LEUPHANA
UNIVERSITÄT LÜNEBURG

Mathematik für Wirtschaftsinformatiker und Informatiker

Universität Lüneburg

Fakultät III

Prof. Dr. rer. nat. Ulrich Hoffmann

September 2007

Vorwort

Das vorliegende Skript dient als Unterlage für Veranstaltungen zur Mathematik für Wirtschaftsinformatiker und Informatiker in den ersten Studiensemestern an der Universität Lüneburg. Es werden Schulkenntnisse der Mathematik und der Wille, sich mit mehr oder weniger abstrakten Sachverhalten auseinandersetzen zu wollen, vorausgesetzt.

Die Themenauswahl erfolgte mit Blick auf die Anwendungen in der Informatik und der Wirtschaftsinformatik. Schwerpunkte dieser Anwendungen sind die in späteren Semestern behandelten Analyseverfahren in der Theorie der Datenstrukturen und Algorithmen, der Theoretischen Informatik, insbesondere der Angewandten Komplexitätstheorie und der Kryptologie. Das für die Vorgehensweise und Argumentationstechnik in der Informatik besonders wichtige Prinzip der vollständigen Induktion wird an vielen auch nichttrivialen Beispielen dargestellt. Darüber hinaus werden Grundlagen vermittelt, die dem Verständnis wirtschaftswissenschaftlicher Zusammenhänge dienen. So wurden Themen aus verschiedenen Gebieten der Mathematik ausgewählt, deren Inhalte im späteren Studium von Belang sind: aus der elementaren Zahlentheorie, der Kombinatorik, der Analysis und der Linearen Algebra. Übungsaufgaben mit Lösungen zu den einzelnen Kapiteln ergänzen das Skript.

Die Durcharbeitung des Skripts ersetzt nicht den Besuch der Veranstaltungen zur Mathematik, da dort zusätzlich wichtige Zusammenhänge und Beispiele erläutert und mathematische Beweise, die dem Verständnis der mathematischen Sätze dienen, und ergänzende Sachverhalte behandelt werden. Daher kann das Skript weitgehend auf die übliche Darstellung der mathematischen Beweise der zitierten Sätze verzichten.

Gegenüber der Vorgängerversion des vorliegenden Skripts (FINAL, 15:1, Oktober 2005, ISSN 0939-8821) wurde einige Fehler beseitigt und Themen ergänzt, insbesondere das Kapitel über die Methode der erzeugenden Funktionen zur Lösung rekursiver Gleichungen.

Literaturauswahl zur begleitenden Lektüre

Aigner, M.: **Diskrete Mathematik**, 5. Aufl., Vieweg, 2004.

Bartholomé, A.; Rung, J.; Kern, H.: **Zahlentheorie für Einsteiger**, Vieweg, 1995.

Beutelspacher, A.: **Lineare Algebra**, Vieweg, 1994.

Beutelspacher, A.; Neumann, H.B.; Schwarzpaul, T.: **Kryptografie in Theorie und Praxis**, Vieweg, 2005.

Brill, M.: **Mathematik für Informatiker**, Hanser, 2001.

[*] Hachenberger, D.: **Mathematik für Informatiker**, Pearson Studium, 2005.

Haggarty, R.: **Diskrete Mathematik für Informatiker**, Pearson Studium, 2004.

[*] Hartmann, P.: **Mathematik für Informatiker**, 2. Aufl., Vieweg, 2003.

[*] Meinel, C.; Mundhenk, M.: **Mathematische Grundlagen der Informatik**, 2. Aufl., Teubner, 2002.

Purkert, W.: **Brückenkurs Mathematik für Wirtschaftswissenschafter**, Teubner, 1995.

[*] Sydsæter, K.; Hammond, P.: **Mathematik für Wirtschaftswissenschaftler**, Pearson Studium, 2004.

Witt, K.-U.: **Algebraische Grundlagen der Informatik**, 2. Aufl., Vieweg, 2005.

Weiterführende mathematische Werke:

[*] Graham, R.L.; Knuth, D.E.; Patashnik, O.: **Concrete Mathematics**, Addison-Wesley, 1995.

Maurer, S.B.; Ralston, A.: **Discrete Algorithmic Mathematics**, Addison-Wesley, 1991.

Yan, S.Y.: **Number Theory for Computing**, Springer, 2000.

[*] Diese Bücher werden als begleitende Lektüre besonders empfohlen.

Inhaltsverzeichnis

1 Grundlegende Definitionen und Bezeichnungen

In diesem Kapitel werden grundlegende Definitionen, Bezeichnungen und Regeln aus verschiedenen Gebieten der Mathematik zusammengestellt. Dabei wird eine gewisse Vertrautheit mit der Symbolik der Mathematik vorausgesetzt.

Die Mathematik begründet ihre Theorien formal jeweils durch ein System von **Axiomen**, d.h. Grundaussagen, die in einem Teil der mathematischen Welt als Basisbausteine dienen, um aus ihnen Aussagen und Erkenntnisse über diesen Teil der mathematischen Welt abzuleiten. Der Ableitungsvorgang wird durch definierte **logische Schlussregeln** gesteuert. So wurde versucht, den Aufbau der gesamten Mathematik, insbesondere den Aufbau des Zahlensystems und der Mengenlehre, streng axiomatisch zu begründen. Die Entwicklung entsprechender Axiomensysteme bzw. die Erkenntnis über Grenzen der Möglichkeiten dieses Ansatzes können als überragende Ergebnisse der mathematischen Forschung des 19. und 20. Jahrhunderts angesehen werden. Leider sprengt eine formale Behandlung dieser Themen den Rahmen einer universitären Anfängerveranstaltung, so dass im folgenden ausgewählte Themen aus einzelnen Teilgebieten der Mathematik, die für die spätere Informatikausbildung von Bedeutung sind, eher anschaulich und intuitiv beschrieben werden. Natürlich werden auch hierbei Präzision in den Begriffen und formale Korrektheit in der Argumentation versucht.

1.1 Mengen

Die Definition des Begriffs **Menge** gehört zu den grundlegenden Bausteinen der Mathematik. Die formale Behandlung ist den Grundlagen der Mathematik vorbehalten. Georg Cantor (1845 – 1918), der Begründer der Mengenlehre, hat den Begriff der Menge anschaulich folgendermaßen definiert:

> *Eine Menge ist eine Zusammenfassung von bestimmten, wohlunterscheidbaren Objekten unserer Anschauung oder unseres Denkens zu einem Ganzen.*

Eine Menge *A* kann durch die **Aufzählung der in ihr enthaltenen Elemente** beschrieben werden, z.B.

$$A = \{ a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, r, t, u, v, w, x, y, z, \}$$

als die Menge der Buchstaben unseres Alphabets in Kleinschreibung ohne Umlaute. Falls die einzelnen Elemente der Aufzählung allgemein geläufig sind, beschränkt man sich häufig auf die Angabe der ersten und letzten Elemente:

$$A = \{ a, b, c, \ldots, x, y, z \},$$

bzw. auf die Angabe der ersten Elemente, z.B.

$$A = \{ 4, 6, 8, \ldots \}$$

als die Menge aller geraden Zahlen, die größer oder gleich 4 sind.

Zu beachten ist, dass es bei der Aufzählung auf die Reihenfolge der Elemente einer Menge nicht ankommt und dass gleiche Elemente in der Aufzählung nur einmal angegeben werden.

Sehr häufig wird eine Menge **durch charakteristische Eigenschaften beschrieben**, die jedem ihrer Elemente zukommen, und zwar in der Form

$$M = \{ x \mid x \text{ hat die Eigenschaften } \ldots \},$$

z.B.

$L = \{ z \mid z \text{ ist Lösung der Gleichung } x^2 + 2x - 3 = 0 \}$ oder in aufzählender Schreibweise $L = \{ -3,\ 1 \}$.

Liegt ein Element *a* in der Menge *A* (ist *a* in der Menge *A* enthalten), so wird $a \in A$ geschrieben; liegt *a* nicht in *A*, so wird $a \notin A$ geschrieben.

Besitzen alle Elemente einer Menge *A* auch die Eigenschaften, durch die die Elemente einer Menge *B* gekennzeichnet sind, so ist *A* **Teilmenge** von *B*, geschrieben $A \subseteq B$. Enthält *B* mindestens ein Element, das nicht in *A* vorkommt, so ist *A* **echte Teilmenge** von *B*, geschrieben $A \subset B$.

Werden also die Elemente von *B* durch die Eigenschaften E_1, E_2, ..., E_n charakterisiert, d.h.

$$B = \{ x \mid x \text{ hat die Eigenschaften } E_1, \ldots, E_n \},$$

und werden die Elemente von *A* durch die Eigenschaften E_1', ..., E_m' charakterisiert, d.h.

$A = \{ x \mid x \text{ hat die Eigenschaften } E_1', ..., E_m' \}$,

wobei alle Eigenschaften E_1, E_2, ..., E_n unter den Eigenschaften E_1', ..., E_m' vorkommen oder sich aus den Eigenschaften E_1', ..., E_m' durch logische Schlüsse ableiten lassen, so ist $A \subseteq B$. Die Elemente einer Teilmenge *A* einer Menge *B* werden also in der Regel durch mehr Eigenschaften charakterisiert als die Elemente der Obermenge *B*.

Zwei Mengen *A* und *B* sind **gleich**, geschrieben $A = B$, wenn für jedes Element $a \in A$ auch $a \in B$ und für jedes $b \in B$ auch $b \in A$ gilt, wenn also sowohl $A \subseteq B$ als auch $B \subseteq A$ gelten.

Die **leere Menge**, bezeichnet mit $\varnothing$, ist diejenige Menge, die kein Element enthält.

Für jede Menge *A* gelten die beiden Teilmengenbeziehungen $\varnothing \subseteq A$ und $A \subseteq A$.

Die **Vereinigung** der Mengen *A* und *B*, geschrieben $A \cup B$, besteht aus den Elementen, die in *A* oder in *B* (oder in beiden Mengen) liegen:

$A \cup B = \{ x \mid x \in A \text{ oder } x \in B \}$.

Der **Schnitt** der Mengen *A* und *B*, geschrieben $A \cap B$, besteht aus den Elementen, die sowohl in *A* als auch in *B* liegen:

$A \cap B = \{ x \mid x \in A \text{ und } x \in B \}$.

Die **Differenz** der Mengen *B* und *A*, geschrieben $B \setminus A$ besteht aus den Elementen, die in *B*, aber nicht in *A* liegen:

$B \setminus A = \{ x \mid x \in B \text{ und } x \notin A \}$.

Ist $A \subseteq B$, so ist das **Komplement** der Menge *A* **bezüglich** der Menge *B*, geschrieben $\overline{A}^B$, definiert durch

$\overline{A}^B = \{ x \mid x \in B \text{ und } x \notin A \}$.

Offensichtlich ist (für $A \subseteq B$) $\overline{A}^B = B \setminus A$.

Für eine Menge A bezeichnet $|A|$ die **Anzahl der Elemente (**oder die **Mächtigkeit)** von A.

Mit $\mathbf{P}(A)$ wird die **Potenzmenge** der Menge A bezeichnet, die aus allen Teilmengen der Menge A besteht, d.h. $\mathbf{P}(A) = \{ L \mid L \subseteq A \}$.

Beispielsweise lautet für $A = \{1, 2, 3\}$ die Potenzmenge

$$\mathbf{P}(A) = \{ \varnothing, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\} \}.$$

In Kapitel 1.6 wird gezeigt, dass die Potenzmenge $\mathbf{P}(A)$ einer endlichen Menge A mit n vielen Elementen 2^n viele Elemente enthält, d.h. es gibt $2^{|A|}$ viele Teilmengen einer endlichen Menge A.

Für Mengen A_1, A_2, ..., A_n wird das **kartesische Produkt** definiert als

$$A_1 \times A_2 \times \ldots \times A_n = \{ (a_1, a_2, \ldots, a_n) \mid a_1 \in A_1, a_2 \in A_2, \ldots, a_n \in A_n \}.$$

Ein Element $(a_1, a_2, \ldots, a_n) \in A_1 \times A_2 \times \ldots \times A_n$ wird als ***n*-Tupel** bezeichnet. Bei einem 2-Tupel spricht man auch von einem **Paar**.

Die grundlegenden Rechenregeln für Operationen mit Mengen werden in folgendem Satz zusammengefasst:

Satz 1.1-1:

Es seien im folgenden A, B und C Mengen. Dann gilt:

(i) $A \cup \varnothing = A,\ A \cap \varnothing = \varnothing$.

(ii) $A \cap B \subseteq A,\ A \cap B \subseteq B,\ A \subseteq A \cup B,\ B \subseteq A \cup B$.

(iii) $A \cup B = B \cup A,\ A \cap B = B \cap A$ (**Kommutativgesetze**).

(iv) $A \cup (B \cup C) = (A \cup B) \cup C,\ A \cap (B \cap C) = (A \cap B) \cap C$ (**Assoziativgesetze**); diese Regeln rechtfertigen die Schreibweisen
$A \cup B \cup C = A \cup (B \cup C) = (A \cup B) \cup C$ und
$A \cap B \cap C = A \cap (B \cap C) = (A \cap B) \cap C$.

(v) $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$, die rechte Seite ist eine **disjunkte Zerlegung** von $A \cup B$. Dabei heißt eine Zerlegung $M = M_1 \cup M_2$ der Menge M disjunkt, wenn $M_1 \cap M_2 = \varnothing$ ist.

(vi) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (**Distributivgesetze**)

(vii) $A \cap (A \cup B) = A,\ A \cup (A \cap B) = A$.

(viii) Ist $A \subseteq C$, so ist $\left(\overline{A^C}\right)^C = A$.

(ix) Sind $A \subseteq C$ und $B \subseteq C$, so gelten
$\overline{(A \cap B)}^C = \overline{A}^C \cup \overline{B}^C$, $\overline{(A \cup B)}^C = \overline{A}^C \cap \overline{B}^C$ (**Regeln von de Morgan**).

(x) Sind $A \subseteq C$ und $B \subseteq C$, so folgt aus $A \subseteq B$ die Beziehung $\overline{B}^C \subseteq \overline{A}^C$ und umgekehrt.

Bemerkung: Die Voraussetzungen $A \subseteq C$ bzw. $B \subseteq C$ in den Aussagen (viii) – (x) wurden nur gemacht, weil das Komplement einer Menge A relativ zu einer die Menge A umfassenden Menge C definiert wurde.

1.2 Aussagen und deren logische Verknüpfung

Logische Aussagen in der Mathematik werden wie Mengen streng axiomatisch definiert. Auf diesen Ansatz soll hier ebenfalls zugunsten eines intuitiven Ansatzes verzichtet werden.

Unter einer **mathematisch logischen Aussage** versteht man einen Satz (in einem logischen System), der entweder WAHR oder FALSCH ist (den Wahrheitswert WAHR oder FALSCH besitzt, umgangssprachlich: wahr oder falsch ist). Beispielsweise ist

- „*13 ist eine Primzahl*“ eine Aussage mit Wahrheitswert WAHR („eine wahre Aussage“)
- „$\sqrt{2}$ *ist eine rationale Zahl*“ eine Aussage mit Wahrheitswert FALSCH („eine falsche Aussage“)
- „*Jede gerade natürliche Zahl, die größer als 2 ist, lässt sich als Summe zweier Primzahlen darstellen*“ eine Aussage, deren Wahrheitswert noch nicht bekannt ist, die aber einen der beiden Wahrheitswerte WAHR oder FALSCH besitzt.

Der Satz „*Dieser Satz hat den Wahrheitswert* FALSCH“ ist keine mathematische Aussage, da er weder den Wahrheitswert WAHR noch FALSCH haben kann. Derartige Sätze, die eine selbstbezogene semantische Aussage versuchen, heißen **Paradoxien**.

Sind P und Q Aussagen, so kann man sie mit Hilfe der **logischen Junktoren** $\wedge$ („und“), $\vee$ („oder“) bzw. $\neg$ („nicht“) zu neuen Aussagen $(P \wedge Q)$, $(P \vee Q)$ bzw. $(\neg P)$ zusammensetzen. Dabei kann man häufig auf die Klammern verzichten, wenn man annimmt, dass der Junktor $\neg$ stärker als der Junktor $\wedge$ und dieser stärker als der Junktor $\vee$ bindet. Die Wahrheitswerte der zusammengesetzten Aussagen ergeben sich aus den Wahrheitswerten der Teile gemäß folgender **Wahrheitstafeln**:

Wahrheitswerte von			
P	Q	$(P \wedge Q)$	$(P \vee Q)$
FALSCH	FALSCH	FALSCH	FALSCH
FALSCH	WAHR	FALSCH	WAHR
WAHR	FALSCH	FALSCH	WAHR
WAHR	WAHR	WAHR	WAHR
Bezeichnung		**Konjunktion**	**Disjunktion**

Wahrheitswerte von	
P	$(\neg P)$
FALSCH	WAHR
WAHR	FALSCH
Bezeichnung	**Negation**

Neben diesen drei Junktoren werden in logischen Aussagen häufig noch die Junktoren $\Rightarrow$ („impliziert“, „hat zur Folge“, „aus ... folgt ...“), $\Leftrightarrow$ („... ist gleichbedeutend mit ...“, „... gilt genau dann wenn ... gilt“) und $\oplus$ („exklusives oder“, in der englischsprachigen Fachliteratur auch XOR) verwendet, die durch folgende Wahrheitstafeln definiert sind:

Wahrheitswerte von				
P	Q	$(P \Rightarrow Q)$	$(P \Leftrightarrow Q)$	$(P \oplus Q)$
FALSCH	FALSCH	WAHR	WAHR	FALSCH
FALSCH	WAHR	WAHR	FALSCH	WAHR
WAHR	FALSCH	FALSCH	FALSCH	WAHR
WAHR	WAHR	WAHR	WAHR	FALSCH
Bezeichnung		**Implikation**	**Äquivalenz**	**Antivalenz**

Um den Wahrheitswert einer komplexen Aussage zu ermitteln, die aus durch Junktoren verbundenen Teilaussagen besteht, werden alle Kombinationen von Wahrheitswerten in die Grundaussagen, d.h. in die Teilaussagen, die keine Junktoren erhalten, eingesetzt und durch Anwendung obiger Wahrheitstafeln die sich ergebenden Wahrheitswerte unter Beachtung der Klammersetzung bzw. Bindung der Junktoren ermittelt.

Beispiele:

P	Q	$(P \Rightarrow Q)$	$\Leftrightarrow$	$((\neg P)$	$\wedge$	$Q)$
FALSCH	FALSCH	WAHR	FALSCH	W	F	F
FALSCH	WAHR	WAHR	WAHR	W	W	W
WAHR	FALSCH	FALSCH	WAHR	F	F	F
WAHR	WAHR	WAHR	FALSCH	F	F	W

P	Q	$(P \Rightarrow Q)$	$\Leftrightarrow$	$((\neg P)$	$\vee$	$Q)$
FALSCH	FALSCH	WAHR	WAHR	W	W	F
FALSCH	WAHR	WAHR	WAHR	W	W	W
WAHR	FALSCH	FALSCH	WAHR	F	F	F
WAHR	WAHR	WAHR	WAHR	F	W	W

Eine zusammengesetzte Aussage, deren Wahrheitswert bei allen möglichen Belegungen mit Wahrheitswerten der Teilaussagen, die keine Junktoren enthalten, stets WAHR ist, heißt **Tautologie**.

Beispielweise sind die Aussagen $(P \vee (\neg P))$ und $(P \Rightarrow Q) \Leftrightarrow ((\neg P) \vee Q)$ Tautologien, nicht aber $(P \Rightarrow Q) \Leftrightarrow ((\neg P) \wedge Q)$.

Das Beispiel $(P \Rightarrow Q) \Leftrightarrow ((\neg P) \vee Q)$ zeigt, dass man den Junktor $\Rightarrow$ durch die Junktoren $\neg$ und $\vee$ ausdrücken kann, indem man in einer Aussage jedes Auftreten einer Teilaussage der Form $(P \Rightarrow Q)$ durch die Teilaussage $((\neg P) \vee Q)$ ersetzt. Der Junktor $\Rightarrow$ ist im Grunde also überflüssig, nur vereinfacht er die Struktur der Aussagen und erhöht damit ihre Lesbarkeit. Der folgende Satz zeigt, dass sich auch die anderen Junktoren $\wedge$, $\Leftrightarrow$ und $\oplus$ mit Hilfe der Junktoren $\vee$ und $\neg$ ausdrücken lassen, da sich jeweils links und rechts des Junktors $\Leftrightarrow$ die gleichen Wahrheitswerte ergeben. Teil (v) zeigt, dass man auch $\wedge$ und $\neg$ nehmen kann, um alle Junktoren $\vee$, $\Rightarrow$, $\Leftrightarrow$ und $\oplus$ auszudrücken.

Satz 1.2-1:

Es seien P und Q Aussagen. Dann sind die folgenden Aussagen Tautologien.

(i) $(P \wedge Q) \Leftrightarrow (\neg((\neg P) \vee (\neg Q)))$,
in vereinfachter Schreibweise: $(P \wedge Q) \Leftrightarrow \neg(\neg P \vee \neg Q)$.

(ii) $(P \Rightarrow Q) \Leftrightarrow ((\neg P) \vee Q)$,
in vereinfachter Schreibweise: $(P \Rightarrow Q) \Leftrightarrow (\neg P \vee Q)$.

(iii) $(P \Leftrightarrow Q) \Leftrightarrow ((P \Rightarrow Q) \wedge (Q \Rightarrow P))$,

$(P \Leftrightarrow Q) \Leftrightarrow (((\neg P) \vee Q) \wedge ((\neg Q) \vee P))$,
in vereinfachter Schreibweise: $(P \Leftrightarrow Q) \Leftrightarrow ((\neg P \vee Q) \wedge (\neg Q \vee P))$,

$(P \Leftrightarrow Q) \Leftrightarrow (\neg((\neg((\neg P) \vee Q)) \vee (\neg((\neg Q) \vee P))))$,
in vereinfachter Schreibweise: $(P \Leftrightarrow Q) \Leftrightarrow \neg((\neg(\neg P \vee Q)) \vee (\neg(\neg Q \vee P)))$.

(iv) $(P \oplus Q) \Leftrightarrow ((P \wedge (\neg Q)) \vee ((\neg P) \wedge Q))$,
in vereinfachter Schreibweise: $(P \oplus Q) \Leftrightarrow ((P \wedge \neg Q) \vee (\neg P \wedge Q))$,

$(P \oplus Q) \Leftrightarrow ((\neg((\neg P) \vee Q)) \vee (\neg((\neg Q) \vee P)))$,
in vereinfachter Schreibweise: $(P \oplus Q) \Leftrightarrow (\neg(\neg P \vee Q) \vee \neg(\neg Q \vee P))$.

(v) $(P \vee Q) \Leftrightarrow (\neg((\neg P) \wedge (\neg Q)))$,
in vereinfachter Schreibweise: $(P \vee Q) \Leftrightarrow \neg(\neg P \wedge \neg Q)$.

(vi) $(P \wedge (P \Rightarrow Q)) \Rightarrow Q$ (**Modus ponens**).

Bemerkung: Man kommt sogar mit nur einem Junktor aus, um alle anderen Junktoren auszudrücken. Dazu wird der Junktor $\uparrow$ durch folgende Wahrheitstafel definiert:

Wahrheitswerte von		
P	Q	$(P \uparrow Q)$
FALSCH	FALSCH	WAHR
FALSCH	WAHR	WAHR
WAHR	FALSCH	WAHR
WAHR	WAHR	FALSCH
Bezeichnung		**Sheffer-Operation**

Es gilt $(P \uparrow Q) \Leftrightarrow (\neg(P \wedge Q))$, und damit ist $(\neg P)$ gleichwertig mit (lässt sich ausdrücken durch) $(P \uparrow P)$, und es ist $(P \wedge Q)$ gleichwertig mit $((P \uparrow Q) \uparrow (P \uparrow Q))$.

Der folgende Satz zeigt strukturelle Äquivalenzen zwischen Sätzen der elementaren Mengenlehre (Satz 1.1-1) und der Aussagenlogik.

Satz 1.2-2:

	Satz 1.1-1:
Es seien P, Q und R Aussagen. Die Aussage W habe den Wahrheitswert `WAHR`, die Aussage F habe den Wahrheitswert `FALSCH`.	Es seien im folgenden A, B und C Mengen.
Dann sind die folgenden Aussagen Tautologien.	Dann gilt:
(i) $(P \vee F) \Leftrightarrow P$, $P \wedge F \Leftrightarrow F$.	(i) $A \cup \varnothing = A$, $A \cap \varnothing = \varnothing$.
(ii) $(P \wedge Q) \Rightarrow P$, $(P \wedge Q) \Rightarrow Q$, $P \Rightarrow (P \vee Q)$, $Q \Rightarrow (P \vee Q)$.	(ii) $A \cap B \subseteq A$, $A \cap B \subseteq B$, $A \subseteq A \cup B$, $B \subseteq A \cup B$.
(iii) $(P \vee Q) \Leftrightarrow (Q \vee P)$, $(P \wedge Q) \Leftrightarrow (Q \wedge P)$ (**Kommutativgesetze**).	(iii) $A \cup B = B \cup A$, $A \cap B = B \cap A$ (Kommutativgesetze).
(iv) $(P \vee (Q \vee R)) \Leftrightarrow ((P \vee Q) \vee R)$, $(P \wedge (Q \wedge R)) \Leftrightarrow ((P \wedge Q) \wedge R)$ (**Assoziativgesetze**); diese Regeln rechtfertigen die Schreibweisen $(P \vee Q \vee R)$ anstelle von $(P \vee (Q \vee R))$ und $(P \wedge Q \wedge R)$ anstelle von $(P \wedge (Q \wedge R))$.	(iv) $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$ (Assoziativgesetze).
(v) $(P \vee Q) \Leftrightarrow ((P \wedge \neg Q) \vee (P \wedge Q) \vee (Q \wedge \neg P))$.	(v) $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$.
(vi) $(P \wedge (Q \vee R)) \Leftrightarrow ((P \wedge Q) \vee (P \wedge R))$, $(P \vee (Q \wedge R)) \Leftrightarrow ((P \vee Q) \wedge (P \vee R))$ (**Distributivgesetze**)	(vi) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (Distributivgesetze)
(vii) $(P \wedge (P \vee Q)) \Leftrightarrow P$, $(P \vee (P \wedge Q)) \Leftrightarrow P$.	(vii) $A \cap (A \cup B) = A$, $A \cup (A \cap B) = A$.
(viii) $(\neg(\neg P)) \Leftrightarrow P$.	(viii) Ist $A \subseteq C$, so ist $\left(\overline{A^C}\right)^C = A$.

../..

(ix)	$(\neg(P \wedge Q)) \Leftrightarrow (\neg P \vee \neg Q)$, $(\neg(P \vee Q)) \Leftrightarrow (\neg P \wedge \neg Q)$ (**Regeln von de Morgan**)	(ix)	Sind $A \subseteq C$ und $B \subseteq C$, so gelten $\overline{(A \cap B)}^C = \overline{A}^C \cup \overline{B}^C$, $\overline{(A \cup B)}^C = \overline{A}^C \cap \overline{B}^C$ (Regeln von de Morgan)
(x)	$(P \Rightarrow Q) \Leftrightarrow (\neg Q \Rightarrow \neg P)$	(x)	Sind $A \subseteq C$ und $B \subseteq C$, so folgt aus $A \subseteq B$ die Beziehung $\overline{B}^C \subseteq \overline{A}^C$ und umgekehrt.

Allgemein gilt, dass ein Satz der elementaren Mengenlehre, der nur die Relationenzeichen = und $\subseteq$ und die Operatoren $\cup, \cap$ bzw. $^{-C}$ (Komplement) verwendet, eine korrespondierende logische Aussage besitzt und umgekehrt, indem man Ersetzungen gemäß folgender Tabelle vornimmt.

Elementare Mengenlehre	=	$\subseteq$	$\cup$	$\cap$	$^{-C}$
Aussagenlogik	$\Leftrightarrow$	$\Rightarrow$	$\vee$	$\wedge$	$\neg$

Die Erweiterung der Aussagenlogik führt auf die **Prädikatenlogik (erster Stufe)**, in der logische Sätze formuliert werden, die die **Quantoren** $\forall$ („für alle ...“) und $\exists$ („es gibt ...“) enthalten können, wobei über „freie Variablen“ in Formeln quantifiziert wird.

Beispiel:

$$\forall x \left(\left((x \in \mathbf{N}) \wedge (x > 1) \right) \Rightarrow \left(\exists p \left((p \text{ ist Primzahl}) \wedge (p \text{ teilt } x) \right) \right) \right).$$

Auf eine weiterführende Einführung in die mathematische Logik soll hier jedoch verzichtet werden.

1.3 Beweistechniken

Um den Wahrheitswert einer komplexen Aussage zu ermitteln, die aus Teilaussagen besteht, die durch Junktoren verbundenen sind, wird ein **mathematischer Beweis** angeführt. Dieser besteht aus einer Aneinanderreihung logischer Schlüsse, die genau spezifizierten Schlussregeln folgen und jederzeit eindeutig nachvollziehbar sind (zumindest sein sollten). Die Grund-

lage aller Beweise in einem theoretischen System ist eine **Menge von Axiomen**, die als wahr angenommen werden und eine „vernünftige“ Basierung der zugrundeliegenden Theorie bilden. Außerdem gibt es eine **endliche Menge von Schlussregeln**, die es erlauben, aus Aussagen, die bereits als wahr erkannt wurden (dazu gehören die Axiome, deren Wahrheitswert als WAHR angenommen wird), neue wahre Aussagen herzuleiten.

Formal ist ein Beweis eines mathematischen Satzes *P* eine Aneinanderreihung

$p_1, p_2, \ldots, p_n$

von Aussagen, wobei am Anfang Axiome oder bereits bewiesene mathematische Sätze, d.h. Sätze mit dem Wahrheitswert WAHR, stehen, und p_n gleich *P* ist. Den Übergang von einer Zeile p_i zur Zeile p_{i+1} erhält man beispielsweise, indem man eine Tautologie, etwa aus Satz 1.2-2, anwendet: Ist p_i die linke Seite einer Tautologie der Form $Q \Leftrightarrow R$, d.h. p_i hat dieselbe Struktur wie *Q*, dann ist p_{i+1} gleich *R*. Eine weitere Möglichkeit der Beweisführung besteht in der Anwendung der als **Modus ponens** bekannten Tautologie $(P \wedge (P \Rightarrow Q)) \Rightarrow Q$ (Satz 1.2-1 (vi)). Ist p_j mit $j < i$ eine Zeile, die mit der Aussage *P* gleichzusetzen ist und hat p_i die Form $(P \Rightarrow Q)$, dann ist p_{i+1} gleich *Q*. Wenn man also die Aussage *P* in Zeile p_j bereits als Aussage mit dem Wahrheitswert WAHR erkannt hat und in Zeile p_i feststellt, dass die Gültigkeit der Aussage *P* die Gültigkeit der Aussage *Q* impliziert, dann ist *Q* eine Aussage mit dem Wahrheitswert WAHR.

Ohne an dieser Stelle genauer auf den formalen Vorgang des Beweisens in der Mathematik einzugehen, werden einige mögliche Vorgehensweisen bei Beweisführung von Aussagen beschrieben.

A. **Direkter Beweis**:

Die oben beschriebene Vorgehensweise ist ein Beispiel für einen direkten Beweis.

Häufig treten mathematische Aussagen der Form $(P \Rightarrow Q)$ auf. Für einen Beweis dieser Aussage kann man folgendermaßen vorgehen:

Man nimmt an, dass *P* den Wahrheitswert WAHR besitzt (dass *P* wahr ist). Durch eine „geeignete“ Argumentation (Anwendung logischer Schlüsse) zeigt man, dass dann auch *Q* den Wahrheitswert WAHR hat (dass *Q* wahr ist).

B. **Indirekter Beweis**:

Zum Beweis der Aussage $(P \Rightarrow Q)$ zeigt man $(\neg Q \Rightarrow \neg P)$, weil eventuell die Argumentation in dieser Richtung einfacher ist. Der indirekte Beweis beruht auf der Tautologie $(P \Rightarrow Q) \Leftrightarrow (\neg Q \Rightarrow \neg P)$ (Satz 1.2-2 (x)).

C. **Beweis einer Äquivalenz**:

Um die Aussage $(P \Leftrightarrow Q)$ zu beweisen, sind zwei „Richtungen" zu zeigen, nämlich ein Beweis für $(P \Rightarrow Q)$ und ein Beweis für $(Q \Rightarrow P)$.

Häufig treten Äquivalenzen auch in der Form $(P \Leftrightarrow Q)$ und $(Q \Leftrightarrow R)$. In diesem Fall kann man anstelle der vier Beweise für $(P \Rightarrow Q)$, $(Q \Rightarrow P)$, $(Q \Rightarrow R)$ und $(R \Rightarrow Q)$ auch drei Beweise, nämlich für $(P \Rightarrow Q)$, $(Q \Rightarrow R)$ und $(R \Rightarrow P)$, erbringen.

D. **Beweis durch Widerspruch**:

Zum Beweis der Gültigkeit einer Aussage P nimmt man an, dass $\neg P$ den Wahrheitswert `WAHR` besitzt. Durch eine „geeignete" Argumentation (Anwendung logischer Schlüsse) zeigt man von einer Aussage Q, deren Wahrheitswert vorher bereits als `WAHR` erkannt wurde, dass sie dann den Wahrheitswert `FALSCH` besitzen muss. Man zeigt also die Gültigkeit von $(\neg P \Rightarrow (Q \wedge \neg Q))$. Diese Aussage kann jedoch aufgrund des Wahrheitswerts einer Implikation und wegen der Tatsache, dass $(Q \wedge \neg Q)$ immer den Wahrheitswert `FALSCH` besitzt, nur dann gültig sein, wenn $\neg P$ den Wahrheitswert `FALSCH` bzw. P den Wahrheitswert `WAHR` besitzt.

E. **Beweis durch vollständige Induktion**:

Zum Beweis von Aussagen über natürliche Zahlen wird häufig die Beweismethode der vollständigen Induktion eingesetzt. Diese Methode beruht auf einer charakteristischen Eigenschaft der natürlichen Zahlen und wird in Kapitel 1.5 behandelt.

1.4 Algebraische Grundstrukturen und Zahlensysteme

Im folgenden werden die üblichen Zahlensysteme beschrieben. Die hier angegebenen Bezeichnungen stellen keine Definitionen im mathematischen Sinne dar, sondern geben einen Überblick über den Aufbau der Zahlensysteme und definieren einige grundlegende algebraische Strukturen.

Die **Menge der natürlichen Zahlen** wird definiert durch

$\mathbf{N} = \{ 0, 1, 2, 3, 4, ... \}$

Ein Axiomensystem für die Menge der natürlichen Zahlen, d.h. ein Regelsystem, das die Menge der natürlichen Zahlen eindeutig durch ihre Eigenschaften definiert, lautet:

Axiom 1: $0 \in \mathbf{N}$

Axiom 2: Jedes $n \in \mathbf{N}$ hat einen Nachfolger $n' \in \mathbf{N}$.

Axiom 3: 0 ist das erste Element in **N**, d.h. es gibt kein $n \in \mathbf{N}$ mit $n' = 0$ bzw. 0 hat keinen Vorgänger.

Axiom 4: Unterschiedliche Elemente n und m haben unterschiedliche Nachfolger. Gleichbedeutend damit ist: Sind die Nachfolger n' und m' zweier natürlicher Zahlen gleich, so sind die Zahlen n und m ebenfalls gleich.

Axiom 5: Eine Menge M natürlicher Zahlen, die die Zahl 0 und mit jeder Zahl n auch ihren Nachfolger n' enthält, ist mit **N** identisch.

N ist das einzige „Modell“ dieses Axiomensystems. Statt n' schreibt man gewöhnlich $n+1$.

Statt $0'$ schreibt man auch 1, statt $0''$ schreibt man 2, statt $0'''$ schreibt man 3 usw. Der n-te Nachfolger der 0 ist die natürliche Zahl n.

Aufbauend auf den so definierten Grundeigenschaften der natürlichen Zahlen werden **arithmetische Operationen** auf den natürlichen Zahlen eingeführt:

Die **Addition** wird über die Nachfolger n' einer natürlichen Zahl *n* definiert durch die Regeln

$m + 0 = m$,
$m + n' = (m + n)'$ für jede natürliche Zahl *m* und jede natürliche Zahl *n*.

Damit ist z.B. $2 + 3 = 2 + 0''' = (2 + 0'')' = ((2 + 0')')' = (((2 + 0)')')' = 2''' = (0'')''' = 5$.

Auf ähnliche Weise und durch Zurückführung auf die Addition wird die **Multiplikation** eingeführt:

$m \cdot 0 = 0$,
$m \cdot n' = (m \cdot n) + m$ für jede natürliche Zahl *m* und jede natürliche Zahl *n*.

Damit ist z.B. $m \cdot 1 = m \cdot 0' = (m \cdot 0) + m = 0 + m = m$ und
$7 \cdot 3 = 7 \cdot 2' = (7 \cdot 2) + 7 = (7 \cdot 1') + 7 = ((7 \cdot 1) + 7) + 7 = (7 + 7 + 7) = 21$.

Die so eingeführten arithmetischen Operationen genügen wichtigen Gesetzmäßigkeiten:

Es gilt für jede natürliche Zahl *n*, für jede natürliche Zahl *m* und für jede natürliche Zahl *k*:
$k + (m + n) = (k + m) + n$,
$k \cdot (m \cdot n) = (k \cdot m) \cdot n$ (**Assoziativgesetze**)

$k + m = m + k$,
$k \cdot m = m \cdot k$ (**Kommutativgesetze**)

$k \cdot (m + n) = k \cdot m + k \cdot n$ (**Distributivgesetz**)

Die natürliche Zahl *n* heißt **kleiner als** die natürliche Zahl *m*, geschrieben $n < m$, wenn die Gleichung $n + x = m$ eine Lösung $x \in \mathbf{N}$ mit $x \neq 0$ besitzt. Man schreibt $n \leq m$, wenn $n < m$ oder $n = m$ gilt. Durch diese Festlegungen wird eine **totale Ordnungsrelation** auf den natürlichen Zahlen definiert.

Erläuterung:

Eine Relation $\lhd$ auf einer Menge M heißt **partielle Ordnungsrelation**, wenn für jedes $a \in M$, jedes $b \in M$ und jedes $c \in M$ gilt:

(i) $a \lhd a$ (**Reflexivität**)

(ii) aus $a \lhd b$ und $b \lhd a$ folgt $a = b$ (**Antisymmetrie**)

(iii) aus $a \lhd b$ und $b \lhd c$ folgt $a \lhd c$ (**Transitivität**).

Eine partielle Ordnungsrelation heißt **totale Ordnungsrelation**, wenn für jedes $a \in M$ und für jedes $b \in M$ zusätzlich gilt:

$a \lhd b$ oder $b \lhd a$ (**Vergleichbarkeit**).

Die so definierten arithmetischen Operationen erlauben nur wenige wirkliche Rechenmanipulationen. Operationen wie Differenzen- oder Quotientenbildung sind nur sehr eingeschränkt möglich, wenn man fordert, dass jeweils das Ergebnis einer dieser Operationen wieder ein Element aus **N** ergibt. Daher erweitert man die Menge **N** um neue Elemente:

Für jedes $n \in \mathbf{N}$ wird ein neues Element n^- definiert, für das $n + n^- = n^- + n = 0$ gilt. Statt n^- schreibt man auch $(-n)$ bzw. $-n$. Die Definition erfolgt formal über die Definition einer geeigneten Äquivalenzrelation auf Paaren von natürlichen Zahlen und Übergang auf die zugehörigen Äquivalenzklassen und Einbettung von **N** in die Menge dieser Äquivalenzklassen (vgl. Aufgabe 1.5). Auf Details soll hier verzichtet werden.

Erläuterung:

Eine Relation $\approx$ auf einer Menge M heißt **Äquivalenzrelation**, wenn für jedes $a \in M$, jedes $b \in M$ und jedes $c \in M$ gilt:

(i) $a \approx a$ (**Reflexivität**)

(ii) aus $a \approx b$ folgt $b \approx a$ (**Symmetrie**)

(iii) aus $a \approx b$ und $b \approx c$ folgt $a \approx c$ (**Transitivität**).

Für $a \in M$ bezeichnet $[a]_{\approx} = \{ b \mid b \approx a \}$ die zu a gehörende **Äquivalenzklasse**.

Das Ergebnis ist die **Menge der ganzen Zahlen:**

$$\begin{aligned}\mathbf{Z} &= \mathbf{N} \cup \{-n \mid n \in \mathbf{N}\} \\ &= \{0, 1, -1, 2, -2, 3, -3, \ldots\}\end{aligned}$$

Offensichtlich ist (in der hier gewählten vereinfachten Definition von **Z**) $\mathbf{Z} \subset \mathbf{N}$.

Die **Ordnungsrelation** $\leq$ und die Operationen **Addition** und **Multiplikation** werden auf **Z** in einer zu **N** kompatiblen Weise erweitert:

$(-m) \leq n$,
$((-m) \leq (-n)) \Leftrightarrow (n \leq m)$ für jede natürliche Zahl *m* und jede natürliche Zahl *n*.

$$m + (-n) = (-n) + m = \begin{cases} \text{Lösung } x \text{ der Gleichung } n + x = m & \text{für } n \leq m \\ -(\text{Lösung } x \text{ der Gleichung } m + x = n) & \text{für } m < n, \end{cases}$$
$(-m) + (-n) = (-n) + (-m) = -(n + m)$ für jede natürliche Zahl *m* und jede natürliche Zahl *n*.

$m \cdot (-n) = (-n) \cdot m = -(m \cdot n)$,
$(-m) \cdot (-n) = (-n) \cdot (-m) = n \cdot m$ für jede natürliche Zahl *m* und jede natürliche Zahl *n*.

Mit diesen Definitionen bildet die algebraische Struktur $(\mathbf{Z}, +, \cdot, 0, 1)$ einen nullteilerfreien kommunikativen Ring mit 1.

Erläuterung:

Eine algebraische Struktur $(G, \circ)$ heißt **Gruppe**, wenn gilt:

(i) $a \circ b \in G$ für jedes $a \in G$ und jedes $b \in G$ (**Abgeschlossenheit der Operation** $\circ$)

(ii) $a \circ (b \circ c) = (a \circ b) \circ c$ für jedes $a \in G$, jedes $b \in G$ und jedes $c \in G$ (**Assozitivität**)

(iii) es gibt ein Element $e \in G$ mit der Eigenschaft $e \circ a = a \circ e = a$ für jedes $a \in G$ (**Existenz eines neutralen Elements**)

(iv) für jedes $a \in G$ gibt es ein eindeutiges Element $a^{-1} \in G$ mit $a \circ a^{-1} = a^{-1} \circ a = e$; dieses Element heißt **inverses Element** zu *a*.

Das neutrale Element *e* einer Gruppe ist eindeutig bestimmt, daher wird es in die Angabe der Gruppe mit aufgenommen: $(G, \circ, e)$

Eine Gruppe $(G, \circ)$ heißt **kommutative Gruppe**, wenn zusätzlich gilt:
$a \circ b = b \circ a$ für jedes $a \in G$ und für jedes $b \in G$.

Eine algebraische Struktur $(R, \oplus, \otimes)$ heißt **Ring**, wenn gilt:

(i) $(R, \oplus, 0)$ ist eine kommutative Gruppe (mit neutralem Element $0 \in R$)

(ii) $a \otimes b \in R$ für jedes $a \in R$ und jedes $b \in R$ (**Abgeschlossenheit der Operation** $\otimes$)

(iii) $a \otimes (b \otimes c) = (a \otimes b) \otimes c$ für jedes $a \in R$, jedes $b \in R$ und jedes $c \in R$ (**Assozitivität der Operation** $\otimes$)

(iv) $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$ und $(b \oplus c) \otimes a = (b \otimes a) \oplus (c \otimes a)$ für jedes $a \in R$, jedes $b \in R$ und jedes $c \in R$ (**Distributivität der Operation** $\otimes$ **über die Operation** $\oplus$)

Ein Ring $(R, \oplus, \otimes)$ heißt **Ring mit 1**, wenn es ein Element $1 \in R$ gibt mit $a \otimes 1 = 1 \otimes a = a$ für jedes $a \in R$ (**Existenz eines neutralen Elements bezüglich der Operation** $\otimes$).

Ein Ring $(R, \oplus, \otimes)$ heißt **kommutativer Ring**, wenn zusätzlich gilt:
$a \otimes b = b \otimes a$ für jedes $a \in R$ und für jedes $b \in R$.

../..

Ein Ring $(R, \oplus, \otimes)$ heißt **nullteilerfreier Ring**, wenn zusätzlich gilt:
Die Gleichung $a \otimes x = 0$ besitzt für jedes $a \in R$ nur die Lösung $x = 0$.

Ist $(R, \oplus, \otimes)$ ein Ring mit 1, so werden die neutralen Elemente mit in die Angabe des Rings aufgenommen: $(R, \oplus, \otimes, 0, 1)$.

Z erlaubt bereits eine Vielzahl interessanter arithmetischer Operationen, jedoch ist „richtiges Rechnen", d.h. auch **Division (Umkehrung der Multiplikation)** nicht immer möglich. Daher wird **Z** auf die **Menge der rationalen Zahlen** erweitert. Auch diese Erweiterung erfolgt wieder formal über die Definition einer geeigneten Äquivalenzrelation auf Paaren dieses Mal von ganzen Zahlen und Übergang auf die zugehörigen Äquivalenzklassen und Einbettung von **Z** in die Menge dieser Äquivalenzklassen:

Die Paare (a, b) und (c, d) ganzer Zahlen mit $b \neq 0$ und $d \neq 0$ werden hierbei als äquivalent definiert, wenn $a \cdot d = c \cdot b$ gilt. Die bezüglich dieser Äquivalenzrelation zum Paar (a, b) mit $b \neq 0$ ganzer Zahlen gehörende Äquivalenzklasse wird mit $\frac{a}{b}$ bezeichnet. Die rationalen Zahlen sind dann genau die Menge dieser Äquivalenzklassen:

$$\mathbf{Q} = \left\{ \frac{m}{n} \;\middle|\; m \in \mathbf{Z} \text{ und } n \in \mathbf{Z} \text{ und } n \neq 0 \right\}.$$

Die Menge der ganzen Zahlen ist in der Menge der rationalen Zahlen eingebettet:

$$\left\{ \frac{m}{1} \;\middle|\; m \in \mathbf{Z} \right\} \subset \mathbf{Q} \text{ und } \left\{ \frac{m}{1} \;\middle|\; m \in \mathbf{Z} \right\} \approx \mathbf{Z}.$$

Daher schreiben wir $\mathbf{Z} \subset \mathbf{Q}$ (obwohl diese Aussage mathematisch nicht korrekt ist).

