere Bezeichnungen:

$\mathbb{R}_{>0} := \{x \in \mathbb{R} \mid x > 0\}$
$\mathbb{R}_{\geq 0} := \{x \in \mathbb{R} \mid x \geq 0\}$
$\mathbb{R}_{\neq 0} := \{x \in \mathbb{R} \mid x \neq 0\}$
$\hat{\mathbb{R}}_{>0} := \{x \in \hat{\mathbb{R}} \mid x > 0\}$
$\hat{\mathbb{R}}_{\geq 0} := \{x \in \hat{\mathbb{R}} \mid x \geq 0\}$
$\hat{\mathbb{R}}_{\neq 0} := \{x \in \hat{\mathbb{R}} \mid x \neq 0\}$

Intervalle definieren Mengen und es gilt z. B. $[a, b] \subset \mathbb{R}$. Die folgenden beiden Definitionen legen Eigenschaften von Teilmengen von $\mathbb{R}$ fest. Diese können Intervalle sein, die Eigenschaften gelten aber auch für andere Teilmengen von $\mathbb{R}$.

Definition 2.15 (Beschränkte Menge)**.** *Sei $A \subseteq \mathbb{R}$ nicht leer. A heißt* nach oben *(bzw.* nach unten*)* beschränkt*, wenn es eine Konstante $K \in \mathbb{R}$ gibt, sodass $x \leq K$ (bzw. $x \geq K$) für alle $x \in A$. Man nennt K dann* obere *(bzw.* untere*)* Schranke *von A. Die Menge A heißt* beschränkt*, wenn sie nach oben und nach unten beschränkt ist.*

Definition 2.16 (Supremum und Infimum)**.** *Sei $A \subset \mathbb{R}$ nicht leer. Eine Zahl $K \in \mathbb{R}$ heißt* Supremum *(bzw.* Infimum*) von A, wenn K die kleinste obere (bzw. größte untere) Schranke von A ist.*
Dabei heißt K kleinste obere Schranke*, falls gilt*

i) K ist eine obere Schranke von A,

ii) für jede obere Schranke K' von A gilt $K \leq K'$

und größte untere Schranke*, falls gilt*

i) K ist eine untere Schranke von A,

ii) für jede untere Schranke K' von A gilt $K \geq K'$.

Es folgen einige Sätze über das Supremum/Infimum.

Satz 2.17 (Eindeutigkeit des Supremums/Infimums)**.** *Jede nichtleere Teilmenge A von $\mathbb{R}$ hat höchstens ein Supremum und höchsten ein Infimum. D. h. das Supremum (bzw. Infimum) von A ist, falls vorhanden, eindeutig und wird mit* $\sup(A)$ *(bzw.* $\inf(A)$*) bezeichnet.*

Beweis: Seien K_1 und K_2 obere Schranken von A, dann gilt $K_1 \leq K_2$ oder $K_1 \geq K_2$. Falls beide Relationen gelten, ist $K_1 = K_2$, ansonsten ist $K_1 < K_2$ oder $K_2 < K_1$. Falls K_1 das Supremum ist, kann nur $K_1 < K_2$ gelten. Analog kann der Beweis für das Infimum geführt werden. □

Wir sagen $[a, b] \subset [c, d]$, falls $c \leq a \wedge d \geq b \wedge (a \neq c \vee d \neq b)$. Eine Folge von immer kleiner werdenden Intervallen kann genutzt werden, um Elemente aus $\mathbb{R}$ festzulegen, die bestimmte Eigenschaften haben. Dazu definieren wir zuerst den zentralen Begriff der Intervallschachtelung.

Definition 2.18 (Intervallschachtelung)**.** *Eine Folge von abgeschlossenen Intervallen $I_1, I_2, I_3, \ldots$ heißt* Intervallschachtelung*, falls gilt:*

i) $I_{n+1} \subset I_n$ für $n = 1, 2, 3, \ldots$

ii) Zu jedem $\varepsilon > 0$ gibt es ein Intervall I_n mit $|I_n| \leq \varepsilon$.

Wir fordern folgendes Axiom:
Für jede Intervallschachtellung gibt es genau ein $x \in \mathbb{R}$, sodass $x \in I_n$ für alle $n = 1, 2, 3, \ldots$.