Im folgenden seien $a \in \mathbf{Z}, b \in \mathbf{Z}, c \in \mathbf{Z}, d \in \mathbf{Z}, b \neq 0$ und $d \neq 0$.

Die Darstellung einer rationalen Zahl als Bruch zweier ganzer Zahlen ist nicht eindeutig. So ist etwa 3/1 = 6/2 = 2712/904 = $(-12)/(-4)$.

Es gilt für $\frac{a}{b} \in \mathbf{Q}$ und $\frac{c}{d} \in \mathbf{Q}$:

$\frac{a}{b} = \frac{c}{d}$ genau dann, wenn $a \cdot d = c \cdot b$.

Damit sind $\frac{a}{b} = \frac{(-a)}{(-b)}$ und

$\frac{a}{b} = 0 = \frac{0}{1}$ genau dann, wenn $a = 0$ ist.

Die additiv inverse rationale Zahl zu $\frac{a}{b} \in \mathbf{Q}$ ist $\frac{(-a)}{b} = \frac{a}{(-b)}$.

Zu jeder rationalen Zahl $r = \frac{a}{b}$ mit $a \neq 0$ gibt es eine (multiplikativ) **inverse Zahl** r^{-1} mit $r \cdot r^{-1} = 1$: es ist $r^{-1} = \left(\frac{a}{b}\right)^{-1} = \left(\frac{b}{a}\right)$.

Die arithmetischen Operationen +, - und $\cdot$ auf rationalen Zahlen $\frac{a}{b} \in \mathbf{Q}$ und $\frac{c}{d} \in \mathbf{Q}$ sind definiert durch

$$\frac{a}{b} \pm \frac{c}{d} = \frac{ad \pm cb}{bd}$$

und

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

Die Division zweier rationaler Zahlen $\frac{a}{b} \in \mathbf{Q}$ und $\frac{c}{d} \in \mathbf{Q}$ wird auf die Multiplikation zurückgeführt:

$$\frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a}{b} \cdot \left(\frac{c}{d}\right)^{-1} = \frac{a}{b} \cdot \frac{d}{c} = \frac{ad}{bc}.$$

Die Ordnungsrelation $\leq$ wird von **Z** auf **Q** erweitert:

Es ist $0 \leq \frac{a}{b}$ genau dann, wenn $((0 \leq a) \wedge (0 < b))$ oder $((a \leq 0) \wedge (b < 0))$ gelten. Außerdem ist $\frac{a}{b} \leq \frac{c}{d}$ genau dann, wenn $\frac{a}{b} + \left(-\frac{c}{d}\right) \leq 0$ gilt.

Mit diesen Festlegungen bildet die algebraische Struktur $(\mathbf{Q}, +, \cdot, 0, 1)$ mit der Ordnungsrelation $\leq$ einen angeordneten Körper.

Erläuterung:

Eine algebraische Struktur $(K, \oplus, \otimes)$ heißt **Körper**, wenn gilt:

(i) $(K, \oplus, \otimes, 0, 1)$ ist ein kommutativer Ring mit 1

(ii) für jedes $a \in K$ mit $a \neq 0$ gibt es ein Element $a^{-1} \in K$ mit $a \otimes a^{-1} = 1$; dieses Element heißt **multipliktives inverses Element** zu *a*.

$(K, \oplus, 0)$ ist also eine kommutative Gruppe, die additive Gruppe des Körpers,
$(K \setminus \{ 0 \}, \otimes, 1)$ ist eine kommutative Gruppe, die multiplikative Gruppe des Körpers, und es gelten die Distributivgesetze.

Der Körper $(K, \oplus, \otimes)$ heißt **angeordneter Körper**, wenn es eine totale Ordnungsrelation $\lhd$ auf *K* gibt mit folgenden Eigenschaften:

(i) für jedes $x \in K$, für jedes $y \in K$ und für jedes $z \in K$ gilt:
aus $x \lhd y$ folgt $x \oplus z \lhd y \oplus z$

(ii) für jedes $x \in K$ und für jedes $y \in K$ gilt:
aus $0 \lhd x$ und $0 \lhd y$ folgt $0 \lhd x \otimes y$.

In $(\mathbf{Q}, +, \cdot, 0, 1)$ sind die wichtigsten arithmetischen Operationen möglich. Jedoch fehlt der Ordnungsrelation auf **Q** eine wichtige Eigenschaft, nämlich die Vollständigkeit. Beispielsweise sind die Elemente der Menge

$$\left\{ q \,\middle|\, q \in \mathbf{Q} \text{ und } q^2 \leq 2 \right\}$$

wohl nach oben beschränkt, z.B. durch $r = 3/2$, es gibt in **Q** aber keine kleinste obere Schranke für die Elemente dieser Menge (denn $\sqrt{2} \notin \mathbf{Q}$). Daher werden die rationalen Zahlen um die irrationalen Zahlen erweitert. Das Resultat ist der Körper $(\mathbf{R}, +, \cdot, 0, 1)$ der **reellen Zahlen**. Der Erweiterungsprozess kann auf verschiedene Weisen unter topologischen Aspekten vollzogen werden (z.B. Dedekind-Schnitte, mittels Fundamentalfolgen, Intervallschachtelung oder durch Dezimalbruchentwicklung). Exemplarisch wird hier die Methode der Dedekind-Schnitte angegeben:

Erläuterung:

Eine Teilmenge $S \subseteq \mathbf{Q}$ heißt (**Dedekind-) Schnitt**, wenn gilt:

(i) $S \neq \varnothing$ und $S \neq \mathbf{Q}$

(ii) Für jedes $r \in S$ ist die Menge $\{ q \mid q \in \mathbf{Q} \text{ und } q \leq r \}$ eine echte Teilmenge von *S*.

Bedingung (ii) beinhaltet zwei Eigenschaften:

(i') Ist $r \in S$ und $q \in \mathbf{Q}$ mit $q \leq r$, so ist auch $q \in S$ (**Abgeschlossenheit nach unten**)

(ii') Ist $r \in S$, so gibt es ein $p \in S$ mit $r < p$ (**Nichtexistenz eines Maximums**).

Die Menge **R** der reellen Zahlen ist die Menge aller (Dedekind-) Schnitte.

Die Menge der rationalen Zahlen lässt sich in **R** einbetten, indem man $r \in \mathbf{Q}$ mit dem Schnitt $\{ x \mid x \in \mathbf{Q} \text{ und } -\infty < x < r \}$ identifiziert. Daher schreiben wir $\mathbf{Q} \subset \mathbf{R}$ (obwohl diese Aussage mathematisch nicht korrekt ist).

Auf der Menge der Schnitte wird durch die Mengeninklusion $\subseteq$ eine totale Ordnungsrelation definiert, die die Ordnungsrelation $\leq$ auf **Q** fortsetzt. Diese Fortsetzung der Ordnungsrelation auf die Schnitte in **Q**, d.h. die reellen Zahlen, hat die Eigenschaft, dass jede nichtleere nach oben beschränkt Teilmenge reeller Zahlen eine kleinste obere Schranke besitzt (**Vollständigkeit der Ordnungsrelation**).

Damit hat der Schnitt $\{ q \mid q \in \mathbf{Q} \text{ und } q^2 < 2 \}$ eine kleinste obere Schranke, die allerdings nicht in **Q** liegt, nämlich die als $\sqrt{2}$ bezeichnete reelle Zahl.

Die Operationen der Addition und der Multiplikation lassen sich von **Q** auf die Menge der Schnitte, d.h. **R**, fortsetzen, so dass insgesamt $(\mathbf{R}, +, \cdot, 0, 1)$ zu einem vollständig angeordneten Körper wird, der sogar „strukturell“ der einzige vollständig angeordneten Körper ist. Die Einzelheiten sollen an dieser Stelle nicht weiter behandelt werden (siehe beispielsweise in Hachenberger, D.: **Mathematik für Informatiker**, Pearson Studium, 2005).

Leider sind auch in $(\mathbf{R}, +, \cdot, 0, 1)$ noch nicht alle arithmetischen Operationen möglich. So besitzt die Gleichung $x^2 + 1 = 0$ keine Lösung $x \in \mathbf{R}$. Daher erweitert man den Zahlbereich **R** (unter Wahrung der arithmetischen Operationen):

Die **imaginäre Zahl** *i* wird durch die Eigenschaft $i^2 = -1$ definiert. Dann ist die **Menge der komplexen Zahlen** definiert durch

$$\mathbf{C} = \{\ a + bi \mid a \in \mathbf{R} \text{ und } b \in \mathbf{R}\ \}.$$

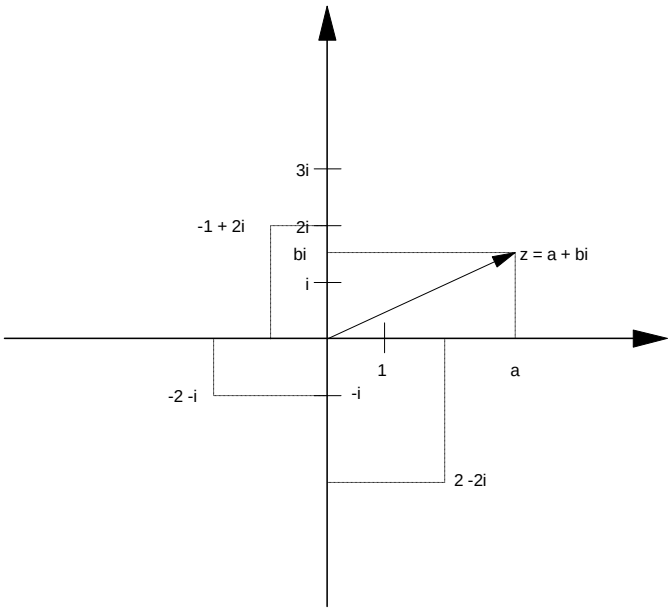
Bei der Zahl $z = a + bi$ heißt *a* der **Realteil** und *b* der **Imaginärteil** von *z*.

Die Menge der reellen Zahlen ist in der Menge der komplexen Zahlen eingebettet:

$$\{\ r + 0i \mid r \in \mathbf{R}\ \} \subset \mathbf{C} \text{ und } \{\ r + 0i \mid r \in \mathbf{R}\ \} \approx \mathbf{R}.$$

Daher schreiben wir $\mathbf{R} \subset \mathbf{C}$ (obwohl diese Aussage mathematisch nicht korrekt ist).

Die komplexen Zahlen lassen sich als Punkte in einer Ebene mit rechtwinkligem Koordinatensystem, der **komplexen Ebene**, darstellen. Dabei wird für eine komplexe Zahl $z = a + bi$ ihr Realteil *a* auf der horizontalen Achse abgetragen, ihr Imaginärteil *b* auf der vertikalen Achse. Die folgende Abbildung zeigt die komplexen Zahlen $z = a + bi$, $-1 + 2i$, $-2 - i$ und $2 - 2i$.



Der **Betrag** $|z|$ **der komplexen Zahl** $z = a + bi$ ist geometrisch durch die Länge der Verbindungslinie des Punkts (0, 0) der komplexen Ebene mit dem Punkt (a, b) definiert:

$$|z| = |\ a + bi\ | = \sqrt{a^2 + b^2}\ .$$

Die arithmetischen Operationen werden definiert durch

$$(a + bi) \pm (c + di) = (a \pm c) + (b \pm d)i \text{ und}$$
$$(a + bi) \cdot (c + di) = (ac - bd) + (ad + bc)i\ .$$

Die zu einer komplexen Zahl $a + bi$ (multiplikativ) inverse Zahl $(a + bi)^{-1}$ lautet

$$(a + bi)^{-1} = \frac{a}{a^2 + b^2} + \frac{-b}{a^2 + b^2}\, i\ .$$

Die Division zweier komplexer Zahlen $a + bi$ und $c + di$ wird auf die Multiplikation zurückgeführt:

$$\begin{aligned}(a + bi) / (c + di) &= (a + bi) \cdot (c + di)^{-1} \\ &= (a + bi) \cdot \left(\frac{c}{c^2 + d^2} + \frac{-d}{c^2 + d^2}\, i \right) \\ &= \frac{ac + bd}{c^2 + d^2} + \frac{bc - ad}{c^2 + d^2}\, i\end{aligned}$$

Mit diesen Operationen bildet auch $(\mathbf{C}, +, \cdot, 0, 1)$ einen Körper. Zu beachten ist, dass die Ordnungsrelation der reellen Zahlen, die $(\mathbf{R}, +, \cdot, 0, 1)$ zu einem vollständig angeordneten Körper macht, nicht auf die komplexen Zahlen fortgesetzt wird; denn der Körper $(\mathbf{C}, +, \cdot, 0, 1)$ lässt sich nicht anordnen (in einem angeordneten Körper *K* gilt $0 \lhd a \otimes a$ für jedes $a \in K$, aber in **C** ist nach Definition $i^2 = -1 < 0$).

Insgesamt gilt (mathematisch nicht korrekt): $\mathbf{N} \subset \mathbf{Z} \subset \mathbf{Q} \subset \mathbf{R} \subset \mathbf{C}$. Die jeweiligen arithmetischen Operationen $+, -, \cdot, /$ und die Ordnungsrelation $\leq$, soweit sie in den einzelnen Zahlensystemen überhaupt definiert sind, werden für alle Zahlensysteme gleich bezeichnet.

1.5 Vollständige Induktion

Das vorliegende Kapitel behandelt eine der wichtigsten Beweismethoden, wenn es um Aussagen über natürliche Zahlen oder um Aussagen über Mengen geht, die strukturell äquivalent zu den natürlichen Zahlen sind: die vollständige Induktion. Wegen der großen Bedeutung dieser Methode für die Informatik werden in diesem Kapitel ausnahmsweise die durchgeführten Beweise in den Beispielen explizit angegeben.

Es sei $A(n)$ eine Aussage über die natürliche Zahl $n \in \mathbf{N}$, d.h. die von *n* abhängt. Es soll gezeigt werden, dass diese Aussage für alle natürlichen Zahlen $n \geq n_0$ gilt.

Häufig ist $n_0 = 0$; dann soll $A(n)$ für alle natürlichen Zahlen *n* bewiesen werden.

Nach der **Beweismethode der vollständigen Induktion** geht man wie folgt vor:

1. Man zeigt die Gültigkeit der Aussage $A(n_0)$ (**Induktionsanfang**)

2. Man beweist die Gültigkeit der Implikation $(A(n) \Rightarrow A(n+1))$ (**Induktionsschluss**).

Aus Axiom 5 der natürlichen Zahlen in Kapitel 1.4 („Eine Menge *M* natürlicher Zahlen, die die Zahl 0 und mit jeder Zahl *n* auch ihren Nachfolger *n*' enthält, ist mit **N** identisch.“) folgt dann, dass $A(n)$ für alle natürlichen Zahlen gilt. Hier soll nur der Spezialfall $n_0 = 0$ gezeigt werden. Dazu setzt man $M = \{ m \mid A(m) \text{ gilt für } m \}$. Der Induktionsanfang besagt $n_0 \in M$;

der Induktionsschluss besagt: wenn $n \in M$ ist, dann ist auch der Nachfolger $n' \in M$; Axiom 5 besagt nun gerade, dass $M = \mathbf{N}$ gilt, d.h. dass $A(n)$ für alle natürlichen Zahlen *n* gilt[1].

Die Beweismethode der vollständigen Induktion wird in unterschiedlichen Varianten an einigen Beispielen erläutert:

Beispiel:

Zu beweisen ist die Aussage

Für alle $n \in \mathbf{N}$ gilt: $0+1+2+...+n = \frac{n \cdot (n+1)}{2}$.

Induktionsanfang: Die Aussage gilt für $n = 0$, denn auf der linken Seite des Gleichheitszeichens steht nur 0, und auf der rechten Seite des Gleichheitszeichens steht $\frac{0 \cdot 1}{2}$; beide Seiten ergeben denselben Wert.

Induktionsschluss: Für $n \in \mathbf{N}$ gelte: $0+1+2+...+n = \frac{n \cdot (n+1)}{2}$. Zu zeigen ist, dass dann diese Formel auch für die natürliche Zahl *n* + 1 gilt. Das lässt sich aber leicht nachrechnen: Auf der linken Seite des Gleichheitszeichens steht

$0+1+2+...+n+(n+1)$.

Für diese Summe gilt (da die Formel für *n* als gültig vorausgesetzt wird):

$$0+1+2+...+n+(n+1) = \frac{n \cdot (n+1)}{2} + (n+1) = \frac{n \cdot (n+1) + 2 \cdot (n+1)}{2} = \frac{n^2 + 3 \cdot n + 2}{2};$$

auf der rechten Seite des Gleichheitszeichens steht $\frac{(n+1) \cdot (n+2)}{2} = \frac{n^2 + 3 \cdot n + 2}{2}$;

beide Seiten sind gleich, also gilt die Formel auch für die natürliche Zahl *n* + 1.

---

[1] Der Fall $n_0 > 0$ verläuft analog, indem man eine Variante (Folgerung) von Axiom 5 verwendet, nämlich: „Eine Menge *M* natürlicher Zahlen, die die Zahl $n_0 \in \mathbf{N}$ und mit jeder Zahl *n* auch ihren Nachfolger *n*' enthält, ist mit $\{ n \mid n \in \mathbf{N} \text{ und } n \geq n_0 \}$ identisch.“.

Beispiel:

Zu beweisen ist die Aussage

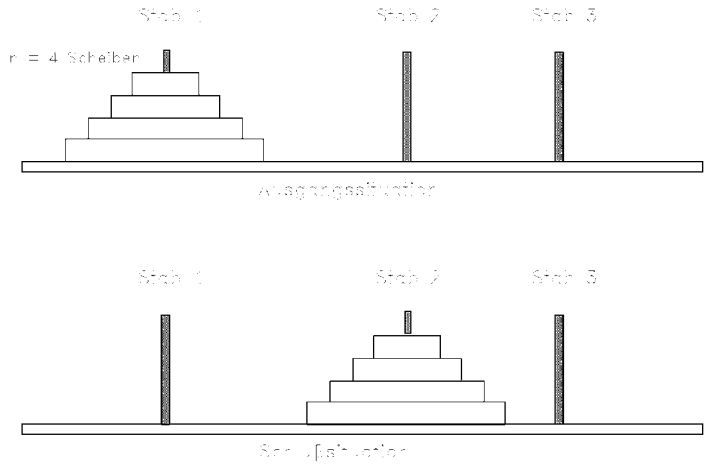
Ist *A* eine endliche Menge mit *n* Elementen, dann enthält die Potenzmenge $\mathbf{P}(A)$ 2^n viele Elemente (d.h. eine Menge mit *n* Elementen besitzt 2^n viele Teilmengen).

Induktionsanfang: Die Aussage gilt für $n = 0$, denn wenn *A* kein Element besitzt, dann ist $A = \varnothing$; andererseits ist $\mathbf{P}(\varnothing) = \{B | B \subseteq \varnothing\} = \{\varnothing\}$, und diese Menge enthält $1 = 2^0$ viele Elemente.

Induktionsschluss: Die Aussage gelte für Mengen *A* mit *n* Elementen, d.h. $|\mathbf{P}(A)| = 2^{|A|} = 2^n$. Es sei *B* eine Menge mit $n+1$ Elementen. Zu zeigen ist, dass die Anzahl der Teilmengen von *B* gleich 2^{n+1} ist: Da $|B| = n+1 > 0$ ist, gibt es ein Element $b \in B$. Die Teilmengen *C* von *B* werden danach klassifiziert, ob sie das Element *b* enthalten oder nicht. Jede Teilmenge $C \subseteq B$, mit $b \notin C$ ist Teilmenge von $B \setminus \{ b \}$. Da diese Menge genau *n* Elemente enthält, gibt es 2^n viele Teilmengen $C \subseteq B$ mit $b \notin C$. Jeder Teilmenge $C \subseteq B$ mit $b \in C$ entspricht genau eine Teilmenge $C' \subseteq B$ mit $b \notin C'$, nämlich $C' = C \setminus \{ b \}$. Umgekehrt entspricht jeder Teilmenge $C' \subseteq B$ mit $b \notin C'$ eine Teilmenge $C \subseteq B$ mit $b \in C$, nämlich $C = C' \cup \{ b \}$. Daher gibt es genauso viele Teilmengen $C \subseteq B$ mit $b \in C$ wie Teilmengen $C \subseteq B$ mit $b \notin C$, nämlich 2^n viele. Insgesamt ist die Anzahl an Teilmengen von *B* gleich $2 \cdot 2^n = 2^{n+1}$.

Beispiel:

Beim Spiel der **Türme von Hanoi** sind drei Stapel mit Bezeichnern *A*, *B* und *C* gegeben. Auf dem Stapel *A* befinden sich *n* Scheiben, die übereinander in absteigender Größe liegen; die beiden anderen Stapel sind leer. Die Aufgabe besteht darin, die Scheiben vom Ausgangsstapel *A* auf einen der anderen Stapel, etwa *B*, zu bewegen, wobei jeweils die Scheiben nur einzeln bewegt werden dürfen. Der dritte Stapel *C* kann als Zwischenablage verwendet werden. Dabei ist die Randbedingung, niemals eine größere auf eine kleinere Scheibe zu legen, einzuhalten. Wieviele Scheiben müssen genau bewegt werden?



Es bezeichne $T(n)$ die minimale Anzahl an Bewegungen, um *n* Scheiben von einem Stapel auf einen anderen Stapel unter Zuhilfenahme des dritten Stapels zu bewegen.

Offensichtlich ist $T(0) = 0$, $T(1) = 1$, $T(2) = 3$.

Folgende Lösungsstrategie führt zum Ziel: Man ignoriere zunächst die größte (unterste) Scheibe auf Stapel *A* und bringe die oberen $n-1$ Scheiben vom Stapel *A* zum Stapel *C* unter Zuhilfenahme des Stapels *B* als Zwischenspeicher unter Beachtung obiger Randbedingung. Anschließend bewege man die auf Stapel *A* verbliebene größte Scheibe zum Stapel *B*, er ja jetzt leer ist. Dann bringe man die $n-1$ Scheiben vom Stapel *C* zum Stapel *B* unter Zuhilfenahme des Stapels *A* als Zwischenspeicher unter Beachtung obiger Randbedingung.

Der folgende (Pascal-) Programmausschnitt realisiert diese Startegie:

```
CONST max = ...;

TYPE disk_typ = 0 .. max;
     stab_typ = 1 .. 3;

PROCEDURE move (anz: disk_typ;
                a  : stab_typ;
                b  : stab_typ;
                c  : stab_typ);
  { move bewegt Scheiben, deren Anzahl in anz angegeben wird,
    vom Stab a zum Stab b, wobei der Stab c als Zwischenspeicher
    verwendet wird                                              }

    BEGIN { move }
      IF anz > 0 THEN BEGIN
                        move (anz - 1, a, c, b);
                        Writeln ('Lege eine Scheibe von ', a,
```

```
                                                    ' nach ', b, '.');
                        move (anz - 1, c, b, a)
                       END
    END   { move };
```

Es gilt $T(0) = 0$ und $T(n) \le 2 \cdot T(n-1) + 1$ für $n > 0$.

Man kommt nicht mit weniger Scheibenbewegungen aus; denn um an die größte Scheibe auf dem Ausgangsstapel heranzukommen und diese vom Ausgangsstapel auf einen anderen Stapel zu bewegen, müssen zuvor $n-1$ kleinere Scheiben auf einem einzigen Stapel liegen. Dann kann die größte Scheibe bewegt werden (mindestens einmal), und dann müssen noch einmal $n-1$ kleinere Scheiben bewegt werden. Das bedeutet $T(n) \ge 2 \cdot T(n-1) + 1$.

Insgesamt gilt $T(0) = 0$ und $T(n) = 2 \cdot T(n-1) + 1$ für $n > 0$.

Es bleibt die Bestimmung von $T(n)$ in alleiniger Abhängigkeit von n (und nicht von $T(n-1)$). Dazu werde einige kleinere Werte für n ausprobiert:

$T(0) = 0$,
$T(1) = 2 \cdot T(0) + 1 = 2 \cdot 0 + 1 = 1$,
$T(2) = 2 \cdot T(1) + 1 = 2 \cdot 1 + 1 = 3$,
$T(3) = 2 \cdot T(2) + 1 = 2 \cdot 3 + 1 = 7$,
$T(4) = 2 \cdot T(3) + 1 = 2 \cdot 7 + 1 = 15$,
$T(5) = 2 \cdot T(4) + 1 = 2 \cdot 15 + 1 = 31$,
$T(6) = 2 \cdot T(5) + 1 = 2 \cdot 31 + 1 = 63$.

Die Vermutung liegt nahe, dass $T(n) = 2^n - 1$ für alle $n \in \mathbf{N}$ gilt (für $n = 0, ..., 6$ wurde es explizit ausgerechnet. Für die übrigen $n \in \mathbf{N}$ wird die Vermutung durch vollständige Induktion nachgewiesen:

Induktionsanfang: Die Aussage gilt für $n = 0, ..., 6$ (siehe oben).

Induktionsschluss: Die Aussage gelte bis $n \in \mathbf{N}$. Dann ist

$$\begin{aligned} T(n+1) &= 2 \cdot T(n) + 1 \qquad \text{(nach Definition von } T(n)\text{)} \\ &= 2 \cdot \left(2^n - 1\right) + 1 \quad \text{(Voraussetzung im Induktionsschluss)} \\ &= 2^{n+1} - 2 + 1 \\ &= 2^{n+1} - 1. \end{aligned}$$

Beispiel:

Ein wichtiges Suchverfahren in der Informatik ist die **Binärsuche**:

Gegeben sei ein Feld `t[1]`, ..., `t[`n`]` mit ganzzahligen Einträgen (allgemeiner: mit bezüglich einer Ordnungsrelation vergleichbaren Einträgen), die nach aufsteigender Größe sortiert sind, d.h. es gilt `t[1]` $\leq$ `t[2]` $\leq$ ... $\leq$ `t[`$n-1$`]` $\leq$ `t[`n`]`. Die Aufgabe besteht darin festzustellen, ob ein vorgegebener Wert a unter `t[1]`, ..., `t[`n`]` vorkommt und in diesem Fall den Index i zu ermitteln, für den a = t[i] gilt. Anstelle das Feld linear von Anfang bis eventuell zum Ende zu durchsuchen, kann man folgendermaßen vorgehen:

Zunächst wird das mittlere Element `t[middle]` geprüft (bei einer geraden Anzahl von Elementen ist das mittlere Element das erste Element der zweiten Feldhälfte). Ist es gleich a, so ist der gesuchte Feldindex gefunden, und die Suche ist beendet. Andernfalls liegt a, wenn es überhaupt im Feld vorkommt, im vorderen Feldabschnitt, falls a < `t[middle]` ist, oder im hinteren Feldabschnitt, falls a > `t[middle]` ist. Die Entscheidung, in welchem Feldabschnitt weiterzusuchen ist, kann jetzt getroffen werden. Gleichzeitig wird durch diese Entscheidung die andere Hälfte aller potentiell auf Übereinstimmung mit a zu überprüfenden Feldelement ausgeschlossen. Im Feldabschnitt, der weiter zu überprüfen ist, wird nach dem gleichen Prinzip (also rekursiv) verfahren. Unter Umständen muss die Suche fortgesetzt werden, bis ein noch zu überprüfender Feldabschnitt nur noch ein Feldelement enthält.

Eine (Pascal-) Implementierung der Binärsuche lautet wie folgt:

```
CONST n = ...;

TYPE Tarray = ARRAY [1..n] OF INTEGER;

FUNCTION Binaersuche (t   : Tarray;
                      a   : INTEGER;
                      von : INTEGER;
                      bis : INTEGER) : INTEGER;

VAR mitte : INTEGER;

BEGIN { Binaersuche }
  Binaersuche := -1;
  IF von < bis
  THEN BEGIN { der Feldausschnitt
               t[von] , ... t[bis]
               enthält mindestens 2 Elemente                }
         mitte := von + ((bis - von + 1) DIV 2);
         IF a = t[mitte]          { ← }
         THEN Binaersuche := mitte
```

```
        ELSE BEGIN
               IF a < t[middle]  { ← }
               THEN Binaersuche := Binaersuche (t, a, von, mitte-1)
               ELSE Binaersuche
                            := Binaersuche (t, a, mitte + 1, bis);
             END
      END
  ELSE BEGIN
        IF a = t[von]
        THEN Binaersuche := von;
      END;
END   { Binaersuche };
```

Der Aufruf zum Durchsuchen des Feldes `t[1]`, ..., `t[`*n*`]` nach dem Element *a* lautet

```
Binaersuche (t, a, 1, n);
```

Der Rechenaufwand der Binärsuche ist proportional zur Anzahl der Vergleiche in der mit ← gekennzeichneten Zeilen. Zur Vereinfachung der Analyse des Rechenaufwands wird

$$n = 2^m - 1$$

angenommen (hat *n* nicht diese Form, dann ergibt sich eine ähnliche Abschätzung). Der Wert *n* beschreibt die Anzahl der Elemente des zu durchsuchenden Felds.

In diesem Fall ist

$$\texttt{mitte = } 1 + \left((n - 1 + 1)\ \texttt{DIV}\ 2\right) = 1 + \left(\left(2^m - 1\right)\ \texttt{DIV}\ 2\right) = 2^{m-1},$$

d.h. wenn *a* nicht in der Mitte des Felds vorkommt, enthält das Anfangsstück des Felds `t[1]`, ..., `t[mitte - 1]` $2^{m-1} - 1$ viele Elemente bzw. das Endstück `t[mitte + 1]`, ..., `t[`*n*`]` ebenfalls $2^{m-1} - 1$ viele Elemente; die Suche wird in einem dieser Abschnitte fortgeführt.

Mit $B(n)$ wird die Anzahl der Vergleiche des Elements *a* mit einem Feldelement bezeichnet, wenn das Feld *n* Elemente enthält. Dann gilt

$$B(n) = B\left(2^m - 1\right) \le B\left(2^{m-1} - 1\right) + 2,$$
$$B(1) = B\left(2^1 - 1\right) = 1.$$

Um diese Ungleichungen nur in Abhängigkeit von $n = 2^m - 1$ bzw. von *m* auszudrücken, werden einige Werte für *m* ausprobiert:

$$B(1) = B\left(2^1 - 1\right) = 1,$$
$$B(3) = B\left(2^2 - 1\right) \le B\left(2^1 - 1\right) + 2 = 3,$$
$$B(7) = B\left(2^3 - 1\right) \le B\left(2^2 - 1\right) + 2 \le 3 + 2 = 5,$$
$$B(15) = B\left(2^4 - 1\right) \le B\left(2^3 - 1\right) + 2 \le 5 + 2 = 7,$$
$$B(31) = B\left(2^5 - 1\right) \le B\left(2^4 - 1\right) + 2 \le 7 + 2 = 9,$$
$$B(63) = B\left(2^6 - 1\right) \le B\left(2^5 - 1\right) + 2 \le 9 + 2 = 11.$$

Die Vermutung liegt nahe, dass folgende Gesetzmäßigkeit gilt:

$B\left(2^m - 1\right) \le 2 \cdot m - 1$ für jedes $m \in \mathbf{N}$ mit $m \ge 1$.

Diese Gesetzmäßigkeit wird durch vollständige Induktion bezogen auf m („über m“) bewiesen:

Induktionsanfang: Die Aussage gilt für $m = 1, ..., 6$ (siehe oben).

Induktionsschluss: Die Aussage gelte bis zur natürlichen Zahl $m \in \mathbf{N}$. Dann ist

$$\begin{aligned} B(2^{m+1} - 1) &\le B(2^m - 1) + 2 && \text{(aus dem Algorithmus)} \\ &\le \left(2 \cdot m - 1\right) + 2 && \text{(Voraussetzung im Induktionsschluss)} \\ &= 2 \cdot \left(m + 1\right) - 1. \end{aligned}$$

Dieses Ergebnis zeigt beispielsweise, dass die Binärsuche in einem Feld mit $n = 2.147.483.647 = 2^{31} - 1$ Elementen maximal nur 61 Feldelementvergleiche benötigt.

Weitere Beispiele für Beweise durch vollständige Induktion finden sich in den folgenden Kapiteln.

Die Methode der vollständigen Induktion erfordert die sorgfältige Formulierung der zu beweisenden Aussage $A(n)$. Dabei muss man die zu beweisende Aussage bereits kennen (raten) und sie dann mit Hilfe des Induktionsanfangs und des Induktionsschlusses beweisen (verifizieren).

Folgende Hinweise sollten beachtet werden:

1. Der Induktionsanfang ist wichtig. Fehlt er, kann der Beweis fehlschlagen.

2. Im Induktionschluss lautet die Annahme „Es gelte $A(n)$“, d.h. die Gültigkeit von $A(n)$ wird nur für ein $n \in \mathbf{N}$ angenommen und nicht für *alle* $n \in \mathbf{N}$.

3. Bei der Durchführung des Induktionsschlusses muss die Annahme der Gültigkeit von $A(n)$ auch verwendet, d.h. in die Argumentation eingebaut werden.

1.6 Endliche Summen

Häufig hat man es mit **Summen mit einer endlichen Anzahl von Summanden** zu tun, die alle jeweils nach einem ähnlichen Schema aufgebaut sind, etwa

$$S = a_1 + a_2 + a_3 + \ \ldots\ a_{n-1} + a_n .$$

Für diese Summe schreibt man abkürzend:

$$S = \sum_{i=1}^{n} a_i .$$

In die „Formel“ a_i wird nacheinander $i = 1$, $i = 2$, ... , $i = n-1$ und $i = n$ eingesetzt, und die so erhaltenen Summanden werden aufsummiert. Die Berechnung von S könnte also in einer Programmiersprache wie folgt formuliert werden (PASCAL-Pseudocode):

```
S := 0;
FOR i := 1 TO n DO
  S := S + ai;
```

bzw.

```
S := 0;
i := 1;
WHILE i <= n DO
  BEGIN
    S := S + ai;
    i := i + 1;
  END;
```

Beispiel:

Es sei $a_i = 3i^2 + 1$. Dann ist

$$\sum_{i=1}^{4} a_i = \sum_{i=1}^{4} (3i^2 + 1) = (3 \cdot \mathbf{1}^2 + 1) + (3 \cdot \mathbf{2}^2 + 1) + (3 \cdot \mathbf{3}^2 + 1) + (3 \cdot \mathbf{4}^2 + 1) = 4 + 13 + 28 + 49 = 94.$$

Häufig beginnt eine Summe nicht mit dem kleinsten Index $i = 1$, sondern mit einer anderen ganzen Zahl (auch negative Zahlen sind zugelassen), so dass man es allgemein mit einer endlichen Summe der Form $S = \sum_{i=k}^{n} a_i$ zu tun hat. Hierin heißt *i* der **Summationsindex**, die Zahl *k* die **Summationsuntergrenze** und die Zahl *n* die **Summationsobergrenze**.

Die Summe $S = \sum_{i=k}^{n} a_i$ enthält $n - k + 1$ viele Summanden.

In der Darstellung der Summe $S = \sum_{i=k}^{n} a_i$ wird deutlich, wie die einzelnen Summanden aufgebaut sind, nämlich gemäß einer Formel $a_i = a(i)$. Die Summe *S* ist nicht nur von den einzelnen Summanden, sondern auch von der Summationsuntergrenze und –obergrenze abhängig, d.h. $S = S(k, n)$. Die Darstellung $S(k, n) = \sum_{i=k}^{n} a_i$ zeigt nicht den Wert der Summe in Abhängigkeit von der Summationsuntergrenze *k* und der Summationsobergrenze *n*. Eine Aufgabe besteht daher in der Berechnung des Werts der Summe in Abhängigkeit von den Summationsgrenzen (**Berechnung der Summe *S*(*k*, *n*) in geschlossener Form**).

Beispiel:

Die Summe $S(1, n) = \sum_{i=1}^{n} (3i^2 + 1)$ hat den Wert $S(1, n) = \frac{n(2 + (2n+1)(n+1))}{2}$. Bei $n = 4$ ergibt sich $S(1, 4) = 94$.

Eine endliche Summe lässt sich in Teilsummen zerlegen, die ihrerseits wieder mit jeweils einem Summenzeichen zusammengefasst werden können, z.B.

$$a_1 + a_2 + \ \dots\ + a_{k-1} + a_k + a_{k+1} + a_{k+2} + \ \dots\ + a_{j-1} + a_j + a_{j+1} + a_{j+2} + \ \dots\ + a_{n-1} + a_n$$
$$= a_1 + a_2 + \ \dots\ + a_{k-1} + \left(\sum_{i=k}^{j} a_i\right) + a_{j+1} + a_{j+2} + \ \dots\ + a_{n-1} + a_n$$
$$= \left(\sum_{i=1}^{k-1} a_i\right) + \left(\sum_{i=k}^{j} a_i\right) + \left(\sum_{i=j+1}^{n} a_i\right).$$

Die *Bezeichnung i* des Summationsindex kann beliebig geändert werden:

$$\sum_{i=k}^{n} a_i = \sum_{\mu=k}^{n} a_\mu \ .$$

Anstelle von $\sum_{i=k}^{n} a_i$ schreibt man auch $\sum_{k \le i \le n} a_i$.

Ist I eine beliebige Menge (**Indexmenge**), so ist $\sum_{i \in I} a_i$ die Summe, die man dadurch enthält, dass man nacheinander a_i für jedes $i \in I$ bildet und die einzelnen Summanden aufaddiert. Auf die Reihenfolge, in der man die einzelnen Indizes $i \in I$ betrachtet, kommt es nicht an.

Beispiel:

Die Summe der Quadrate aller geraden Zahlen zwischen 4 und 12 ist

$$\sum_{i \in \{4,6,8,10,12\}} i^2 = 4^2 + 6^2 + 8^2 + 10^2 + 12^2$$
$$= (2 \cdot 2)^2 + (2 \cdot 3)^2 + (2 \cdot 4)^2 + (2 \cdot 5)^2 + (2 \cdot 6)^2$$
$$= \sum_{i=2}^{6} (2i)^2$$
$$= \sum_{i=2}^{6} 4i^2 = 4 \cdot 2^2 + 4 \cdot 3^2 + 4 \cdot 4^2 + 4 \cdot 5^2 + 4 \cdot 6^2$$
$$= 4 \cdot \sum_{i=2}^{6} i^2 = 360.$$

Vereinbarungsgemäß ist die Summe über eine leere Anzahl von Summanden gleich 0:

$\sum_{i\in\varnothing} a_i = 0$ und $\sum_{i=k}^{n} a_i = 0$ für $k > n$.

Satz 1.6-1:

(i) Ist c eine Konstante, die vom Summationsindex nicht abhängt, so ist

$$\sum_{i\in I}(c\cdot a_i) = c\cdot\sum_{i\in I} a_i \ .$$

(ii) $$\sum_{i\in I}(a_i \pm b_i) = \left(\sum_{i\in I} a_i\right) \pm \left(\sum_{i\in I} b_i\right).$$

(iii) Ist c eine Konstante, die vom Summationsindex nicht abhängt, so ist

$$\sum_{i=1}^{n} c = n\cdot c \text{ und } \sum_{i=k}^{n} c = (n-k+1)\cdot c \ .$$

(iv) Es sei $k \in \mathbf{N}$ mit $1 \le k \le n$. Dann ist

$$\sum_{i=1}^{n} a_i = \sum_{i=k}^{n+k-1} a_{i-k+1} \text{ (**Indexverschiebung**)}.$$

(v)
$$\left(\sum_{i=1}^{n} a_i\right)\cdot\left(\sum_{j=1}^{m} b_j\right) = (a_1 + \ldots a_n)\cdot(b_1 + \ldots b_m)$$
$$= a_1\cdot(b_1 + \ldots b_m) + \ldots a_n\cdot(b_1 + \ldots b_m)$$
$$= a_1\cdot\left(\sum_{j=1}^{m} b_j\right) + \ldots + a_n\cdot\left(\sum_{j=1}^{m} b_j\right)$$
$$= \sum_{i=1}^{n}\left(a_i\cdot\left(\sum_{j=1}^{m} b_j\right)\right)$$
$$= \sum_{i=1}^{n}\left(\left(\sum_{j=1}^{m} a_i\cdot b_j\right)\right),$$

$$\left(\sum_{i\in I} a_i\right)\cdot\left(\sum_{j\in J} b_j\right) = \sum_{i\in I, j\in J}(a_i\cdot b_j).$$

Satz 1.6-2:

(i) Die Summe aller natürlichen Zahlen bis zur Zahl *n* ist gleich

$$\sum_{i=0}^{n} i = 1 + 2 + \ \ldots\ (n-1) + n = \frac{n(n+1)}{2}.$$

Die Summe aller *geraden* natürlichen Zahlen bis zur Zahl *n* ist gleich

$$\sum_{\substack{0 \le i \le n \\ \text{und} \\ i \text{ ist gerade}}} i \ = \lfloor n/2 \rfloor \cdot (\lfloor n/2 \rfloor + 1).$$

Die Summe aller *ungeraden* natürlichen Zahlen bis zur Zahl *n* ist gleich

$$\sum_{\substack{0 \le i \le n \\ \text{und} \\ i \text{ ist ungerade}}} i \ = \left\lfloor \frac{n+1}{2} \right\rfloor^2.$$

(ii) Es sei $q \in \mathbf{R}$ eine Konstante. Dann ist

$$\sum_{i=0}^{n} q^i = 1 + q + q^2 + q^3 + \ldots + q^{n-1} + q^n$$

$$= \begin{cases} n+1 & \text{für } q = 1 \\ \dfrac{1-q^{n+1}}{1-q} = \dfrac{q^{n+1}-1}{q-1} & \text{für } q \neq 1 \end{cases}$$

Spezialfall: $q = 2$: $\quad \sum_{i=0}^{n} 2^i = 1 + 2 + 4 + \ \ldots\ 2^n = 2^{n+1} - 1.$

$$\sum_{i=0}^{n} i q^i = q + 2q^2 + 3q^3 + \ldots + (n-1)q^{n-1} + nq^n$$

$$= \begin{cases} \dfrac{n(n+1)}{2} & \text{für } q = 1 \\ \dfrac{q - (n+1)q^{n+1} + nq^{n+2}}{(1-q)^2} & \text{für } q \neq 1 \end{cases}$$

Spezialfall: $q = 2$: $\quad \sum_{i=0}^{n} i 2^i = (n-1)2^{n+1} + 2.$

../..

(iii) $$\sum_{i=2}^{n} \frac{1}{i(i-1)} = \frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \ \ldots \ + \frac{1}{n(n-1)} = 1 - \frac{1}{n} \quad .$$

$$\sum_{i=1}^{n} \frac{1}{i(i+1)} = \frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \ \ldots \ + \frac{1}{n(n+1)} = 1 - \frac{1}{n+1} \quad .$$

(iv) $$\sum_{i=0}^{n} i^2 = 1 + 4 + 9 + \ \ldots \ + (n-1)^2 + n^2 = \frac{n(n+1)(2n+1)}{6} \quad .$$

2 Abbildungen

Abbildungen stellen Beziehungen zwischen Mengen *A* und *B* her. Sie können als Spezialisierung des Konzepts der Relationen zwischen Mengen definiert werden.

2.1 Allgemeines

In Kapitel 1.4 wurden die Begriffe Äquivalenzrelation und der Ordnungsrelation eingeführt. Diese sind Spezialisierungen des allgemeineren Begriffs der Relation:

Es seien *A* und *B* zwei Mengen. Eine Teilmenge $R \subseteq A \times B$ heißt **Relation** zwischen *A* und *B*.

Eine Relation besteht also aus Paaren (a, b) mit $a \in A$ und $b \in B$.

In Kapitel 1.4 wird eine Ordnungsrelation auf einer Menge *M* mit $\lhd$ bezeichnet und Eigenschaften dieser Relation beschrieben, nämlich Reflexivität, Antisymmetrie und Transitivität. Im Sinne der obigen Definition schreibt man anstelle von $a \lhd b$ auch $(a, b) \in \lhd$ und $\lhd \subseteq M \times M$. Entsprechendes gilt für eine Äquivalenzrelation $\approx$ auf einer Menge *M*; diese besitzt die Eigenschaften Reflexivität, Symmetrie und Transitivität. Anstelle von $a \approx b$ schreibt man auch $(a, b) \in \approx$ und $\approx \subseteq M \times M$.

Eine Relation $R \subseteq A \times B$ heißt **linkstotal**, wenn es zu jedem $a \in A$ ein $b \in B$ mit $(a, b) \in R$ gibt. Bei einer linkstotalen Relation *R* kommen alle Elemente von *A* als erste Komponenten in den Paaren in *R* vor. Zu $a \in A$ kann es auch mehrere $b_1 \in B$, ..., $b_m \in B$ geben mit $(a, b_1) \in R$, ..., $(a, b_m) \in R$.

Eine Relation $R \subseteq A \times B$ heißt **rechtstotal**, wenn es zu jedem $b \in B$ ein $a \in A$ mit $(a, b) \in R$ gibt. Bei einer rechtstotalen Relation *R* kommen alle Elemente von *B* als zweite Komponenten in den Paaren in *R* vor. Das Element $a \in A$, das es zu $b \in B$ mit $(a, b) \in R$ gibt, muss auch hier nicht eindeutig bestimmt sein, d.h. es kann zu $b \in B$ mehrere $a_1 \in A$, ..., $a_n \in A$ geben mit $(a_1, b) \in R$, ..., $(a_n, b) \in R$.

Eine Relation $R \subseteq A \times B$ heißt **linkseindeutig**, wenn gilt: aus $(a_1, b) \in R$ und $(a_2, b) \in R$ folgt $a_1 = a_2$. Bei einer linkseindeutigen Relation gilt dann: Sind $(a_1, b_1) \in R$ und $(a_2, b_2) \in R$ und gilt $a_1 \neq a_2$, so ist auch $b_1 \neq b_2$.

Eine Relation $R \subseteq A \times B$ heißt **rechteindeutig**, wenn gilt: aus $(a, b_1) \in R$ und $(a, b_2) \in R$ folgt $b_1 = b_2$. Bei einer rechtseindeutigen Relation gilt dann: Sind $(a_1, b_1) \in R$ und $(a_2, b_2) \in R$ und gilt $b_1 \neq b_2$, so ist auch $a_1 \neq a_2$.

Eine Relation $f \subseteq A \times B$ heißt **Abbildung** von *A* nach *B*, wenn f linkstotal und rechtseindeutig ist. Gleichbedeutend damit ist folgende Formulierung:

$f \subseteq A \times B$ ist eine Abbildung, wenn es zu jedem $a \in A$ genau ein $b \in B$ gibt mit $(a, b) \in f$.

Da dieses eindeutig bestimmt Element $b \in B$ „zu $a \in A$ gehört“, schreibt man anstelle von $(a, b) \in f$ auch $b = f(a)$ und bezeichnet es als **Bild** von *a* unter *f*. Häufig gibt es eine Rechenvorschrift, nach der zu gegebenem $a \in A$ das Bild $f(a)$ zu bestimmen ist, etwa

$f(a) = a^3 - 3a^2 + 2$.

Dann wird eine Abbildung *f* von *A* nach *B* beschrieben durch

$$f: \begin{cases} A \to B \\ a \to f(a) \end{cases},$$

oder auch in der Form

$f: A \to B,\ f(a) = \ ...$

Die Menge *A* heißt **Definitionsbereich** von *f*, die Menge

$W(f) = \{ b \mid b \in B, \text{und es gibt } a \in A \text{ mit } f(a) = b \}$

heißt **Wertebereich** von *f*. Es ist $W(f) \subseteq B$.

Anstelle von $W(f)$ schreibt man auch $f(A)$.

Für eine Funktion $f: A \to B$ ist also der Wertebereich gleich

$f(A) = \{ b \mid b \in B, \text{ und es gibt } a \in A \text{ mit } f(a) = b \}$.

Die Angabe $f: A \to B$ legt fest, dass einem Element vom (Daten-) Typ, der „charakteristisch“ für *A* ist, jeweils genau ein Element vom (Daten-) Typ, der „charakteristisch“ für *B* ist, zugeordnet wird. Beispielsweise könnte die Menge *A* aus Objekten vom Objekttyp *T* und die Menge *B* aus natürlichen Zahlen bestehen. Dann legt die Angabe $f: A \to B$ fest, dass jedem Objekt vom Objekttyp *T* in der Menge *A* durch *f* eine natürliche Zahl, die beispielsweise als Primärschlüsselwert interpretierbar ist, zugeordnet wird. Die Angabe $f(a) =$ ... beschreibt, wie diese Zuordnung für jedes Element $a \in A$ geschieht.

Das Bild eines Elements $a \in A$ unter *f* ist eindeutig bestimmt, und es gilt $|f(a)| = 1$ für jedes $a \in A$. Andererseits kann es durchaus Werte a_1 und a_2 mit $a_1 \neq a_2$ und $f(a_1) = f(a_2)$ geben; beispielsweise ist für die durch $f(x) = x^2$ für $x \in \mathbf{R}$ definierte Abbildung $f(-2) = f(2) = 4$.

Die Menge $f^{-1}(b) = \{ a \mid a \in A \text{ und } f(a) = b \}$ wird als **Urbild** von *b* unter *f* bezeichnet.

Eine Abbildung $f: A \to B$ mit $A \subseteq \mathbf{R}$ und $W(f) \subseteq \mathbf{R}$ heißt **reelle Funktion einer Veränderlichen**.

Beispiele:

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

$$g: \begin{cases} \mathbf{R} \setminus \{0\} & \to & \mathbf{R} \\ x & \to & 3/x \end{cases}$$

$$h: \begin{cases} [0, \infty[& \to & \mathbf{R} \\ x & \to & 1 - e^{-x} \end{cases}$$

$$id_A: \begin{cases} A & \to & A \\ x & \to & x \end{cases}$$ **Identität auf *A***

$$par_a: \begin{cases} \mathbf{R} & \rightarrow & \mathbf{R} \\ x & \rightarrow & ax(x-1) \end{cases} \quad \textbf{Parabel}$$

$$F: \begin{cases}]-10,\infty[& \rightarrow & \mathbf{R} \\ x & \rightarrow & \begin{cases} x^2 & \text{für } -10 < x \le 2 \\ x^3 - x & \text{für } 2 < x \le 20 \\ 2x+8 & \text{für } x > 20 \end{cases} \end{cases}$$

$$f_!: \begin{cases} \mathbf{N} & \rightarrow & \mathbf{N} \\ n & \rightarrow & \begin{cases} 1 & \text{für } n = 0 \\ n \cdot f_!(n-1) & \text{für } n > 0 \end{cases} \end{cases} \quad \textbf{Fakultätsfunktion}$$

Die hier aufgeführte Definition der Fakultätsfunktion zeigt die Form einer **rekursiven Definition**. Rekursive Funktionsdefinitionen werden häufig angewandt, wenn der Definitionsbereich der Funktion die natürlichen Zahlen oder eine Teilmenge der natürlichen Zahlen ist. Für den kleinsten Wert *n* des Definitionsbereich bzw. für mehrere der kleinsten Werte wird $f(n)$ direkt angegeben. Für größere Werte *n* wird $f(n)$ als arithmetischer Ausdruck, der *n*, eventuell kleinere Werte *m* und Funktionswerte $f(m)$ mit $m < n$ enthält.

Die Fakultätsfunktion kann auch nicht-rekursiv definiert werden:

$$f_!: \begin{cases} \mathbf{N} & \rightarrow & \mathbf{N} \\ n & \rightarrow & \begin{cases} 1 & \text{für } n = 0 \\ 1 \cdot \ldots \cdot (n-1) \cdot n & \text{für } n > 0 \end{cases} \end{cases}$$

Ein weiteres Beispiel einer rekursiven Funktion mit der zugehörigen nicht-rekursiven Definition ist die Fibonnacci-Funktion, die einen nichttrivialen Zusammenhang zwischen beiden Formen der Definition zeigt (siehe Kapitel 5.10):

$$fib: \begin{cases} \mathbf{N} & \rightarrow & \mathbf{N} \\ n & \rightarrow & \begin{cases} n & \text{für } n = 0 \text{ und } n = 1 \\ fib(n-1) + fib(n-2) & \text{für } n \ge 2 \end{cases} \end{cases} \quad \text{bzw.}$$

$$fib(n) = \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right) \text{ für } n \ge 0 .$$

Im folgenden werden hauptsächlich reelle Funktionen betrachtet. Gelegentlich wird auf die Angabe des Definitionsbereichs einer Abbildung verzichtet; dann wird implizit immer die größte Teilmenge von **R** genommen, für die die Abbildungsvorschrift definiert ist.

Für eine Abbildung $f: A \to B$ heißt die Menge $\{(a, f(a)) \mid a \in A\}$ **Graph** der Abbildung *f*.

Sind $f: A \to B$ und $g: B \to C$ zwei Abbildungen, dann heißt die Abbildung $h: A \to C$ mit $h(a) = g(f(a))$ die **Komposition (Zusammensetzung)** der Abbildungen *f* und *g*, geschrieben $h = g \circ f$.

Es ist $W(g \circ f) \subseteq W(g) \subseteq C$, und i.a. gilt $g \circ f \neq f \circ g$.

Beispiel:

$$f: \begin{cases} \mathbf{R} \setminus \{-1\} & \to & \mathbf{R} \\ x & \to & 1/(1+x) \end{cases} \qquad g: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

$$g \circ f: \begin{cases} \mathbf{R} \setminus \{-1\} & \to & \mathbf{R} \\ x & \to & 1/(1+x)^2 \end{cases}$$

$$f \circ g: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & 1/(1+x^2) \end{cases}$$

2.2 Grundlegende Eigenschaften von Abbildungen

Eine Abbildung $f: A \to B$ heißt **surjektive Abbildung (Surjektion)**, wenn sie rechtstotal ist.

Die Abbildung $f: A \to B$ sei surjektiv. Dann ist $f(A) = B$, d.h. der Wertebereich von *f* umfasst ganz *B*. Für jedes $b \in B$ ist also $\left|f^{-1}(b)\right| \geq 1$, d.h. es gibt mindestens ein $a \in A$ mit $f(a) = b$ (eventuell gibt es mehrere Werte $a \in A$, die auf *b* abgebildet werden).

Beispiel:

Die Abbildung

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

ist nicht surjektiv, da es zu keiner negativen Zahl $y \in \mathbf{R}$ einen Wert $x \in \mathbf{R}$ gibt mit $f(x) = x^2 = y < 0$. Durch Einschränkung der Zielmenge kann man jedoch die Surjektivität erzwingen. Beispielsweise ist

$$f_0: \begin{cases} \mathbf{R} & \to & \mathbf{R}_{\geq \mathbf{0}} \\ x & \to & x^2 \end{cases}$$

surjektiv.

Eine Abbildung $f : A \to B$ heißt **injektive Abbildung (Injektion)**, wenn sie linkseindeutig ist.

Beispiel:

Die Abbildung

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

ist nicht injektiv, da für $x_1 = -1$ und $x_2 = 1$ offensichtlich $x_1 \neq x_2$ ist, aber $f(x_1) = f(-1) = (-1)^2 = 1 = f(1) = f(x_2)$ ist.

Die Injektivität kann man durch Einschränkung des Definitionsbereichs erzwingen. Beispielsweise ist die Funktion

$$f_1: \begin{cases} \mathbf{R}_{\geq \mathbf{0}} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

injektiv.

Die Injektivität einer Funktion kann an ihrem Graphen abgelesen werden: Jede Parallele zur *x*-Achse schneidet den Graphen einer injektiven Funktion in höchstens einem Punkt.

Eine Abbildung heißt **bijektive Abbildung (Bijektion)**, wenn sie sowohl surjektiv als auch injektiv ist.

Die Begriffe bezüglich Relationen und Abbildungen fasst folgende Tabelle zusammen.

Typ der Relation				
linkstotal	x	x	x	x
rechtstotal		x		x
linkseindeitig			x	x
rechtseindeutig	x	x	x	x
	Abbildung	Surjektion	Injektion	Bijektion

Satz 2.2-1:

Es sei $f : A \to B$ eine bijektive Abbildung. Dann gilt:

(i) Für jedes $b \in B$ gibt es genau ein $a_b \in A$ mit $b = f(a_b)$.

(ii) Es gibt eine eindeutig bestimmte Abbildung $g : B \to A$ mit $g(b) = a_b$; außerdem gilt für jedes $a \in A$: $g(f(a)) = a$ und für jedes $b \in B$: $f(g(b)) = b$.

Die Aussage in Satz 2.2-1 (i) kann man so interpretieren, dass es eine Eins-zu-eins-Beziehung zwischenden Elementen der Menge *A* und der Menge *B* gibt.

Ist $f : A \to B$ eine bijektive Abbildung, so heißt die gemäß Satz 2.2-1 (ii) existierende Funktion $g : B \to A$ die **Umkehrabbildung** von *f* und wird mit f^{-1} bezeichnet. Es gilt:

$f^{-1} \circ f = id_A$ und $f \circ f^{-1} = id_B$, d.h. $\left(f^{-1} \circ f\right)(a) = a$ und $\left(f \circ f^{-1}\right)(b) = b$.

Beim Graph einer bijektiven Abbildung $f : \mathbf{R} \to \mathbf{R}$ vollzieht sich der Übergang zur Umkehrfunktion $\overset{-1}{f}$ durch Spiegelung an der Winkelhalbierenden (45°-Linie).

Beispiele:

Die Abbildung

$$F : \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & ax+b \end{cases}$$

ist für festes $a \in \mathbf{R}$ mit $a \neq 0$ und festem $b \in \mathbf{R}$ bijektiv und hat die Umkehrfunktion

$$\overset{-1}{F}: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ y & \to & \frac{1}{a} y - \frac{b}{a} \end{cases} .$$

Im allgemeinen ist eine Abbildung der Form

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & ax^3 + bx^2 + cx + d \end{cases}$$

mit festen rellen Werten *a*, *b*, *c* und *d* nicht bijektiv.

Die Abbildung

$$f_1 : \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$$

ist weder injektiv noch surjektiv. Jedoch sind die Abbildungen

$$f_2 : \begin{cases} \mathbf{R}_{\geq 0} & \to & \mathbf{R}_{\geq 0} \\ x & \to & x^2 \end{cases}$$

und

$$f_3 : \begin{cases} \mathbf{R}_{\leq 0} & \to & \mathbf{R}_{\geq 0} \\ x & \to & x^2 \end{cases}$$

jeweils bijektiv mit den durch $\overset{-1}{f}_2(y) = +\sqrt{y}$ bzw. $\overset{-1}{f}_3(y) = -\sqrt{y}$ definierten Umkehrabbildungen.

Häufig wird bereits für eine injektive Abbildung $f : A \to B$, die nicht notwendigerweise surjektiv ist, die Umkehrabbildung $\overset{-1}{f}$ definiert, und zwar nur für die Werte $b \in B$ aus dem Wertebereich von *f*:

$\overset{-1}{f} : f(X) \to X$.

Satz 2.2-2:

Es seien $f : A \to B$ und $g : B \to C$ Abbildungen. Dann gilt:

(i) Sind *f* und *g* surjektiv, dann ist auch $g \circ f$ surjektiv.

(ii) Sind *f* und *g* injektiv, dann ist auch $g \circ f$ injektiv.

(iii) Sind f und g bijektiv, dann ist auch $g \circ f$ bijektiv. In diesem Fall gilt $\left(\overset{-1}{g \circ f}\right) = \overset{-1}{f} \circ \overset{-1}{g}$.

Für **endliche** Mengen A und B kann man die Existenz surjektiver, injektiver und bijektiver Abbildungen zwischen A und B folgendermaßen entscheiden:

Satz 2.2-3:

Die Mengen A und B seien endliche Mengen mit $|A| = n$ und $|B| = m$. Dann gilt:

(i) Es gibt genau dann eine surjektive Abbildung $f : A \to B$, wenn $n \geq m$ ist.

(ii) Es gibt genau dann eine injektive Abbildung $f : A \to B$, wenn $n \leq m$ ist.

(iii) Es gibt genau dann eine bijektive Abbildung $f : A \to B$, wenn $n = m$ ist.

Satz 2.2-3 lässt den Schluss zu, dass bei Abbildungen zwischen endlichen Mengen mit derselben Elementanzahl die Begriffe Injektivität, Surjektivität und Bijektivität zusammenfallen:

Satz 2.2-4:

Die Mengen A und B seien endliche Mengen mit derselben Elementanzahl $|A| = |B|$. Es sei $f : A \to B$ eine Abbildung. Dann gilt:

Die folgenden drei Aussagen (a), (b) und (c) sind äquivalent:

(a) f ist eine surjektive Abbildung.

(b) f ist eine injektive Abbildung.

(c) f ist eine bijektive Abbildung.

Satz 2.2-3 (iii) besagt für endliche Mengen, dass sie genau dann gleichmächtig sind, wenn es zwischen ihnen eine bijektive Abbildung gibt. Dieser Ansatz lässt sich auf unendliche Mengen übertragen:

Zwei unendliche Mengen A und B heißen **gleichmächtig**, wenn es eine bijektive Abbildung $f : A \to B$ gibt. Wegen der Existenz der bijektiven Umkehrfunktion g zu f ist diese Definition gleichbedeutend mit der Existenz einer bijektiven Abbildung $g : B \to A$.

Bei unendlichen Mengen *A* und *B* tritt die folgende Situation auf, die sich am Beispiel der Nachfolgerfunktion *succ* auf den natürlichen Zahlen verdeutlichen lässt:

$$succ: \begin{cases} \mathbf{N} & \to & \mathbf{N}_{>0} \\ n & \to & n+1 \end{cases}$$

Wie man leicht nachrechnet, handelt es sich hierbei um eine bijektive Abbildung, d.h. die Menge **N** der natürlichen Zahlen ist gleichmächtig mit der echten Teilmenge $\mathbf{N}_{>0} = \mathbf{N} \setminus \{ 0 \}$. Dieses Phänomen erlaubt es, die Endlichkeit bzw. Unendlichkeit einer Menge exakt zu definieren:

Eine Menge *A* ist **endlich von der Mächtigkeit** *n*, wenn es eine bijektive Abbildung $f : \{ 0, ..., n-1 \} \to A$ gibt, d.h. man kann die Elemente in *A* mit den natürlichen Zahlen 0, ..., $n-1$ durchnumerieren: $A = \{ f(0), ..., f(n-1) \} = \{ a_0, ..., a_{n-1} \}$. Hierbei ist $f(i) \neq f(j)$ bzw. $a_i \neq a_j$ für $i \neq j$.

Eine Menge *A* ist **von der Mächtigkeit unendlich**, wenn es eine bijektive Abbildung $f : B \to A$ zwischen einer echten Teilmenge $B \subset A$ und *A* gibt.

Eine Menge heißt **abzählbar**, wenn sie entweder endlich oder gleichmächtig zu den natürlichen Zahlen ist. Eine unendliche Menge, die nicht abzählbar ist, heißt **überabzählbar**.

Satz 2.2-5:

(i) Es gibt eine bijektive Abbildung $h_{\mathbf{Z}} : \mathbf{N} \to \mathbf{Z}$, und es gibt eine bijektive Abbildung $h_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}$. Die Mengen **N**, **Z** und **Q** sind daher abzählbar.

(ii) Die Menge **N** der natürlichen Zahlen lässt sich nicht auf die Menge **R** der reellen Zahlen bijektiv abbilden. Die Menge **R** ist daher überabzählbar.

Für Satz 2.2-5 (i) sind zwei bijektive Abbildungen $h_{\mathbf{Z}} : \mathbf{N} \to \mathbf{Z}$ und $h_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}$ anzugeben.

Die bijektive Abbildung $h_{\mathbf{Z}} : \mathbf{N} \to \mathbf{Z}$ wird definiert durch

$$h_{\mathbf{Z}} : \begin{cases} \mathbf{N} & \to & \mathbf{Z} \\ n & \to & \begin{cases} -n/2 & \text{falls } n \text{ gerade ist} \\ (n+1)/2 & \text{falls } n \text{ ungerade ist} \end{cases} \end{cases}.$$

Einige Werte dieser Abbildung sind

n	0	1	2	3	4	5	6	7	8	9	10	11
$h_{\mathbf{Z}}(n)$	0	1	-1	2	-2	3	-3	4	-4	5	-5	6

Die bijektive Abbildung $h_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}$ wird in zwei Schritten konstruiert. Zunächst wird eine bijektive Abbildung $f_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}_{\geq 0}$ angegeben, die dann zu einer bijektiven Abbildung $h_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}$ erweitert wird:

Die rationalen Zahlen $\mathbf{Q}_{>0} = \left\{ \frac{r}{t} \,\middle|\, r \in \mathbf{N}_{>0} \text{ und } t \in \mathbf{N}_{>0} \right\}$ kann man sich in ein unendliches Zahlenschema eingetragen denken, das aus Zeilen und Spalten besteht. In der ersten Zeile stehen alle Zahlen $1/1$, $1/2$, $1/3$, $1/4$, ... ; die zweite Zeile enthält $2/1$, $2/2$, $2/3$, $2/4$, ...; die i-te Zeile enthält $i/1$, $i/2$, $i/3$, $i/4$, ... Dann steht die Zahl $\frac{r}{t}$ in der r-ten Zeile und t-ten Spalte. Dieses Zahlenschema wird durchnumeriert: $1/1$ erhält die Nummer 1 ($1/1$ ist die einzige rationale Zahl $\frac{r}{t} \in \mathbf{Q}_{>0}$ mit $r + t = 2$). Dann kommen alle Zahlen $\frac{r}{t} \in \mathbf{Q}_{>0}$ mit $r + t = 3$, nach aufsteigenden Zählern geordnet (das sind $1/2$ und $2/1$). Anschließend kommen alle $\frac{r}{t} \in \mathbf{Q}_{>0}$ mit $r + t = 4$, nach aufsteigenden Zählern geordnet (das sind $1/3$, $2/2$ und $3/1$) usw. Die folgende Tabelle zeigt einige kleine rationale Zahlen mit ihren Nummern.

m	r/t mit $r \geq 1$, $t \geq 1$ und $r + t = m$	Anzahl	Nummern
2	$1/1$	1	1
3	$1/2$ $2/1$	2	2 3
4	$1/3$ $2/2$ $3/1$	3	4 5 6
5	$1/4$ $2/3$ $3/2$ $4/1$	4	7 8 9 10
6	$1/5$ $2/4$ $3/3$ $4/2$ $5/1$	5	11 12 13 14 15

Es gibt genau $m-1$ rationale Zahlen $\frac{r}{t} \in \mathbf{Q}_{>0}$ mit $r + t = m$, nämlich $1/(m-1)$, $2/(m-2)$, $3/(m-3)$, ..., $(m-1)/1$. Vor diesem „Block“ von Zahlen liegen alle Zahlen $\frac{u}{v} \in \mathbf{Q}_{>0}$ mit $u \geq 1$, $v \geq 1$ und $u + v = i$ mit $2 \leq i \leq m-1$. Daher bekommt $1/(m-1)$ die Nummer

$$\sum_{i=2}^{m-1} (i-1) + 1 = \sum_{i=1}^{m-2} i + 1 = \frac{(m-1)\cdot(m-2)}{2} + 1 .$$

Die Zahl $k/(m-k)$ für $k = 1, ..., m-1$ bekommt die Nummer $\frac{(m-1)\cdot(m-2)}{2} + k$, d.h. die natürlichen Zahlen

$$\frac{(m-1)\cdot(m-2)}{2}+1,\ \frac{(m-1)\cdot(m-2)}{2}+2,\ ...,\ \frac{(m-1)\cdot(m-2)}{2}+m-1=\frac{(m-1)\cdot m}{2}$$

numerieren die Zahlen $\frac{r}{t} \in \mathbf{Q}_{>0}$ mit $r+t=m$.

Die Abbildung $f_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}_{\geq 0}$ wird nun wie folgt definiert:

$f_{\mathbf{Q}}(0) = 0$.

Für $n > 0$ gibt es eine eindeutig bestimmte Zahl $m \in \mathbf{N}_{>0}$ mit $\frac{(m-1)\cdot(m-2)}{2} < n \leq \frac{(m-1)\cdot m}{2}$.

Beispiel:

m	1	2	3	4	5	6	7	8	9	10	11	12
$\frac{(m-1)\cdot(m-2)}{2}$	0	0	1	3	6	10	15	21	28	36	45	55
$\frac{(m-1)\cdot m}{2}$	0	1	3	6	10	15	21	28	36	45	55	66

Für $n = 17$ ist $m = 7$, für $n = 21$ ist $m = 7$, für $n = 22$ ist $m = 8$.

Die Anzahl der Werte *n* mit $\frac{(m-1)\cdot(m-2)}{2} < n \leq \frac{(m-1)\cdot m}{2}$ beträgt $m-1$.

Es wird

$r = n - \frac{(m-1)\cdot(m-2)}{2}$ und

$t = m - r$

gesetzt. Da $(m-1)\cdot(m-2)$ gerade ist und $\frac{(m-1)\cdot(m-2)}{2} < n$ gilt, ist $r \in \mathbf{N}_{>0}$. Außerdem ist $r = n - \frac{(m-1)\cdot(m-2)}{2} \leq \frac{(m-1)\cdot m}{2} - \frac{(m-1)\cdot(m-2)}{2} = m-1$ und damit auch $t \in \mathbf{N}_{>0}$.

Es wird $f_{\mathbf{Q}}(n)$ durch

$f_{\mathbf{Q}}(n) = \frac{r}{t}$ (in dieser ungekürzten Darstellung)

definiert. Die folgende Tabelle zeigt die Ergebnisse $f_{\mathbf{Q}}(n)$ für die $m-1$ Werte n mit

$$\frac{(m-1)\cdot(m-2)}{2} < n \le \frac{(m-1)\cdot m}{2} :$$

n	$\frac{(m-1)\cdot(m-2)}{2}+1$	$\frac{(m-1)\cdot(m-2)}{2}+2$	...	$\frac{(m-1)\cdot m}{2}-1$	$\frac{(m-1)\cdot m}{2}$
r	1	2	...	$m-2$	$m-1$
t	$m-1$	$m-2$	...	2	1
$f_{\mathbf{Q}}(n)$	$1/(m-1)$	$2/(m-2)$		$(m-2)/2$	$(m-1)/1$

Offensichtlich gilt für diese n jeweils $f_{\mathbf{Q}}(n) = \frac{r}{t}$ mit $1 \le r \le m-1$, $1 \le t \le m-1$ und $r+t=m$.

Man kann zeigen, dass $f_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}_{\ge 0}$ bijektiv ist. Die gesuchte Bijektion $h_{\mathbf{Q}} : \mathbf{N} \to \mathbf{Q}$ ergibt sich zu

$$h_{\mathbf{Q}} : \begin{cases} \mathbf{N} & \to & \mathbf{Q} \\ n & \to & \begin{cases} -f_{\mathbf{Q}}(n/2) & \text{falls } n \text{ gerade ist} \\ f_{\mathbf{Q}}((n+1)/2) & \text{falls } n \text{ ungerade ist} \end{cases} \end{cases}.$$

Für Satz 2.2-5 (ii) wird angenommen, dass es eine bijektive Abbildungen $h_{\mathbf{R}} : \mathbf{N} \to \mathbf{R}$ gibt. Diese Annahme muss auf einen Widerspruch führen. Da dieses Vorgehen eine „klassische“ Beweismethode auch besonders der Theoretischen Informatik ist und auch in die populärwissenschaftliche mathematische Literatur Eingang gefunden hat, soll die Beweisidee hier skizziert werden:

Es seien $n_1 < n_2 < n_3 < ...$ diejenigen $n_i \in \mathbf{N}$, für die $h_{\mathbf{R}}(n_i) \in \mathbf{R}$ mit $0 \le h_{\mathbf{R}}(n_i) \le 1$ ist. Jedes $r \in \mathbf{R}$ mit $0 \le r \le 1$ hat eine eindeutige Nummer n_i und lautet in Dezimalschreibweise

$$r = 0, d_{n_i,-1} d_{n_i,-2} d_{n_i,-3} ...$$

mit den Dezimalziffern $d_{n_i,-j} \in \{0,1,2,3,4,5,6,7,8,9\}$. Dabei ist $1 = 0,\overline{9}...$ (siehe Kapitel 5.1).

Es wird eine reelle Zahl $\bar{r}$ mit $0 \le \bar{r} \le 1$ durch folgende Vorschrift konstruiert:

Die erste Dezimalziffer von $\bar{r}$ nach dem Komma lautet $9 - d_{n_1,-1}$ (es wird aus der reellen Zahl mit der Nummer n_1 die erste Dezimalziffer nach dem Komma genommen und von 9 abgezogen; dadurch entsteht wieder eine Ziffer zwischen 0 und 9); es ist $9 - d_{n_1,-1} \neq d_{n_1,-1}$. Die zweite Dezimalziffer von $\bar{r}$ nach dem Komma lautet $9 - d_{n_2,-2}$ (es wird aus der reellen Zahl mit der Nummer n_2 die zweite Dezimalziffer nach dem Komma genommen und von 9 abgezogen); es ist $9 - d_{n_2,-2} \neq d_{n_2,-2}$. Allgemein: die *j*-te Dezimalziffer von $\bar{r}$ nach dem Komma lautet $9 - d_{n_j,-j}$; es ist $9 - d_{n_j,-j} \neq d_{n_j,-j}$.

Da $\bar{r}$ eine reelle Zahl mit $0 \leq \bar{r} \leq 1$ ist und $h_{\mathbf{R}}$ als bijektiv angenommen wurde, gibt es einen Wert $n_k \in \mathbf{N}$ mit $h_{\mathbf{R}}(n_k) = \bar{r}$ ($\bar{r}$ ist die reelle Zahl mit der Nummer n_k):

$$\bar{r} = 0, d_{n_k,-1} d_{n_k,-2} d_{n_k,-3} \ldots d_{n_k,-k} \ldots$$

Die *k*-te Dezimalziffer von $\bar{r}$ nach dem Komma ist $d_{n_k,-k}$. Nach Konstruktion von $\bar{r}$ lautet die *k*-te Dezimalziffer von $\bar{r}$ nach dem Komma jedoch $9 - d_{n_k,-k}$, und es ist $9 - d_{n_k,-k} \neq d_{n_k,-k}$. Daher kann es keinen Wert $n_k \in \mathbf{N}$ mit $h_{\mathbf{R}}(n_k) = \bar{r}$ geben, und die Annahme der Existenz einer bijektiven Abbildung $h_{\mathbf{R}} : \mathbf{N} \to \mathbf{R}$ ist falsch.

3 Ausgewählte Themen der elementaren Zahlentheorie

In diesem Kapitel werden einige für die Informatik grundlegende und wichtige Themen der elementaren Zahlentheorie behandelt. Neben der Tatsache, dass sie zum mathematischen Basiswissen in jeder Disziplin gehören, haben diese Themen in den letzten Jahren insbesondere in der Kryptologie zunehmende Bedeutung erlangt.

3.1 Primzahlen

Es seien $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ ganze Zahlen mit $b \neq 0$. Die Zahl *a* heißt durch *b* **teilbar** (*b* **teilt** *a*), geschrieben $b \mid a$, wenn es ein $d \in \mathbf{Z}$ gibt mit $a = d \cdot b$.

Der folgende Satz führt einige wichtige Teilbarkeitsregeln ganzer Zahlen auf und lässt sich durch Zurückführen auf obige Definition leicht verifizieren:

Satz 3.1-1:

(i) Gilt $c \mid b$ und $b \mid a$, so gilt auch $c \mid a$.

(ii) Gilt $b_1 \mid a_1$ und $b_2 \mid a_2$, so gilt auch $b_1 b_2 \mid a_1 a_2$.

(iii) Gilt $b \mid a_1$ und $b \mid a_2$, so gilt für jedes $x \in \mathbf{Z}$ und für jedes $y \in \mathbf{Z}$: $b \mid (x a_1 + y a_2)$.

(iv) Gilt $b \mid a$ und $a \mid b$, so ist $a = b$ oder $a = -b$.

Bemerkung: Da trivialerweise immer $a \mid a$ gilt, definiert wegen Satz 3.1-1 (i) und (iv) die durch

„$(n, m) \in R$ genau dann, wenn $n \mid m$ gilt“

definierte Relation eine partielle Ordnungsrelation (siehe Kapitel 1.4) auf $\mathbf{N} \times \mathbf{N}$.

Eine wichtige Teilmenge der natürlichen Zahlen ist die Menge **P** der **Primzahlen**:

$\mathbf{P} = \{\ p \mid\ p \in \mathbf{N},\ p \geq 2,$ und die einzigen Teiler von p sind 1 und $p\ \}$
$= \{\ 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, \ldots\ \}.$

Der folgende Satz zeigt, dass die Primzahlen als Grundbausteine der natürlichen Zahlen und damit des gesamten Zahlensystems angesehen werden können.

Satz 3.1-2:

Jedes $n \in \mathbf{N}$ lässt sich in ein **Produkt aus Primzahlpotenzen** zerlegen, d.h.

$$n = p_1^{e_1} \cdot p_2^{e_2} \cdot \ldots \cdot p_r^{e_r}$$

mit Primzahlen $p_1,\ p_2,\ \ldots,\ p_r$ und natürlichen Zahlen $e_1 \geq 1,\ e_2 \geq 1,\ \ldots,\ e_r \geq 1$. Diese Zerlegung ist (bis auf die Reihenfolge der Primzahlpotenzen) eindeutig.

Beispielsweise ist $600 = 2^3 \cdot 3 \cdot 5^2$.

Der folgende Satz fasst einige wichtige Eigenschaften von Primzahlen zusammen:

Satz 3.1-3:

(i) Es gibt unendlich viele Primzahlen.

(ii) Es gibt beliebig große Abstände zwischen zwei aufeinanderfolgenden Primzahlen.

(iii) Es gibt unendlich viele Paare $(p,\ p+2)$, die beide Primzahlen sind (**Primzahlzwillinge**)

(iv) Ist $2^n + 1$ eine Primzahl, so ist n eine Zweierpotenz.

(v) Ist $2^n - 1$ eine Primzahl, so ist n eine Primzahl.

In der Praxis der Kryptographie werden ständig große Primzahlen benötigt (mit einer Stellenzahl von mehr als 150 Dezimalstellen). Dabei sind neben Primzahlen, deren Ziffernfolgen keinen festen Gesetzmäßigkeiten unterliegen, Primzahlen der Form $2^n + 1$ und $2^n - 1$ besonders interessant. Diese haben nämlich eine sehr einfache Binärdarstellung ($2^n + 1 = 1\ \underbrace{0 \ldots 0}_{(n-1)\text{-mal}}\ 1$,

$2^n - 1 = \underbrace{1 \ldots 1}_{n-\text{mal}}$). Wegen Satz 3.1-3 (iv) kann man die Suche nach sehr großen Primzahlen der Form $2^n + 1$ auf diejenigen *n* beschränken, die die Form $n = 2^m$ haben, d.h. auf Zahlen der Form $2^n + 1 = 2^{2^m} + 1$. Zahlen der Form $2^{2^m} + 1$ heißen **Fermat**-Zahlen. Beispielsweise sind die Fermat-Zahlen

$$2^{2^0} + 1 = 3,\ 2^{2^1} + 1 = 5,\ 2^{2^2} + 1 = 17,\ 2^{2^3} + 1 = 257,\ 2^{2^4} + 1 = 65.537$$

Primzahlen. Nicht jede Fermat-Zahl ist jedoch eine Primzahl, wie das Beispiel

$$2^{2^5} + 1 = 641 \cdot 6\,700\,417$$

zeigt.

Satz 3.1-1(v) sagt *nicht*, dass jede Zahl der Form $2^p - 1$ mit einer Primzahl *p* selbst Primzahl ist. Die Zahlen der Form $2^p - 1$ mit einer Primzahl *p* heißen **Mersenne**-Zahlen. Nicht jede Mersenne-Zahl ist Primzahl. Beispielsweise sind die Zahlen

$$2^2 - 1 = 3,\ 2^3 - 1 = 7,\ 2^5 - 1 = 31,\ 2^7 - 1 = 127,\ 2^{13} - 1 = 8191$$

Primzahlen, nicht aber $2^{11} - 1 = 2047 = 23 \cdot 89$. Die bisher bekannten größten Primzahlen sind Mersenne-Zahlen (1998 stand der Rekord bei $2^{3.021.377} - 1$, 2001 bei $2^{13.466.917} - 1$, eine Zahl mit 4.053.946 Dezimalstellen, und im Februar 2005 bei $2^{25.964.951} - 1$, eine Zahl mit 7.816.230 Dezimalstellen).

Einer der wichtigsten Sätze der Zahlentheorie beschreibt die Anzahl der Primzahlen unterhalb einer vorgegebenen Grenze *x*:

Es sei $\pi(x)$ die Anzahl der Primzahlen, die $\le x$ sind, d.h. $\pi(x) = \sum_{\substack{p \in \mathbf{P} \\ p \le x}} 1$.

Mit p_n werde die *n*-te Primzahl bezeichnet: $p_1 = 2,\ p_2 = 3,\ p_3 = 5, \ldots$

Satz 3.1-4:

(i) Es gilt $\lim_{x\to\infty}\frac{\pi(x)\cdot\ln(x)}{x}=1$, d.h. $\pi(x)\sim\frac{x}{\ln(x)}$ (für große x).

(ii) Für $x\geq 67$ ist $\ln(x)-3/2<\frac{x}{\pi(x)}<\ln(x)-1/2$.

(iii) Für $n\geq 20$ ist $n\cdot\left(\ln(n)+\ln(\ln(n))-3/2\right)<p_n<n\cdot\left(\ln(n)+\ln(\ln(n))-1/2\right)$.

Auf der Grundlage dieser Sätze lässt sich ein sehr effizientes Verfahren zur Erzeugung von (großen) Zahlen angeben, die mit beliebig großer Wahrscheinlichkeit Primzahlen sind. Dabei wird in Kauf genommen, dass das Verfahren eine Zahl eventuell als Primzahl einstuft, die keine Primzahl ist. Die Fehlerwahrscheinlichkeit dieser falschen Entscheidung kann jedoch auf einfache Weise beliebig klein gehalten werden. Man spricht hier von einem **probabilistischen Verfahren (nach dem Monte-Carlo-Prinzip)**.

3.2 Modulare Artihmetik

Es sei $n\in\mathbf{N}$ eine natürliche Zahl mit $n\geq 1$. Auf den ganzen Zahlen **Z** wird durch die folgende Festlegung eine Relation $\equiv$ definiert:

Die Zahlen $a\in\mathbf{Z}$ und $b\in\mathbf{Z}$ heißen **kongruent modulo** *n*, geschrieben

$$a\equiv b\pmod n$$

genau dann, wenn gilt: die Zahl *n* teilt $a-b$.

Anders ausgedrückt: $a\equiv b\pmod n$ gilt genau dann, wenn es ein $k\in\mathbf{Z}$ mit $a-b=k\cdot n$ gibt.

Beispielsweise gilt

$21\equiv 0\,(\mathrm{mod}\,7)$, $22\equiv 1\,(\mathrm{mod}\,7)$, $23\equiv 2\,(\mathrm{mod}\,7)$, $24\equiv 3\,(\mathrm{mod}\,7)$, $25\equiv 4\,(\mathrm{mod}\,7)$,
$26\equiv 5\,(\mathrm{mod}\,7)$, $27\equiv 6\,(\mathrm{mod}\,7)$,
$28\equiv 0\,(\mathrm{mod}\,7)$, $28\equiv 21\,(\mathrm{mod}\,7)$,
$29\equiv 1\,(\mathrm{mod}\,7)$, $29\equiv 22\,(\mathrm{mod}\,7)$.

Satz 3.2-1:

Es sei $n \in \mathbf{N}$ eine natürliche Zahl mit $n \geq 1$. Die Relation $\equiv$ ist eine Äquivalenzrelation auf den ganzen Zahlen **Z**, d.h. es gilt:

(i) $a \equiv a \pmod{n}$ für jedes $a \in \mathbf{Z}$

(ii) Aus $a \equiv b \pmod{n}$ folgt $b \equiv a \pmod{n}$

(iii) Aus $a \equiv b \pmod{n}$ und $b \equiv c \pmod{n}$ folgt $a \equiv c \pmod{n}$.

Für $a \in \mathbf{Z}$ bezeichnet $[a]_n = \{ b \mid b \in \mathbf{Z} \text{ und } a \equiv b \pmod{n} \}$ die zu a gehörende **Restklasse (mod n)**.

Beispielsweise ist

$$[3]_7 = \{3, 10, 17, 24, 31, \ldots\} \cup \{-4, -11, -18, -25, \ldots\}$$
$$= \{ m \mid \text{es gibt } k \in \mathbf{Z} \text{ mit } m = k \cdot 7 + 3 \}.$$

Allgemein ist für $a \in \mathbf{Z}$

$$[a]_n = \{ b \mid b \in \mathbf{Z} \text{ und } a \equiv b \pmod{n} \}$$
$$= \{ b \mid \text{es gibt } k \in \mathbf{Z} \text{ mit } m = k \cdot n + a \}.$$

Satz 3.2-2:

Es sei $n \in \mathbf{N}$ eine natürliche Zahl mit $n \geq 1$. Dann gilt:

(i) Es gilt $a \equiv b \pmod{n}$ genau dann, wenn $[a]_n = [b]_n$ ist.

(ii) Jeweils zwei Restklassen $[a]_n$ und $[b]_n$ sind entweder gleich oder disjunkt.

(iii) Es gibt genau n disjunkte Restklassen modulo n, nämlich $[0]_n$, $[1]_n$, $[2]_n$, ..., $[n-1]_n$, und es gilt $\bigcup_{a=0}^{n-1} [a]_n = \mathbf{Z}$.

Jede Restklasse $[a]_n$ besteht aus unendlich vielen Elementen, nämlich aus allen Elementen der Form $k \cdot n + a$ mit $k \in \mathbf{Z}$. Für ein festes $k \in \mathbf{Z}$ sind (wegen Satz 3.2-2 (i)) die Restklassen $[a]_n$ und $[k \cdot n + a]_n$ gleich, d.h. jede Zahl der Form $(k \cdot n + a) \in [a]_n$ repräsentiert die Restklasse $[a]_n$. Man kann daher in jeder Restklasse $[a]_n$ eine Zahl a' mit folgenden Eigenschaften (i) und (ii) finden:

(i) $0 \le a' < n$
(ii) $a' \equiv a \pmod n$, d.h. $[a']_n = [a]_n$.

Für positives $a \in \mathbf{Z}$ erhält man dieses Element a' beispielsweise dadurch, dass man von a so lange den Wert n abzieht, bis die Bedingung $0 \le a' < n$ erfüllt ist. Für negatives $a \in \mathbf{Z}$ wird der Wert n sukzessive addiert. Dieser kleinste Wert a' mit $0 \le a' < n$ heißt **Rest bei der ganzzahligen Division** von a durch n und wird mit

$a \bmod n$

bezeichnet.

Beispielsweise ist wegen $3 = 45 - 7 - 7 - 7 - 7 - 7 - 7$: $45 \bmod 7 = 3$ und $[45]_7 = [3]_7$ und $5 = -16 + 7 + 7 + 7$: $-16 \bmod 7 = 5$ und $[-16]_7 = [5]_7$.

Es gilt also:

$0 \le (a \bmod n) \le n - 1$ und $[(a \bmod n)]_n = [a]_n$.

Für positives $a \in \mathbf{Z}$ ist nach Konstruktion $(a \bmod n) = a - \lfloor a/n \rfloor \cdot n$;
für negatives $a \in \mathbf{Z}$ ist $(a \bmod n) = a + \lfloor a/n \rfloor \cdot n$.

Beispiele:

$(21 \bmod 7) = 0$, $(28 \bmod 7) = 0$,
$(22 \bmod 7) = 1$, $(29 \bmod 7) = 1$,
$(27 \bmod 7) = 6$, $(6 \bmod 7) = 6$, $(-1 \bmod 7) = 6$.

Für zwei Restklassen $[a]_n$ und $[b]_n$ gilt:

Sind $a_1 \in [a]_n$ und $a_2 \in [a]_n$ bzw. $b_1 \in [b]_n$ und $b_2 \in [b]_n$, dann ist
$a_1 + b_1 \equiv a_2 + b_2 \equiv a + b \pmod{n}$, d.h. $[a_1 + b_1]_n = [a_2 + b_2]_n = [a + b]_n$.
Entsprechend gilt $a_1 \cdot b_1 \equiv a_2 \cdot b_2 \equiv a \cdot b \pmod{n}$.

Daher kann man auf eindeutige Weise **arithmetische Operationen auf den Restklassen** (modulo *n*) definieren:

$[a]_n +_n [b]_n = [a + b]_n$ und $[a]_n \cdot_n [b]_n = [a \cdot b]_n$.

Man nimmt also aus jeder Restklasse $[a]_n$ bzw. $[b]_n$ ein beliebiges Element $a_1 \in [a]_n$ bzw. $b_1 \in [b]_n$ und bildet $[a_1 + b_1]_n = [a + b]_n$. Entsprechendes gilt für die Multiplikation. Insbesondere folgt hieraus:

Satz 3.2-3:

(i) $(a \bmod n) + (b \bmod n) \equiv (a + b) \bmod n$,
$[a]_n +_n [b]_n = [(a + b) \bmod n]_n$.

(ii) $(a \bmod n) \cdot (b \bmod n) \equiv (a \cdot b) \bmod n$,
$[a]_n \cdot_n [b]_n = [(a \cdot b) \bmod n]_n$.

(iii) $b \cdot (a \bmod n) \equiv (a \cdot b) \bmod n$.

Beispiele:

$[3]_7 +_7 [6]_7 = [(3 + 6) \bmod 7]_7 = [2]_7$,
$[3]_7 \cdot_7 [5]_7 = [(3 \cdot 5) \bmod 7]_7 = [1]_7$,
$[3]_7 \cdot_{12} [4]_{12} = [(3 \cdot 4) \bmod 12]_{12} = [0]_{12}$.

Mit $\mathbf{Z}/n\mathbf{Z}$ wird die Menge $\{[0]_n, [1]_n, ..., [n-1]_n\}$ bezeichnet. Häufig findet man auch die Bezeichnung $\mathbf{Z}/n\mathbf{Z} = \{0, 1, ..., n-1\}$ und meint damit die Restklassen modulo *n*. Zusammen mit den oben definierten artihmetischen Operationen auf Restklassen weist die endliche Menge $\mathbf{Z}/n\mathbf{Z}$ sehr ähnliche Eigenschaften zu der unendlichen Menge $\mathbf{Z}$ auf:

Satz 3.2-4:

$(\mathbf{Z}/n\mathbf{Z}, +_n, \cdot_n)$ bildet einen kommutativen Ring mit 1. Das neutrale Element der Addition ist $[0]_n = \{ a \mid a = k \cdot n \text{ mit } k \in \mathbf{Z} \}$, das neutrale Element der Multiplikation ist $[1]_n = \{ a \mid a = k \cdot n + 1 \text{ mit } k \in \mathbf{Z} \}$. Das bezüglich der Addition $+_n$ inverse Element zur Restklasse $[a]_n$ ist die Restklasse $-[a]_n = [-a]_n = [n-a]_n$.

Eine Restklasse $[a]_n$ besitzt bezüglich der Multiplikation $\cdot_n$ genau dann ein inverses Element $[a]_n^{-1}$, wenn $ggT(a,n) = 1$ ist (zur Definition von $ggT(a,n)$ und zur Bestimmung der inversen Restklasse in diesem Fall siehe Kapitel 3.3).