Satz 2.19 (Existenz des Supremums/Infimums)**.** *Jede nichtleere nach oben (bzw. unten) beschränkte Menge A besitzt ein Supremum (bzw. Infimum).*

Beweis: Wir zeigen den Beweis für das Supremum und konstruieren das Supremum per Intervallschachtelung $[a_n, b_n]$.
Wir beginnen mit einem beliebigen $a_0 \in A$ und einer beliebigen oberen Schranke b_0 von A. Aus $[a_n, b_n]$ konstruieren wir $[a_{n+1}, b_{n+1}]$ wie folgt:
Sei $m = \frac{a_n+b_n}{2}$ (der Mittelpunkt des Intervalls)

$$[a_{n+1}, b_{n+1}] = \begin{cases} [a_n, m] & \text{falls } m \text{ obere Schranke von } A \\ [m, b_n] & \text{sonst } (m \text{ keine obere Schranke von } A) \end{cases}$$

Falls m keine obere Schranke von A ist, so existiert mindestens ein $x \in [m, b_n] \cap A$ mit $m < x$, da b_n durch die Konstruktion des Intervalls immer eine obere Schranke von A ist.
Sei $s \in [a_n, b_n]$ für alle $n = 0, 1, 2, \cdots$. s ist eine obere Schranke von A, sonst gäbe es ein $x \in A$ mit $x > s$ und ein Intervall $[a_n, b_n]$ mit $|[a_n, b_n]| < x - s$. Da $s \in [a_n, b_n]$ gilt $b_n - s < x - s$ also auch $b_n < x$, was aber im Widerspruch zur Wahl von b_n steht. Damit ist s obere Schranke.
Gibt es eine kleinere obere Schranke? Gäbe es eine obere Schranke $s' < s$, so gäbe es ein Intervall $[a_n, b_n]$ mit $|[a_n, b_n]| < s - s'$ und damit $s' < a_n$, was im Widerspruch zur Wahl von a_n steht.
Der Beweis für das Infimum ist analog zu führen! □

Der Beweis des vorherigen Satzes zeigt, dass die Intervallschachtelung auch als Beweisprinzip eingesetzt werden kann.

Beispiel 2.7.

i) *Für das abgeschlossene Intervall* $A = [a, b]$ *gilt*
$\inf(A) = \min(A) = a$
$\sup(A) = \max(A) = b$

ii) *Für das halboffene Intervall* $A = (a, b]$ *gilt*
$\sup(A) = \max(A) = b$
$\inf(A) = a$ *aber* $\min(A)$ *existiert nicht!*

iii) *Für* $A := \left\{ \frac{n}{n+1} \mid n \in \mathbb{N} \right\}$ *gilt* $\sup(A) = 1$

iv) *Für* $A := \left\{ \frac{n^2}{2^n} \mid n \in \mathbb{N} \right\}$ *gilt* $\sup(A) = \max(A) = \frac{9}{8}$

Die Ergebnisse für die letzten beiden Beispiele werden mit den Methoden aus Kapitel 3 hergeleitet werden.

Satz 2.20 (Existenz von Wurzeln)**.** *Zu jedem* $x \in \mathbb{R}_{>0}$ *und jedem* $k \in \mathbb{N}$ *gibt es genau ein* $y \in \mathbb{R}_{>0}$ *mit* $y^k = x$*, d. h.* $y = x^{\frac{1}{k}}$ *oder* $y = \sqrt[k]{x}$ *(k-te Wurzel von x).*

Beweis: Es genügt $x > 1$ zu behandeln, denn den Fall $x < 1$ führt man durch den Übergang $x' = \frac{1}{x}$ auf den Fall $x > 1$ zurück. Für $x = 1$ ist $y = 1$ die Lösung.
Wir konstruieren per vollständige Induktion eine Intervallschachtelung mit Intervallen $I_n = [a_n, b_n]$, für die gilt:

$$\left.\begin{array}{ll} i) & a_n^k \le x \le b_n^k \\ ii) & |I_n| = \left(\frac{1}{2}\right)^{n-1} \cdot |I_1| \end{array}\right\} \text{ für } n = 1, 2, \ldots$$