Beispiele:

Die Additions- und Multiplikationstabellen von

$(\mathbf{Z}/7\mathbf{Z}, +_7, \cdot_7) = (\{ [0]_7, [1]_7, [2]_7, [3]_7, [4]_7, [5]_7, [6]_7 \}, +_7, \cdot_7)$

lauten (statt $[a]_7$ wird zur Vereinfachung *a* geschrieben):

$+_7$	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

$\cdot_7$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	1	3	5
3	3	6	2	5	1	4
4	4	1	5	2	6	3
5	5	3	1	6	4	2
6	6	5	4	3	2	1

In $(\mathbf{Z}/7\mathbf{Z}, +_7, \cdot_7)$ ist das inverse Element bezüglich der Addition zum Element $[3]_7$ das Element $-[3]_7 = [4]_7$ und das inverse Element bezüglich der Multiplikation zum Element $[3]_7$ das Element $[3]_7^{-1} = [5]_7$.

Die Additions- und Multiplikationstabellen von $(\mathbf{Z}/12\mathbf{Z}, +_{12}, \cdot_{12})$ lauten (statt $[a]_{12}$ wird zur Vereinfachung wieder *a* geschrieben):

$+_{12}$	0	1	2	3	4	5	6	7	8	9	10	11
0	0	1	2	3	4	5	6	7	8	9	10	11
1	1	2	3	4	5	6	7	8	9	10	11	0
2	2	3	4	5	6	7	8	9	10	11	0	1
3	3	4	5	6	7	8	9	10	11	0	1	2
4	4	5	6	7	8	9	10	11	0	1	2	3
5	5	6	7	8	9	10	11	0	1	2	3	4
6	6	7	8	9	10	11	0	1	2	3	4	5
7	7	8	9	10	11	0	1	2	3	4	5	6
8	8	9	10	11	0	1	2	3	4	5	6	7
9	9	10	11	0	1	2	3	4	5	6	7	8
10	10	11	0	1	2	3	4	5	6	7	8	9
11	11	0	1	2	3	4	5	6	7	8	9	10

$\cdot_{12}$	1	2	3	4	5	6	7	8	9	10	11
1	1	2	3	4	5	6	7	8	9	10	11
2	2	4	6	8	10	0	2	4	6	8	10
3	3	6	9	0	3	6	9	0	3	6	9
4	4	8	0	4	8	0	4	8	0	4	8
5	5	10	3	8	1	6	11	4	9	2	7
6	6	0	6	0	6	0	6	0	6	0	6
7	7	2	9	4	11	6	1	8	3	10	5
8	8	4	0	8	4	0	8	4	0	8	4
9	9	6	3	0	9	6	3	0	9	6	3
10	10	8	6	4	2	0	10	8	6	4	2
11	11	10	9	8	7	6	5	4	3	2	1

In $(\mathbf{Z}/12\mathbf{Z}, +_{12}, \cdot_{12})$ hat das Element $[3]_{12}$ wegen $[3]_7 \cdot_{12} [4]_{12} = [(3 \cdot 4) \bmod 12]_{12} = [0]_{12}$ kein inverses Element bezüglich der Multiplikation.

3.3 Der Euklidische Algorithmus

Es seien $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$. Besitzt $d \in \mathbf{Z}$ die Eigenschaften $d \mid a$ und $d \mid b$, dann heißt d **gemeinsamer Teiler** von a und b. Besitzt jeder gemeinsame Teiler c von a und b die Eigenschaft $c \mid d$, dann heißt d **größter gemeinsamer Teiler** von a und b und wird mit $ggT(a,b)$ bezeichnet.

Zur Bestimmung des größten gemeinsamen Teilers zweier Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ könnte man diese gemäß Satz 3.1-2 in ihre Primfaktoren zerlegen und alle gemeinsamen Primfaktoren in ihrer gemeinsamen Vielfachheit herausziehen. Beispielsweise ist $792 = 2^3 \cdot 3^2 \cdot 11$ und $84 = 2^2 \cdot 3 \cdot 7$, d.h. $ggT(792,84) = 2^2 \cdot 3 = 12$. Dieses Verfahren (Schulmethode) ist höchstens für kleine Zahlen praktisch einsetzbar; denn für große Zahlen (in der Praxis mit mehr als 100 Dezimalstellen) stößt man auf Effizienzgrenzen. Als äußerst effizient zur Bestimmung des größten gemeinsamen Teilers zweier ganzer Zahlen hat sich der **Euklidische Algorithmus** erwiesen (Euklid, um 325 v. Chr.). Dieses Verfahren geht läuft folgendermaßen ab:

Für die beiden Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ kann $a \geq b$ angenommen werden. Da es bei der Bestimmung von Teilern nicht auf das Vorzeichen ankommt, kann weiterhin $b > 0$ angenommen werden, so dass insgesamt $a \geq b > 0$ ist. Es werden ganze Zahlen m_1 und r_1 bestimmt mit

$a = m_1 \cdot b + r_1$ und $0 \leq r_1 < b$.

Durch die Festlegung $0 \leq r_1 < b$ sind m_1 und r_1 eindeutig bestimmt: $r_1 = a \bmod b$ und $m_1 = \lfloor a/b \rfloor$. Für $r_1 = 0$ endet das Verfahren hier, und es ist $ggT(a,b) = b$. Ansonsten werden ganze Zahlen m_2 und r_2 bestimmt mit

$b = m_2 \cdot r_1 + r_2$ und $0 \leq r_2 < r_1$.

Man sieht, dass b die Rolle von a und r_1 die Rolle von b übernimmt. Wieder sind durch die Festlegung $0 \leq r_2 < r_1$ die Werte m_2 und r_2 eindeutig bestimmt. Für $r_2 = 0$ endet das Verfahren hier, und es ist $ggT(a,b) = r_1$ (das muss man natürlich mathematisch beweisen). Ansonsten werden ganze Zahlen m_3 und r_3 bestimmt mit

$r_1 = m_3 \cdot r_2 + r_3$ und $0 \leq r_3 < r_2$.

Man sieht, dass r_1 die Rolle von b und r_2 die Rolle von r_1 übernimmt. Das Verfahren wird so lange fortgesetzt, bis zum ersten Mal der Rest $r_n = 0$ entsteht. Insgesamt lassen sich die einzelnen Schritte wie folgt zusammenfassen:

Man bestimmt ganze Zahlen m_1 und r_1 mit	$a = m_1 \cdot b + r_1$ und $0 < r_1 < b$.
Man bestimmt ganze Zahlen m_2 und r_2 mit	$b = m_2 \cdot r_1 + r_2$ und $0 < r_2 < r_1$.
Man bestimmt ganze Zahlen m_3 und r_3 mit	$r_1 = m_3 \cdot r_2 + r_3$ und $0 < r_3 < r_2$.
usw.	
Man bestimmt ganze Zahlen m_n und r_n mit	$r_{n-2} = m_n \cdot r_{n-1} + r_n$ und $0 < r_n < r_{n-1}$.
Fortsetzung des Verfahrens, bis	$r_{n-1} = m_{n+1} \cdot r_n + 0$ gilt.

Es gilt $b > r_1 > r_2 > ... > r_{n-1} > r_n > 0$, d.h. die Reste $r_1, r_2, ..., r_n$ werden immer kleiner, so dass das Verfahren abbricht.

Satz 3.3-1:

Das beschriebene Verfahren bestimmt den größten gemeinsamen Teiler zweier ganzer Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ mit $b \neq 0$, und zwar gilt

$ggT(a,b) = r_n$,

d.h. der größte gemeinsame Teiler von *a* und *b* ist der letzte von 0 verschiedene Rest.

Die folgende PASCAL-Funktion `ggT` ist eine Implementierung des Verfahrens; sie bestimmt den größten gemeinsamen Teiler der als Parameter übergebenen ganzen Zahlen *a* und *b*. Die Anzahl der von ihm ausgeführten arithmetischen Operationen ist proportional zur Länge der Zahlendarstellung von *a* und *b*.

```
FUNCTION ggT (a : INTEGER; b : INTEGER) : INTEGER;

VAR   r : INTEGER;
      s : INTEGER;
      t : INTEGER;
      m : INTEGER;

BEGIN { ggT }
   r := b;
   s := a;

   WHILE r <> 0 DO
     BEGIN
       { t und s aus der vorherigen Iteration neu besetzen }
       t := s;
       s := r;
       { bilde t =  m * s + r }
       m := t DIV s;
       r := t - m*s
     END;

   { der größte gemeinsame Teiler ist der letzte von 0
     verschiedene Rest                                  }

   ggT := s

END   { ggT };
```

Beispiele:

a: 28 *b*: 15					*a*: 198 *b*: 84					*a*: 84 *b*: 198				
t	=	m *	s +	r	t	=	m *	s +	r	t	=	m *	s +	r
28	=	1 *	15 +	13	198	=	2 *	84 +	30	84	=	0 *	198 +	84
15	=	1 *	13 +	2	84	=	2 *	30 +	24	198	=	2 *	84 +	30
13	=	6 *	2 +	1	30	=	1 *	24 +	6	84	=	2 *	30 +	24
2	=	2 *	1 +	0	24	=	4 *	6 +	0	30	=	1 *	24 +	6
										24	=	4 *	6 +	0
ggT (28, 15) =				1	ggT (198, 84) =				6					
										ggT (84, 198) =				6

Das obige Zahlenschema (eine typische Zeile *i* ist hinzugefügt)

$a = m_1 \cdot b + r_1$ und $0 < r_1 < b$, Zeile 1

$b = m_2 \cdot r_1 + r_2$ und $0 < r_2 < r_1$, Zeile 2

$r_1 = m_3 \cdot r_2 + r_3$ und $0 < r_3 < r_2$, Zeile 3

...

$r_{i-2} = m_i \cdot r_{i-1} + r_i$ und $0 < r_i < r_{i-1}$, Zeile *i*

...

$r_{n-2} = m_n \cdot r_{n-1} + r_n$ und $0 < r_n < r_{n-1}$, Zeile *n*

$r_{n-1} = m_{n+1} \cdot r_n + 0$, Zeile $n+1$

$ggT(a,b) = r_n$

mit $r_1 = a \bmod b$ ergibt unmittelbar den

Satz 3.3-2:

Für zwei Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ ist $ggT(a,b) = \begin{cases} a & \text{für } b = 0 \\ ggT(b, a \bmod b) & \text{für } b \neq 0 \end{cases}$.

Löst man in dem Zahlenschema die Zeilen *i* für $i = 1, ..., n$ nach r_i auf, so lassen sich ganze Zahlen $a_1, ..., a_n$ und $b_1, ..., b_n$ definieren, für die gilt:

Zeile 1: $r_1 = a - m_1 \cdot b$
$= 1 \cdot a + (-m_1) \cdot b$
$= a_1 \cdot a + b_1 \cdot b$ mit $a_1 = 1$, $b_1 = -m_1$.

Zeile 2: $r_2 = b - m_2 \cdot r_1$
$= b - m_2 \cdot (a - m_1 \cdot b)$
$= -m_2 \cdot a + (1 + m_1 \cdot m_2) \cdot b$
$= a_2 \cdot a + b_2 \cdot b$ mit $a_2 = -m_2$, $b_2 = 1 + m_1 \cdot m_2$.

Angenommen, in allen Zeilen $l = 1, ..., i-1$ ließe sich der jeweilige Rest r_l in der Form

$r_l = a_l \cdot a + b_l \cdot b$

schreiben. Dann geht das auch in Zeile *i*:

Zeile *i*: $r_i = r_{i-2} - m_i \cdot r_{i-1}$
$= a_{i-2} \cdot a + b_{i-2} \cdot b - m_i \cdot (a_{i-1} \cdot a + b_{i-1} \cdot b)$
$= (a_{i-2} - m_i \cdot a_{i-1}) \cdot a + (b_{i-2} - m_i \cdot b_{i-1}) \cdot b$
$= a_i \cdot a + b_i \cdot b$ mit $a_i = a_{i-2} - m_i \cdot a_{i-1}$, $b_i = b_{i-2} - m_i \cdot b_{i-1}$.

Insbesondere

Zeile *n*: $ggT(a,b) = r_n = a_n \cdot a + b_n \cdot b$.

Die Folgen $a_1, ..., a_n$ und $b_1, ..., b_n$ werden also rekursiv definiert durch:

$a_{-1} = 1$, $a_0 = 0$, $a_i = a_{i-2} - m_i \cdot a_{i-1}$ für $i = 1, ..., n$
$b_{-1} = 0$, $b_0 = 1$, $b_i = b_{i-2} - m_i \cdot b_{i-1}$ für $i = 1, ..., n$.

Die Berechnung dieser beiden Folgen kann in den Euklidischen Algorithmus direkt eingebaut werden. Die PASCAL-Funktion `ggt` wird erweitert zur Funktion `invers` (die Wahl des Prozedurbezeichners ergibt sich aus den anschließenden Bemerkungen zu Satz 3.3-5).

```
PROCEDURE  invers (a : LONGINT; b : LONGINT;
                   VAR a_inv : LONGINT;
                   VAR b_inv : LONGINT;
                   VAR ggt   : LONGINT);

{ die Funktion berechnet zu a und b ganze
  Zahlen a_inv und b_inv mit a*a_inv + b*b_inv = ggT(a, b) }

VAR    r        : LONGINT;
       s        : LONGINT;
       t        : LONGINT;
       m        : LONGINT;
       a_min_2  : LONGINT;
       a_min_1  : LONGINT;
       b_min_2  : LONGINT;
       b_min_1  : LONGINT;
       store    : LONGINT;

BEGIN
  r       := b;
  s       := a;
  a_min_2 := 1;
  a_min_1 := 0;
  b_min_2 := 0;
  b_min_1 := 1;

WHILE r <> 0 DO
  BEGIN
    { t und s aus der vorigen Iteration neu besetzen }
    t := s;
    s := r;

    { bilde t = m * s + r }
    m := t DIV s;
    r := t - m*s;

    store   := a_min_2;
    a_min_2 := a_min_1;
    a_min_1 := store - m * a_min_1;
    store   := b_min_2;
    b_min_2 := b_min_1;
    b_min_1 := store - m * b_min_1
   END;

{ der ggT (a, m) ist der letzte von 0 verschiedene Rest,
   d. h. der gegenwärtige Wert von s                    }
   ggT := s;
   a_inv := a_min_2;
   b_inv := b_min_2

END { invers };
```

Satz 3.3-3:

Zu zwei Zahlen $a \in \mathbf{Z}$ und $b \in \mathbf{Z}$ gibt es eindeutig bestimmte Zahlen $a' \in \mathbf{Z}$ und $b' \in \mathbf{Z}$ mit

$a \cdot a' + b \cdot b' = ggT(a,b)$.

Die Zahlen a' und b' lassen sich mit der PASCAL-Funktion `invers`, einer Erweiterung des Euklidischen Algorithmus, bestimmen.

Der folgende Satz stellt einige wichtige Eigenschaften des *ggT* zusammen:

Satz 3.3-4:

(i) Es sei $d = ggT(a,b)$. Dann gibt es Zahlen $a_1 \in \mathbf{Z}$ und $b_1 \in \mathbf{Z}$ mit $a = d \cdot a_1$ und $b = d \cdot b_1$ und $ggT(a_1,b_1) = 1$.

(ii) Es gilt $ggT(a,b) = 1$ genau dann, wenn es Zahlen $x \in \mathbf{Z}$ und $y \in \mathbf{Z}$ gibt mit $a \cdot x + b \cdot y = 1$.

(iii) Es sei $ggT(a,b) = 1$. Falls *a* das Produkt $b \cdot c$ teilt, dann teilt *a* die Zahl *c*.

(iv) Es sei $ggT(a,b) = 1$. Falls *a* die Zahl *c* teilt und *b* die Zahl *c* teilt, dann teilt $a \cdot b$ die Zahl *c*.

In vielen Anwendungen spielen **lineare Kongruenzen** eine wichtige Rolle. Dabei handelt es sich um Gleichungen der Form $a \cdot x \equiv b \pmod{n}$, wobei *a* und *b* vorgegebene ganze Zahlen sind und $n > 1$ eine natürliche Zahl ist. Gesucht wird nach einer ganzzahligen Lösung *x*. Die folgenden Sätze sagen aus, wann eine lineare Kongruenz lösbar ist. In diesem Fall lassen sich die Lösungen mit Hilfe der angegebenen Prozedur `invers`, d.h. im wesentlichen mit Hilfe des Euklidischen Algorithmus bestimmen.

Satz 3.3-5:

Es sei $ggT(a,n)=1$.

Dann hat die lineare Kongruenz $a \cdot x \equiv b \pmod{n}$ eine Lösung. Alle Lösungen sind kongruent modulo *n*. Man sagt daher, dass die lineare Kongruenz $a \cdot x \equiv b \pmod{n}$ in diesem Fall genau eine Lösung modulo *n* besitzt.

Nach Satz 3.3-3 lassen sich zu *a* und *n* mit der Prozedur `invers` Zahlen a' und n' finden, für die $a \cdot a' + n \cdot n' = ggT(a,n) = 1$ gilt. Die gesuchte Lösung lautet dann $x = a' \cdot b \pmod{n}$. Diese Lösung ist modulo *n* eindeutig.

In Satz 3.2-4 wird behauptet, dass eine Restklasse $[a]_n$ genau dann ein bezüglich der Multiplikation $\cdot_n$ inverses Element $[a]_n^{-1}$ besitzt, wenn $ggT(a,n)=1$ gilt. Dazu bestimmt man wie oben die Zahlen a' und n' mit $a \cdot a' + n \cdot n' = ggT(a,n) = 1$. Wegen $a \cdot a' \equiv 1 \pmod{n}$ gilt $[a \cdot a']_n = [a]_n \cdot_n [a']_n = [1]_n$. Daher kann man $[a]_n^{-1} = [a']_n = [a' \bmod n]_n$ setzen.

Eine Verallgemeinerung von Satz 3.3-5 ist der folgende Satz.

Satz 3.3-6:

Es sei $ggT(a,n)=d$. Dann hat die lineare Kongruenz $a \cdot x \equiv b \pmod{n}$ genau dann Lösungen, wenn *d* ein Teiler von *b* ist.

Ist *d* ein Teiler von *b*, so gilt $b = d \cdot b_1$ mit einer ganzen Zahl b_1. Alle Lösungen der linearen Kongruenz $a \cdot x \equiv b \pmod{n}$ erhält man folgendermaßen:

Nach Satz 3.3-4 (i) gibt es Zahlen $a_1 \in \mathbf{Z}$ und $n_1 \in \mathbf{Z}$ mit $a = d \cdot a_1$ und $n = d \cdot n_1$ und $ggT(a_1,n_1)=1$. Nach Satz 3.3-5 wird die modulo n_1 eindeutige Lösung *y* der linearen Kongruenz $a_1 \cdot y \equiv b_1 \pmod{n_1}$ bestimmt. Alle Lösungen (es sind genau *d* viele) der linearen Kongruenz $a \cdot x \equiv b \pmod{n}$ lauten dann:

$y,\ y+n_1,\ y+2 \cdot n_1,\ ...,\ y+(d-1) \cdot n_1$.

3.4 Weitere ausgewählte Ergebnisse der elementaren Zahlentheorie

Die **Eulersche ϕ-Funktion (phi-Funktion)** wird für jede natürliche Zahl $n \geq 1$ definiert durch die Anzahl der natürlichen Zahlen a zwischen 1 und n (einschließlich) mit $ggT(a,n) = 1$, d.h.

$$\phi(n) = \sum_{\substack{a \\ 1 \leq a \leq n, \\ ggT(a,n)=1}} 1 \ .$$

Satz 3.4-1:

(i) Ist p eine Primzahl, dann ist $\phi(p) = p - 1$. Gilt umgekehrt $\phi(n) = n - 1$, dann ist n eine Primzahl.

(ii) Ist p eine Primzahl, dann ist $\phi(p^k) = p^k - p^{k-1}$.

(iii) Für natürliche Zahlen n und m mit $ggT(n,m) = 1$ ist $\phi(n \cdot m) = \phi(n) \cdot \phi(m)$.

(iv) $\phi(n) = n \cdot \prod_{p \text{ teilt } n} \left(1 - 1/p\right)$.

Der folgende Satz (Satz von Euler) ist wichtig für die Korrektheit des Public Key Encryption-Verfahrens RSA:

Satz 3.4-2:

Es seien a und n natürliche Zahlen mit $ggT(a,n) = 1$. Dann ist $a^{\phi(n)} \equiv 1 \pmod{n}$.

Der folgende Satz (Satz von Fermat) ist ein Spezialfall von Satz 3.4-2:

Satz 3.4-3:

(i) Es sei a eine natürliche Zahl und p eine Primzahl mit $ggT(a,p)=1$. Dann ist $a^{p-1} \equiv 1 \pmod{p}$.

(ii) Es sei a eine natürliche Zahl und n eine ungerade natürliche Zahl mit $ggT(a,n)=1$. Gilt *nicht* $a^{n-1} \equiv 1 \pmod{n}$, dann ist n keine Primzahl.

Mit Satz 3.3-5 wurde das zu einer Restklasse $[a]_n$ bezüglich der Multiplikation $\cdot_n$ inverse Element $[a]_n^{-1}$ bestimmt, falls $ggT(a,n)=1$ gilt. Nach Satz 3.4-2 gilt (bei $ggT(a,n)=1$): $a \cdot a^{\phi(n)-1} = a^{\phi(n)} \equiv 1 \pmod{n}$. Daher ist $[a]_n^{-1} = \left[a^{\phi(n)-1}\right]_n$. Dieses Ergebnis führt (in Erweiterung von Satz 3.3-5) auf

Satz 3.4-4:

Es sei $ggT(a,n)=1$. Dann hat die lineare Kongruenz $a \cdot x \equiv b \pmod{n}$ genau eine Lösung modulo n, nämlich $x = \left[b \cdot a^{\phi(n)-1}\right]_n$, d.h. für alle Lösungen x der linearen Kongruenz $a \cdot x \equiv b \pmod{n}$ gilt $x \equiv b \cdot a^{\phi(n)-1} \pmod{n}$.

3.5 Anwendung in der Kryptologie

Die folgende Abbildung zeigt das grundlegende Szenario, in dem kryptographische Verfahren zur Datenverschlüsselung und –entschlüsselung eingesetzt werden.

Vertrauliche Daten werden von einem Sender *A* zu einem Empfänger *B* gesandt. Fragen der korrekten Datenübertragung sollen in diesem Zusammenhang ausgeklammert werden. Es soll lediglich garantiert werden, dass ein unberechtigter Dritter, der die Daten während der Übertragungsphase eventuell mithört, diese inhaltlich nicht interpretieren kann. Dieser „Angreifer“ auf das Übertragungssystem wird als **Kryptoanalytiker** bezeichnet; seine Tätigkeit heißt **Kryptoanalyse**. Zum Schutz werden die Daten vor ihrer Übertragung vom Sender verschlüsselt. Die unverschlüsselten Daten werden als **Klartext** bezeichnet, die verschlüsselten Daten als **Schlüsseltext (Chiffretext)**. Der Schlüsseltext wird zum Empfänger gesendet und dort von diesem entschlüsselt, so dass er wieder den Klartext erhält. Zwischen Sender und Emp-

fänger sind also **Absprachen** über das verwendete Verschlüsselungs- und Entschlüsselungsverfahren notwendig.

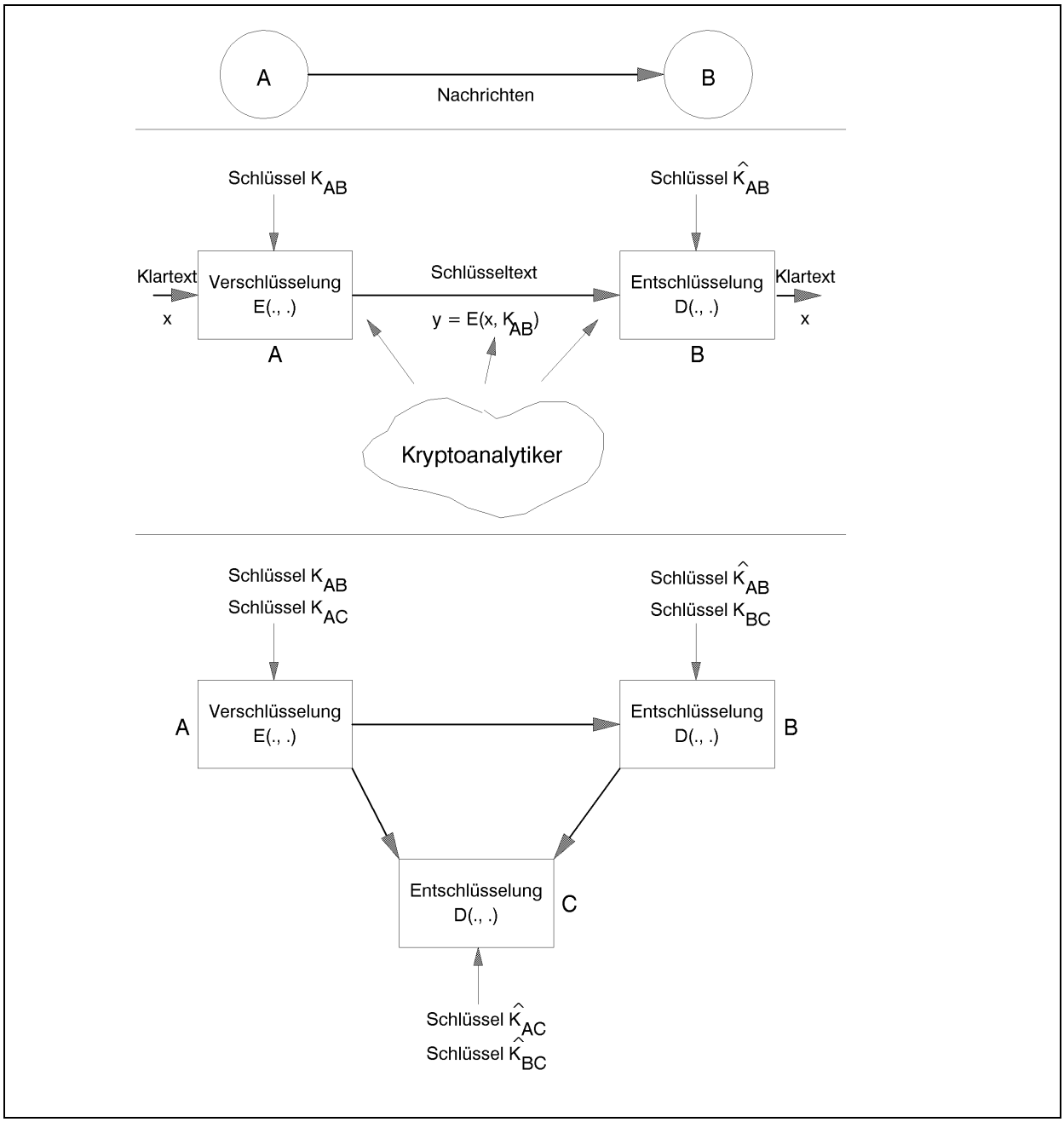


Abbildung: Datenverschlüsselung und -entschlüsselung

Die im folgenden beschriebenen Verfahren zur Verschlüsselung und Entschlüsselung von Daten zwischen einem Sender *A* und einem Empfänger *B* bestehen formal aus mehreren Teilen:

1. Mit dem **Verschlüsselungsalgorithmus** *E* (encryption, Verschlüsselung) verschlüsselt der Sender *A* Klartexte. Der Verschlüsselungsalgorithmus *E* hat zwei Eingabeparameter, nämlich einen Klartext *x* und einen **Schlüssel** (key) K_{AB}, mit dem **alle Klartexte**, die **von *A* nach *B* laufen**, verschlüsselt werden. Für eine Kommunikationsbeziehung von *A*

an einen Empfänger *C* mit $C \neq B$ wird ein Schlüssel $K_{AC} \neq K_{AB}$, aber derselbe Verschlüsselungsalgorithmus *E* verwendet.

Der aus einem Klartext *x* entstehende Schlüsseltext ist $y = E(x, K_{AB})$.

2. Beim Empfänger *B* ankommende Schlüsseltexte werden von ihm mit Hilfe des **Entschlüsselungsalgorithmus** *D* (decryption, Entschlüsselung) entschlüsselt. Der Entschlüsselungsalgorithmus *D* hat ebenfalls zwei Eingabeparameter, nämlich einen Schlüsseltext *y* und einen Schlüssel $\hat{K}_{AB}$, mit dem alle Nachrichten, die von *A* nach *B* laufen, entschlüsselt werden. Zwischen den Algorithmen *E* und *D* und den Schlüsseln K_{AB} und $\hat{K}_{AB}$ besteht die Beziehung

   $$D\left(E\left(x, K_{AB}\right), \hat{K}_{AB}\right) = x,$$

   d.h. der gesendete Klartext kann aus dem empfangenen Schlüsseltext bei korrekter Verwendung der Verfahren wieder gewonnen werden. Der Sender *A* muss den vom Empfänger *B* eingesetzten Schlüssel $\hat{K}_{AB}$ zum Entschlüsseln eines Schlüsseltextes nicht notwendigerweise kennen.

3. Das Schlüsselpaar $\left(K_{AB}, \hat{K}_{AB}\right)$ wird für die Ver- bzw. Entschlüsselung aller Klartexte verwendet, die von *A* nach *B* laufen. Für die umgekehrte Kommunikationsrichtung ist eventuell ein anderes Schlüsselpaar erforderlich ebenso für die Kommunikation zwischen anderen Teilnehmern.

Einige **grundlegende Anforderungen an ein kryptographisches Verfahren** sind:

(i) Die Berechnung von $y = E\left(x, K_{AB}\right)$ aus *x* und K_{AB} (Verschlüsselung) muss vom Sender auf einfache Weise, d.h. mit geringem algorithmischen Aufwand, durchführbar sein. Außerdem sollte der Schlüsseltext $y = E\left(x, K_{AB}\right)$ nicht wesentlich länger als der zugehörige Klartext *x* sein. Natürlich wird dabei vorausgesetzt, dass der Sender über geeignete Rechenkapazität verfügt.

(ii) Die Berechnung von *x* aus einer empfangenen Nachricht der Form $y = E\left(x, K_{AB}\right)$ mit Hilfe von $\hat{K}_{AB}$ (Entschlüsselung) muss vom Empfänger ebenfalls auf einfache Weise, d.h. mit geringem algorithmischen Aufwand, durchführbar sein. Auch hier wird vorausgesetzt, dass der Empfänger über geeignete Rechenkapazität verfügt.

(iii) Ohne Kenntnis des Schlüssels $\hat{K}_{AB}$ zum Entschlüsseln ist es „unmöglich“, aus $y = E\left(x, K_{AB}\right)$ auf den Klartext *x* zu schließen. Systematisches Probieren aller Werte,

die als eventuelle Schlüssel $\hat{K}_{AB}$ in Frage kommen, ist mit einem derart großen algorithmischen Aufwand verbunden, dass diese experimentelle Suche praktisch nicht durchführbar ist.

(iv) Die Verschlüsselungs- bzw. Entschlüsselungsalgorithmen *E* bzw. *D* sollten nicht geheim gehalten werden. Abgesehen davon, dass eine Geheimhaltung wahrscheinlich nur temporär möglich ist, wird durch eine Offenlegung von *E* und *D* erreicht, dass die Verfahren mathematisch analysiert und eventuelle Schwachstellen aufgedeckt und behoben werden können. Zusätzlich lässt sich eine korrekte Implementierung der Verfahren verifizieren.

Die **Angriffe** durch einen unbefugten Kryptoanalytiker **auf ein Verschlüsselungsverfahren** lassen sich in verschiedene Gruppen einteilen:

- Der Kryptoanalytiker versucht, aus der Kenntnis von $y = E(x, K_{AB})$ den Klartext *x* zu erhalten. Man nennt diesen Angriff **Cipher-text-only-Attacke**. Diese Form der Attacke ist die schwierigste, denn im Normalfall hat man wenig Informationen darüber, welchen Inhalt der Klartext *x* aufweist. Gleichzeitig ist sie aber auch diejenige, die in der Praxis am häufigsten vorkommt. Ein anderes Ziel eines Kryptoanalytikers bei einer Cipher-text-only-Attacke ist die Ermittlung des Schlüssels $\hat{K}_{AB}$ zum Entschlüsseln aus der Kenntnis einer oder mehrerer verschlüsselter Nachrichten. Damit können dann spätere verschlüsselte Texte entschlüsselt werden.

- Der Kryptoanalytiker kennt eine von ihm nicht beeinflußte Auswahl von Klartexten x_1, ..., x_n mit den zugehörigen Schlüsseltexten $E(x_1, K_{AB})$, ..., $E(x_n, K_{AB})$ und versucht daraus, das Schlüsselpaar $(K_{AB}, \hat{K}_{AB})$ abzuleiten. Man nennt diesen Angriff **Known-plaintext-Attacke**. Eine derartige Attacke ist häufig dann möglich, wenn sich Nachrichten oder Teile davon wiederholen. Wenn Klartexte beispielsweise immer denselben Briefkopf oder dieselbe Anrede verwenden, sind zumindest Teile eines Klartextes bekannt.

- Der Kryptoanalytiker kann selbst eine Auswahl von Klartexten x_1, ..., x_n vorschlagen und sieht die zugehörigen Schlüsseltexte $E(x_1, K_{AB})$, ..., $E(x_n, K_{AB})$. Er wählt die Klartexte so, dass er daraus eventuell leicht auf das verwendete Schlüsselpaar $(K_{AB}, \hat{K}_{AB})$ schließen kann. Man nennt diesen Angriff **Chosen-plaintext-Attacke**. Ein gutes kryptographisches Verfahren muss gegen Chosen-plaintext-Attacke resistent sein.

- Der Kryptoanalytiker kennt das Verschlüsselungsverfahren *E* und das Entschlüsselungsverfahren *D* einschließlich des verwendeten Schlüssels K_{AB} zum Verschlüsseln eines Klartextes (eine typische Situation der Public-Key-Encryption-Verfahren). Er ver-

fügt über viel Zeit und Rechnerleistung, um aus dieser Kenntnis den Schlüssel $\hat{K}_{AB}$ zu ermitteln.

Bei einer Chosen-Plaintext-Attacke kann man versuchen, systematisch alle möglichen Schlüssel $\hat{K}_{AB}$ auszuprobieren (ein in der Praxis durchaus gängiger Ansatz). Dabei hofft man natürlich, schon nach wenigen Versuchen auf den richtigen Schlüssel zu stoßen. Eine derartige Attacke heißt **Brute–Force–Attacke**. Man muss sich jedoch darüber im klaren sein, dass die Anzahl auszuprobierender Schlüssel exponentiell wächst. Geht man davon aus, dass der Schlüssel $\hat{K}_{AB}$ eine Binärzahl der Länge *n* ist, so gibt es 2^n viele Kandidaten für $\hat{K}_{AB}$. Um eine Vorstellung von der Größenordnung dieser Zahl zu bekommen, wird angenommen, dass die Erzeugung und das Ausprobieren eines einzigen Schlüssels nur 10^{-9} Sekunden benötigt. Dann dauert eine Brute–Force–Attacke bei einer Schlüssellänge von 56 Bits (eine heute nicht mehr als sicher angesehene Schlüssellänge), d.h. das Durchprobieren sämtlicher $2^{56} \approx 7{,}20576 \cdot 10^{16}$ verschiedener Schlüssel, insgesamt mehr als 8,34 Tage benötigt. Bei einer Schlüssellänge von 64 Bits braucht man dann bereits etwa 584 Jahre, um alle Schlüssel zu erzeugen. Nimmt man an, dass bei einer Schlüssellänge von 56 Bits alle Schlüssel in nur 1 Sekunde ausprobiert werden können, dann ergeben sich die folgenden Werte:

Schlüssellänge *n*	Anzahl an Schlüsseln	Aufwand zur Erzeugung aller 2^n Schlüssel
56 Bits	$7{,}20576 \cdot 10^{16}$	1 Sekunde (angenommen)
64 Bits	$1{,}84467 \cdot 10^{19}$	4 Minuten 16 Sekunden
80 Bits	$1{,}20893 \cdot 10^{24}$	194 Tage
112 Bits	$5{,}19230 \cdot 10^{33}$	$\approx 2{,}285 \cdot 10^{9}$ Jahre
128 Bits	$3{,}40282 \cdot 10^{38}$	$\approx 1{,}497 \cdot 10^{14}$ Jahre
192 Bits	$6{,}27710 \cdot 10^{57}$	$\approx 2{,}7623 \cdot 10^{33}$ Jahre
256 Bits	$1{,}15792 \cdot 10^{77}$	$\approx 5{,}0956 \cdot 10^{52}$ Jahre

Nimmt man an, dass die Erzeugung eines Schlüssels in einem Rechner die Zeit *t* benötigt, so beträgt der Aufwand zur Erzeugung aller Schlüssel der Länge *n* die Zeit $t \cdot 2^n$. Die minimale Zeit für einen Schaltvorgang in einem Rechner beträgt aus physikalischen Gründen (u.a. weil sich Elektronen mit einer Geschwindigkeit bewegen, die die Lichtgeschwindigkeit nicht überschreitet) mindestens $t_m \approx 5{,}6 \cdot 10^{-33}$ Sekunden. Setzt man für *t* diesen Wert ein, so beträgt die Dauer in einer Brute-Force-Attacke bei einer Schlüssellänge von 128 Bits allein zur Erzeugung aller 2^{128} Schlüssel immer noch mehr als 22 Tage. Das zeigt, dass eine Brute-Force-Attacke auf die Güte der Verschlüsselung nur unter massiver Parallelisierung sinnvoll ist, indem die Menge aller zu probierender Schlüssel auf eine Vielzahl gleichzeitig agierender Kryptoanalytiker aufgeteilt wird.

Bei den **symmetrischen Kryptologieverfahren** werden für jede Kommunikationsbeziehung zwischen einem Sender *A* und einem Empfänger *B* zum Ver- und Entschlüsseln dieselben Schlüssel verwendet, d.h. es gilt $K_{AB} = \hat{K}_{AB}$. Der Sender verwendet den Schlüssel, um die Nachricht zu verschlüsseln und der Empfänger, um diese zu entschlüsseln. Folglich muss sowohl der Sender als auch der Empfänger denselben Schlüssel K_{AB} kennen und gegenüber Dritten, auch anderen Kommunikationsteilnehmern, geheimhalten. Aus diesem Grund bietet sich an, den Schlüssel K_{AB} auch für die Kommunikationsrichtung von *B* nach *A* zu verwenden. Es gilt dann $K_{AB} = \hat{K}_{AB} = K_{BA} = \hat{K}_{BA}$.

Bei den **asymmetrischen Kryptologieverfahren** werden verschiedene Schlüssel zum Verschlüsseln und Entschlüsseln der Nachrichten verwendet. Eine für die Praxis bedeutende Klasse asymmetrischer Verschlüsselungsverfahren bilden die **öffentlichen Verschlüsselungsverfahren (PKE-Verfahren, public key encryption)**. Zunächst soll das allgemeine Prinzip eines PKE-Verfahrens am Nachrichtenaustausch zwischen einem Sender *A* und einem Empfänger *B* und weiteren Teilnehmern *C* erläutert werden.

Jeder Kommunikationsteilnehmer *B*, der von anderen Kommunikationsteilnehmern verschlüsselte Nachrichten empfangen möchte, stellt einen Schlüssel c_B in einem **öffentlichen Register** bereit, auf das alle Kommunikationsteilnehmer zugreifen können. Der Schlüssel c_B („codieren“) dient allen Kommunikationsteilnehmern zur Verschlüsselung von Nachrichten, die an *B* gesendet werden. Zusätzlich besitzt jeder Empfänger *B* einen **geheimen Schlüssel** d_B („decodieren“), mit dem er alle Nachrichten entschlüsselt, die an ihn gesandt wurden.

Oben wurde der Schlüssel zum Verschlüsseln einer Nachricht von *A* nach *B* mit K_{AB} bezeichnet. Um auszudrücken, dass alle Kommunikationsteilnehmer denselben Schlüssel für Nachrichten an *B* verwenden, wird er hier c_B (anstelle von K_{AB} bzw. K_{CB}) geschrieben. Entsprechend wird hier nicht die allgemeine Bezeichnung $\hat{K}_{AB}$ für den Schlüssel zum Entschlüsseln einer Nachricht verwendet, die *B* von einem Kommunikationsteilnehmer *A* empfangen hat, sondern d_B, da *B* den Schlüssel d_B zum Entschlüsseln aller Nachrichten an ihn, unabhängig vom Absender, verwendet.

Das Eintragen des öffentlichen Schlüssels c_B in das Register unterliegt keiner Geheimhaltung, da dieser Schlüssel ja sowieso öffentlich ist. Das Problem der Schlüsselverteilung wie bei symmetrischen Verfahren stellt sich hier nicht.

Ein Klartext *x*, der von *A* nach *B* verschlüsselt gesandt werden soll, wird von *A* in den Schlüsseltext $y = E(x, c_B)$ transformiert. Eine von *B* empfangene verschlüsselte Nachricht *y* wird von *B* in $D(y, d_B)$ entschlüsselt.

Die folgende Abbildung zeigt drei Kommunikationsteilnehmer *A*, *B* und *C* mit den jeweiligen Schlüsseln.

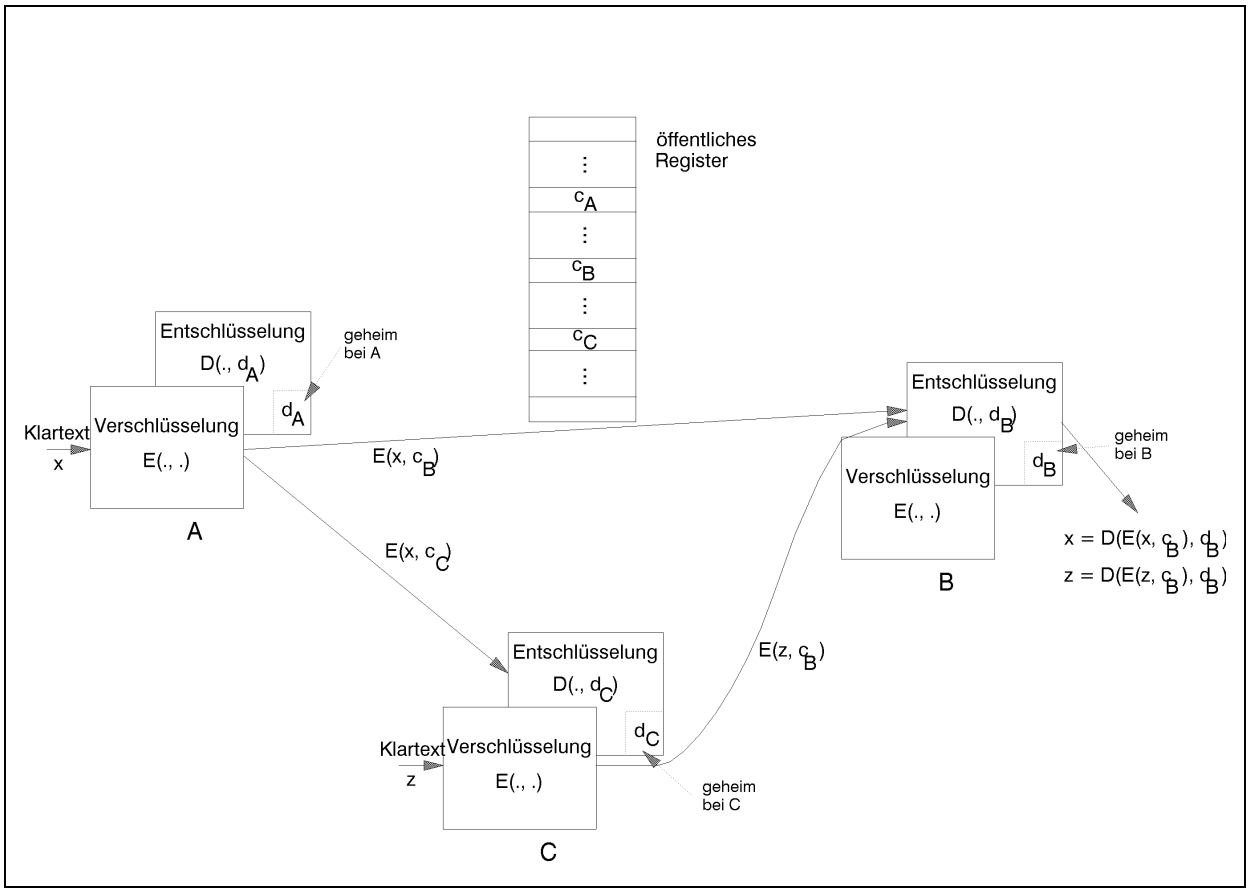


Abbildung: Verschlüsselung und Entschlüsselung mit asymmetrischen Verfahren

Verschlüsselungs- und Entschlüsselungsverfahren müssen folgende Bedingungen erfüllen:

(i) Ein Empfänger *B* kann eine empfangene verschlüsselte Nachricht mit seinem Schlüssel korrekt entschlüsseln, d.h. $D(E(x, c_B), d_B) = x$.

(ii) Die Verschlüsselung einer Nachricht, d.h. die Berechnung von $E(x, c_B)$, und die Entschlüsselung einer Nachricht bei Kenntnis des Schlüssels d_B, d.h. die Berechnung von $D(y, d_B)$, sind mit geringem Rechenaufwand durchzuführen.

(iii) Aus der Kenntnis eines öffentlichen Schlüssels c_B zum Verschlüsseln der Nachrichten an einen Empfänger *B* kann man „nicht leicht“ auf den bei *B* geheim gehaltenen Schlüssel d_B schließen. Die Forderung, eine Berechnung „nicht leicht“ durchführen zu können, wird mathematisch exakt durch den Begriff „**intractable**“ umschrieben, der ausdrückt, dass es zur Berechnung (beweisbar) keinen schnell ausführbaren Algorithmus gibt.

(iv) Ohne d_B zu kennen, kann ein Kryptoanalytiker aus einem Schlüsseltext $E(x, c_B)$ nicht leicht x ermitteln. Die Verschlüsselungsfunktion $E(.,.)$ stellt eine sogenannte **Einwegfunktion mit Falltür** dar. Erst wenn man die geheime Zusatzinformation d_B (die **Falltürinformation**) kennt, kann man die zu E inverse Funktion leicht berechnen.

(v) Zur Realisierung eines Unterschriftenprotokolls wird zusätzlich die Vertauschbarkeit der Verschlüsselung und Entschlüsselung gefordert. Neben der in (i) formulierten Bedingung $D(E(x, c_B), d_B) = x$ gilt auch $E(D(y, d_B), c_B) = y$.

In der Literatur sind eine Reihe von PKE-Verfahren veröffentlicht. Ihre Sicherheit ist mit Einschränkungen mathematisch beweisbar und hat sich in der Praxis bewährt; die Einschränkung bezieht sich auf eine bisher unbewiesene mathematische Vermutung bezüglich der Komplexität nichtdeterministischer Rechenverfahren (das sogenannte P-NP-Problem bzw. gewisser zahlentheoretischer Problemstellungen).

Zur Beschreibung eines PKA-Verfahrens muss angegeben werden, wie ein Kommunikationsteilnehmer B seinen geheimen Schlüssel d_B und seinen öffentlichen Schlüssel c_B festlegt, und wie die Verschlüsselungs- bzw. Entschlüsselungsalgorithmen $E(.,.)$ bzw. $D(.,.)$ definiert sind.

Das bekannteste PKE-Verfahren wird nach seinen Entdeckern Rivest, Shamir und Adleman ***RSA*-Verfahren** genannt[2]. Es bietet bei sorgfältiger Auswahl einiger im Verfahren frei wählbarer Parameter und entsprechender Implementierung eine sehr hohe Sicherheit. Es ist ein rein **softwaremäßig implementiertes Verfahren**. Dadurch ist seine Verschlüsselungs- bzw. Entschlüsselungsgeschwindigkeit etwa um den Faktor 1.000 langsamer als beispielsweise bei *DES* (gängiges symmetrisches Verfahren, das hardwaremäßig implementierbar ist). Es erfordert eine besondere Arithmetik natürlicher Zahlen mit sehr großen Stellenzahlen. Daher eignet es sich zur Verschlüsselung langer Nachrichten bzw. zur online-Verschlüsselung nur begrenzt. Ein „Hybrid"-Verfahren, das den Vorteil der Schnelligkeit von *Tripel-DES* bzw. *IDEA* beim Verschlüsseln und Entschlüsseln mit der Sicherheit von *RSA* verbinden, ist das seit 1991 über das Internet als Shareware verbreitete ***PGP*-Verfahren** (**pretty good privacy**), das sich besonders im E-Mail-Bereich bewährt hat.

Im folgenden werden einige Details des *RSA*-Verfahrens beschrieben. Das Verfahren beruht auf mathematischen, insbesondere zahlentheoretischen Grundlagen, Erkenntnissen der Komplexitätstheorie und dem Einsatz sehr großer Zahlen (mehr als 300 Dezimalstellen).

---

[2] Rivest, M.; Shamir, A.; Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems, Comm. ACM, 21, S.120-126, 1978.

Im RSA-Verfahren wird die Modulo-Arithmetik für ganze Zahlen (im folgenden nur natürliche Zahlen, d.h. nicht-negative ganze Zahlen) eingesetzt. Neben Additionen und Multiplikationen wird auch die Exponentiation verwendet.

Eine Zahlenpotenz $a^e \bmod n$ lässt sich durch wiederholte Multiplikation und sofortigem Übergang zum ganzzahligen Rest während des Rechenvorgangs berechnen:

$$a^e \bmod n \equiv \underbrace{((((a \cdot a \bmod n) \cdot a \bmod n) \cdot a \bmod n) \cdot ...) \bmod n}_{e \text{ viele } a\text{'s}}$$

Die Festlegung des öffentlichen Schlüssels c_B für die Verschlüsselung von Klartexten an einen Empfänger *B* und des geheimen Schlüssels d_B zum Entschlüsseln eines Schlüsseltextes beim Empfänger *B* verläuft im *RSA*-Verfahren wie folgt.

1. Es werden zwei verschiedene sehr große Primzahlen *p* und *q* ausgewählt, z.B. in der Größenordnung von 150 Dezimalstellen. Dann wird die Zahl $n = p \cdot q$ gebildet. Die Zahl *n* hat dann mindestens 300 Dezimalstellen, d.h. etwa 1.000 Binärstellen; in der Praxis wählt man *p* und *q* so, dass die Zahl *n* eine Binärstellenzahl von 1.024 aufweist.

   Zum Auffinden von Primzahlen in dieser Größenordnung und zum Testen auf Primzahleigenschaft kennt man schnelle Verfahren.

2. Für *B* wird eine Zufallszahl $d > \max\{p, q\}$ ausgewählt, die mit $\phi(n) = (p-1) \cdot (q-1)$ keinen gemeinsamen Teiler ausser 1 besitzt. Der Wert *d* darf nicht zu klein sein, da er Teil des geheim gehalten Schlüssels zum Entschlüsseln ist und daher von einem Kryptoanalytiker nicht durch systematisches Probieren gefunden werden darf.

3. Mit Hilfe der Erweiterung des Euklidischen Algorithmus (Funktion `invers`) ermittelt man eine eindeutig bestimmte Zahl *e* und eine für das Verfahren nicht weiter verwendete Zahl *f* mit den Eigenschaften

   $0 < e < \phi(n)$ und $e \cdot d + f \cdot \phi(n) = 1$.

   Da nach Konstruktion $ggT(d, \phi(n)) = 1$ gilt, findet man eine derartige Zahl *e* immer. Diese hat die Eigenschaft

   $e \cdot d \equiv 1 \,(\bmod \phi(n))$ bzw. $(e \cdot d \bmod \phi(n)) = 1$.

4. Der von *B* **geheim gehaltene Schlüssel zum Entschlüsseln** von Nachrichten, **die an *B* gesendet werden**, besteht aus der Zahlenfolge $d_B = [d, p, q, \phi(n)]$. Zum Entschlüsseln wird nur *d* verwendet, es ist jedoch unbedingt erforderlich, die Werte *p*, *q* und $\phi(n)$ ebenfalls geheim zu halten, da ein Kryptoanalytiker aus der Kenntnis des öffentlichen

benfalls geheim zu halten, da ein Kryptoanalytiker aus der Kenntnis des öffentlichen Schlüssels (siehe 5.) und aus der Kenntnis eines der Werte *p*, *q* oder $\phi(n)$ den geheimen Schlüsselteil *d* leicht ermitteln kann (siehe unten). Die Zahlenfolge $d_B = [d, p, q, \phi(n)]$ stellt die Falltürinformation dar.

5. Der in das öffentliche Register eingetragene Schlüssel zum Verschlüsseln aller Nachrichten an *B* ist die Zahlenfolge $c_B = [e,\ n]$.

Die **Vorschrift zur Verschlüsselung** von Nachrichten an *B* lautet:

Ein eventuell sehr lange Klartext *x* wird als Binärmuster aufgefasst und in Blöcke x_i mit jeweils $\lfloor \log_2(n) \rfloor$ vielen Stellen aufgeteilt: $x = x_1 x_2 \ldots x_r$. Eventuell wird dabei der letzte Teilblock x_r mit binären Nullen aufgefüllt. Jeder Teilblock x_i kann als Binärzahl interpretiert werden, die einen Wert $0 \le x_i < 2^{\log_2(n)} = n$ hat. Der Klartext *x* wird blockweise verschlüsselt; die einzelnen verschlüsselten Klartextblöcke werden dann wieder zu einem Schlüsseltext *y* zusammengesetzt:

Die Verschlüsselung eines Blockes x_i lautet

$$y_i = E(x_i, c_B) = E(x_i, [e, n]) = \left(x_i^e \bmod n\right).$$

Diese Zahl kann wieder als Binärmuster mit $\lfloor \log_2(n) \rfloor$ vielen Stellen aufgefasst werden.

Die Hintereinanderreihung aller so entstandenen Binärmuster $y_1, y_2, \ldots, y_r$ ergibt den zu *x* gehörenden Schlüsseltexte $y = E(x, c_B) = E(x_1, c_B) E(x_2, c_B) \ldots E(x_r, c_B)$.

Es gibt sehr effiziente Algorithmen zur Berechnung von $y_i = \left(x_i^e \bmod n\right)$, so dass die Verschlüsselung schnell erfolgen kann.

Eine bei *B* ankommende verschlüsselte Nachricht *y* wird zur **Entschlüsselung** als Binärmuster interpretiert und in einzelne Blöcke mit $\lfloor \log_2(n) \rfloor$ vielen Stellen zerlegt, d.h. $y = y_1 y_2 \ldots y_r$. Jeder Block y_i wird einzeln nach folgender Vorschrift entschlüsselt:

$$D(y_i, d_B) = D(y_i, [d, p, q, \phi(n)]) = \left(y_i^d \bmod n\right).$$

Die so entstehenden Zahlen werden als Bitmuster mit jeweils mit $\lfloor \log_2(n) \rfloor$ vielen Stellen interpretiert und durch Hintereinanderreihung zum entschlüsselten Text zusammengesetzt. Der algorithmische Aufwand zur Entschlüsselung ist wie bei der Verschlüsselung klein.

Die Korrektheit des Verfahrens, nämlich

$$D\left(E\left(x_i, c_B\right), d_B\right) = \left(\left(x_i^e\right)^d \bmod n\right) = x_i ,$$

folgt aus Satz 3.4-2:

Dazu werden 3 Fälle unterschieden:

1. Fall: Weder p noch q teilen x_i. Dann gilt $ggT\left(x_i, n\right) = 1$ und mit Satz 3.4-2: $x_i^{\phi(n)} \equiv 1 \left(\bmod n\right)$. Nach Konstruktion ist $e \cdot d + f \cdot \phi(n) = 1$ bzw. $e \cdot d = 1 + \left(-f\right) \cdot \phi(n)$. Also $\left(x_i^e\right)^d \equiv x_i^{1+(-f)\phi(n)} \equiv x_i \cdot \left(x_i^{\phi(n)}\right)^{(-f)} \equiv x_i \left(\bmod n\right)$. Da $x_i < n$ ist, ergibt sich $\left(\left(x_i^e\right)^d \bmod n\right) = x_i$.

2. Fall: Genau eine der Zahlen p oder q teilt x_i; es sei dieses p. Dann gilt (wieder mit Satz 3.4-2): $x_i^{q-1} \equiv 1 \left(\bmod q\right)$. Damit folgt nacheinander $x_i^{\phi(q)} \equiv 1 \left(\bmod q\right)$, $x_i^{(-f)\phi(q)} \equiv 1 \left(\bmod q\right)$, $x_i^{e \cdot d} \equiv x_i \left(\bmod q\right)$, d.h. q teilt $x_i^{e \cdot d} - x_i$. Da nach Fallannahme die Zahl p den Wert x_i teilt, teilt p auch $x_i^{e \cdot d}$, und daher teilt p den Wert $x_i^{e \cdot d} - x_i$. Mit Satz 3.3-4 (iv) folgt: $n = p \cdot q$ teilt $x_i^{e \cdot d} - x_i$, d.h. $x_i^{e \cdot d} \equiv x_i \left(\bmod n\right)$. Da $x_i < n$ ist, ergibt sich wie im 1. Fall: $\left(\left(x_i^e\right)^d \bmod n\right) = x_i$.

3. Fall: Beide Zahlen p und q teilen x_i. Dann teilen p und q den Wert $x_i^{e \cdot d} - x_i$ und mit Satz 3.3-4 (iv) folgt: $n = p \cdot q$ teilt $x_i^{e \cdot d} - x_i$, d.h. $x_i^{e \cdot d} \equiv x_i \left(\bmod n\right)$. Da $x_i < n$ ist, ergibt sich wie im 1. Fall: $\left(\left(x_i^e\right)^d \bmod n\right) = x_i$.

Es gilt außerdem die Symmetriegleichung $E\left(D\left(y_i, d_B\right), c_B\right) = \left(\left(y_i^d\right)^e \bmod n\right) = y_i$, so dass das *RSA*-Verfahren für ein digitales Unterschriftenprotokoll geeignet ist.

Bei der Konstruktion des geheimen Schlüssels $d_B = \left[d, p, q, \phi(n)\right]$ besteht eine gewisse Freiheit bezüglich der Wahl der einzelnen Komponenten. Beispielsweise kann man den Wert *d* so groß wählen, dass er von einem Kryptoanalytiker nicht leicht durch systematisches Testen gefunden werden kann. Der Exponent *e* zum Verschlüsseln eines Klartextes an *B* ist nach Wahl von *d* eindeutig bestimmt. Der umgekehrte Weg, nämlich erst *e* zu wählen, und zwar so, dass *e* und $\phi(n) = (p-1) \cdot (q-1)$ teilerfremd sind, und dann mit Hilfe des Euklidischen Algorithmus *d* zu ermitteln, ist ebenfalls möglich. Auf diese Weise kann man für *e* einen „günstigen" Wert nehmen. Als günstige Werte haben sich die Primzahlen $e = 3$, $e = 17$ und $e = 65.537$

erwiesen, da diese Fermat-Zahlen in ihrer Binärdarstellung nur jeweils zwei binäre Einsen haben und damit die in der Verschlüsselung auszuführende Exponentiation sehr schnell abläuft.

Die folgenden **Empfehlungen bezüglich der im Verfahren auszuwählenden Zahlen** zielen auf die Gewährung eines hohen Sicherheitsniveaus des *RSA*-Verfahrens.

- Die Primzahlen *p* und *q* sollten „zufällig" gewählt und nicht etwa einer Primzahltabelle entnommen werden und auch keine spezielle funktionale Form (etwa $2^{2^k}-1$) aufweisen.

- Die Primzahlen *p* und *q* sollten nicht zu dicht zusammenliegen.

- Die Primzahlen *p* und *q* sollten so gewählt werden, dass $p-1$ und $q-1$ keine großen gemeinsamen Faktoren besitzen.

- Die Primzahlen *p* und *q* sollten so gewählt werden, dass $\phi(n)=(p-1)\cdot(q-1)$ nicht nur kleine Primfaktoren enthält.

- Der Wert *d* sollte nicht zu klein sein, damit man ihn nicht durch systematisches Testen ermitteln kann.

- Verschiedene Kommunikationspartner sollten nicht denselben Wert oder einen zu kleinen Wert für *e* nehmen.

- Die Klartexte (hier als numerische Werte aufgefasst) $x=1$ und $x=n-1$ werden auf sich selbst verschlüsselt, d.h. in diesen Fällen gilt $E(x,c_B)=x$. Dasselbe Fixpunktverhalten der Funktion *E* zeigt sich, wenn $e-1$ ein gemeinsames Vielfaches von $p-1$ und $q-1$ ist, etwa $e-1=\phi(n)/2$. Dann gilt sogar für jeden Klartext *x* die Gleichung $E(x,c_B)=x$. In diesem Fall ist eine andere Wahl von *d* angeraten.

Da beim *RSA*-Verfahren alle Komponenten bis auf den geheimen Schlüssel $d_B=[d,p,q,\phi(n)]$ öffentlich sind, bietet es für einen Kryptoanalytiker Angriffspunkte. Ein Kryptoanalytiker ist prinzipiell nur an der Kenntnis des Schlüsselteils *d* des geheimen Schlüssels d_B interessiert, wobei er beide Teile *e* und *n* des öffentlichen Schlüssels c_B kennt. Folgende Überlegungen zeigen, dass es erforderlich ist, neben *d* auch die Werte *p*, *q* und $\phi(n)$ geheimzuhalten.

Kennt der Kryptoanalytiker die Werte *e, n* (aus dem öffentlichen Register) und $\phi(n)$, dann kann er mit Hilfe des Euklidischen Algorithmus zwei Zahlen *a* und *b* berechnen, für die die Beziehung $e \cdot a + \phi(n) \cdot b = 1$ gilt (hierbei ist zu beachten, dass nach Konstruktion des Verfahrens *e* und $\phi(n)$ teilerfremd sind). Einfache zahlentheoretische Überlegungen zeigen, dass man *d* aus der Gleichung $d = (a \bmod \phi(n))$ erhält. Kennt der Kryptoanalytiker die Werte *e, n* (aus dem öffentlichen Register) und mindestens einen der Werte *p* oder *q*, etwa *p*, dann kann er wegen $\phi(n) = (p-1) \cdot (q-1)$ und $n = p \cdot q$ bzw. $\phi(n) = (p-1) \cdot (n/p - 1)$ sofort $\phi(n)$ und damit *d* ermitteln.

Offensichtlich ist die Geheimhaltung von $\phi(n)$ wesentlich. Natürlich könnte der Kryptoanalytiker versuchen, den Wert $\phi(n)$ direkt aus dem öffentlichen Schlüssel $c_B = [e,\ n]$ zu gewinnen. Falls ihm dieses mit geringem Rechenaufwand gelänge, hätte er gleichzeitig einen schnellen Algorithmus, um die Zahl *n* in ihre Primfaktoren *p* und *q* zu zerlegen: Er berechnet nacheinander $z = \phi(n) - n - 1$, $y = \sqrt{z^2 - 4n}$, $p = 1/2 \cdot (-z - y)$ und $q = n/p$. Daher ist die schnelle Berechnung von $\phi(n)$ aus $c_B = [e,\ n]$ gleichbedeutend mit der schnellen Primfaktorisierung von *n*. Andererseits kennt man bis heute kein schnelles Verfahren, um *n* zu faktorisieren. Die schnellsten bisher bekannten Verfahren zur Zerlegung einer Zahl *n* in ihre Primfaktoren haben eine Laufzeit, die proportional zu $L(n) = e^{\sqrt{\ln(n) \cdot \ln(\ln(n))}}$ ist. Die folgende Tabelle zeigt einige Werte von $L(n)$.

n	10^{50}	10^{100}	10^{150}	10^{200}	10^{250}	10^{300}
$L(n)$	$1{,}42 \cdot 10^{10}$	$2{,}34 \cdot 10^{15}$	$3{,}26 \cdot 10^{19}$	$1{,}20 \cdot 10^{23}$	$1{,}86 \cdot 10^{26}$	$1{,}53 \cdot 10^{29}$

Wäre man heute technisch in der Lage, Rechengeschwindigkeit von 10^{12} Operationen pro Sekunde zu realisieren, würde die Faktorisierung einer 200-stellige Zahl immer noch etwa 1.000 Jahre erfordern, die Faktorisierung einer 300-stelligen Zahl sogar mehr als 10^6 viele Jahrtausende. Heutige Schlüssellängen von 1.024 Bits bzw. ca. 300 Dezimalstellen erscheinen daher heute sicher.

Des weiteren könnte der Kryptoanalytiker versuchen, den Wert *d* direkt aus dem öffentlichen Schlüssel $c_B = [e,\ n]$ zu ermitteln. Es lässt sich zeigen, dass ein schneller Algorithmus zur Ermittlung von *d* aus $c_B = [e,\ n]$ in einen schnellen (probabilistischen) Algorithmus umgewandelt werden kann, der mit beliebig großer Wahrscheinlichkeit die Zahl *n* in ihre Primfaktoren *p* und *q* korrekt zerlegt. Eine Brute-Force-Attacke, in der alle möglichen Werte für *d* systematisch probiert werden, verspricht darüber hinaus wegen der großen Schlüssellänge (Stellenzahl von *n*) keinen Erfolg.

Zusammenfassend kann man feststellen, dass die Garantie der Sicherheit des *RSA*-Verfahrens darauf zurückzuführen ist, dass kein schnelles Verfahren bekannt ist, das eine gegebene natürliche Zahl in ihre Primfaktoren zerlegt. Sollte ein derartiges Verfahren für das Faktorisierungsproblem gefunden werden, ist das *RSA*-Verfahren nicht mehr sicher.

Die bisher in diesem Kapitel beschriebenen Methoden beruhen auf der Anwendung zahlentheoretischer Erkenntnisse, die im wesentlichen im 18. Jahrhundert entdeckt wurden. Die Überlegungen zum Laufzeitverhalten der beteiligten Algorithmen stammen aus den letzten 30 Jahren des 20. Jahrhunderts. Seit etwa 1987 findet eine Theorie, deren Grundlagen zum Ende des 19. Jahrhunderts gelegt wurden, beim Entwurf kryptographischer Verfahren verstärkt Anwendung. Diese Kryptographie-Verfahren setzen zur Verschlüsselung die **Arithmetik elliptischer Kurven über endlichen Körpern** ein. Da zum Verständnis dieser Methoden jedoch weitergehende mathematische Kenntnisse erforderlich sind, wird auf deren Darstellung hier verzichtet.

4 Ausgewählte Themen der Kombinatorik

Die Kombinatorik befasst sich im wesentlichen mit dem Abzählen endlicher Mengen und damit verwandter Fragestellungen. Die in diesem Kapitel behandelte Themenauswahl gehört zum mathematischen Handwerkszeug, das in vielen Teilgebieten der Mathematik und Informatik benötigt wird. Insbesondere in der Wahrscheinlichkeitsrechnung (diskrete Wahrscheinlichkeitsverteilungen) werden die Themen weiter vertieft.

4.1 Binomialkoeffizienten

Es seien $n \in \mathbf{N}$, $x \in \mathbf{R}$ und $y \in \mathbf{R}$. Der aus der Schule bekannte binomische Lehrsatz besagt

$$(x+y)^2 = x^2 + 2 \cdot x \cdot y + y^2 .$$

Wie man leicht nachrechnet, ist

$$(x+y)^3 = x^3 + 3 \cdot x^2 \cdot y + 3 \cdot x \cdot y^2 + y^3 ,$$
$$(x+y)^4 = x^4 + 4 \cdot x^3 \cdot y + 6 \cdot x^2 \cdot y^2 + 4 \cdot x \cdot y^3 + y^4 .$$

Es soll nun die allgemeine Form von $(x+y)^n$ als ausgeschriebene Summe hergeleitet werden (das könnte wieder formal nach dem Induktionsprinzip geschehen; hier soll die Herleitung etwas informeller beschrieben werden). Durch vollständige Induktion kann man zeigen, dass die Summanden in der ausgeschriebenen Summe von $(x+y)^n$ die Form $k_{i,j} \cdot x^i \cdot y^j$ mit $i + j = n$ haben. Der Faktor $k_{i,j}$ im Summanden $k_{i,j} \cdot x^i \cdot y^{n-i}$ der ausgeschriebenen Summe von $(x+y)^n$ heißt **Binomialkoeffizient** $\binom{n}{i}$, gesprochen „*n* über *i*“, d.h.

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} \cdot x^i \cdot y^{n-i} = \binom{n}{0} \cdot y^n + \binom{n}{1} \cdot x \cdot y^{n-1} + \binom{n}{2} \cdot x^2 \cdot y^{n-2} + \ldots + \binom{n}{n} \cdot x^n .$$

Das vorliegende Kapitel untersucht Eigenschaften und Interpretationen der Binomialkoeffizienten.

Offensichtlich gilt $\binom{n}{0}=1$ und $\binom{n}{n}=1$.

Es ist $(x+y)^n=(x+y)^{n-1}\cdot(x+y)=(x+y)^{n-1}\cdot x+(x+y)^{n-1}\cdot y$. Aus dieser Darstellung kann man ablesen, wie der Faktor $\binom{n}{i}$ im Summanden $\binom{n}{i}\cdot x^i\cdot y^{n-i}$ der ausgeschriebenen Summe von $(x+y)^n$ entsteht: man sucht in der ausgeschriebenen Summe von $(x+y)^{n-1}$ denjenigen Summanden, in dem *x* mit der Potenz $i-1$ und *y* mit der Potenz $n-i$ steht, und denjenigen Summanden in der ausgeschriebenen Summe von $(x+y)^{n-1}$, in dem *x* mit der Potenz i und *y* mit der Potenz $n-i-1$ steht. Diese Summanden sind

$$\binom{n-1}{i-1}\cdot x^{i-1}\cdot y^{n-i}=\binom{n-1}{i-1}\cdot x^{i-1}\cdot y^{(n-1)-(i-1)} \quad \text{und} \quad \binom{n-1}{i}\cdot x^i\cdot y^{n-i-1}=\binom{n-1}{i}\cdot x^i\cdot y^{(n-1)-i}.$$

Wird der erste Summand mit *x* und der zweite Summand mit *y* multipliziert und anschließend beide Summanden addiert, entsteht

$$\begin{aligned}\binom{n-1}{i-1}\cdot x^{i-1}\cdot y^{n-i}\cdot x+\binom{n-1}{i}\cdot x^i\cdot y^{n-i-1}\cdot y&=\left(\binom{n-1}{i-1}+\binom{n-1}{i}\right)\cdot x^i\cdot y^{n-i}\\&=\binom{n}{i}\cdot x^i\cdot y^{n-i}.\end{aligned}$$

Damit ergibt sich

Satz 4.1-1:

Für jedes $n\in\mathbf{N}$ und für jedes $i\in\mathbf{N}$ ist

$\binom{n}{0}=1,\ \binom{n}{n}=1,$

$\binom{n}{i}=\binom{n-1}{i-1}+\binom{n-1}{i}$ mit $0<i<n$.

Mit Hilfe der Rekursionsformel in Satz 4.1-1 lassen sich die Binomialkoeffizienten berechnen. Die einzelnen Werte können in einem Schema in Dreicksform (**Pascal'sche Dreieck**)

angeordnet werden. Dabei steht in der *n*-ten Zeile und der *i*-ten Spalte der Wert $\binom{n}{i}$ für $n \geq 0$ und $0 \leq i \leq n$. Dieser Eintrag ist die Summe, die sich aus dem direkt drüber stehenden Eintrag $\binom{n-1}{i}$ und dem Eintrag $\binom{n-1}{i-1}$ links davon ergibt. Der Anfang des Pascal'schen Dreiecks lautet:

	Spalte 0					Spalte $i = 5$				
Zeile 0	1									
	1	1								
	1	2	1							
	1	3	3	1						
	1	4	6	4	1					
	1	5	10	10	5	1				
	1	6	15	20	15	6	1			
	1	7	21	35	35	21	7	1		
Zeile $n = 8$	1	8	28	56	70	56	28	8	1	
	1	9	36	84	126	126	84	36	9	1
					...	...				

Beispiel:

$\binom{8}{5} = 56$.

Der folgende Satz beschreibt, wie der Wert eines Binomialkoeffizienten $\binom{n}{i}$ direkt in Abhängigkeit von *n* und *i* ausgedrückt werden kann:

Satz 4.1-2:

Für jedes $n \in \mathbf{N}$ und für jedes $i \in \mathbf{N}$ mit $0 \leq i \leq n$ ist

$$\binom{n}{i} = \frac{n!}{i! \cdot (n-i)!}.$$

Der Beweis soll hier als Beispiel eines Beweises durch vollständige Induktion angegeben werden:

Für $n = 0$ ist $\binom{n}{i} = \binom{0}{0} = 1$ und $\frac{n!}{i!\cdot(n-i)!} = \frac{0!}{0!\cdot 0!} = 1$.

Es wird angenommen, dass die Formel in Satz 4.1-2 für ein $n \in \mathbf{N}$ und für jedes $i \in \mathbf{N}$ mit $0 \le i \le n$ gilt. Zu zeigen ist, dass aus dieser Annahme die Gültigkeit der Formel auch für $n+1$ und für jedes $i \in \mathbf{N}$ mit $0 \le i \le n+1$ folgt.

Für $i = 0$ ist $\binom{n+1}{i} = \binom{n+1}{0} = 1$ und $\frac{(n+1)!}{i!\cdot(n+1-i)!} = \frac{(n+1)!}{0!\cdot(n+1)!} = 1$.

Für $i = n+1$ ist $\binom{n+1}{i} = \binom{n+1}{n+1} = 1$ und $\frac{(n+1)!}{(n+1)!\cdot 0!} = 1$.

Für $0 < i < n+1$ verwendet man die Rekursionsformel aus Satz 4.1-1:

$$\begin{aligned}
\binom{n+1}{i} &= \binom{n}{i-1} + \binom{n}{i} && \text{nach Satz 4.1-1} \\
&= \frac{n!}{(i-1)!\cdot(n-i+1)!} + \frac{n!}{i!\cdot(n-i)!} && \text{nach Induktionsannahme} \\
&= \frac{n!\cdot i + n!\cdot(n-i+1)}{i!\cdot(n-i+1)!} \\
&= \frac{n!\cdot(i+n-i+1)}{i!\cdot(n-i+1)!} \\
&= \frac{(n+1)!}{i!\cdot(n+1-i)!}.
\end{aligned}$$

Die Formel gilt also auch für $n+1$.

Im Pascal'schen Dreieck kann man einige Gesetzmäßigkeiten der Binomialkoeffizienten verifizieren, die direkt aus der Definition $(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} \cdot x^i \cdot y^{n-i}$ bzw. aus den Formeln in den vorherigen Sätzen folgen. Beispielsweise hat die Summe aller Binomialkoeffizienten in der *n*-ten Zeile des Pascal'schen Dreiecks den Wert 2^n; die Summe der Binomialkoeffizienten $\binom{n}{i}$ in Zeile *n* mit $n \ge 1$ und geradem *i* ist gleich der Summe der Binomialkoeffizienten $\binom{n}{i}$ in Zeile *n* mit ungeradem *i*; summiert man in der *i*-ten Spalte alle Binomialkoeffizienten bis zur Zeile *n*, so erhält man wieder einen Binomialkoeffizienten, nämlich $\binom{n+1}{i+1}$; summiert man

alle Binomialkoeffizienten ab Zeile $n-i$ und Spalte 0 diagonal (von links oben nach rechts unten) bis zur Zeile *n* und Spalte *i*, so ist das Ergebnis der Binomialkoeffizient $\binom{n+1}{i}$. Diese und weitere Eigenschaften der Binomialkoeffizienten werden im folgenden Satz zusammengefasst.

Satz 4.1-3:

Es sei $n \in \mathbf{N}$. Dann gilt:

(i) $\sum_{i=0}^{n} \binom{n}{i} \cdot x^i = (1+x)^n$ für jedes $x \in \mathbf{R}$.

(ii) $\sum_{i=0}^{n} \binom{n}{i} = 2^n$ (Summe über die *n*-te Zeile im Pascal'schen Dreieck).

(iii) $\sum_{\substack{i=0 \\ (i \bmod 2)=0}}^{n} \binom{n}{i} = \sum_{\substack{i=0 \\ (i \bmod 2)=1}}^{n} \binom{n}{i}$ bzw. $\sum_{i=0}^{n} (-1)^i \binom{n}{i} = 0$ für $n \geq 1$

(Die Summe der Binomialkoeffizienten mit geradem *i* ist gleich der Summe der Binomialkoeffizienten mit ungeradem *i*).

(iv) $\sum_{k=i}^{n} \binom{k}{i} = \binom{n+1}{i+1}$ für $i \in \mathbf{N}$ mit $0 \leq i \leq n$

(Summe der *i*-ten Spalte bis zur Zeile *n* im Pascal'schen Dreieck).

(v) $\sum_{k=0}^{i} \binom{n-i+k}{k} = \binom{n+1}{i}$ für $i \in \mathbf{N}$ mit $0 \leq i \leq n$

(Summe der Binomialkoeffizienten ab Zeile $n-i$ und Spalte 0 im Pascal'schen Dreieck diagonal bis zur Zeile *n* und Spalte *i*).

(vi) $\binom{n}{i} = \binom{n}{n-i}$ für $i \in \mathbf{N}$ mit $0 \leq i \leq n$

(Symmetrie der Binomialkoeffizienten).

(vii) $\binom{n}{i} = \frac{n}{i} \cdot \binom{n-1}{i-1}$ für $i \in \mathbf{N}$ mit $0 < i \leq n$

und $(n-i) \cdot \binom{n}{i} = n \cdot \binom{n-1}{i}$ für $i \in \mathbf{N}$ mit $0 \leq i \leq n$.

Die Formel in Satz 4.1-3 (i) erhält man, indem man in der Definitionsgleichung $(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} \cdot x^i \cdot y^{n-i}$ den Wert $y = 1$ setzt. Die Formel in Satz 4.1-3 (ii) erhält man aus der Definitionsgleichung für $x = y = 1$. Die Formel $\sum_{i=0}^{n} (-1)^i \binom{n}{i} = 0$ in Satz 4.1-3 (iii) erhält man aus der Definitionsgleichung für $x = -1$ und $y = 1$; bringt man in dieser Formel die Summanden $(-1)^i \binom{n}{i}$ mit ungeradem *i* auf die rechte Seite der Gleichung, so ergibt sich die erste Formel in Satz 4.1-3 (iii). Die Formeln in Satz 4.1-3 (vi) und (vii) ergeben sich unmittelbar aus Satz 4.1-2. Die Formeln in Satz 4.1-3 (iv) und (v) können durch vollständige Induktion bewiesen werden oder durch direkte wiederholte Anwendung der Rekursionsgleichung aus Satz 4.1-1. Für die Formel aus Satz 4.1-3 (iv) ergibt sich ausgehend von der rechten Seite der Gleichung:

$$\begin{aligned}
\binom{n+1}{i+1} &= \binom{n}{i+1} + \binom{n}{i} && \text{mit Satz 4.1-1} \\
&= \left(\binom{n-1}{i+1} + \binom{n-1}{i} \right) + \binom{n}{i} && \text{mit Satz 4.1-1, angewandt auf den ersten Binomialkoeffizienten} \\
&= \left(\binom{n-2}{i+1} + \binom{n-2}{i} \right) + \binom{n-1}{i} + \binom{n}{i} \\
&= \\
&\ldots \\
&= \binom{n-l}{i+1} + \sum_{k=0}^{l} \binom{n-k}{i} && \text{allgemeine Form; mit } l = n-i-1: \\
&= \binom{i+1}{i+1} + \sum_{k=0}^{n-i-1} \binom{n-k}{i} \\
&= 1 + \binom{n}{i} + \binom{n-1}{i} + \ldots + \binom{i+1}{i} \\
&= \binom{i}{i} + \binom{n}{i} + \binom{n-1}{i} + \ldots + \binom{i+1}{i} \\
&= \sum_{k=i}^{n} \binom{k}{i}.
\end{aligned}$$

Für die Formel aus Satz 4.1-3 (v) ergibt sich wieder ausgehend von der rechten Seite der Gleichung:

$$\binom{n+1}{i} = \binom{n}{i} + \binom{n}{i-1} \qquad \text{mit Satz 4.1-1}$$

$$= \binom{n}{i} + \left(\binom{n-1}{i-1} + \binom{n-1}{i-2} \right) \quad \text{mit Satz 4.1-1, angewandt auf den zweiten Binomialkoeffizienten}$$

$$= \binom{n}{i} + \binom{n-1}{i-1} + \left(\binom{n-2}{i-2} + \binom{n-2}{i-3} \right)$$

$$=$$

$$\ldots$$

$$= \sum_{k=0}^{l} \binom{n-k}{i-k} + \binom{n-l}{i-l-1} \qquad \text{allgemeine Form; mit } l = i-1:$$

$$= \sum_{k=0}^{i-1} \binom{n-k}{i-k} + \binom{n-i+1}{0}$$

$$= 1 + \binom{n}{i} + \binom{n-1}{i-1} + \ldots + \binom{n-i+1}{1}$$

$$= \binom{n-i}{0} + \binom{n}{i} + \binom{n-1}{i-1} + \ldots + \binom{n-i+1}{1}$$

$$= \sum_{k=0}^{i} \binom{n-i+k}{k}.$$

Die Binomialkoeffizienten kommen nicht nur als Faktor $k_{i,j}$ im Summanden $k_{i,j} \cdot x^i \cdot y^{n-i}$ der ausgeschriebenen Summe von $(x+y)^n$ vor, sondern auch in vielen praktischen Abzählproblemen. In Kapitel 1.5 wird gezeigt, dass eine endliche Menge *A* mit *n* Elementen genau 2^n viele Teilmengen besitzt. Es soll nun untersucht werden, wie viele Teilmengen *A* hat, die aus genau *k* Elementen (mit $0 \le k \le n$) bestehen.

Es sei $A = \{a_1, \ldots, a_n\}$. Unter einer **Permutation** von *A* versteht man eine feste Anordnung der Elemente von *A*. Beispielsweise sind alle Permutationen der Menge $A = \{1,2,3,4\}$ die Anordnungen

1 2 3 4	4 1 3 2	3 1 2 4	4 3 2 1	2 3 1 4	4 2 1 3
1 2 4 3	1 4 3 2	3 1 4 2	3 4 2 1	2 3 4 1	2 4 1 3
1 4 2 3	1 3 4 2	3 4 1 2	3 2 4 1	2 4 3 1	2 1 4 3
4 1 2 3	1 3 2 4	4 3 1 2	3 2 1 4	4 2 3 1	2 1 3 4.

Eine Permutation der Menge $A = \{a_1, \ldots, a_n\}$ ist also ein Tupel $(a_{i_1}, \ldots, a_{i_n})$ mit $a_{i_j} \in A$ für $j = 1, \ldots, n$ und $a_{i_j} \neq a_{i_l}$ für $i_j \neq i_l$. Um die Anzahl aller Permutationen von *A* zu bestim-

men, betrachtet man alle derartigen Tupel $\left(a_{i_1},...,a_{i_n}\right)$: Für a_{i_1} gibt es *n* mögliche Werte, nämlich $a_1,...,a_n$. Hat man sich für eine Möglichkeit entschieden, bleiben für a_{i_2} noch $n-1$ Möglichkeiten. Für a_{i_3} bleiben nach Auswahl für a_{i_1} und a_{i_2} noch $n-2$ Möglichkeiten usw. Für a_{i_n} bleibt nach Festlegung der ersten $n-1$ Elemente nur noch 1 Möglichkeit. Insgesamt hat gibt es also

$$n\cdot(n-1)\cdot...\cdot 1 = n!$$

viele Möglichkeiten zur Bildung einer Permutation einer *n*-elementigen Menge.

Unter einer ***k*-Permutation** von *A* versteht man ein Tupel $\left(a_{i_1},...,a_{i_k}\right)$ mit $a_{i_j}\in A$ für $j=1,...,k$ und $a_{i_j}\neq a_{i_l}$ für $i_j\neq i_l$. Beispielsweise sind alle 2-Permutationen von $A=\{1,2,3,4\}$ die Anordnungen

1 2	2 1	3 1	4 1
1 3	2 3	3 2	4 2
1 4	2 4	3 4	4 3

Um die Anzahl aller *k*-Permutationen von *A* zu bestimmen, betrachtet man wieder alle derartigen Tupel $\left(a_{i_1},...,a_{i_k}\right)$: Für a_{i_1} gibt es *n* mögliche Werte, nämlich $a_1,...,a_n$. Hat man sich für eine Möglichkeit entschieden, bleiben für a_{i_2} noch $n-1$ Möglichkeiten. Für a_{i_3} bleiben nach Auswahl für a_{i_1} und a_{i_2} noch $n-2$ Möglichkeiten usw. Für a_{i_k} bleibt nach Festlegung der ersten $k-1$ Elemente noch $n-(k-1)=n-k+1$ Möglichkeit. Insgesamt hat gibt es also

$$n\cdot(n-1)\cdot...\cdot(n-k+1)=\frac{n!}{(n-k)!}$$

viele Möglichkeiten zur Bildung einer *k*-Permutation einer *n*-elementigen Menge.

Für jede Teilmengen $B\subseteq A$ mit $|B|=k$, etwa $B=\{a_{i_1},...,a_{i_k}\}$, gilt $a_{i_j}\neq a_{i_l}$ für $i_j\neq i_l$. Jede Permutation von *B* (davon gibt es *k*! viele) ist eine *k*-Permutation von *A*. Verschiedene Teilmengen $B_1\subseteq A$ und $B_2\subseteq A$ mit $|B_1|=|B_2|=k$ ergeben paarweise verschiedene *k*-Permutation von *A*, da die *k*-Permutationen, die aus B_1 entstanden sind, mindestens ein unterschiedliches Element zu den *k*-Permutationen enthalten, die aus B_2 entstanden sind, und die Permutationen von B_1 bzw. von B_2 sind untereinander paarweise verschieden. Umgekehrt

gibt es zu jeder *k*-Permutation von *A* eine *k*-elementige Teilmengen von *A*, nämlich die Menge ihrer Elemente. Bezeichnet $C_{n,k}$ die Anzahl *k*-elementiger Teilmengen von *A*, so gilt daher:

$$C_{n,k} \cdot k! = \frac{n!}{(n-k)!}.$$

Satz 4.1-4:

Es sei *A* eine endliche Menge mit $|A| = n$. Die Anzahl der Teilmengen $B \subseteq A$ mit $|B| = k$ für $0 \le k \le n$ ist

$$\binom{n}{k} = \frac{n!}{k!\cdot(n-k)!}.$$

Damit kann man auch noch einmal die Formel aus Satz 4.1-3 (ii)

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n$$

verifizieren: Links steht die Anzahl aller Teilmengen einer *n*-elementigen Menge, aufgeteilt nach den *i*-elementigen Teilmengen für $0 \le i \le n$, von der bereits früher gezeigt wurde, dass sie gleich 2^n ist.

4.2 Abbildungen zwischen endlichen Mengen

In diesem Kapitel seien *A* und *B* endliche Mengen mit $|A| = n$ und $|B| = m$:

$$A = \{a_1, ..., a_n\},$$

$$B = \{b_1, ..., b_m\}.$$

Es sollen die Anzahlen der Abbildungen, der injektiven, surjektiven und bijektiven Abbildungen $f : A \to B$ ermittelt werden.

A. Anzahl der Abbildungen $f : A \to B$

Jede Abbildung $f : A \to B$ kann in Form einer endlichen Tabelle notiert werden:

a_i	a_1	...	a_n
$f(a_i)$	$f(a_1)$	...	$f(a_n)$

Dabei genügt das Notieren der Funktionswerte in der Reihenfolge $f(a_1)$, ..., $f(a_n)$ bzw. als *n*-Tupel

$(f(a_1), ..., f(a_n))$.

Verschiedene Abbildungen führen zu verschiedenen *n*-Tupeln. Umgekehrt kann man jedes *n*-Tupel

$(b_1, ..., b_n)$ mit $b_i \in B$ für $1 \le i \le n$

als eine Abbildung $f : A \to B$ auffassen, nämlich als die durch $f(a_i) = b_i$ definierte Abbildung, und unterschiedliche *n*-Tupel beschreiben unterschiedliche Abbildungen. Daher ist die Anzahl der Abbildungen $f : A \to B$ gleich der Anzahl der *n*-Tupeln $(b_1, ..., b_n)$ mit $b_i \in B$ für $1 \le i \le n$.

Satz 4.2-1:

Es seien *A* und *B* endliche Mengen mit $|A| = n$ und $|B| = m$. Dann ist die Anzahl der Abbildungen $f : A \to B$ gleich

m^n.

B. Anzahl bijektiver Abbildungen $f : A \to B$

Nach Satz 2.2-3 gilt im Falle bijektiver Abbildungen $f : A \to B$ bezüglich der Mächtigkeiten: $n = m$. Jede bijektive Abbildung $f : A \to B$ bzw. in Tupel-Schreibweise $(f(a_1), ..., f(a_n))$ beschreibt daher eine Permutation der Menge *B*. Verschiedene bijektive Abbildungen führen

zu verschiedenen Permutationen. Umgekehrt kann jede Permutation $(b_{i_1}, ..., b_{i_n})$ von B mit einer bijektiven Abbildung $f : A \to B$, nämlich

$$f : \begin{cases} A & \to & B \\ a_j & \to & b_{i_j} \end{cases}$$

gleichgesetzt werden.

Die Anzahl bijektiver Abbildungen $f : A \to B$ ist daher gleich der Anzahl der Permutationen der Menge B. Mit den Überlegungen am Ende von Kapitel 4.1 folgt

Satz 4.2-2:

Es seien A und B endliche Mengen mit $|A| = |B| = n$. Dann ist die Anzahl der bijektiven Abbildungen $f : A \to B$ gleich

$n!$.

C. Anzahl injektiver Abbildungen $f : A \to B$

Nach Satz 2.2-3 gilt im Falle injektiver Abbildungen $f : A \to B$ bezüglich der Mächtigkeiten: $n \le m$. Der Wertebereich $W(f) = \{ b \mid b \in B, \text{und es gibt } a \in A \text{ mit } f(a) = b \}$ ist eine n-elementige Teilmenge von B, und die Abbildung

$$g : \begin{cases} A & \to & W(f) \\ a & \to & f(a) \end{cases}$$

ist eine Bijektion zwischen A und W(f). Daher ist W(f) eine n-Permutation von B. Umgekehrt ist kann man jede n-Permutation von B, etwa $(b_{i_1}, ..., b_{i_n})$ mit der injektiven Abbildung

$$f : \begin{cases} A & \to & B \\ a_j & \to & b_{i_j} \end{cases}$$

identifizieren. Mit den Überlegungen am Ende von Kapitel 4.1 folgt

Satz 4.2-3:

Es seien A und B endliche Mengen mit $|A| = n$ und $|B| = m$ und $n \le m$. Dann ist die Anzahl injektiver Abbildungen $f : A \to B$ gleich

$$\frac{m!}{(m-n)!} = n! \cdot \binom{m}{n}.$$

D. Anzahl surjektiver Abbildungen $f : A \to B$

Nach Satz 2.2-3 gilt im Falle surjektiver Abbildungen $f : A \to B$ bezüglich der Mächtigkeiten: $n \ge m$. Die Bestimmung der Anzahl surjektiver Abbildungen zwischen A und B ist schwieriger und benötigt Hilfsmittel, die in Kapitel 4.3 bereitgestellt werden. Das Ergebnis soll aber der Vollständigkeit halber hier bereits zitiert werden.