Sei $I_1 = [1, x]$. Offensichtlich gelten $i)$ und $ii)$ da $x > 1$.
I_{n+1} wird aus I_n, indem wir $m = \frac{1}{2} \cdot (a_n + b_n)$ wählen und dann

$$I_{n+1} = [a_{n+1}, b_{n+1}] = \begin{cases} [a_n, m] & \text{falls } m^k \geq x \\ [m, b_n] & \text{falls } m^k < x \end{cases}$$

wählen.
Offensichtlich gilt $i)$ und $|I_{n+1}| = \frac{1}{2} \cdot |I_n|$. Damit folgt aus $|I_2| = \left(\frac{1}{2}\right)^{2-1} \cdot |I_1|$ auch $ii)$. Die Folge der Intervalle bildet eine Intervallschachtelung, da $I_{n+1} \subset I_n$ und es für jedes $\varepsilon > 0$ ein n gibt, sodass $\left(\frac{1}{2}\right)^{n-1} < \varepsilon \cdot |I_1|^{-1} \Rightarrow |I_n| < \varepsilon$.
Sei y die in allen Intervallen I_n liegende Zahl.
Zu zeigen $y^k = x$:
Zunächst zeigen wir, dass auch die Intervalle $I_n^k = [a_n^k, b_n^k]$ eine Intervallschachtelung bilden.

$i')$ $I_{n+1}^k \subset I_n^k$ gilt wegen $I_{n+1} \subset I_n$

$ii')$ Für die Länge jedes Intervalls I_n^k gilt

$$\begin{aligned} |I_n^k| &= (b_n - a_n) \cdot (b_n^{k-1} + b_n^{k-2} \cdot a_n + \ldots + a_n^{k-1}) \\ &< |I_n| \cdot k \cdot b_1^{k-1} \end{aligned}$$

da $b_n > a_n \geq 1$ und $1 < b_n \leq b_1$.

Sei nun $\varepsilon > 0$ gegeben, so existiert ein Index υ, so dass $|I_\upsilon| < \varepsilon' = \frac{\varepsilon}{k \cdot b_1^{k-1}}$ und damit $|I_\upsilon^k| < \varepsilon$.
Da y in I_n liegt, liegt y^k in I_n^k für $n = 1, 2, \ldots$. Ferner liegt x in I_n^k für alle $n \in \mathbb{N}$, da $i)$ gilt.
Da es nur genau eine Zahl gibt, die in allen Intervallen I_n^k liegt, gilt $y^k = x$.
Zu zeigen bleibt die Eindeutigkeit von y.
Sei z eine weitere Zahl mit $z^k = x$ und $z \neq y$. Dann muss gelten $z > y$ oder $z < y$, woraus $z^k > y^k = x$ bzw. $z^k < y^k = x$ folgen würde, da $y > 1$ vorausgesetzt wurde. □

Wir können die Definition von Wurzeln auf $\mathbb{R}_{\geq 0}$ erweitern, indem wir $\sqrt[k]{0} = 0$ für alle $k \in \mathbb{N}$ definieren.

Zum Abschluss des Kapitels wollen wir die Frage beantworten, ob $\mathbb{R}$ mächtiger als $\mathbb{Q}$ ist? Da $Q \subseteq \mathbb{R}$ ist, lautet die Frage anders ausgedrückt: Ist auch $\mathbb{R}$ abzählbar? Die Antwort liefert das folgende Resultat.

Satz 2.21 (Überabzählbarkeit von $\mathbb{R}$)**.** *Die Menge der reellen Zahlen $\mathbb{R}$ ist nicht abzählbar.*

Beweis: Wir nehmen an, dass es eine Aufzählung von $\mathbb{R} = \{x_1, x_2, x_3, \ldots\}$ gibt.
Wir konstruieren eine Intervallschachtelung I_n, sodass $x_n \notin I_n$ für jedes $n \in \mathbb{N}$. I_n sei dazu rekursiv wie folgt definiert:

1. $I_1 = [x_1 + 1, x_1 + 2]$ (offensichtlich $x_1 \notin I_1$)

2. I_{n+1} entsteht aus I_n, indem I_n in 3 gleich lange Intervalle unterteilt wird, von denen mindestens eins x_{n+1} nicht enthalten kann. Es wird das bzw. ein Teilintervall gewählt, das x_{n+1} nicht enthält.

Offensichtlich konstruiert die Vorschrift eine Intervallschachtelung und es gilt $x_n \notin I_n$. Sei y die Zahl, die in allen Intervallen liegt. Nach obiger Aufzählung hätte y eine Nummer k, also $y = x_k$. Damit wäre aber $y = y_k \in I_k$, da y in allen Intervallen enthalten ist. Dies widerspricht aber der Konstruktion der Intervalle und hier insbesondere der Konstruktion von I_k. Damit kann unsere Annahme, dass $\mathbb{R}$ abzählbar ist, nicht gelten. □