Satz 4.2-4:

Es seien A und B endliche Mengen mit $|A| = n$ und $|B| = m$ und $n \ge m$. Dann ist die Anzahl surjektiver Abbildungen $f : A \to B$ gleich

$$\sum_{i=0}^{m} (-1)^{m-i} \cdot \binom{m}{i} \cdot i^n .$$

4.3 Das Prinzip von Inklusion und Exklusion

Es seien A und B wieder endliche Mengen mit $|A| = n$ und $|B| = m$. Nach Satz 1.1-1 (v) ist $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$ eine disjunkte Zerlegung von $A \cup B$. Daraus folgt für die Mächtigkeit von $A \cup B$:

$$|A \cup B| = |A \setminus B| + |A \cap B| + |B \setminus A|.$$

Es sei $|A \cap B| = i$. Dann ist

$$\begin{aligned}|A \cup B| &= |A \setminus B| + |A \cap B| + |B \setminus A| \\ &= (n-i) + i + (m-i) \\ &= n + m - i \\ &= |A| + |B| - |A \cap B|.\end{aligned}$$

Bei drei Mengen *A*, *B* und *C* mit $|A| = n$, $|B| = m$ und $|C| = k$ lauten die entsprechenden Formeln:

Die Menge $A \cup B \cup C$ lässt sich disjunkt zerlegen in

$$\begin{aligned}A \cup B \cup C = &((A \setminus B) \setminus C) \cup ((B \setminus A) \setminus C) \cup ((C \setminus A) \setminus B) \\ &\cup ((A \cap B) \setminus C) \\ &\cup ((B \cap C) \setminus A) \\ &\cup ((A \cap C) \setminus B) \\ &\cup (A \cap B \cap C).\end{aligned}$$

Mit $i_1 = |(A \cap B) \setminus C|$, $i_2 = |(B \cap C) \setminus A|$, $i_3 = |(A \cap C) \setminus B|$ und $i_4 = |A \cap B \cap C|$ folgt daraus:

$$\begin{aligned}|A \cup B \cup C| &= n - (i_1 + i_3 + i_4) + m - (i_1 + i_2 + i_4) + k - (i_2 + i_3 + i_4) \\ &\quad + i_1 + i_2 + i_3 + i_4 \\ &= n + m + k - (i_1 + i_4) - (i_2 + i_4) - (i_3 + i_4) + i_4 \\ &= |A| + |B| + |C| - |A \cap B| - |B \cap C| - |A \cap C| + |A \cap B \cap C|.\end{aligned}$$

Zur Berechnung von $|A \cup B \cup C|$ wäre $|A| + |B| + |C|$ korrekt, wenn die drei Mengen disjunkt wären (Inklusion). Andernfalls zählt man die Elemente, die in jeweils zwei Mengen liegen, doppelt, und man muss die Anzahl der zwei Mengen gemeinsamen Elemente wieder abziehen (Exklusion). Allerdings hat man dadurch die Elemente „vergessen“, die in allen drei Mengen liegen, so dass diese wieder hinzugezählt werden müssen (Inklusion). Dieses **Prinzip der Inklusion und Exklusion** lässt sich auf eine endliche Anzahl von Mengen erweitern.

Es seien $A_1, ..., A_l$ Teilmengen einer endlichen Menge *M*. Mit $\bigcup_{i=1}^{l} A_i$ wird $A_1 \cup ... \cup A_l$ abgekürzt; mit $\bigcap_{i \in I} A_i$ mit $I \subseteq \{1, ..., l\}$ wird der Schnitt derjenigen Mengen A_i bezeichnet, für deren Index $i \in I$ gilt.

Satz 4.3-1:

Es seien A_1, ..., A_l Teilmengen einer endlichen Menge M, $l \geq 1$ Dann gilt:

$$\left|\bigcup_{i=1}^{l} A_i\right| = \sum_{\substack{I \subseteq \{1,...,l\} \\ \text{und } I \neq \varnothing}} (-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right|.$$

Die Summation geht über alle nichtleeren Teilmengen I der Indexmenge $\{1,...,l\}$.

Der Satz kann durch vollständige Induktion bewiesen werden. Zuvor soll er auf den obigen Fall dreier Mengen angewendet werden ($A_1 = A$, $A_2 = B$ und $A_3 = C$). Hierbei ist $\{1,...,l\} = \{1,2,3\}$. Alle nichtleeren Teilmengen der Indexmenge sind

$\{1\}$, $\{2\}$, $\{3\}$, $\{1,2\}$, $\{1,3\}$, $\{2,3\}$ und $\{1,2,3\}$.

Die einzelnen Summanden lauten

für $I = \{1\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right| = (-1)^{1+1} \cdot |A_1| = |A_1|$,

für $I = \{2\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right| = (-1)^{1+1} \cdot |A_2| = |A_2|$,

für $I = \{3\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right| = (-1)^{1+1} \cdot |A_3| = |A_3|$,

für $I = \{1,2\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right| = (-1)^{1+2} \cdot |A_1 \cap A_2| = -|A_1 \cap A_2|$,

für $I = \{1,3\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right| = (-1)^{1+2} \cdot |A_1 \cap A_3| = -|A_1 \cap A_3|$,

für $I = \{2,3\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right| = (-1)^{1+2} \cdot |A_2 \cap A_3| = -|A_2 \cap A_3|$,

für $I = \{1,2,3\}$: $(-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right| = (-1)^{1+3} \cdot |A_1 \cap A_2 \cap A_3| = |A_1 \cap A_2 \cap A_3|$.

Der Beweis der Formel in Satz 4.3-1 erfolgt durch vollständige Induktion über die Anzahl l der beteiligten Teilmengen:

Der Induktionsanfang für $l = 1$, $l = 2$ und $l = 3$ ist offensichtlich bzw. wurde in den obigen Beispielen gezeigt. Die Formel gelte für $l \geq 3$. Zu zeigen ist, dass aus dieser Annahme ihre Gültigkeit auch für $l + 1$ folgt.

Es sei $A = \bigcup_{i=1}^{l} A_i$. Dann ist gemäß Induktionsanfang

$$\left|\bigcup_{i=1}^{l+1} A_i\right| = \left|A \cup A_{l+1}\right| = \left|A\right| + \left|A_{l+1}\right| - \left|A \cap A_{l+1}\right|.$$

Nach Induktionsannahme ist

$$\left|A\right| = \left|\bigcup_{i=1}^{l} A_i\right| = \sum_{\substack{I \subseteq \{1,\ldots,l\} \\ \text{und } I \neq \varnothing}} (-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right|.$$

Die in Satz 1.1-1 (vi) formulierten Distributivgesetze gelten auch für mehr als drei Mengen, so dass gilt:

$$A \cap A_{l+1} = \left(\bigcup_{i=1}^{l} A_i\right) \cap A_{l+1} = \bigcup_{i=1}^{l} \left(A_i \cap A_{l+1}\right).$$

Hierbei handelt es sich also um die Vereinigung von *l* Mengen der Form $A_i \cap A_{l+1}$, so dass die Induktionsannahme anwendbar ist:

$$\left|A \cap A_{l+1}\right| = \left|\bigcup_{i=1}^{l} \left(A_i \cap A_{l+1}\right)\right| = \sum_{\substack{I \subseteq \{1,\ldots,l\} \\ \text{und } I \neq \varnothing}} (-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} \left(A_i \cap A_{l+1}\right)\right|.$$

Insgesamt ergibt sich

$$\left|\bigcup_{i=1}^{l+1} A_i\right| = \sum_{\substack{I \subseteq \{1,\ldots,l\} \\ \text{und } I \neq \varnothing}} (-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} A_i\right| + \left|A_{l+1}\right| - \sum_{\substack{I \subseteq \{1,\ldots,l\} \\ \text{und } I \neq \varnothing}} (-1)^{1+|I|} \cdot \left|\bigcap_{i \in I} \left(A_i \cap A_{l+1}\right)\right|.$$

In der ersten Summe lautet die Bedingung des Laufindex „$I \subseteq \{1,\ldots,l\}$ und $I \neq \varnothing$“. Man verändert nichts, wenn man diese durch „$I \subseteq \{1,\ldots,l+1\}$ und $I \neq \varnothing$ und $l+1 \notin I$“ ersetzt. In der zweiten Summe lautet die Bedingung des Laufindex ebenfalls „$I \subseteq \{1,\ldots,l\}$ und $I \neq \varnothing$“, allerdings taucht als Summand nicht $\left|\bigcap_{i \in I} A_i\right|$, sondern $\left|\bigcap_{i \in I} \left(A_i \cap A_{l+1}\right)\right|$ auf, d.h. die durch den jeweiligen Laufindex *I* bestimmten Mengen A_i werden noch mit A_{l+1} geschnitten; man kann also zu jedem Laufindex *I* noch den Index $l+1$ hinzunehmen. In der zweiten Summe wird die Bedingung „$I \subseteq \{1,\ldots,l\}$ und $I \neq \varnothing$“ des Laufindex durch „$I \subseteq \{1,\ldots,l+1\}$ und $I \neq \varnothing$ und $l+1 \in I$“ ersetzt und die Summanden entsprechend ange-

passt; der Summand $\left|A_{l+1}\right|$ kann in die Summe mit aufgenommen werden ($I = \{l+1\}$). Damit wird die letzte Gleichung zu

$$\left|\bigcup_{i=1}^{l+1} A_i\right| = \sum_{\substack{I\subseteq\{1,\ldots,l+1\}\\ \text{und } I\neq\emptyset\\ \text{und } l+1\notin I}} (-1)^{1+|I|}\cdot\left|\bigcap_{i\in I} A_i\right| + \sum_{\substack{I\subseteq\{1,\ldots,l+1\}\\ \text{und } I\neq\emptyset\\ \text{und } l+1\in I}} (-1)^{1+|I|}\cdot\left|\bigcap_{i\in I} A_i\right|$$

$$= \sum_{\substack{I\subseteq\{1,\ldots,l+1\}\\ \text{und } I\neq\emptyset}} (-1)^{1+|I|}\cdot\left|\bigcap_{i\in I} A_i\right|.$$

Die Formel in Satz 4.3-1 gilt also auch für $l+1$ Teilmengen.

Eine direkte Folgerung aus Satz 4.3-1 ist der folgende

Satz 4.3-2:

Es seien $A_1, \ldots, A_l$ Teilmengen einer endlichen Menge M, $l \geq 1$ Dann ist die Anzahl der $x \in M$, die in keiner der Mengen $A_1, \ldots, A_l$ liegen, gleich

$$|M| + \sum_{\substack{I\subseteq\{1,\ldots,l\}\\ \text{und } I\neq\emptyset}} (-1)^{|I|}\cdot\left|\bigcap_{i\in I} A_i\right|.$$

Die Formel ergibt sich wie folgt:

$$\left|M \setminus \bigcup_{i=1}^{l} A_i\right| = |M| - \left|\bigcup_{i=1}^{l} A_i\right|$$

$$= |M| - \sum_{\substack{I\subseteq\{1,\ldots,l\}\\ \text{und } I\neq\emptyset}} (-1)^{1+|I|}\cdot\left|\bigcap_{i\in I} A_i\right| \quad \text{nach Satz 4.3-1}$$

$$= |M| + \sum_{\substack{I\subseteq\{1,\ldots,l\}\\ \text{und } I\neq\emptyset}} (-1)^{|I|}\cdot\left|\bigcap_{i\in I} A_i\right|.$$

Beide Sätze finden ihre Anwendung in vielen Teilen der Mathematik.

Als Beispiel dient der Beweis von Satz 4.2-4:

Es seien A und B endliche Mengen mit $|A| = n$ und $|B| = m$ und $n \geq m$, $A = \{a_1, \ldots, a_n\}$, $B = \{b_1, \ldots, b_m\}$. Die Mengen A_i werden definiert durch

$A_i = \{ f \mid f : A \to B$ und es gibt kein $a \in A$ mit $f(a) = b_i \}$ für $i = 1, ..., m$.

Eine Abbildung $f : A \to B$ ist damit genau dann surjektiv, wenn f in keiner der Mengen A_i für $i = 1, ..., m$ enthalten ist.

Es sei $M = \{ f \mid f : A \to B \}$. Nach Satz 4.3-1 ist $|M| = m^n$.

In der Terminologie von Satz 4.3-2 wird die Anzahl der $f \in M$ gesucht, die in keiner der Mengen A_1, ..., A_m liegen. Diese Anzahl ist

$$|M| + \sum_{\substack{I \subseteq \{1,...,m\} \\ \text{und } I \neq \varnothing}} (-1)^{|I|} \cdot \left| \bigcap_{i \in I} A_i \right| = m^n + \sum_{\substack{I \subseteq \{1,...,m\} \\ \text{und } I \neq \varnothing}} (-1)^{|I|} \cdot \left| \bigcap_{i \in I} A_i \right|.$$

Für $I \subseteq \{1, ..., m\}$ mit $I \neq \varnothing$ wird $\bigcap_{i \in I} A_i$ betrachtet. Für jedes $f \in \bigcap_{i \in I} A_i$ gilt: es gibt kein $a \in A$, das durch f auf b_i abgebildet wird, wobei $i \in I$ ist. Also ist f eine Abbildung

$$f : A \to B \setminus \{ b_i \mid i \in I \}.$$

Ist umgekehrt f eine Abbildung mit $f : A \to B \setminus \{ b_i \mid i \in I \}$, so ist $f \in \bigcap_{i \in I} A_i$. Daher ist nach Satz 4.2-1

$$\left| \bigcap_{i \in I} A_i \right| = \left| \{ f \mid f : A \to B \setminus \{ b_i \mid i \in I \} \} \right| = (m - |I|)^n.$$

Nach Satz 4.1-4 gibt es $\binom{m}{|I|}$ viele Teilmengen $I \subseteq \{1, ..., m\}$ mit Mächtigkeit $|I|$. Damit wird

$$
\begin{aligned}
|M| + \sum_{\substack{I \subseteq \{1,\ldots,m\} \\ \text{und } I \neq \emptyset}} (-1)^{|I|} \cdot \left| \bigcap_{i \in I} A_i \right| &= m^n + \sum_{\substack{I \subseteq \{1,\ldots,m\} \\ \text{und } I \neq \emptyset}} (-1)^{|I|} \cdot \left| \bigcap_{i \in I} A_i \right| \\
&= m^n + \sum_{\substack{I \subseteq \{1,\ldots,m\} \\ \text{und } I \neq \emptyset}} (-1)^{|I|} \cdot \left(m - |I|\right)^n \\
&= m^n + \sum_{k=1}^{m} (-1)^k \cdot \binom{m}{k} (m-k)^n \\
&= \sum_{k=0}^{m} (-1)^k \cdot \binom{m}{k} (m-k)^n \qquad \text{mit der Indextransformation } i = m-k \\
&= \sum_{i=0}^{m} (-1)^{m-i} \cdot \binom{m}{m-i} i^n \qquad \text{mit Satz 4.1-3 (vi)} \\
&= \sum_{i=0}^{m} (-1)^{m-i} \cdot \binom{m}{i} i^n .
\end{aligned}
$$

Das ist das Ergebnis aus Satz 4.2-4.

5 Ausgewählte Themen der Analysis

Die für praktische Anwendungen wichtigsten Themen der Analysis werden im mathematischen Schulunterricht behandelt. Daher kann das vorliegende Kapitel als Wiederholung und Vertiefung dieser Themen betrachtet werden.

5.1 Folgen und Reihen

Wird jedes $n \in \mathbf{N}$ nach einer bestimmten Vorschrift eine reelle Zahl $a_n \in \mathbf{R}$ zugeordnet, so entsteht eine reellwertige Zahlenfolge $a_0, a_1, a_2, \ldots$. Sie wird mit $(a_n)_{n\in\mathbf{N}}$ bezeichnet. a_n heißt auch ***n*-tes Folgenglied von** $(a_n)_{n\in\mathbf{N}}$.

In der Regel stellt a_n eine von n abhängige Formel dar. Beispielsweise ist für $a_n = \sqrt{n+1} - \sqrt{n}$:

$$(a_n)_{n\in\mathbf{N}} = \left(\sqrt{n+1} - \sqrt{n}\right)_{n\in\mathbf{N}}.$$

Die Definition einer Folge kann auf unterschiedliche Weise geschehen, wie das folgende Beispiel zeigt:

$$a_0 = 1,\ a_1 = 2,\ a_2 = 4,\ a_3 = 8, \ldots, a_n = 2^n, \ldots$$

oder

$$(a_n)_{n\in\mathbf{N}} = \left(2^n\right)_{n\in\mathbf{N}}$$

oder als **rekursive Definition**

$$a_0 = 1,\ a_n = 2 \cdot a_{n-1} \text{ für } n \geq 1.$$

Eine Zahl $a \in \mathbf{R}$ heißt **Grenzwert (Limes)** der Folge $(a_n)_{n\in\mathbf{N}}$, wenn folgender Sachverhalt gilt:

Für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ gibt es eine (eventuell von ε abhängige) natürliche Zahl $n_0 = n_0(\varepsilon)$ mit der Eigenschaft:

Für jedes $n \in \mathbf{N}$ mit $n \geq n_0$ gilt $|a_n - a| < \varepsilon$.

Die Folge heißt dann **gegen** *a* **konvergent** (sie **konvergiert gegen** *a*), und man schreibt $a = \lim\limits_{n\to\infty} a_n$ bzw. $a_n \to a$ für $n \to \infty$.

Die Konvergenz der Folge $(a_n)_{n\in\mathbf{N}}$ gegen den Wert $a \in \mathbf{R}$ bedeutet anschaulich, dass bei Vorgabe eines beliebig kleinen Werts $\varepsilon > 0$ *alle* Folgenglieder a_n, bis auf höchstens endlich viele Ausnahmen („*fast alle*“ Folgenglieder), „dicht“ bei *a*, genauer einen Abstand von *a* haben, der kleiner als ε ist. Verkleinert man ε auf den Wert $\varepsilon' < \varepsilon$, so steigt eventuell die Anzahl der Ausnahmen, die nicht dicht bei *a* liegen; es bleiben aber weiterhin höchstens endlich viele Ausnahmen.

Beispiel:

Die Folge $(a_n)_{n\in\mathbf{N}} = \left(\sqrt{n+1} - \sqrt{n}\right)_{n\in\mathbf{N}}$ konvergiert gegen 0: Gibt man $\varepsilon > 0$ vor und setzt $n_0 = n_0(\varepsilon) = \left\lceil \left(\frac{1}{2\cdot\varepsilon}\right)^2 \right\rceil + 1$, so gilt für $n \geq n_0$:

$$\begin{aligned}
\left|\left(\sqrt{n+1} - \sqrt{n}\right) - 0\right| &= \sqrt{n+1} - \sqrt{n} \\
&= \frac{\left(\sqrt{n+1} - \sqrt{n}\right)\left(\sqrt{n+1} + \sqrt{n}\right)}{\sqrt{n+1} + \sqrt{n}} \\
&= \frac{1}{\sqrt{n+1} + \sqrt{n}} \\
&< \frac{1}{2\cdot\sqrt{n}} \qquad \text{(wegen } \sqrt{n+1} > \sqrt{n}\text{)} \\
&\leq \frac{1}{2\cdot\sqrt{n_0}} \\
&< \frac{\sqrt{(2\cdot\varepsilon)^2}}{2} = \varepsilon \qquad \text{(nach Wahl von } n_0\text{)}.
\end{aligned}$$

Wie man sieht, ist es nicht immer ganz leicht, bei Vorgabe von $\varepsilon > 0$ die passende Zahl $n_0(\varepsilon)$ zu finden, von der an alle Folgenglieder dicht beim Grenzwert *a* liegen, den man zudem be-

reits kennen oder vermuten muss. Häufig berechnet man die Werte a_n einiger Folgenglieder für wachsende Werte von *n*. Falls man sieht, dass sich diese Werte einer Größe *a* annähern, stellt man die allgemeine Ungleichung $|a_n - a| < \varepsilon$ auf und versucht diese nach *n* aufzulösen. Ergibt sich dabei eine Ungleichung der Form „$n > \ldots$“, dann kann man als $n_0(\varepsilon)$ die kleinste natürliche Zahl wählen, von der an die Ungleichung „$n > \ldots$“ gilt:

Beispiel:

Für die Folge $(a_n)_{n\in\mathbf{N}} = \left(\frac{2^n + (-1)^n}{2^n}\right)_{n\in\mathbf{N}}$ werden zunächst einige Werte berechnet:

n	0	1	2	3	4	5	10	15	100
a_n	2	1/2	5/4	7/8	17/16	31/32	1025/1024	32 767/32 768	1,0000000000000000000000000000008

Die Vermutung liegt nahe, dass diese Folge gegen $a = 1$ konvergiert, also wird versucht, die Ungleichung $|a_n - 1| < \varepsilon$ nach *n* aufzulösen:

$$|a_n - 1| = \left|\frac{2^n + (-1)^n}{2^n} - 1\right| = \left|\frac{(-1)^n}{2^n}\right| = \frac{1}{2^n} < \varepsilon \text{, d.h. } 2^n > 1/\varepsilon .$$

Man kann die nächstgrößere 2er-Potenz oberhalb $1/\varepsilon$ wählen, und n_0 ist dann der Exponent in dieser 2er-Potenz. Für jedes $n \geq n_0$ ist dann $2^n \geq 2^{n_0} > 1/\varepsilon$ bzw. $|a_n - 1| = 1/2^n \leq 1/2^{n_0} < \varepsilon$.

Eine Folge, die nicht konvergent ist, also keinen Grenzwert hat, heißt **divergent**.

Für eine divergente Folge gilt
entweder:

(i) es gibt mindestens zwei reelle Zahlen $a \in \mathbf{R}$ und $b \in \mathbf{R}$ mit $a \neq b$, so dass es sowohl beliebig dicht bei *a* als auch beliebig dicht bei *b* unendlich viele Folgenglieder gibt

oder:

(ii) die Werte a_n werden mit wachsendem *n* beliebig groß bzw. beliebig klein (Schreibweise dann $a_n \to \infty$, $\lim\limits_{n\to\infty} a_n = \infty$, bzw. $a_n \to -\infty$, $\lim\limits_{n\to\infty} a_n = -\infty$).

Im Fall (i) sagt man, die Folge habe mindestens zwei verschiedene (reelle) **Häufungspunkte**.

Beispielsweise besitzt die Folge $(a_n)_{n\in\mathbf{N}} = \left((-1)^n\right)_{n\in\mathbf{N}}$ die beiden Häufungspunkte $+1$ und -1 und ist daher nicht konvergent.

Eine gegen eine reelle Zahl a konvergente Folge besitzt nur den einzigen Häufungspunkt a, der dann auch Grenzwert der Folge ist.

Satz 5.1-1:

(i) Jede Folge $(a_n)_{n\in\mathbf{N}}$, die gegen einen Grenzwert konvergiert, ist beschränkt, d.h. es gibt eine Konstante $C \in \mathbf{R}$ mit $|a_n| < C$ für alle $n \in \mathbf{N}$.

(ii) Konvergiert die Folge $(a_n)_{n\in\mathbf{N}}$ gegen $a \in \mathbf{R}$, so ist a eindeutig bestimmt.

Die Limesbildung und das Rechnen mit arithmetischen Ausdrücken ist häufig miteinander vertauschbar. So kann man den Grenzwert einer Folge, deren Folgenglieder sich als arithmetischer Ausdruck (gebildet mit den Operatoren +, − , · und /) von Folgenglieder konvergenter Folgen darstellen lassen, dadurch berechnen, dass man die Limesbildung in den arithmetischen Ausdruck hineinzieht: Man berechnet die Grenzwerte der einzelnen Teile und verknüpft diese dann gemäß dem arithmetischen Ausdruck. Konstante Faktoren, die nicht von n abhängen, kann man jeweils vor den Limes ziehen. Die Grundlage dieses Kalküls liefert der folgende Satz.

Satz 5.1-2:

Es seien $(a_n)_{n\in\mathbf{N}}$ bzw. $(b_n)_{n\in\mathbf{N}}$ zwei konvergente Folgen mit den Grenzwerten *a* bzw. *b*. Dann gilt:

(i) $\lim\limits_{n\to\infty}(a_n \pm b_n) = a \pm b$,
$\lim\limits_{n\to\infty}(a_n \cdot b_n) = a \cdot b$.

(ii) Für $r \in \mathbf{R}$ ist $\lim\limits_{n\to\infty}(r \cdot a_n) = r \cdot \lim\limits_{n\to\infty} a_n = r \cdot a$.

(iii) Gilt $b \neq 0$ und $b_n \neq 0$ für alle $n \in \mathbf{N}$, so ist $\lim\limits_{n\to\infty}\left(\frac{a_n}{b_n}\right) = \frac{a}{b}$.

(iv) Aus $\lim\limits_{n\to\infty} a_n = a$ kann man $\lim\limits_{n\to\infty}|a_n| = |a|$ schließen.
Für $a = 0$ gilt auch die Umkehrung: $\lim\limits_{n\to\infty}|a_n| = 0$ impliziert $\lim\limits_{n\to\infty} a_n = 0$.

(v) Jede (fast überall) konstante Folge $(a_n)_{n\in\mathbf{N}}$ konvergiert, genauer:
Ist $a_n = a$ für (fast) alle $n \in \mathbf{N}$, so ist
$\lim\limits_{n\to\infty} a_n = a$.

„Ähnlich“ aussehende Folgen verhalten sich bezüglich der Konvergenz häufig sehr unterschiedlich: So ist die durch

$$a_n = \frac{2^n + (-2)^n}{2^n}$$

definierte Folge $(a_n)_{n\in\mathbf{N}}$ nicht konvergent. Dagegen konvergiert die durch

$$b_n = \frac{2^n + (-2)^n}{3^n}$$

definierte Folge $(b_n)_{n\in\mathbf{N}}$ gegen 0.

Häufig ist das Konvergenzverhalten einer Folge $(a_n)_{n\in\mathbf{N}}$ zu untersuchen. Ist eine Zahl $a \in \mathbf{R}$ „verdächtig“, Grenzwert der Folge $(a_n)_{n\in\mathbf{N}}$ zu sein, so lässt sich durch Rückgriff auf die Definition des Limesbegriffs nachprüfen, ob die Folge tatsächlich konvergiert, und zwar gegen a, d.h. ob $\lim\limits_{n\to\infty} a_n = a$ gilt. Kann man einer Folge, ohne von ihrem möglichen Grenzwert etwas zu wissen, ansehen, dass sie konvergiert? Die folgenden Sätze liefern einige **Konvergenzkriterien für Folgen**. Daneben gibt es eine Reihe weiterer Konvergenzkriterien, die mit Hinweis auf die angegebene Literatur hier nicht angeführt werden.

Satz 5.1-3:

Es seien $(a_n)_{n\in\mathbf{N}}$ bzw. $(b_n)_{n\in\mathbf{N}}$ zwei konvergente Folgen mit den Grenzwerten a bzw. b. Dann gilt:

(i) Es seien $(a_n)_{n\in\mathbf{N}}$ und $(b_n)_{n\in\mathbf{N}}$ zwei konvergente Folgen mit demselben Grenzwert $a \in \mathbf{R}$. Für fast alle Folgenglieder c_n der Folge $(c_n)_{n\in\mathbf{N}}$ gelte $a_n \le c_n \le b_n$. Dann konvergiert auch die Folge $(c_n)_{n\in\mathbf{N}}$, und zwar zum selben Grenzwert a.

(ii) Jede beschränkte und monoton wachsende Folge konvergiert, und ihr Limes ist gleich der kleinsten oberen Schranke (**Supremum**) ihrer Wertemenge.

Jede beschränkte und monoton fallende Folge konvergiert, und ihr Limes ist gleich der größten unteren Schranke (**Infimum**) ihrer Wertemenge.

Eine unbeschränkte, monoton wachsende bzw. monoton fallende Folge strebt gegen ∞ bzw. $-\infty$.

Bemerkung: Nicht jede beschränkte Folge ist konvergent.

(iii) Jede Umordnung und jede Teilfolge einer konvergenten Folge ist ebenfalls konvergent mit demselben Grenzwert. Dasselbe gilt, wenn man endlich viele Folgenglieder einer konvergenten Folge abändert.

Das folgende **Konvergenzkriterium von Cauchy** gibt eine notwendige und hinreichende Eigenschaft der Konvergenz einer Folge an:

Satz 5.1-4:

Eine Folge $(a_n)_{n\in\mathbf{N}}$ ist genau dann konvergent, wenn es zu jedem $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ eine natürliche Zahl $n_0 = n_0(\varepsilon)$ gibt, so dass gilt:

$|a_n - a_m| < \varepsilon$ für jedes $n \in \mathbf{N}$ und jedes $m \in \mathbf{N}$ mit $n \geq n_0$ und $m \geq n_0$.

Der folgende Satz liefert einige Beispiele konvergenter und divergenter Folgen.

Satz 5.1-5:

(i) Es sei $q \in \mathbf{R}$. Dann gilt

$$\lim_{n\to\infty} q^n = \begin{cases} 0 & \text{für } -1 < q < 1 \\ 1 & \text{für } q = 1 \end{cases};$$

für $q > 1$ und $q \le -1$ ist $\left(q^n\right)_{n\in\mathbf{N}}$ divergent.

Ist $k \in \mathbf{N}$ und $|q| < 1$, so ist $\lim_{n\to\infty}(n^k q^n) = 0$.

(ii) Es sei $a \in \mathbf{R}$. Dann ist

$$\lim_{n\to\infty} n^a = \begin{cases} 0 & \text{für } a < 0 \\ 1 & \text{für } a = 0 \\ \infty & \text{für } a > 0 \end{cases}.$$

(iii) Für jedes $a \in \mathbf{R}$ ist

$$\lim_{n\to\infty} \frac{a^n}{n!} = 0.$$

(iv) $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e = 2.7182818284590\ldots;$

e heißt **Eulersche Konstante**.

(vi) Die Folge $\left(\sqrt{n}\right)_{n\in\mathbf{N}}$ divergiert: mit wachsendem *n* werden die Folgenglieder beliebig groß. Hingegen werden die Zuwächse von einem Folgenglied zum nächsten mit wachsendem *n* beliebig klein; denn die Folge

$$\left(\sqrt{n+1} - \sqrt{n}\right)_{n\in\mathbf{N}}$$

konvergiert gegen 0.

(vii) Die Folge $\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}\right)_{n\in\mathbf{N}}$ divergiert.

Das Beispiel in Satz 5.1-5 (vii) zeigt eine Folge in einer speziellen Form: Das *n*-te Folgenglied ist selbst eine Summe aus einer endlichen Anzahl von Summanden, die aus einer Folge stammen, die nach einer einheitlichen Gesetzmäßigkeit aufgebaut ist. Derartige Folgen sollen nun genauer betrachtet werden.

Zur Zahlenfolge $(a_n)_{n\in\mathbf{N}}$ wird eine neue Zahlenfolge $(s_n)_{n\in\mathbf{N}}$ durch

$$s_n = \sum_{i=0}^{n} a_i$$

definiert. Der Wert s_n heißt ***n*-te Partialsumme** von $(a_n)_{n\in\mathbf{N}}$:

$$s_0 = a_0,$$
$$s_1 = a_0 + a_1,$$
...
$$s_n = a_0 + a_1 + \ \ldots\ a_n = s_{n-1} + a_n.$$

Falls der Grenzwert $\lim\limits_{n\to\infty} s_n = \lim\limits_{n\to\infty}\left(\sum\limits_{i=0}^{n} a_i\right)$ existiert, so heißt er **unendliche Reihe der Folge** $(a_n)_{n\in\mathbf{N}}$ und wird mit

$$\sum_{i=0}^{\infty} a_i$$

bezeichnet. Gelegentlich schreibt man auch

$$\sum_{i=0}^{\infty} a_i = a_0 + a_1 + a_2 + \ \ldots$$

Satz 5.1-6:

Existiert der Grenzwert $\sum_{i=0}^{\infty} a_i$, so konvergiert die Folge $(a_n)_{n\in\mathbf{N}}$ gegen 0, d.h. aus $\sum_{i=0}^{\infty} a_i < \infty$ folgt $\lim_{n\to\infty} a_n = 0$.

Bemerkung: Die Umkehrung dieser Aussage gilt i.a. nicht, d.h. aus der Konvergenz der Folge $(a_n)_{n\in\mathbf{N}}$ gegen 0 folgt i.a. nicht die Existenz von $\sum_{i=0}^{\infty} a_i$, wie das Beispiel der Folge $(a_n)_{n\in\mathbf{N}} = \left(1/n\right)_{n\in\mathbf{N}}$ zeigt: die Folge $\left(1/n\right)_{n\in\mathbf{N}}$ konvergiert gegen 0, aber die Folge der *n*-ten Partialsummen $(s_n)_{n\in\mathbf{N}} = \left(\sum_{i\_1}^{n} 1/i\right)$ konvergiert nicht (Satz 5.1-5 (vii)).

Satz 5.1-7:

Es seien $\sum_{i=0}^{\infty} a_i$ und $\sum_{i=0}^{\infty} b_i$ konvergent. Dann gilt:

(i) $\sum_{i=0}^{\infty}(c\cdot a_i) = c\cdot\sum_{i=0}^{\infty} a_i$ für jedes $c \in \mathbf{R}$.

(ii) $\sum_{i=0}^{\infty}(a_i \pm b_i) = \sum_{i=0}^{\infty} a_i \pm \sum_{i=0}^{\infty} b_i$.

Zur **Berechnung** von $\sum_{i=0}^{\infty} a_i$ sind folgende Schritte erforderlich:

1. Schritt:

Man bildet die *n*-te Partialsumme $s_n = \sum_{i=0}^{n} a_i$. Falls möglich, findet man hierfür einen geschlossenen Ausdruck, der von *n* abhängt.

2. *Schritt:*

Man vollzieht den Grenzübergang $n \to \infty$ und erhält $\lim\limits_{n\to\infty} s_n = \lim\limits_{n\to\infty}\left(\sum\limits_{i=0}^{n} a_i\right) = \sum\limits_{i=0}^{\infty} a_i$.

Es sei $m \in \mathbf{N}$. Unter $\sum\limits_{i=m}^{\infty} a_i$ versteht man den Grenzwert der Folge $\left(\sum\limits_{i=m}^{n} a_i\right)_{n\in\mathbf{N},n\geq m}$.

Satz 5.1-8:

Falls $\sum\limits_{i=0}^{\infty} a_i$ existiert, so existiert auch $\sum\limits_{i=m}^{\infty} a_i$ für jedes $m \in \mathbf{N}$, und es gilt

$$\sum_{i=m}^{\infty} a_i = \sum_{i=0}^{\infty} a_i - \sum_{i=0}^{m-1} a_i \; .$$

Der folgende Satz liefert einige Beispiele:

Satz 5.1-9:

(i) Für $q \in \mathbf{R}$ mit $-1 < q < 1$ ist

$$\sum_{i=0}^{\infty} q^i = \frac{1}{1-q},$$

$$\sum_{i=1}^{\infty} q^i = \frac{q}{1-q},$$

$$\sum_{i=0}^{\infty} i q^i = \frac{q}{(1-q)^2}.$$

(ii) Die Reihen $\sum_{i=0}^{\infty} (-1)^i$ und $\sum_{i=1}^{\infty} \frac{1}{i}$ existieren nicht (sind divergent).

(iii) $$\sum_{i=1}^{\infty} (-1)^{i-1} \cdot \frac{1}{i} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots + (-1)^{n-1} \cdot \frac{1}{n} \pm \ldots = \ln(2)$$

$$\approx 0{,}6931471805599$$

$$\sum_{i=1}^{\infty} (-1)^{i-1} \cdot \frac{1}{2i-1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ \ldots\ + (-1)^{n-1} \cdot \frac{1}{2n-1} \pm \ \ldots\ = \frac{\pi}{4}$$

$$\approx\ 0{,}7853981633974$$

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \ \ldots\ \frac{1}{n^2} + \ \ldots\ = \frac{\pi^2}{6} \approx\ 1{,}644934066848$$

$$\sum_{i=1}^{\infty} \frac{1}{i \cdot (i+1)} = \sum_{i=2}^{\infty} \frac{1}{i \cdot (i-1)} = 1.$$

(iv) $\sum_{i=1}^{\infty} \frac{1}{i^{\alpha}}$ konvergiert für jedes $\alpha \in \mathbf{R}$ mit $\alpha > 1$.

Die rationalen Zahlen werden in Kapitel 1.4 durch

$$\mathbf{Q} = \left\{ m/n \mid m \in \mathbf{Z} \text{ und } n \in \mathbf{Z} \text{ und } n \neq 0 \right\}$$

definiert. Im folgenden wird eine Charakterisierung mit Hilfe ihrer Dezialbruchentwicklung gegeben.

Es sei $r \in \mathbf{R}$ eine reelle Zahl mit $0 \leq r < 1$, deren Dezimalbruchentwicklung nach endlich vielen Stellen nur noch aus den Ziffern 0 besteht, d.h. nach endlich vielen Stellen abbricht:

$$r = [0, d_{-1} d_{-2} \ldots d_{-m}]_{10} = \sum_{i=1}^{m} d_{-i} \cdot 10^{-i} \text{ mit } d_{-i} \in \{0,1,2,3,4,5,6,7,8,9\} \text{ für } i = 1, \ldots, m .$$

Da jeder Summand $d_{-i} \cdot 10^{-i} = \frac{d_{-i}}{10^i}$ eine rationale Zahl ist, gilt auch $r \in \mathbf{Q}$. In diesem Fall ist

$$r = [0, d_{-1} d_{-2} \ldots d_{-m}]_{10} = \frac{[d_{-1} d_{-2} \ldots d_{-m}]_{10}}{10^m}$$

(im Zähler steht die natürliche Zahl, deren Dezimalzahldarstellung aus der Ziffernfolge der Dezimalbruchentwicklung besteht, und im Nenner steht die Zahl, deren Dezimaldarstellung von links gelesen aus einer Ziffer 1, gefolgt von *m* Ziffern 0 besteht).

Es sei nun $r \in \mathbf{R}$ eine reelle Zahl mit $0 \leq r < 1$, deren Dezimalbruchentwicklung aus einem nichtperiodischen und einem periodischen Teil besteht, etwa

$$r = [0, d_{-1} d_{-2} \ldots d_{-m} \overline{d_{-m-1} d_{-m-2} \ldots d_{-m-k}} \ldots]_{10}$$

Hierbei wiederholt sich die Ziffernfolge $d_{-m-1} d_{-m-2} \ldots d_{-m-k}$ beliebig oft, und es kommen keine nichtperiodischen Abschnitte in der Dezimalziffernfolge mehr vor. Zu beachten ist, dass man mit der Periodennotation nur endlich viele Dezimalziffern notieren muss, obwohl die Zahl in ihrer Dezimaldarstellung unendlich viele Ziffern benötigt. Es ist

$$\begin{aligned} r &= [0, d_{-1} d_{-2} \ldots d_{-m} \overline{d_{-m-1} d_{-m-2} \ldots d_{-m-k}} \ldots]_{10} \\ &= [0, d_{-1} d_{-2} \ldots d_{-m}]_{10} + \left[0, \underbrace{00 \ldots 0}_{m-mal} \overline{d_{-m-1} d_{-m-2} \ldots d_{-m-k}} \ldots \right]_{10} . \end{aligned}$$

Der zweite Summand hat den Wert

$$\left[0,\underbrace{00\ldots0}_{m-mal}\overline{d_{-m-1}d_{-m-2}\ldots d_{-m-k}}\ldots\right]_{10}$$

$$= 1/10^m \cdot \left[0,\overline{d_{-m-1}d_{-m-2}\ldots d_{-m-k}}\ldots\right]_{10}$$

$$= 1/10^m \cdot \sum_{i=1}^{\infty} \frac{\left[d_{-m-1}d_{-m-2}\ldots d_{-m-k}\right]_{10}}{\left(10^k\right)^i} \qquad \text{(im Zähler die Dezimalzahl } d_{-m-1}d_{-m-2}\ldots d_{-m-k}\text{)}$$

$$= 1/10^m \cdot \left[d_{-m-1}d_{-m-2}\ldots d_{-m-k}\right]_{10} \cdot \sum_{i=1}^{\infty} \frac{1}{\left(10^k\right)^i}$$

$$= 1/10^m \cdot \left[d_{-m-1}d_{-m-2}\ldots d_{-m-k}\right]_{10} \cdot \frac{(1/10)^k}{1-(1/10)^k} \qquad \text{(nach Satz 5.1-9 (i))}$$

$$= 1/10^m \cdot \left[d_{-m-1}d_{-m-2}\ldots d_{-m-k}\right]_{10} \cdot \frac{1}{10^k-1}.$$

Insgesamt ergibt sich

$$r = \left[0,d_{-1}d_{-2}\ldots d_{-m}\overline{d_{-m-1}d_{-m-2}\ldots d_{-m-k}}\ldots\right]_{10}$$
$$= \frac{\left[d_{-1}d_{-2}\ldots d_{-m}\right]_{10}}{10^m} + \frac{\left[d_{-m-1}d_{-m-2}\ldots d_{-m-k}\right]_{10}}{10^m\cdot\left(10^k-1\right)},$$

(die Dezimaldarstellung der Zahl $10^m \cdot \left(10^k - 1\right)$ besteht von links gelesen aus *k* Ziffern 9, gefolgt von *m* Ziffern 0). Auch hier gilt wieder $r \in \mathbf{Q}$.

Bemerkung: Die Zahl $0,\overline{9}\ldots$ hat den Wert 1; denn mit $m=1$ und $k=1$ ist

$$0,\overline{9}\ldots = 0,9\overline{9}\ldots = \frac{9}{10} + \frac{9}{10\cdot(10-1)} = 1.$$

Es sei umgekehrt die rationale Zahl $r \in \mathbf{Q}$ mit $0 \le r < 1$ in gekürzter Form $r = \frac{a}{b}$ mit $a \in \mathbf{N}$ und $b \in \mathbf{N}$ mit $b \ge 1$ gegeben. Es ist $ggT(a,b) = 1$. Es werden drei Fälle unterschieden:

1. Fall: *b* hat die Form $b = 2^i \cdot 5^j$ mit $i + j = m \ge 1$.

   Dann teilt *b* die Zahl $2^{i+j} \cdot 5^{i+j} = 10^{i+j} = 10^m$, d.h. $10^m = b \cdot b'$ mit einer natürlichen Zahl b'. Der Bruch $\frac{a}{b}$ lässt sich erweitern zu $\frac{a}{b} = \frac{a \cdot b'}{10^m}$. Da $r = \frac{a}{b} < 1$ gilt, ist $0 \le a \cdot b' < 10^m$. Die Darstellung von $a \cdot b'$ als Dezimalzahl hat endliche Länge und lautet $a \cdot b' = \sum_{l=0}^{m-1} d_l \cdot 10^l$ mit den Dezimalziffern d_0, ..., d_{m-1}; zu beachten ist hierbei, dass $a \cdot b'$ als Dezimalziffernfolge $a \cdot b' = \left[d_{m-1}\ldots d_0\right]_{10}$ lautet. Damit ergibt sich

$$\frac{a}{b}=\frac{a\cdot b'}{10^m}=\frac{\sum_{l=0}^{m-1} d_l\cdot 10^l}{10^m}=\sum_{l=0}^{m-1} d_l\cdot 10^{l-m}=\sum_{l=1}^{m} d_{m-l}\cdot 10^{-l},$$

also die endliche nichtperiodische Dezimalbruchentwicklung $\frac{a}{b}=[0,d_{m-1}\ldots d_0]_{10}$.

2. Fall: $ggT(b,10)=1$.

Nach Satz 3.4-2 gilt in diesem Fall $10^{\phi(b)}\equiv 1 \pmod b$. Es gibt also eine kleinste natürliche Zahl $m\le\phi(b)$, so dass b die Zahl 10^m-1 teilt. Wie im 1. Fall lässt sich der Bruch $\frac{a}{b}$ erweitern zu $\frac{a}{b}=\frac{a\cdot b'}{10^m-1}$. Wegen $0\le a\cdot b'<10^m-1$ hat die Darstellung von $a\cdot b'$ als Dezimalzahl wieder endliche Länge und lautet $a\cdot b'=\sum_{l=0}^{m-1} d_l\cdot 10^l$ mit den Dezimalziffern d_0, ..., d_{m-1}; wieder ist zu beachten, dass $a\cdot b'$ als Dezimalziffernfolge $a\cdot b'=[d_{m-1}\ldots d_0]_{10}$ lautet. Nach Satz 5.1-9 (i) gilt $\frac{1}{10^m-1}=\sum_{i=1}^{\infty}\left((1/10)^m\right)^i$ und damit

$$\begin{aligned}
\frac{a}{b}&=\frac{a\cdot b'}{10^m-1}=\left(\sum_{l=0}^{m-1} d_l\cdot 10^l\right)\cdot\left(\sum_{i=1}^{\infty}10^{-m\cdot i}\right)\\
&=d_0\cdot\left(\sum_{i=1}^{\infty}10^{-m\cdot i}\right)+d_1\cdot 10\cdot\left(\sum_{i=1}^{\infty}10^{-m\cdot i}\right)+\ldots+d_{m-1}\cdot 10^{m-1}\cdot\left(\sum_{i=1}^{\infty}10^{-m\cdot i}\right)\\
&=\left(\sum_{i=1}^{\infty}d_0\cdot 10^{-m\cdot i}\right)+\left(\sum_{i=1}^{\infty}d_1\cdot 10^{-m\cdot i+1}\right)+\ldots+\left(\sum_{i=1}^{\infty}d_{m-1}10^{-m\cdot i+m-1}\right)\\
&=\sum_{i=1}^{\infty}\left(d_0\cdot 10^{-m\cdot i}+d_1\cdot 10^{-m\cdot i+1}+\ldots+d_{m-1}10^{-m\cdot i+m-1}\right)\\
&=\sum_{i=1}^{\infty}\left(\sum_{l=0}^{m-1} d_l\cdot 10^{l-m\cdot i}\right)\\
&=\sum_{i=1}^{\infty}\left(\sum_{l=0}^{m-1} d_l\cdot 10^{l-m-(i-1)\cdot m}\right)\\
&=\sum_{i=1}^{\infty}10^{-(i-1)\cdot m}\cdot\left(\sum_{l=0}^{m-1} d_l\cdot 10^{l-m}\right)\\
&=\sum_{i=1}^{\infty}10^{-(i-1)\cdot m}\cdot\left(\sum_{l=1}^{m} d_{m-l}\cdot 10^{-l}\right).
\end{aligned}$$

Diese Zahl in der Darstellung als Dezimalbruchentwicklung lautet

$$\frac{a}{b} = \left[0, \underbrace{d_{m-1} \dots d_0}_{i=1} \underbrace{d_{m-1} \dots d_0}_{i=2} \underbrace{d_{m-1} \dots d_0}_{i=3} \dots\right]_{10} = \left[0, \overline{d_{m-1} \dots d_0} \dots\right]_{10},$$

ist also ein rein periodischer Dezimalbruch.

3. Fall: b hat die Form $b = 2^i \cdot 5^j \cdot b'$ mit $i + j = m \geq 1$, $b' > 1$ und $ggT(b', 10) = 1$. Wie im 1. Fall lässt sich der Bruch $\frac{a}{b}$ erweitern zu $\frac{a}{b} = \frac{a \cdot 2^j \cdot 5^i}{2^{i+j} \cdot 5^{i+j} \cdot b'} = \frac{a \cdot 2^j \cdot 5^i}{10^m \cdot b'}$. Gilt $\frac{a \cdot 2^j \cdot 5^i}{b'} \geq 1$, dann kann den Zähler in der Form $a \cdot 2^j \cdot 5^i = k \cdot b' + a'$ mit $k \in \mathbf{N}$ und $0 \leq a' = (a \cdot 2^j \cdot 5^i) \bmod b' < b'$ schreiben, d.h. $\frac{a \cdot 2^j \cdot 5^i}{b'} = k + \frac{a'}{b'}$, $0 \leq \frac{a'}{b'} < 1$. Ist $\frac{a \cdot 2^j \cdot 5^i}{b'} < 1$, dann hat der Bruch ebenfalls die Form $\frac{a \cdot 2^j \cdot 5^i}{b'} = k + \frac{a'}{b'}$ (mit $k = 0$ und $a' = a \cdot 2^j \cdot 5^i$). In beiden Fällen ergibt sich aus $ggT(a, b) = 1$ und $ggT(b', 10) = 1$, dass auch $ggT(a', b') = 1$ ist; denn für $\frac{a \cdot 2^j \cdot 5^i}{b'} \geq 1$ folgt mit Satz 3.3-2: $ggT(a', b') = ggT((a \cdot 2^j \cdot 5^i) \bmod b', b') = ggT(a \cdot 2^j \cdot 5^i, b') = 1$, und für $\frac{a \cdot 2^j \cdot 5^i}{b'} < 1$ ist bereits $ggT(a', b') = ggT(a \cdot 2^j \cdot 5^i, b') = 1$. Die Dezimalzahldarstellung von k sei $k = \sum_{l=0}^{h-1} d_l \cdot 10^l = [d_{h-1} \dots d_0]_{10}$, die Darstellung von $\frac{a'}{b'}$ als Dezimalbruch gemäß dem 2. Fall sei $\frac{a'}{b'} = \left[0, \overline{d'_{n-1} \dots d'_0} \dots\right]_{10}$ Dann hat $\frac{a \cdot 2^j \cdot 5^i}{b'} = k + \frac{a'}{b'}$ die Dezimaldarstellung $\frac{a \cdot 2^j \cdot 5^i}{b'} = k + \frac{a'}{b'} = \left[d_{h-1} \dots d_0, \overline{d'_{n-1} \dots d'_0} \dots\right]_{10}$. Die Darstellung von $\frac{a}{b}$ als Dezimalbruch enthält genau dieselbe Ziffernfolge, nur ist das Komma um m Stellen nach links geschoben, d.h. die Dezimalbruchdarstellung von $\frac{a}{b}$ besteht aus einer endlichen nichtperiodischen Ziffernfolge, gefolgt von einer endlichen periodischen Ziffernfolge.

Die Betrachtung bezieht sich auf rationale Zahlen $r \in \mathbf{Q}$ mit $0 \leq r < 1$. Sie ist natürlich auf ganz **Q** erweiterbar:

$$\mathbf{Q} = \left\{z + r \mid z \in \mathbf{Z} \text{ und } (r = [0, d_{-1} d_{-2} \dots d_{-m}]_{10} \text{ oder } r = \left[0, d_{-1} d_{-2} \dots d_{-m} \overline{d_{-m-1} d_{-m-2} \dots d_{-m-k}} \dots\right]_{10})\right\},$$

d.h. **Q** besteht aus allen Zahlen, deren gebrochener Anteil in Dezimaldarstellung entweder nach endlich vielen Dezimalziffern abbricht (es folgen nur noch Nullen) oder unendlich periodisch endet. Irrationalen Zahlen haben demzufolge einen gebrochenen Anteil in Dezimaldarstellung, der unendlich und nichtperiodisch ist.

Eine Reihe $\sum_{i=0}^{\infty} a_i$ heißt **absolut konvergent**, wenn die Reihe $\sum_{i=0}^{\infty} |a_i|$ konvergiert.

Es lässt sich zeigen, dass jede absolut konvergente Reihe auch konvergiert, d.h. aus der Konvergenz von $\sum_{i=0}^{\infty} |a_i|$ folgt die Konvergenz von $\sum_{i=0}^{\infty} a_i$.

Unter bestimmten Voraussetzungen kann man Reihen miteinander multiplizieren, jedenfalls dann, wenn sie absolut konvergent sind. Die folgende (nichtmathematische) Darstellung liefert die Motivation für die etwas komplizierte Indizierung in den auftretenden Reihen. Zwei Reihen $\sum_{i=0}^{\infty} a_i$ und $\sum_{i=0}^{\infty} b_i$ werden miteinander multipliziert, indem die einzelnen Summanden nach der Summe ihrer Indizes sortiert werden:

$$(a_0 + a_1 + a_2 + a_3 + \ldots a_n + \ldots) \cdot (b_0 + b_1 + b_2 + b_3 + \ldots b_n + \ldots)$$

$$= \underbrace{a_0 \cdot b_0}_{\text{Indexsumme } 0} + \underbrace{a_0 \cdot b_1 + a_1 \cdot b_0}_{\text{Indexsumme } 1} + \underbrace{a_0 \cdot b_2 + a_1 \cdot b_1 + a_2 \cdot b_0}_{\text{Indexsumme } 2} + \ldots + \underbrace{\sum_{j=0}^{n} a_j \cdot b_{n-j}}_{\text{Indexsumme } n} + \ldots$$

Satz 5.1-10:

Sind die Reihen $\sum_{i=0}^{\infty} a_i = a$ und $\sum_{i=0}^{\infty} b_i = b$ absolut konvergent, so ist auch die Reihe $\sum_{i=0}^{\infty} \left(\sum_{k=0}^{i} a_k \cdot b_{i-k} \right)$ absolut konvergent, und es gilt

$$\sum_{i=0}^{\infty} \left(\sum_{k=0}^{i} a_k \cdot b_{i-k} \right) = \left(\sum_{i=0}^{\infty} a_i \right) \cdot \left(\sum_{i=0}^{\infty} b_i \right) = a \cdot b \, .$$

Im allgemeinen ist die Bestimmung des Konvergenzverhaltens einer Reihe nicht einfach, so dass sich die Frage nach **Konvergenzkriterien für Reihen** stellt. Ein „Negativkriterium“ liefert Satz 5.1-6: Falls für eine Reihe $\sum_{i=0}^{\infty} a_i$ die Folge $(a_n)_{n\in\mathbf{N}}$ *nicht* gegen 0 konvergiert, existiert auch der Grenzwert $\sum_{i=0}^{\infty} a_i$ nicht. Ein „Positivkriterium“ ergibt sich aus Satz 5.1-3: Ist die Folge $\left(\sum_{i=0}^{n} |a_i|\right)_{n\in\mathbf{N}}$ monoton wachsend und nach oben beschränkt, so konvergiert $\sum_{i=0}^{\infty} a_i$ absolut. Zwei weitere Konvergenzkriterien sind im folgenden Satz zusammengefasst.

Satz 5.1-11:

(i) (**Majorantenkriterium**)

Ist $\sum_{i=0}^{\infty} b_i$ absolut konvergent und gilt $|a_i| \le |b_i|$ für (fast) alle $i \in \mathbf{N}$, so ist auch die Reihe $\sum_{i=0}^{\infty} a_i$ absolut konvergent.

(ii) (**Quotientenkriterium**)

Die Reihe $\sum_{i=0}^{\infty} a_i$ besitze die Eigenschaft, dass ab einem Index $n_0 \in \mathbf{N}$ stets $\left|\frac{a_{n+1}}{a_n}\right| \le q$ für einen Wert q mit $0 < q < 1$ gilt. Dann ist die Reihe $\sum_{i=0}^{\infty} a_i$ absolut konvergent. Ist ab einem Index stets $\left|\frac{a_{n+1}}{a_n}\right| > 1$, so ist die Reihe divergent.

Es sei $x \in \mathbf{R}$. Die Reihe $\sum_{i=0}^{\infty} \frac{x^i}{i!}$ ist absolut konvergent; denn nach Satz 5.1-11 (ii) ist

$$\left|\frac{x^{n+1}/(n+1)!}{x^n/n!}\right| = \left|\frac{x}{n+1}\right| = \frac{|x|}{n+1} < 1/2$$

für jedes $n \in \mathbf{N}$ mit $n > 2\cdot|x| - 1$. Daher ist die Definition der folgenden Funktion sinnvoll.

Die Funktion

$$\exp : \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & \sum_{i=0}^{\infty} \frac{x^i}{i!} \end{cases}$$

heißt **Exponentialfunktion**. Wichtige Eigenschaften der Exponentialfunktion und verwandter Funktionen werden in Kapitel 5.5 behandelt.

In Satz 5.1-5 (iv) wurde die Eulersche Konstante als $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e = 2.7182818284590\ldots$ definiert. Wie man direkt nachrechnet, gilt $\binom{n}{k}\cdot\frac{1}{n^k} \le \frac{1}{k!}$. Daher ist (siehe Kapitel 4.1)

$$\left(1+\frac{1}{n}\right)^n = \sum_{k=0}^{n}\binom{n}{k}\cdot\frac{1}{n^k} \le \sum_{k=0}^{n}\frac{1}{k!} .$$

Daraus folgt $e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \le \lim_{n\to\infty}\left(\sum_{k=0}^{n}\frac{1}{k!}\right) = \sum_{k=0}^{\infty}\frac{1}{k!}$. Es sei umgekehrt $n > m \ge 1$. Dann ist

$$\left(1+\frac{1}{n}\right)^n > \sum_{k=0}^{m}\binom{n}{k}\cdot\frac{1}{n^k} = \sum_{k=0}^{m}\left(\frac{1}{k!}\cdot\frac{n}{n}\cdot\frac{n-1}{n}\cdot\ldots\cdot\frac{n-k+1}{n}\right) = \sum_{k=0}^{m}\left(\frac{1}{k!}\cdot 1\cdot\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\right).$$

Durch Limesbildung folgt $e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \ge \sum_{k=0}^{m}\frac{1}{k!}$ und durch erneute Limesbildung $e \ge \lim_{m\to\infty}\sum_{k=0}^{m}\frac{1}{k!} = \sum_{k=0}^{\infty}\frac{1}{k!}$. Insgesamt ergibt sich

$$e = \sum_{i=0}^{\infty}\frac{1}{i!} = \exp(1) .$$

Die Exponentialfunktion, die durch die Reihe $\sum_{i=0}^{\infty}\frac{x^i}{i!}$ definiert wird, kann man durch eine endliche Summe und einen Restfehlerterm, dessen Größe man abschätzen kann, darstellen:

Satz 5.1-12:

Es sei $n \in \mathbf{N}$ und $x \in \mathbf{R}$ mit $|x| \leq 1 + \frac{n}{2}$. Dann gilt:

$$\exp(x) = \sum_{i=0}^{\infty} \frac{x^i}{i!} = \sum_{i=0}^{n} \frac{x^i}{i!} + R_{n+1}(x) \text{ mit } |R_{n+1}(x)| \leq \frac{2 \cdot |x|^{n+1}}{(n+1)!}.$$

Für $x = 1$ und $n = 10$ liefert Satz 5.1-12 die Abschätzung $R_{11}(1) < 0{,}00000006$ und damit

$2{,}7182816 \leq e \leq 2{,}718282$.

5.2 Eigenschaften reeller Funktionen einer Veränderlichen

Im vorliegenden Kapitel werden eine Reihe wichtiger Definitionen vorgestellt, die Eigenschaften reeller Funktionen einer Veränderlichen beschreiben.

Zur Erinnerung:

Es seien $a \in \mathbf{R}$ und $b \in \mathbf{R}$ reelle Zahlen mit $a \leq b$. Die Menge

$[a,b] = \{ x \mid x \in \mathbf{R} \text{ und } a \leq x \leq b \}$ heißt **abgeschlossenes Intervall** von a bis b,
$]a,b[= \{ x \mid x \in \mathbf{R} \text{ und } a < x < b \}$ heißt **offenes Intervall** von a bis b,
$[a,b[= \{ x \mid x \in \mathbf{R} \text{ und } a \leq x < b \}$ heißt **halboffenes Intervall** von a bis b,
$]a,b] = \{ x \mid x \in \mathbf{R} \text{ und } a < x \leq b \}$ heißt **halboffenes Intervall** von a bis b.

Unter einem **Intervall** wird ein abgeschlossenes oder offenes oder halboffenes Intervall verstanden. Zusätzlich zu den Intervallen mit reellwertigen Begrenzungspunkten werden folgende Intervalle definiert:

$]-\infty,a] = \{ x \mid x \in \mathbf{R} \text{ und } x \leq a \}$,
$]-\infty,a[= \{ x \mid x \in \mathbf{R} \text{ und } x < a \}$,
$[a,\infty[= \{ x \mid x \in \mathbf{R} \text{ und } a \leq x \}$,
$]a,\infty[= \{ x \mid x \in \mathbf{R} \text{ und } a < x \}$ und

$]-\infty, \infty[= \mathbf{R}$.

Im folgenden sei $f: X \to \mathbf{R}$, $X \subseteq \mathbf{R}$, und $I \subseteq \mathbf{R}$ sei ein Intervall.

f heißt **auf *I* monoton steigend** (bzw. **monoton fallend)**, wenn für $x_1 \in I$ und $x_2 \in I$ gilt:

Ist $x_1 < x_2$, so ist $f(x_1) \leq f(x_2)$
(bzw.
ist $x_1 < x_2$, so ist $f(x_1) \geq f(x_2)$).

Der Graph einer monoton steigenden Funktion fällt also mit wachsenden *x*-Werten nicht ab; der Graph einer monoton fallenden Funktion steigt also mit wachsenden *x*-Werten nicht.

f heißt **auf *I* streng monoton steigend** (bzw. **streng monoton fallend**), wenn für $x_1 \in I$ und $x_2 \in I$ gilt:

Ist $x_1 < x_2$, so ist $f(x_1) < f(x_2)$
(bzw.
ist $x_1 < x_2$, so ist $f(x_1) > f(x_2)$).

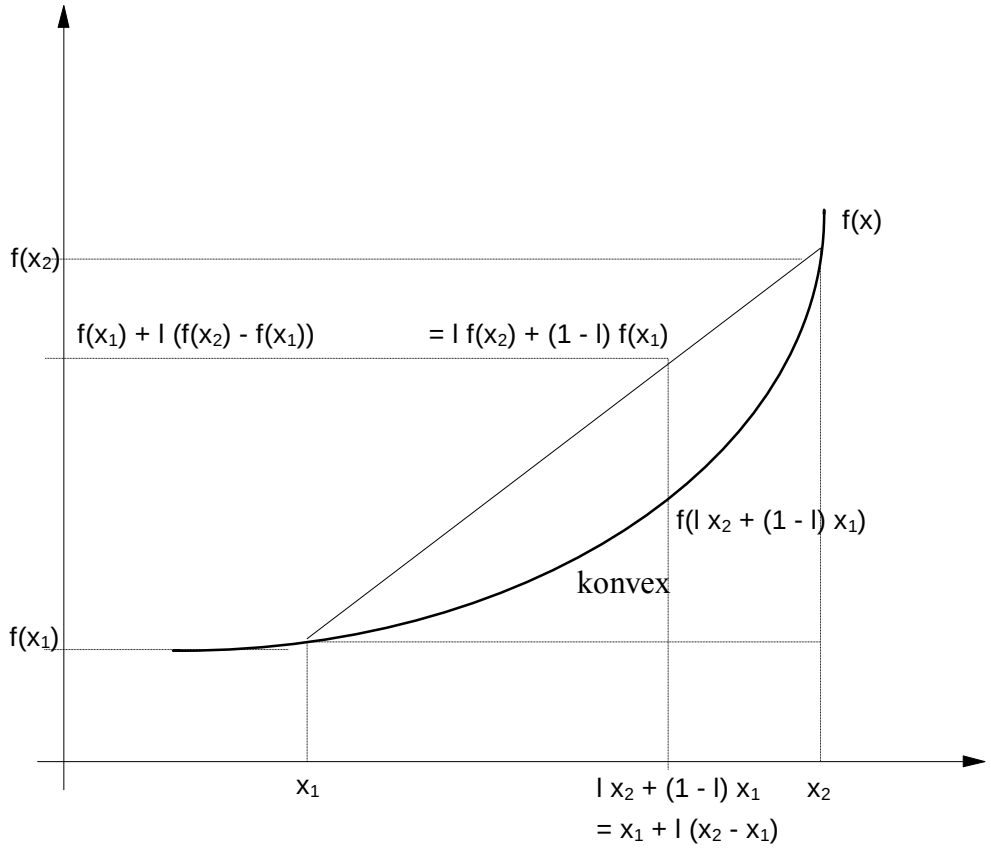
f heißt **auf *I* beschränkt**, wenn es ein $c \in \mathbf{R}_{\geq 0}$ gibt, so dass für jedes $x \in I$ gilt: $|f(x)| \leq c$.

Der Graph einer beschränkten Funktion verläuft also weder oberhalb von *c* noch unterhalb von $-c$.

f heißt **auf *I* nach oben beschränkt** (bzw. **nach unten beschränkt**), wenn es ein $c \in \mathbf{R}$ gibt, so dass für jedes $x \in I$ gilt: $f(x) \leq c$ bzw. $f(x) \geq c$.

f heißt **konvex über *I***, wenn für $x_1 \in I$ und $x_2 \in I$ mit $x_1 \neq x_2$ und für jedes $l \in \mathbf{R}$ mit $0 < l < 1$ gilt:

$$f(l \cdot x_2 + (1-l) \cdot x_1) \leq l \cdot f(x_2) + (1-l) \cdot f(x_1).$$



Nimmt man also zwei beliebige verschiedene Werte $x_1 \in I$ und $x_2 \in I$ und verbindet die Punkte $(x_1, f(x_1))$ und $(x_2, f(x_2))$ des Graphen einer über *I* konvexen Funktion durch eine gerade Linie, so verläuft der Graph zwischen $(x_1, f(x_1))$ und $(x_2, f(x_2))$ unterhalb dieser Verbindungslinie. Betrachtet man diese Verbindungslinie als Annäherung an den Graphen der Funktion zwischen $(x_1, f(x_1))$ und $(x_2, f(x_2))$, so macht man einen Approximationsfehler in Richtung größerer Werte, d.h. die Approximation liefert zu große Werte.

f heißt **konkav über *I***, wenn für $x_1 \in I$ und $x_2 \in I$ mit $x_1 \neq x_2$ und für jedes $l \in \mathbf{R}$ mit $0 < l < 1$ gilt:

$$f(l \cdot x_2 + (1-l) \cdot x_1) \geq l \cdot f(x_2) + (1-l) \cdot f(x_1).$$

Bei einer konkaven Funktion verläuft der Graph oberhalb der entsprechenden Verbindungslinie. Betrachtet man auch hier wieder die Verbindungslinie zwischen den Punkten $(x_1, f(x_1))$ und $(x_2, f(x_2))$ als Annäherung an den Graphen der Funktion, so liefert sie hier zu kleine Werte.

f heißt **streng konvex über *I***, wenn für $x_1 \in I$ und $x_2 \in I$ mit $x_1 \neq x_2$ und für jedes $l \in \mathbf{R}$ mit $0 < l < 1$ gilt:

$$f(l \cdot x_2 + (1-l) \cdot x_1) < l \cdot f(x_2) + (1-l) \cdot f(x_1).$$

f heißt **streng konkav über *I***, wenn für $x_1 \in I$ und $x_2 \in I$ mit $x_1 \neq x_2$ und für jedes $l \in \mathbf{R}$ mit $0 < l < 1$ gilt:

$$f(l \cdot x_2 + (1-l) \cdot x_1) > l \cdot f(x_2) + (1-l) \cdot f(x_1).$$

Eine Funktion kann in Teilintervallen ihres Definitionsbereichs (streng) konvex und in anderen Teilintervallen (streng) konkav sein.

Die Funktion $f: X \to \mathbf{R}$, $X \subseteq \mathbf{R}$ heißt **stetig im Punkt** $x_0 \in X$, wenn gilt:
Für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ gibt es ein $\delta > 0$, das von ε und x_0 abhängen kann (d.h. $\delta = \delta(\varepsilon, x_0)$), mit folgender Eigenschaft:

Für jedes $x \in X$ mit $|x - x_0| < \delta$ ist $|f(x) - f(x_0)| < \varepsilon$.

Die Funktion $f: X \to \mathbf{R}$ heißt **stetig in** $D \subseteq X$, wenn f in jedem Punkt $x_0 \in D$ stetig ist.

Für einen Wert $x \in \mathbf{R}$ und ein $\varepsilon > 0$ bezeichnet man das offene Intervall

$$U_\varepsilon(x) = \{ z \mid x - \varepsilon < z < x + \varepsilon \} = \{ z \mid |x - z| < \varepsilon \}$$

als ε**-Umgebung** von *x*. Mit dieser Bezeichnung bedeutet die Stetigkeit einer Funktion $f: X \to \mathbf{R}$ in einem Punkt $x_0 \in X$:

Zu jeder ε-Umgebung $U(f(x_0), \varepsilon)$ von $f(x_0)$ gibt es eine (von ε und x_0 abhängige) δ-Umgebung $U(x_0, \delta)$ von x_0, die durch f ganz nach $U(f(x_0), \varepsilon)$ abgebildet wird, d.h. für die

$$f(U(x_0, \delta)) \subseteq U(f(x_0), \varepsilon)$$

gilt. Anschaulich heißt dieses, dass für ein Argument *x*, das sich „nahe bei“ x_0 befindet (in der δ-Umgebung $U(x_0, \delta)$ von x_0), der Funktionswert $f(x)$ „nahe bei“ $f(x_0)$ liegt (in der ε-Umgebung $U(f(x_0), \varepsilon)$ von $f(x_0)$). Eine „sehr kleine Änderung“ des Arguments, d.h. der Übergang von x_0 zu *x* mit $|x - x_0| < \delta$, führt zu einer „sehr kleinen Änderung“ von $f(x_0)$, d.h. der Funktionswert $f(x)$ erfüllt $|f(x) - f(x_0)| < \varepsilon$. Insbesondere macht der Graph der Funktion an der Stelle bzw. „nahe“ der Stelle x_0 keinen Sprung. Graphen stetiger Funktionen lassen sich in einem Zuge zeichnen, ohne den Zeichenstift abzusetzen. Der Graph einer in $x_0 \in X$ stetigen Funktion weist in $(x_0, f(x_0))$ ***keine Sprungstelle*** auf.

Ändert sich die Funktion *f* in der Nähe von x_0 langsam, so wird man keine Mühe haben, zu vorgegebenem $\varepsilon > 0$ ein passendes $\delta > 0$ zu finden; ändert sie sich rasch, so wird man δ entsprechend klein wählen müssen.

Beispiele:

Die Funktion

$$f: \begin{cases} \mathbf{R}_{>0} & \to & \mathbf{R} \\ x & \to & 1/x \end{cases}$$

ist stetig in jedem Punkt $x_0 \in \mathbf{R}_{>0}$: Zu $\varepsilon > 0$ kann man $\delta = \delta(\varepsilon, x_0) = \dfrac{\varepsilon \cdot x_0^2}{1 + \varepsilon \cdot x_0} > 0$ nehmen. Ist nämlich $x \in X$ mit $|x - x_0| < \delta$, so ist

$$|f(x) - f(x_0)| = |1/x - 1/x_0| = \left|\frac{x_0 - x}{x \cdot x_0}\right| < \frac{\delta}{x \cdot x_0} < \frac{\delta}{x_0 \cdot (x_0 - \delta)}.$$

Die letzte Ungleichung ergibt sich aus der Annahme $|x - x_0| < \delta$, die gleichbedeutend ist mit $x_0 - \delta < x < x_0 + \delta$ (also gilt insbesondere $x_0 - \delta < x$ bzw. $1/x < 1/(x_0 - \delta)$). Setzt man $\delta = \dfrac{\varepsilon \cdot x_0^2}{1 + \varepsilon \cdot x_0}$ ein, so sieht man $\dfrac{\delta}{x_0 \cdot (x_0 - \delta)} = \varepsilon$, also insgesamt $|f(x) - f(x_0)| < \varepsilon$.

Die Funktion

$$f:\begin{cases}\mathbf{R}_{\geq 0} & \to & \mathbf{R} \\ x & \to & \sqrt{x}\end{cases}$$

ist stetig in jedem Punkt $x_0 \in \mathbf{R}_{\geq 0}$. Zu $\varepsilon > 0$ und $x_0 \geq 0$ wähle man z.B. $\delta = \delta(\varepsilon, x_0) = \varepsilon^2$. Man beachte, dass δ hier nur von ε und nicht von x_0 abhängt. Ist nämlich $x \in X$ mit $|x - x_0| < \delta$ und $x \neq x_0$ (für $x = x_0$ gilt sowieso $|f(x) - f(x_0)| = 0 < \varepsilon$):

$$|f(x) - f(x_0)| = \left|\sqrt{x} - \sqrt{x_0}\right| = \left|\frac{\left(\sqrt{x} - \sqrt{x_0}\right)\cdot\left(\sqrt{x} + \sqrt{x_0}\right)}{\left(\sqrt{x} + \sqrt{x_0}\right)}\right| = \frac{|x - x_0|}{\sqrt{x} + \sqrt{x_0}} \leq \frac{|x - x_0|}{\sqrt{|x - x_0|}}.$$

Die letzte Ungleichung folgt aus der binomischen Formel: Sind $a \in \mathbf{R}_{\geq 0}$ und $b \in \mathbf{R}_{\geq 0}$ reelle Zahlen, so gilt:

$$a + b \leq a + b + 2 \cdot \sqrt{a} \cdot \sqrt{b} = \left(\sqrt{a} + \sqrt{b}\right)^2;$$

außerdem gilt $|a - b| \leq a + b$ und damit $\sqrt{|a - b|} \leq \sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$.

Die obige Ungleichung wird fortgesetzt:

$$|f(x) - f(x_0)| \leq \frac{|x - x_0|}{\sqrt{|x - x_0|}} = \sqrt{|x - x_0|} < \sqrt{\delta} = \sqrt{\varepsilon^2} = \varepsilon.$$

Wie im Fall der Konvergenz ist das zu vorgegebenem $\varepsilon > 0$ „passende“ $\delta > 0$ nicht immer leicht zu finden; hier ist häufig mathematische Phantasie gefragt. Man kann beispielsweise die konkrete Angabe von δ zunächst offen lassen und versuchen, die Ungleichung $|f(x) - f(x_0)| < \varepsilon$ so umzuformen, dass dort der Ausdruck $|x - x_0|$ vorkommt, von dem man dann ja annimmt, dass er kleiner als δ ist. Dann versucht man, die so entstandene Ungleichung nach δ aufzulösen. Das folgende Beispiel soll die Vorgehensweise erläutern.

Die Funktion

$$f:\begin{cases}\mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2\end{cases}$$

ist in jedem Punkt $x_0 \in \mathbf{R}$ stetig. Zu $\varepsilon > 0$ und $x_0 \in \mathbf{R}$ setzt man beispielsweise $\delta = \delta(\varepsilon, x_0) = ?$ Es wird zunächst $|f(x) - f(x_0)|$ in Abhängigkeit von $|x - x_0|$ und x_0 bestimmt:

$$|f(x) - f(x_0)| = |x^2 - x_0^2| = |(x - x_0) \cdot (x + x_0)| = |x - x_0| \cdot |x + x_0|$$
$$\leq |x - x_0| \cdot (|x| + |x_0|) < \delta \cdot (2 \cdot |x_0| + \delta).$$

In der letzten Ungleichung wurden die später zu treffende Annahme $|x - x_0| < \delta$ und die aus dieser Ungleichung leicht nachzurechnende Folgerung $|x| < |x_0| + \delta$ bereits verwendet. Wie ist also δ zu wählen, damit $\delta \cdot (2 \cdot |x_0| + \delta) < \varepsilon$ ist (hier reicht auch „=“ anstelle von „<“, da ja in der Ungleichungskette bereits „<“ vorkommt)?

$$\delta \cdot (2 \cdot |x_0| + \delta) = \delta^2 + 2 \cdot \delta \cdot |x_0| + |x_0|^2 - |x_0|^2 = (\delta + |x_0|)^2 - |x_0|^2 = \varepsilon$$

ergibt

$$\delta = \delta(\varepsilon, x_0) = \sqrt{|x_0|^2 + \varepsilon} - |x_0|.$$

Da der Ausdruck unter dem Wurzelzeichen größer als $|x_0|^2$ ist, ist auch $\delta > 0$. Wählt man den so angegebenen Wert von δ, so folgt aus $|x - x_0| < \delta$ die Ungleichung $|f(x) - f(x_0)| < \varepsilon$.

Die Funktion $f: X \to \mathbf{R}$ heißt **gleichmäßig stetig in** $D \subseteq X$, wenn es für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ ein $\delta > 0$ gibt, das höchstens von ε abhängt (d.h. $\delta = \delta(\varepsilon)$), mit folgender Eigenschaft:

Für jedes $x \in D$ und für jedes $y \in D$ mit $|x - y| < \delta$ ist $|f(x) - f(y)| < \varepsilon$.

Jede in $D \subseteq X$ gleichmäßig stetige Funktion ist dort natürlich auch stetig. Es gibt jedoch stetige Funktionen, die nicht gleichmäßig stetig sind, wie das Beispiel der Funktion $f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & x^2 \end{cases}$ zeigt:

Angenommen, *f* wäre gleichmäßig stetig. Es sei $\varepsilon > 0$ vorgegeben, und $\delta > 0$ sei der zugehörige Wert aus der Definition der gleichmäßigen Stetigkeit. Es wird ein Wert δ' gewählt mit $0 < \delta' < \delta$ und eine große reelle Zahl $x \in \mathbf{R}$ mit $\delta' \cdot (2 \cdot x + \delta') \geq \varepsilon$ (einen derartigen Wert *x* findet man immer, da der Ausdruck links beliebig groß gemacht werden kann). Außerdem wird $y = x + \delta'$ gesetzt. Dann ist $|x - y| = |x - (x + \delta')| = \delta' < \delta$, aber

$$|f(x) - f(y)| = |x^2 - y^2| = |y^2 - x^2| = |(x + \delta')^2 - x^2| = |2 \cdot x \cdot \delta' + \delta'^2| = \delta' \cdot (2 \cdot x + \delta') \geq \varepsilon .$$

Es lässt sich zeigen, dass eine auf **R** gleichmäßig stetige Funktion „nicht zu schnell wächst“, nämlich höchstens wie eine lineare Funktion (genauer: Ist $f: \mathbf{R} \to \mathbf{R}$ gleichmäßig stetig, so gibt es eine Konstante $C > 0$ mit $|f(x)| \leq C \cdot (1 + |x|)$).

Satz 5.2-1:

Sind $f: X \to \mathbf{R}$ und $g: X \to \mathbf{R}$ mit $X \subseteq \mathbf{R}$ stetig, so auch die folgenden Abbildungen:

(i) $f + g: \begin{cases} X & \to & \mathbf{R} \\ x & \to & f(x) + g(x) \end{cases}$

(ii) $f \cdot g: \begin{cases} X & \to & \mathbf{R} \\ x & \to & f(x) \cdot g(x) \end{cases}$

(iii) $|f|: \begin{cases} X & \to & \mathbf{R} \\ x & \to & |f(x)| \end{cases}$

(iv) $c \cdot f: \begin{cases} X & \to & \mathbf{R} \\ x & \to & c \cdot f(x) \end{cases}$ mit $c \in \mathbf{R}$

../..

(v) Ist $g(x_0) \neq 0$, so ist $f / g: \begin{cases} X & \to & \mathbf{R} \\ x & \to & \dfrac{f(x)}{g(x)} \end{cases}$ in $x_0 \in \mathbf{R}$ stetig.

(vi) Mit *f* und *g* ist auch $g \circ f$ stetig.

(vii) Ist *f* eine bijektive Funktion mit der Umkehrfunktion f^{-1}, so ist auch f^{-1} stetig.

Satz 5.2-2:

Die folgenden beiden Aussagen (a) und (b) sind gleichbedeutend:

(a) $f: X \to \mathbf{R}$ ist an der Stelle $x_0 \in X$ stetig.

und

(b) Für jede Folge $(x_n)_{n\in\mathbf{N}}$ mit $\lim\limits_{n\to\infty} x_n = x_0$ gilt $\lim\limits_{n\to\infty} f(x_n) = f(x_0)$.

Mit Hilfe des Satzes 5.2-2 lässt sich häufig nachweisen, dass eine Funktion in einem Punkt $x_0 \in X$ *nicht* stetig ist. Dazu braucht man nur eine einzige Folge $(x_n)_{n\in\mathbf{N}}$ anzugeben, die gegen $x_0 \in X$ konvergiert, deren Bildwerte unter f aber nicht gegen $f(x_0)$ gehen.

Beispiel:

Die Funktion

$$f: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & \begin{cases} x & \text{für } x < 1 \\ x+1 & \text{für } x \geq 1 \end{cases} \end{cases}$$

ist in $x_0 = 1$ nicht stetig. Dazu betrachte man die Folge $(x_n)_{n\in\mathbf{N}}$ mit $x_n = 1 - 1/(n+1)$. Es gilt $\lim\limits_{n\to\infty} x_n = 1$. Andererseits ist $f(x_n) = 1 - 1/(n+1)$ und $f(x_0) = f(1) = 2$, also $\lim\limits_{n\to\infty} f(x_n) \neq f(x_0)$.

Der folgende Satz drückt noch einmal aus, dass der Graph einer stetigen Funktion keine Sprünge aufweist.

Satz 5.2-3: (Zwischenwertsatz)

(i) Es seien $a \in \mathbf{R}$ und $b \in \mathbf{R}$ reelle Zahlen mit $a \le b$. Die Funktion $f : [a,b] \to \mathbf{R}$ sei im Intervall $[a,b]$ stetig. Außerdem gelte $f(a) \le 0 \le f(b)$. Dann gibt es ein $\gamma \in [a,b]$ mit $f(\gamma) = 0$.

(ii) Es seien $a \in \mathbf{R}$ und $b \in \mathbf{R}$ reelle Zahlen mit $a < b$. Die Funktion $f : [a,b] \to \mathbf{R}$ sei im Intervall $[a,b]$ stetig. Es gelte $f(a) < f(b)$. Außerdem sei $c \in \mathbf{R}$ mit $f(a) \le c \le f(b)$. Dann gibt es ein $\gamma \in [a,b]$ mit $f(\gamma) = c$.

Es seien $X \subseteq \mathbf{R}$ und $f : X \to \mathbf{R}$ eine Funktion. Das Element $x_0 \in X$ heißt **Nullstelle** von *f*, wenn $f(x_0) = 0$ gilt.

An einer Nullstelle schneidet der Graph von *f* die *x*-Achse.

Die Funktion $f : X \to \mathbf{R}$ besitzt im Punkt $x_0 \in \mathbf{R}$ den (endlichen) **Grenzwert** $f_0 \in \mathbf{R}$, wenn gilt:

Für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ gibt es ein (von ε abhängiges) $\delta = \delta(\varepsilon)$ mit folgender Eigenschaft:

Für jedes $x \in X$ mit $|x - x_0| < \delta$ ist $|f(x) - f_0| < \varepsilon$.

Zu beachten ist, dass der Wert x_0 nicht zum Definitionsbereich von *f* gehören muss.

Wählt man eine beliebig kleine ε-Umgebung von f_0, so findet man immer eine δ-Umgebung von x_0, die durch *f* komplett in diese ε-Umgebung abgebildet wird. In jeder beliebig kleinen ε-Umgebung von f_0 findet man Bildpunkte (unter *f*), deren Urbilder nahe bei x_0 liegen (zu beachten ist, dass x_0 nicht zu *X* zu gehören braucht).

Schreibweise: $\lim\limits_{x \to x_0} f(x) = f_0$.

Die Funktion $f : X \to \mathbf{R}$ besitzt in $x_P \in \mathbf{R}$ einen **Pol**, wenn gilt:

Für jedes $K \in \mathbf{R}$ mit $K > 0$ gibt es ein (von K abhängiges) $\delta = \delta(K)$ mit folgender Eigenschaft:

Für jedes $x \in X$ mit $|x - x_P| < \delta$ ist $|f(x)| > K$.

Die Funktionswerte wachsen über jede Grenze, wenn man sich dem Wert x_P nähert. Dabei ist zu beachten , dass x_P nicht zu X gehört.

Schreibweise: $\lim\limits_{x \to x_P} f(x) = \pm\infty$.

Satz 5.2-4:

Die folgenden beiden Aussagen (a) und (b) sind gleichbedeutend:

(a) $\lim\limits_{x \to x_P} f(x) = 0$

und

(b) Die durch $\left(\frac{1}{f}\right)(x) = \frac{1}{f(x)}$ definierte Funktion besitzt bei x_P einen Pol.

Die Funktion $f: X \to \mathbf{R}$ hat für $x \to \infty$ die **Asymptote** $s: X \to \mathbf{R}$, wenn gilt:
Für jedes $\varepsilon \in \mathbf{R}$ mit $\varepsilon > 0$ gibt es eine von ε abhängige Konstante $C = C(\varepsilon) > 0$ mit folgender Eigenschaft:

Für jedes $x \in X$ mit $x > C(\varepsilon)$ ist $|f(x) - s(x)| < \varepsilon$.

Der Funktionsverlauf von f nähert sich beliebig dicht dem Funktionsverlauf von s an, wenn man x nur genügend groß wählt.

Schreibweise: $\lim\limits_{x \to \infty} |f(x) - s(x)| = 0$ bzw. $\lim\limits_{x \to \infty} f(x) = s(x)$.

5.3 Polynome

Ein Polynom ist eine Funktion $p : \mathbf{R} \to \mathbf{R}$, zu deren Berechnung man mit den Rechenoperationen Addition, Subtraktion und Multiplikation auskommt.

Beispielsweise wird durch $p(x) = (x-1)\cdot(x^2+5) - \sqrt{2}\cdot x^7 = -\sqrt{2}\cdot x^7 + x^3 - x^2 + 5\cdot x - 5$ ein Polynom definiert.

Polynome lassen sich immer auf eine „standardisierte“ Form bringen:

Eine Funktion

$$p : \begin{cases} \mathbf{R} \to \mathbf{R} \\ x \to a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + \ldots a_1 \cdot x + a_0 \end{cases}$$

mit reellen Konstanten $a_n, a_{n-1}, \ \ldots \ , a_1, a_0$ und $a_n \neq 0$ heißt **Polynom vom Grad *n***.

Für $p(x) = a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + \ldots a_1 \cdot x + a_0$ schreibt man wie üblich $p(x) = \sum_{i=0}^{n} a_i \cdot x^i$.

Beispiele:

Die durch $p(x) = (x-1)\cdot(x^2+5) - \sqrt{2}\cdot x^7 = -\sqrt{2}\cdot x^7 + x^3 - x^2 + 5\cdot x - 5$ definierte Funktion ist ein Polynom vom Grad 7.

Die durch $p(x) = 5\cdot(x^2 - x)\cdot x^2 + \sqrt{2{,}5}$ definierte Funktion ist ein Polynom vom Grad 4.

Die durch $p(x) = 3\cdot x^2 - \sqrt{x} + 9$ definierte Funktion ist kein Polynom.

Polynome vom Grad 0:

$p(x) = a_0 = const.$

Hier wird auch $a_0 = 0$ zugelassen.

Der Graph eines Polynoms vom Grad 0 ist eine Gerade, die im (x, y)-Koordinatensystem parallel zur x-Achse verläuft und die y-Achse im Punkt $(0, a_0)$ schneidet.

Polynome vom Grad 1:

$p(x) = a_1 \cdot x + a_0$ mit $a_1 \neq 0$

Die einzige Nullstelle ist $x_0 = -\frac{a_0}{a_1}$.

Der Graph eines Polynoms 1. Grades ist eine Gerade und schneidet im (x, y)-Koordinatensystem die y-Achse im Punkt $(0, a_0)$.

Polynome vom Grad 2:

$p(x) = a_2 \cdot x^2 + a_1 \cdot x + a_0$ mit $a_2 \neq 0$

Es gibt zwei oder eine oder keine reelle Nullstelle. Die Nullstellen berechnen sich zu

$$x_{01,02} = -\frac{a_1}{2 \cdot a_2} \pm \sqrt{\frac{{a_1}^2}{4 \cdot {a_2}^2} - \frac{a_0}{a_2}}.$$

Diese sind nur dann reellwertig, wenn ${a_1}^2 \geq 4 \cdot a_2 \cdot a_0$ ist.

Ein häufig auftretender Spezialfall ist das Polynom der Form $p(x) = x^2 + p \cdot x + q$ mit $p \in \mathrm{R}$ und $q \in \mathbf{R}$. Dieses Polynom hat die Nullstellen

$$x_{01,02} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}.$$

Die Bedingung für die Reellwertigkeit der Nullstellen lautet $p^2 - 4 \cdot q \geq 0$.

Der Graph eines Polynoms 2.Grades ist eine Parabel, die für $a_2 > 0$ nach oben und für $a_2 < 0$ nach unten geöffnet ist.

Ist $a_2 > 0$ (bzw. $a_2 < 0$), so wird der minimale (bzw. maximale) Wert des Polynoms $p(x) = a_2 \cdot x^2 + a_1 \cdot x + a_0$ an der Stelle

$$x_S = -\frac{a_1}{2 \cdot a_2}$$

angenommen; der Funktionswert lautet dabei $p(x_S) = a_0 - \frac{{a_1}^2}{4 \cdot a_2}$.

Im Spezialfall $p(x) = x^2 + p \cdot x + q$ lauten die entsprechenden Werte

$x_S = -p/2$ und $p(x_S) = q - \left(p/2\right)^2$.

Polynome vom Grad ≥ 3:

Für Polynome 3. und 4. Grades gibt es noch eine geschlossene Formel zur Nullstellenbestimmung, für Polynome höheren Grades i.a. nicht.

Satz 5.3-1:

(i) $p(x) = a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + ... + a_1 \cdot x + a_0 = \sum_{i=0}^{n} a_i \cdot x^i$ sei ein Polynom vom Grad n und x_0 eine Nullstelle von p (d.h. $p(x_0) = 0$). Dann gibt es ein Polynom p_1 vom Grad $n-1$ mit

$$p(x) = (x - x_0) \cdot p_1(x).$$

Man kann also den **Linearfaktor** $x - x_0$ aus $p(x)$ ausklammern.

Im Spezialfall $p(x) = x^n - a^n$ mit $a \neq 0$ lautet eine Nullstelle $x_0 = a$. Es ist

$$p(x) = x^n - a^n = (x - a) \cdot \sum_{i=0}^{n-1} a^{n-i-1} \cdot x^i .$$

(ii) Ein Polynom vom Grad n hat höchstens n viele reelle Nullstellen.

(iii) Ein Polynom von *ungeradem* Grad hat mindestens eine Nullstelle.

Satz 5.3-2:

Das Polynom $p(x) = \sum_{i=0}^{n} a_i \cdot x^i$ vom Grad n habe die reellen Nullstellen $x_{01}, \ldots, x_{0m}$; hierbei werden mehrfache reelle Nullstellen jeweils auch mehrfach aufgeführt. Dann gilt

$$p(x) = (x - x_{01}) \cdot ... \cdot (x - x_{0m}) \cdot p_g(x)$$

mit einem Polynom $p_g(x)$ von geradem Grad $2 \cdot k$, das keine reellen Nullstellen hat. Außerdem ist $n = m + 2 \cdot k$.

5.4 Gebrochen rationale Funktionen

Eine Funktion der Form

$$f: \begin{cases} X & \to & \mathbf{R} \\ x & \to & \dfrac{p(x)}{q(x)} \end{cases}$$

mit $X \subseteq \mathbf{R}$ und den Polynomen $p(x) = \sum_{i=0}^{n} a_i \cdot x^i$ und $q(x) = \sum_{j=0}^{m} b_j \cdot x^j$ und $b_m \neq 0$ heißt **gebrochen rationale Funktion**.

An den Nullstellen von *q* ist *f* nicht definiert, d.h. der Definitionsbereich von f lautet

$$D(f) = \mathbf{R} \setminus \{ x_0 \mid q(x_0) = 0 \}.$$

Skizzierung einer gebrochen rationalen Funktion *f*:

1. Schritt:

Bestimmung aller Nullstellen von *p* und aller Nullstellen von *q*. Alle diese Nullstellen seien $x_{01}, \ldots, x_{0l}$.

Die Nullstellen von *q* gehören nicht zum Definitionsbereich von *f*.

Für jede dieser Nullstellen x_{0i} von *p* und *q* wird der 2. Schritt durchgeführt.

2. Schritt:

Es werden 3 mögliche Fälle unterschieden:

1. Fall: x_{0i} ist eine Nullstelle von *p*, aber nicht von *q*:

$$p(x_{0i}) = 0 \text{ und } q(x_{0i}) \neq 0$$

Es gilt

$$f(x_{0i}) = \frac{p(x_{0i})}{q(x_{0i})} = \frac{0}{q(x_{0i})} = 0,$$

d.h. x_{0i} ist eine Nullstelle von *f*.

2. Fall: x_{0i} ist keine Nullstelle von *p*, aber eine Nullstelle von *q*:

$p(x_{0i}) \neq 0$ und $q(x_{0i}) = 0$

Zu beachten ist, dass *f* für x_{0i} nicht definiert ist.

Es gilt

$$\lim_{x \to x_{0i}} 1 \Big/ f(x) = \frac{\lim\limits_{x \to x_{0i}} q(x)}{\lim\limits_{x \to x_{0i}} p(x)} = \frac{0}{p(x_{0i})} = 0\,,$$

d.h. f besitzt bei x_{0i} einen Pol.

3. Fall: x_{0i} ist sowohl eine Nullstelle von *p*, als auch eine Nullstelle von *q*:

$p(x_{0i}) = 0$ und $q(x_{0i}) = 0$

Zu beachten ist, dass *f* für x_{0i} nicht definiert ist.

Ist x_{0i} eine *r*-fache Nullstelle von *p* und eine *s*-fache Nullstelle von *q*, dann gilt

$$f(x) = \frac{(x - x_{0i})^r \cdot p_1(x)}{(x - x_{0i})^s \cdot q_1(x)} \text{ mit } p_1(x_{0i}) \neq 0 \text{ und } q_1(x_{0i}) \neq 0\,.$$

Fall 3a: $r > s$

$$\lim_{x \to x_{0i}} f(x) = \lim_{x \to x_{0i}} (x - x_{0i})^{r-s} \cdot \frac{p_1(x_{0i})}{q_1(x_{0i})} = 0$$

Fall 3b: $r = s$

$$\lim_{x \to x_{0i}} f(x) = \frac{p_1(x_{0i})}{q_1(x_{0i})} \neq 0$$

In beiden Fällen nennt man x_{0i} eine **behebbare Unstetigkeitsstelle** von *f*, da man *f* stetig nach x_{0i} fortsetzen kann.

Fall 3c: $r < s$

$$\lim_{x \to x_{0i}} 1/f(x) = \lim_{x \to x_{0i}} \frac{(x - x_{0i})^{s-r} \cdot q_1(x)}{p_1(x)} = 0$$, d.h. *f* hat bei x_{0i} einen Pol.

3. Schritt:

Es wird das Verhalten von $f(x)$ bei $x \to \pm\infty$ untersucht.

Es ist $f(x) = \frac{p(x)}{q(x)}$, $p(x) = \sum_{i=0}^{n} a_i \cdot x^i$, $q(x) = \sum_{j=0}^{m} b_j \cdot x^j$ und $b_m \neq 0$.

Fall 4a: Der Grad von *q* ist größer als der Grad von *p*, d.h. $m > n$.

$$\begin{aligned}\lim_{x \to \infty} f(x) &= \lim_{x \to \infty} \frac{a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + \ldots + a_1 \cdot x + a_0}{b_m \cdot x^m + b_{m-1} \cdot x^{m-1} + \ldots + b_1 \cdot x + b_0} \\ &= \lim_{x \to \infty} \frac{a_n \cdot 1/x^{m-n} + a_{n-1} \cdot 1/x^{m-(n-1)} + \ldots + a_1 \cdot 1/x^{m-1} + a_0 \cdot 1/x^m}{b_m \cdot 1 + b_{m-1} \cdot 1/x + \ldots + b_1 \cdot 1/x^{m-1} + b_0 \cdot 1/x^m} \\ &= 0,\end{aligned}$$

d.h. *f* hat bei $x \to \pm\infty$ die Asymptote $s(x) = 0$ (*x*-Achse).

Fall 4b: Der Grad von *q* ist nicht größer als der Grad von *p*, d.h. $m \leq n$.

Durch Ausdividieren (**Polynomdivision**) von $p(x)/q(x)$ erhält man auf eindeutige Weise Polynome $s(x)$ und $r(x)$ mit $f(x) = s(x) + \frac{r(x)}{q(x)}$. Hierbei hat $s(x)$ den Grad $n - m$ und $r(x)$ einen kleineren Grad als $q(x)$, und es gilt:

$\lim_{x\to\pm\infty} f(x) = s(x)$, d.h. f hat bei $x \to \pm\infty$ die Asymptote $s(x)$.

5.5 Exponential- und Logarithmusfunktion

In Kapitel 5.1 wird die Exponentialfunktion

$$\exp: \begin{cases} \mathbf{R} & \to & \mathbf{R} \\ x & \to & \sum_{i=0}^{\infty} \frac{x^i}{i!} \end{cases}$$

definiert. Zunächst werden einige Eigenschaften dieser Funktion untersucht.

Ein wichtiges Ergebnis, nämlich

$$\exp(1) = \sum_{i=0}^{\infty} \frac{1}{i!} = e = 2{,}718281\ldots,$$

wird in Kapitel 5.1 hergeleitet. Außerdem gilt $\sum_{i=0}^{\infty} \frac{x^i}{i!} = \underbrace{1}_{i=0} + \sum_{i=1}^{\infty} \frac{x^i}{i!}$, und daher $\exp(0) = 1$.

Mit Satz 5.1-10 ergibt sich

$$\begin{aligned} \exp(x)\cdot\exp(y) &= \left(\sum_{i=0}^{\infty} \frac{x^i}{i!}\right)\cdot\left(\sum_{i=0}^{\infty} \frac{y^i}{i!}\right) \\ &= \sum_{i=0}^{\infty}\left(\sum_{k=0}^{i} \frac{x^k}{k!}\cdot\frac{y^{i-k}}{(i-k)!}\right) \\ &= \sum_{i=0}^{\infty}\left(\frac{1}{i!}\cdot\sum_{k=0}^{i} \frac{i!}{k!\cdot(i-k)!}\cdot x^k\cdot y^{i-k}\right) \\ &= \sum_{i=0}^{\infty}\left(\frac{1}{i!}\cdot(x+y)^i\right) \\ &= \exp(x+y) \end{aligned}$$

Für jedes $r \in \mathbf{R}$ gilt daher: $\exp(r) = \exp\left(\frac{r}{2}+\frac{r}{2}\right) = \left(\exp\left(\frac{r}{2}\right)\right)^2 \geq 0$. Wegen

$1 = \exp(0) = \exp(r - r) = \exp(r) \cdot \exp(-r)$ ist sogar $\exp(r) > 0$. Daher kann der Wertebereich der Exponentialfunktion auf $\mathbf{R}_{>0}$ eingeschränkt werden. Es lässt sich zeigen, dass die so definierte Funktion $\exp : \mathbf{R} \to \mathbf{R}_{>0}$ bijektiv und stetig ist.

Für jedes $n \in \mathbf{N}$ lässt sich der Wert der Exponentialfunktion folgendermaßen berechnen:

$$\exp(n) = \exp(\underbrace{1+1+\ldots+1}_{n\text{–mal}}) = \left(\exp(1)\right)^n = e^n .$$

Dieses Ergebnis bedeutet, dass man zur Ermittlung des Werts von $\exp(n)$ anstelle der Grenzwertberechnung $\sum_{i=0}^{\infty} \frac{n^i}{i!}$ in $\mathbf{R}$ das n-fache Produkt der reellen Zahl $e \in \mathbf{R}$ bildet. Zur Berechnung des Werts von $\exp(n)$ mit Hilfe eines Computers, in dem reelle Zahlen nur approximiert werden können, wird man eher die Reihenentwicklung $\sum_{i=0}^{\infty} \frac{n^i}{i!}$ verwenden und diese nach einer endlichen Anzahl Summanden, entsprechend der vorgegebenen Genauigkeit zur Darstellung reeller Zahlen, abbrechen.

Für eine negative ganze Zahl $m \in \mathbf{Z}$ ist wegen $\exp(m) \cdot \exp(-m) = \exp(0) = 1$:

$$\exp(m) = \frac{1}{\exp(-m)} = \frac{1}{e^{-m}} = e^m .$$

Für eine rationale Zahl $q = \frac{n}{m}$ mit $n \in \mathbf{N}$ und $m \in \mathbf{N}_{>0}$ ist

$$\left(\exp\left(\frac{n}{m}\right)\right)^m = \exp\left(\underbrace{\frac{n}{m}+\frac{n}{m}+\ldots+\frac{n}{m}}_{m\text{-mal}}\right) = \exp(n) = e^n, \text{ also}$$

$$\exp\left(\frac{n}{m}\right) = e^{\frac{n}{m}} .$$

Für eine rationale Zahl $q < 0$ ist wegen $\exp(q) \cdot \exp(-q) = \exp(q - q) = \exp(0) = 1$:

$$\exp(q) = \frac{1}{\exp(-q)} = \frac{1}{e^{-q}} = e^q .$$

Insgesamt ist also für jedes $x \in \mathbf{Q}$ gezeigt: $\exp(x) = e^x$.

Aufgrund dieses Ergebnisses verwendet man für alle $x \in \mathbf{R}$ anstelle von exp(x) die Bezeichnung e^x; zu beachten ist, dass dieses für $x \in \mathbf{Q}$ bewiesen wurde, für $x \in \mathbf{R} \setminus \mathbf{Q}$ stellt es eine abkürzende Schreibweise für den Grenzwert der Reihe $\sum_{i=0}^{\infty} \frac{x^i}{i!}$ dar.

Die Ergebnisse fasst folgender Satz zusammen:

Satz 5.5-1:

(i) Die Exponentialfunktion

$$\exp : \begin{cases} \mathbf{R} & \to &]0, \infty[\\ x & \to & \sum_{i=0}^{\infty} \frac{x^i}{i!} \end{cases}$$

ist bijektiv und stetig und erfüllt die Funktionalgleichung

$$\exp(x+y) = \exp(x) \cdot \exp(y) \text{ bzw. } e^{x+y} = e^x \cdot e^y .$$

(ii) Es seien $x \in \mathbf{R}$ und $y \in \mathbf{R}$. Dann gilt

$$\exp(0) = 1,$$
$$\exp(1) = e,$$
$$\exp(x-y) = \frac{\exp(x)}{\exp(y)} \text{ bzw. } e^{x-y} = \frac{e^x}{e^y}.$$

Aufgrund von Satz 2.2-1 gibt es zur Exponentialfunktion eine eindeutig bestimmte Umkehrfunktion, die stetig und bijektiv ist (Satz 2.2-1 und Satz 5.2-1). Diese Funktion heißt **natürlicher Logarithmus** und wird mit ln bezeichnet:

$$\ln : \begin{cases}]0, \infty[& \to & \mathbf{R} \\ x & \to & \ln(x) \end{cases}$$

Es ist $y = \ln(x)$ genau dann, wenn $x = e^y = \exp(y)$ gilt. Weiter gilt:

Mit $z = \ln(x)$ und $z' = \ln(y)$ ist $\exp(z+z') = \exp(z) \cdot \exp(z') = x \cdot y$, also

$$\ln\big(\exp(z+z')\big) = z + z' = \ln(x) + \ln(y) = \ln\big(x \cdot y\big).$$

Die Ergebnisse und weitere Eigenschaften des natürlichen Logarithmus sind im folgenden Satz zusammengefasst.

Satz 5.5-2:

(i) Die natürliche Logarithmusfunktion

$$\ln : \begin{cases}]0,\infty[& \to & \mathbf{R} \\ x & \to & \ln(x) \end{cases}$$

ist bijektiv und stetig und erfüllt die Funktionalgleichung

$$\ln\big(x \cdot y\big) = \ln(x) + \ln(y)\,.$$

(ii) Es seien $x \in \mathbf{R}_{>0}$ und $y \in \mathbf{R}_{>0}$. Dann gilt

$$\ln(1) = 0\,,$$
$$\ln(e) = 1\,,$$
$$\ln\big(x/y\big) = \ln(x) - \ln(y)\,,$$
$$\ln\big(x^n\big) = n \cdot \ln(x) \text{ für jedes } n \in \mathbf{Z}\,,$$
$$\ln\big(e^x\big) = x \text{ und } e^{\ln(x)} = x\,.$$

Es sei $a \in \mathbf{R}$ mit $a > 0$. Dann heißt die Funktion

$$\exp_a : \begin{cases} \mathbf{R} & \to &]0,\infty[\\ x & \to & \exp\big(\ln(a) \cdot x\big) = e^{\ln(a)\cdot x} \end{cases}$$

Exponentialfunktion zur Basis *a*. Statt $\exp_a(x)$ schreibt man a^x.

Satz 5.5-3:

Es sei $a \in \mathbf{R}$ mit $a > 0$.

(i) Die Exponentialfunktion zur Basis a

$$\exp_a : \begin{cases} \mathbf{R} & \to &]0, \infty[\\ x & \to & \exp(\ln(a) \cdot x) = e^{\ln(a) \cdot x} \end{cases}$$

ist stetig. Für $a > 1$ ist sie streng monoton steigend, für $a < 1$ ist sie streng monoton fallend, für $a = 1$ ist sie konstant 1.

Für $a \neq 1$ ist die Exponentialfunktion zur Basis a bijektiv.

(ii) Es seien $x \in \mathbf{R}$ und $y \in \mathbf{R}$. Dann gilt

$$\exp_a(0) = 1 \quad \text{bzw.} \quad a^0 = 1,$$
$$\exp_a(1) = a \quad \text{bzw.} \quad a^1 = a,$$
$$\exp_a(x+y) = \exp_a(x) \cdot \exp_a(y) \quad \text{bzw.} \quad a^{x+y} = a^x \cdot a^y,$$
$$\exp_a(x-y) = \frac{\exp_a(x)}{\exp_a(y)} \quad \text{bzw.} \quad a^{x-y} = \frac{a^x}{a^y},$$
$$(\exp_a(x))^y = \exp_a(x \cdot y) \quad \text{bzw.} \quad (a^x)^y = a^{x \cdot y}.$$

Die letzte Gleichung soll verifiziert werden:

$$\begin{aligned} (a^x)^y = (\exp_a(x))^y = (e^{\ln(a) \cdot x})^y \quad & \text{(nach Definition der Exponentialfunktion zur Basis } a) \\ = e^{\ln(e^{\ln(a) \cdot x}) \cdot y} \quad & \text{(nach Definition der Exponentialfunktion zur Basis } e^{\ln(a) \cdot x}) \\ = e^{\ln(a) \cdot x \cdot y} \quad & \text{(da der natürliche Logarithmus und die Exponentialfunktion} \\ & \text{zueinander invers sind)} \\ = a^{x \cdot y} \quad & \text{(nach Definition der Exponentialfunktion zur Basis } a). \end{aligned}$$

Für $x \in \mathbf{R}$ und $y \in \mathbf{R}$ mit $y > 0$ lässt sich nun auch der Ausdruck y^x sinnvoll definieren:

$$y^x = e^{\ln(y) \cdot x}.$$

Beispielsweise ist $1^x = e^{\ln(1)\cdot x} = e^{0\cdot x} = 1$, und Werte wie $\pi^{\sqrt{2}}$ oder 2^e machen einen Sinn.

Für $a \in \mathbf{R}$ mit $a > 0$ und $a \neq 1$ heißt die zur Exponentialfunktion $\exp_a$ existierende Umkehrfunktion $\overset{-1}{\exp}_a$, die **Logarithmusfunktion zur Basis *a*** und wird mit $\log_a$ bezeichnet:

$$\log_a : \begin{cases}]0,\infty[& \to & \mathbf{R} \\ x & \to & \log_a(x) \end{cases}.$$

Satz 5.5-4:

Es sei $a \in \mathbf{R}$ mit $a > 0$ und $a \neq 1$.

(i) Die Logarithmusfunktion zur Basis *a*

$$\log_a : \begin{cases}]0,\infty[& \to & \mathbf{R} \\ x & \to & \log_a(x) \end{cases}$$

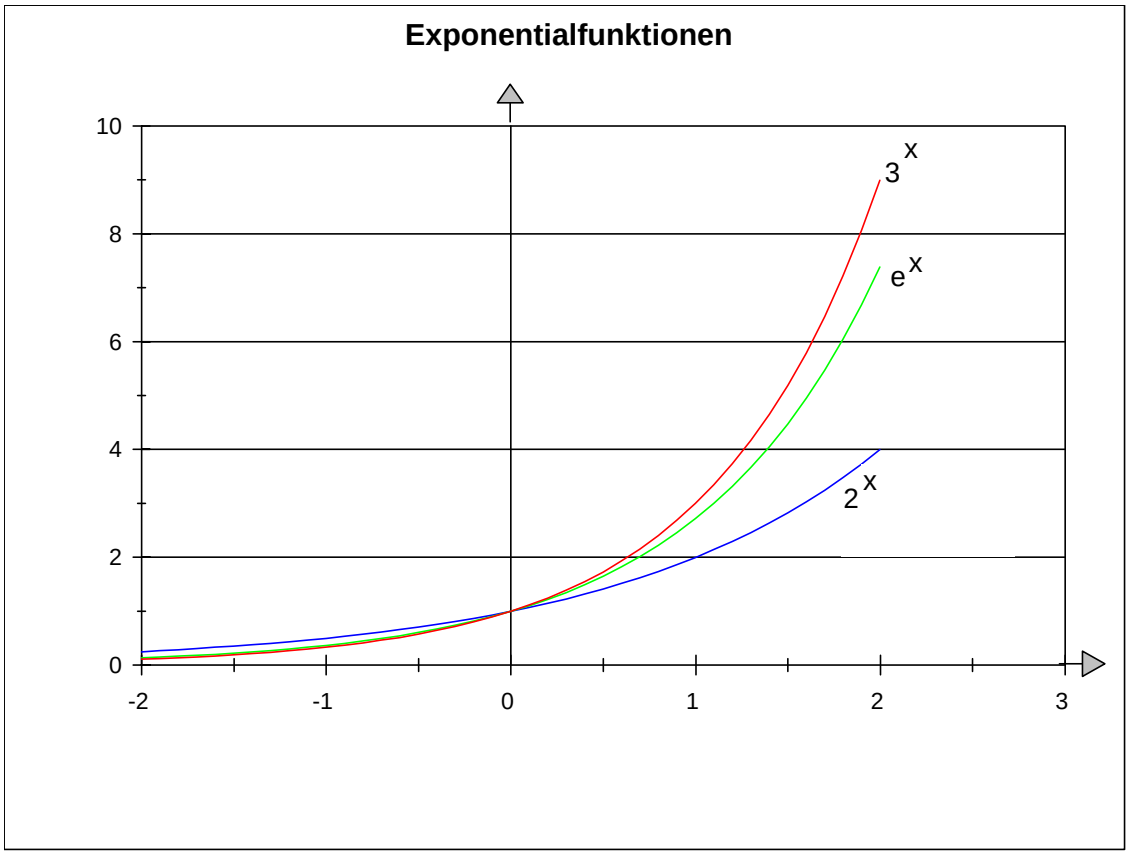
ist bijektiv und stetig und erfüllt die Funktionalgleichung

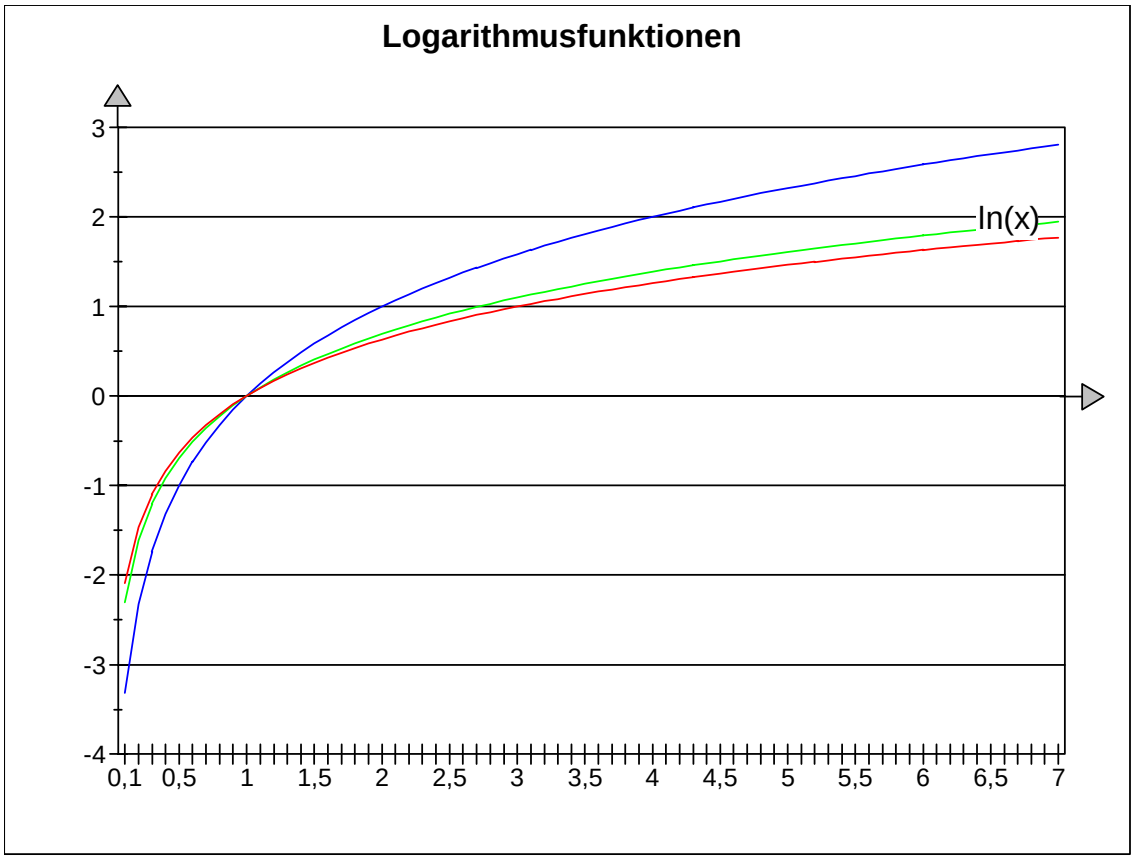
$$\log_a(x \cdot y) = \log_a(x) + \log_a(y).$$

(ii) Es seien $x \in \mathbf{R}_{>0}$ und $y \in \mathbf{R}_{>0}$. Dann gilt

$$\log_a(1) = 0,$$
$$\log_a(a) = 1,$$
$$\log_a(x/y) = \log_a(x) - \log_a(y),$$
$$\log_a(x^m) = m \cdot \log_a(x) \text{ für jedes } m \in \mathbf{Z},$$
$$\log_a(a^x) = x \text{ und } a^{\log_a(x)} = x.$$

Die folgenden Abbildungen zeigen die Verläufe einiger Exponential- und Logarithmusfunktionen.


Exponentialfunktionen
10
8
6
4
2
0
-2
-1
0
1
2
3
3 x
e x
2 x


Logarithmusfunktionen
3
2
1
0
-1
-2
-3
-4
0,1
0,5
1
1,5
2
2,5
3
3,5
4
4,5
5
5,5
6
6,5
7
ln(x)

Die Exponential- und Logarithmusfunktionen zu unterschiedlichen Basen $a \in \mathbf{R}$ mit $a > 0$ und $a \neq 1$ und $b \in \mathbf{R}$ mit $b > 0$ und $b \neq 1$ lassen sich ineinander umrechnen.

Satz 5.5-5:

Es seien $a \in \mathbf{R}$ und $b \in \mathbf{R}$ mit $a > 0$, $a \neq 1$, $b > 0$ und $b \neq 1$.

(i) Den Zusammenhang zwischen verschiedenen Exponentialfunktionen stellt die Gleichung

$$\exp_a(x) = \left(\exp_b(x)\right)^{\log_b(a)} \quad \text{bzw.} \quad a^x = \left(b^x\right)^{\log_b(a)}$$

her. Verschiedene Exponentialfunktionen unterscheiden sich also durch potenzierte Werte.

(ii) Der Zusammenhang zwischen verschiedenen Logarithmusfunktionen wird durch die Gleichung

$$\log_a(x) = \frac{1}{\log_b(a)} \cdot \log_b(x) = \frac{\ln(x)}{\ln(a)}$$

beschrieben. Verschiedene Logarithmusfunktionen unterscheiden sich also durch konstante Faktoren.

(iii) $\log_a\left(b^x\right) = x \cdot \log_a(b)$.

Die Herleitung dieser Gleichungen kann als gute Übung zum Umgang mit Exponential- und Logarithmusausdrücken angesehen werden:

Zunächst wird der zweite Teil der Gleichung in (ii) gezeigt. Diese beschreibt, wie sich $z = \log_a(x)$ mit Hilfe des natürlichen Logarithmus ausdrücken lässt. Aus $z = \log_a(x)$ folgt nacheinander

$a^z = x$,
$e^{\ln(a)\cdot z} = x$ (Definition der Exponentialfunktion zur Basis *a*),
$\ln(a) \cdot z = \ln(x)$ (Übergang zur Umkehrfunktion, dem natürlichen Logarithmus),

$$z = \log_a(x) = \frac{\ln(x)}{\ln(a)}.$$

Daraus folgt direkt der erste Teil der Gleichung in (ii):

$$\log_a(x) = \frac{\ln(x)}{\ln(a)} = \frac{\ln(x)\cdot\ln(b)}{\ln(a)\cdot\ln(b)} = \frac{\ln(x)}{\ln(b)}\cdot\frac{\ln(b)}{\ln(a)} = \log_b(x)\cdot\frac{1}{\log_b(a)}.$$

Die letzte Gleichung entsteht durch Setzen von $x = a$ und $a = b$ in der Formel $\log_a(x) = \frac{\ln(x)}{\ln(a)}$.

Gleichung (i) beschreibt, wie sich die Exponentialfunktion zur Basis *a* durch die Exponentialfunktion zur Basis *b* ausdrücken lässt. Dazu wird die Gleichung

$$a^x = b^y$$

zunächst auf die ursprüngliche Definition zurückgeführt, dann nach *y* aufgelöst und die Gleichung in (ii) verwendet:

$a^x = e^{\ln(a)\cdot x} = b^y = e^{\ln(b)\cdot y}$ (Definition der Exponentialfunktion zur Basis *a* bzw. zur Basis *b*),
$\ln(a)\cdot x = \ln(b)\cdot y$ (Übergang zur inversen Funktion, dem natürlichen Logarithmus),
$y = \frac{\ln(a)}{\ln(b)}\cdot x = \log_b(a)\cdot x$ (aus (ii)),
$a^x = b^y = b^{\log_b(a)\cdot x} = \left(b^x\right)^{\log_b(a)}$ (Satz 5.5-3 (ii)).

Gleichung (iii) ist eine Verallgemeinerung der Gleichung in 5.5-4 (ii) auf alle reellen Zahlen. Mit der Gleichung aus (ii) ergibt sich:

$$\log_a\left(b^x\right) = \frac{\ln\left(b^x\right)}{\ln(a)} = \frac{\ln\left(e^{\ln(b)\cdot x}\right)}{\ln(a)} = \frac{\ln(b)\cdot x}{\ln(a)} = x\cdot\log_a(b).$$

Im folgenden sei $a > 1$. Die Exponentialfunktion zur Basis *a* steigt bei wachsendem *x* schnell an. Es gilt nämlich $\exp_a(x+1) = a\cdot\exp_a(x)$ bzw. $a^{x+1} = a\cdot a^x$, d.h. bei Vergrößerung des Argumentwerts um 1 vergrößert sich der Funktionswert um den Faktor *a*.

Hingegen wachsen die entsprechenden Logarithmusfunktionen sehr langsam. Es gilt nämlich $\lim_{x\to\infty}\left(\log_a(x+1) - \log_a(x)\right) = 0$, d.h. obwohl die Logarithmusfunktion bei wachsendem Argumentwert gegen ∞ strebt, nehmen die Funktionswerte letztlich nur noch geringfügig zu:

Es gilt nämlich wegen der Stetigkeit der Logarithmusfunktion (mit Satz 5.2-2):

$$\lim_{n\to\infty}\left(\log_a(n+1)-\log_a(n)\right)=\lim_{n\to\infty}\left(\log_a\left(\frac{n+1}{n}\right)\right)=\log_a\left(\lim_{n\to\infty}\left(\frac{n+1}{n}\right)\right)=\log_a(1)=0\,.$$

Das Wachstumsverhalten der Exponential- und Logarithmusfunktionen im Vergleich mit Polynomen und Wurzelfunktionen zeigt der folgende Satz, dessen Beweis sich aus Überlegungen ergibt, die in Kapitel 5.7 angestellt werden.

Satz 5.5-6:

Es sei $a\in\mathbf{R}, a>1$.

(i) Es sei $p(x)$ ein Polynom. Dann gilt:

$$\lim_{x\to\infty}\frac{|p(x)|}{a^x}=0\,,$$

d.h. die Exponentialfunktionen wachsen schneller als alle Polynome.

(ii) Für jedes $m\in\mathbf{N}$ ist

$$\lim_{x\to\infty}\frac{(\log_a(x))^m}{x}=0\,.$$

Man sieht, dass selbst Potenzen von Logarithmusfunktionen im Verhältnis zu Polynomen (sogar zu Polynomen 1. Grades) langsamer wachsen.

(iii) Für jedes $m\in\mathbf{N}$ ist

$$\lim_{x\to\infty}\frac{\log_a(x)}{\sqrt[m]{x}}=0\,.$$

Man sieht, dass Logarithmusfunktionen im Verhältnis zu Wurzelfunktionen langsamer wachsen.

Die folgende Tabelle zeigt fünf Funktionen $h_i : \mathbf{R}_{>0} \to \mathbf{R}$, $i = 1, \ldots, 5$ und einige ausgewählte (gerundete) Funktionswerte.

Spalte 1	Spalte 2	Spalte 3	Spalte 4	Spalte 5
i	$h_i(x)$	$h_i(10)$	$h_i(100)$	$h_i(1000)$
1	$\log_2(x)$	3,3219	6,6439	9,9658
2	$\sqrt{x}$	3,1623	10	31,6228
3	x	10	100	1000
4	x^2	100	10.000	1.000.000
5	2^x	1024	$1{,}2676506 \cdot 10^{30}$	$> 10^{693}$

Die folgende Tabelle zeigt noch einmal die fünf Funktionen $h_i : \mathbf{R}_{>0} \to \mathbf{R}$, $i = 1, \ldots, 5$. Es sei $y_0 > 0$ ein fester Wert. Die dritte Spalte zeigt für jede der fünf Funktionen *x*-Werte x_i mit $h_i(x_i) = y_0$. In der vierten Spalte sind diejenigen *x*-Werte $\overline{x_i}$ aufgeführt, für die $h_i\left(\overline{x_i}\right) = 10 \cdot y_0$ gilt, d.h. dort ist angegeben, auf welchen Wert man x_i vergrößern muss, damit der Funktionswert auf den 10-fachen Wert wächst. Wie man sieht, muss bei der Logartihmusfunktion wegen ihres langsamen Wachstums der *x*-Wert stark vergrößert werden, während bei der schnell anwachsenden Exponentialfunktion nur eine additive konstante Steigerung um ca. 3,3 erforderlich ist.

Spalte 1	Spalte 2	Spalte 3	Spalte 4
i	$h_i(x)$	x_i mit $h_i(x_i) = y_0$	$\overline{x_i}$ mit $h_i\left(\overline{x_i}\right) = 10 \cdot y_0$
1	$\log_2(x)$	x_1	$\left(x_1\right)^{10}$
2	$\sqrt{x}$	x_2	$100 \cdot x_2$
3	x	x_3	$10 \cdot x_3$
4	x^2	x_4	$\approx 3{,}162 \cdot x_4$
5	2^x	x_5	$\approx x_5 + 3{,}322$

Die Logarithmusfunktion zu einer Basis $B > 1$ gibt u.a. näherungsweise an, wieviele Ziffern benötigt werden, um eine natürliche Zahl im Zahlensystem zur Basis *B* darzustellen:

Gegeben sei die Zahl $n \in \mathbf{N}$ mit $n > 0$. Sie benötige $m = m(n, B)$ signifikante Stellen zur Darstellung im Zahlensystem zur Basis *B*, d.h.

$$n = \sum_{i=0}^{m-1} a_i \cdot B^i \text{ mit } a_i \in \{0,\ 1,\ \ldots,\ B-1\} \text{ und } a_{m-1} \neq 0.$$

Es ist $B^{m-1} \leq n < B^m$ und folglich $m-1 \leq \log_B(n) < m$. Daraus ergibt sich für die Anzahl der benötigten Stellen, um eine Zahl *n* im Zahlensystem zur Basis *B* darzustellen,

$$m(n,B) = \lfloor \log_B(n) \rfloor + 1 = \lceil \log_B(n+1) \rceil.$$

Die Anzahl an Dezimalziffern zur Darstellung einer Zahl *n* beträgt demnach $\lfloor \log_{10}(n) \rfloor + 1$, an Binärziffern $\lfloor \log_2(n) \rfloor + 1$ und an Sedezimalziffern $\lfloor \log_{16}(n) \rfloor + 1$.

Die folgende Tabelle zeigt die Zusammenhänge an benötigten Stellen zur Darstellung einer Zahl *n* in den in der Informatik üblichen Zahlensystemen.

	Stellenzahl im	
Dezimalsystem $B = 10$	Binärsystem $B = 2$	Sedezimalsystem $B = 16$
m	zwischen $\lfloor c_{10,2} \cdot m \rfloor - 3$ und $\lceil c_{10,2} \cdot m \rceil$ mit $c_{10,2} = 1/\log_{10}(2) \approx 3{,}3219281$	Zwischen $\lfloor c_{10,16} \cdot m \rfloor - 1$ und $\lceil c_{10,16} \cdot m \rceil$ mit $c_{10,16} = 1/\log_{10}(16) \approx 0{,}830482$
Zwischen $\lfloor c_{2,10} \cdot m \rfloor - 1$ und $\lceil c_{2,10} \cdot m \rceil$ mit $c_{2,10} = \log_{10}(2) \approx 0{,}30103$	m	$\left\lceil \frac{m}{4} \right\rceil$
zwischen $\lfloor c_{16,10} \cdot m \rfloor - 1$ und $\lceil c_{16,10} \cdot m \rceil$ mit $c_{16,10} = \log_{10}(16) \approx 1{,}20412$	$4m$	m

Hierbei ist $\lceil x \rceil$ der nach oben auf die nächstgrößere ganze Zahl aufgerundete Wert von *x* und $\lfloor x \rfloor$ der auf die nächstkleinere ganze Zahl abgerundete Wert von *x*.

Werden zwei Zahlen $n_1 \in \mathbf{N}$ und $n_2 \in \mathbf{N}$ mit $n_1 \geq n_2$ addiert, vergrößert sich u.U. die Stellenzahl der Summe im Vergleich zur Stellenzahl von n_1. Ohne die Logarithmusfunktion bemü-

hen zu müssen, kann man die Stellenzahl der Summe $n_1 + n_2$ im Verhältnis zur Stellenzahl von n_1 abschätzen:

Die Zahl n_1 besitze *m* signifikante Stellen, d.h.

$$n_1 = \sum_{i=0}^{m-1} a_i \cdot B^i \text{ mit } a_i \in \{0,1,...,B-1\} \text{ und } a_{m-1} \neq 0 .$$

Der ungünstigste Fall liegt vor, wenn n_1 und n_2 möglichst groß sind, wenn also in n_1 alle Ziffern den Wert $B-1$ haben und $n_1 = n_2$ ist. Dann ist

$$n_1 = \sum_{i=0}^{m-1} (B-1) \cdot B^i = (B-1) \cdot \sum_{i=0}^{m-1} B^i = (B-1) \cdot \frac{B^m - 1}{B-1} = B^m - 1 \text{ und}$$

$$n_1 + n_2 = 2 \cdot \left(B^m - 1\right) = 1 \cdot B^m + \left(B^m - 2\right).$$

Diese Zahl belegt (in der Darstellung im Zahlensystem zur Basis *B*) $m+1$ Stellen:

$$n_1 + n_2 = \left[1 \underbrace{(B-1)...(B-1)}_{(m\text{-}1)\text{-mal}} (B-2) \right]_B .$$

Bei der Addition zweier natürlicher Zahlen nimmt also die Stellenzahl der Summe um höchstens eine Stelle (bezüglich der Stellenzahl der größeren Zahl) zu.

Bei der Multiplikation kann man eine ähnliche Betrachtung durchführen. Wieder liegt der ungünstigste Fall vor, wenn $n_1 = n_2 = B^m - 1$ ist. Dann ist

$$n_1 \cdot n_2 = \left(B^m - 1\right)^2 = B^{2 \cdot m} - 2 \cdot B^m + 1 = B^m \cdot \left(B^m - 2\right) + 1 .$$

Diese Zahl hat folgende Darstellung im Zahlensystem zur Basis *B*:

$$n_1 \cdot n_2 = \left[\underbrace{(B-1)...(B-1)}_{(m\text{-}1)\text{-mal}} (B-2) \underbrace{0...0}_{(m\text{-}1)\text{-mal}} 1 \right]_B ,$$

belegt also $2 \cdot m$ viele Stellen. Bei der Multiplikation zweier natürlicher Zahlen verdoppelt sich die Stellenzahl also höchstens (bezogen auf die Stellenzahl der größeren Zahl).

5.6 Einführung in die Differentialrechnung

Bei der Untersuchung des Kurvenverlaufs einer Funktion *f* ist es häufig notwendig zu wissen, wie sich der Wert von $f(x)$ ändert, wenn man sich von einem festen Wert x_0 „um einen kleinen Betrag“ bis zum Wert $x > x_0$ entfernt. Man vergleicht dabei die Änderung von

$$\Delta y = f(x) - f(x_0)$$

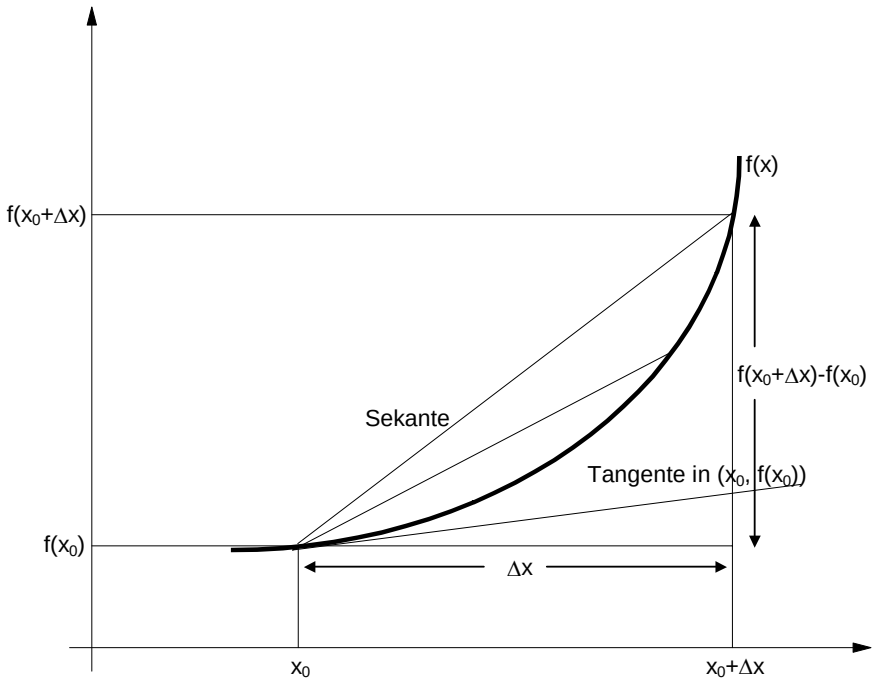
mit der Änderung von

$$\Delta x = x - x_0$$

und bildet den **Differenzenquotienten**

$$\frac{\Delta y}{\Delta x} = \frac{f(x) - f(x_0)}{x - x_0} = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}.$$

Geht man „nahe genug“ an x_0 heran, so wird bei vielen Funktionen der Differenzenquotient unabhängig von Δx und beschreibt dann eine charakteristische quantitative Eigenschaft der Funktion *f* im Punkt x_0: die **Steigung der Funktion *f* im Punkt** x_0 .



Im folgenden sei wieder $X \subseteq \mathbf{R}$ und $f: X \to \mathbf{R}$ eine Funktion.

Die Funktion $f: X \to \mathbf{R}$ heißt **an der Stelle** $x_0 \in X$ **differenzierbar**, wenn der Grenzwert

$$\lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

existiert. Dieser Grenzwert heißt **Ableitung von *f* an der Stelle** x_0.

Übliche Schreibweisen für die Ableitung von *f* an der Stelle x_0 sind:

$$\lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x},$$

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

$$\left.\frac{df(x)}{dx}\right|_{x=x_0},$$

$$f'(x_0).$$

Existiert dieser Grenzwert für jedes $x_0 \in X$, so heißt *f* (nach *x*) **differenzierbar**. $f'(x)$ ist eine Funktion von *x*.

Der Differenzenquotient

$$\frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

gibt die durchschnittliche Veränderung im Intervall $[x_0, x_0 + \Delta x]$ an und ist von x_0 und Δx abhängig. Er ist gleich der Steigung der Sekante zwischen den Punkten $(x_0, f(x_0))$ und $(x_0 + \Delta x, f(x_0 + \Delta x))$ des Graphen von *f*. Nach dem Grenzübergang $\Delta x \to 0$ ist der Quotient gleich der Steigung der Tangente an den Graphen von *f* im Punkt $(x_0, f(x_0))$ und ist nur von x_0 abhängig.

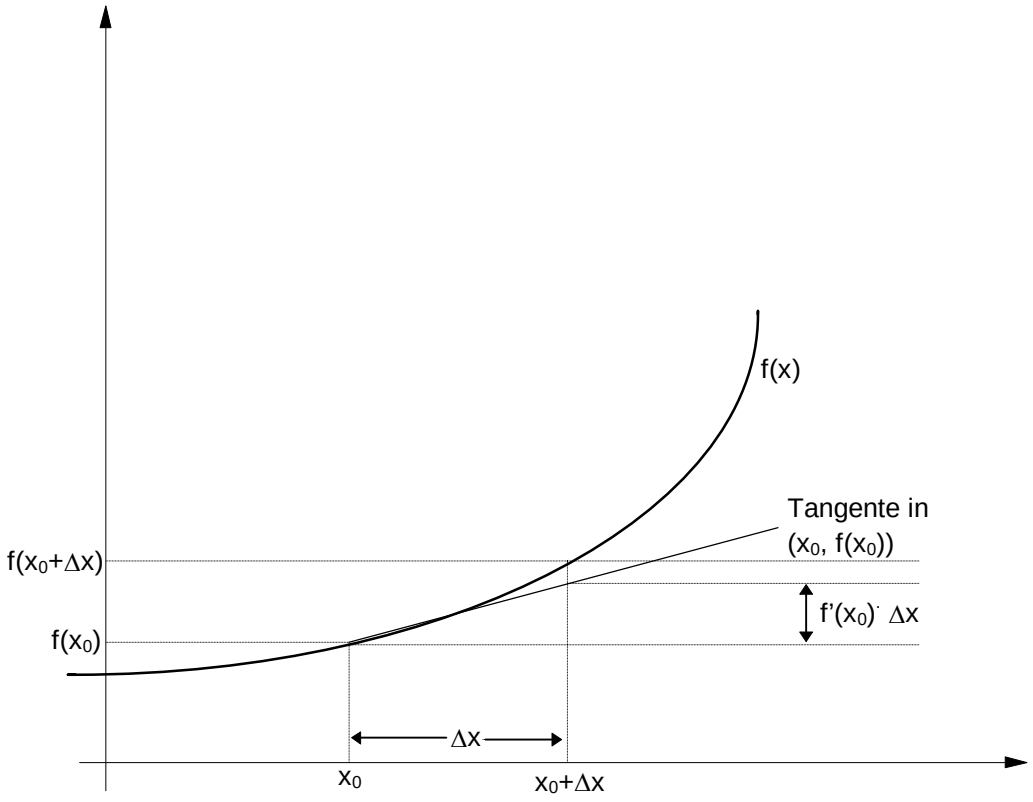
Die Tangente an den Graphen von *f* im Punkt $\left(x_0, f(x_0)\right)$ hat die Geradengleichung

$$y_T(x) = f'(x_0)\cdot(x - x_0) + f(x_0).$$

Im Punkt $x_0 + \Delta x$ hat die Tangente also den Wert

$$y_T(x_0 + \Delta x) = f'(x_0)\cdot\Delta x + f(x_0).$$

Der Wert $f'(x_0)\cdot\Delta x$ gibt also eine ***gute Näherung für die Veränderung von f*** von $f(x_0)$ bis zu $f(x_0 + \Delta x)$, wenn sich x_0 um einen kleinen Wert Δx ändert; diese Änderung ist proportional zu Δx (mit dem Proportionalitätsfaktor $f'(x_0)$).



Satz 5.6-1:

Ist $f: X \to \mathbf{R}$ in $x_0 \in X$ differenzierbar, so ist f in x_0 stetig.

Die Umkehrung gilt im allgemeinen nicht, d.h. aus der Stetigkeit einer Funktion in einem Punkt x_0 folgt i.a. nicht die Differenzierbarkeit in x_0.

Ein Beispiel für eine Funktion, die überall stetig, aber nicht überall differenzierbar ist, ist die Betragsfunktion

$$f_{|\,|}: \begin{cases} \mathbf{R} \to \mathbf{R} \\ x \to |x| \end{cases} .$$

Es gilt

$$\lim_{h\to 0} \frac{f_{|\,|}(x+h) - f_{|\,|}(x)}{h}\bigg|_{x=0} = \lim_{h\to 0} \frac{f_{|\,|}(h)}{h} = \lim_{h\to 0} \frac{|h|}{h} = \begin{cases} +1 & \text{für } h > 0 \\ -1 & \text{für } h < 0 \end{cases}$$

Der Grenzwert existiert also nicht, d.h. f ist in $x_0 = 0$ nicht differenzierbar (aber stetig).

Satz 5.6-2:

Die Funktionen $f: X \to \mathbf{R}$ und $g: X \to \mathbf{R}$ seien differenzierbar. Dann gilt:

(i) $\dfrac{d}{dx}\left(a \cdot f(x) + b \cdot g(x)\right) = a \cdot \dfrac{df(x)}{dx} + b \cdot \dfrac{dg(x)}{dx}$,

$\left(a \cdot f(x) + b \cdot g(x)\right)' = a \cdot f'(x) + b \cdot g'(x)$.

Hierbei sind *a* und *b* Konstanten, die insbesondere nicht von *x* abhängig sind.

(ii) $\dfrac{d}{dx}\left(f(x) \cdot g(x)\right) = \dfrac{df(x)}{dx} \cdot g(x) + f(x) \cdot \dfrac{dg(x)}{dx}$,

$\left(f(x) \cdot g(x)\right)' = f'(x) \cdot g(x) + f(x) \cdot g'(x)$

(**Produktregel**)

(iii) Für $g(x) \neq 0$ gilt:

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{\dfrac{df(x)}{dx} \cdot g(x) - f(x) \cdot \dfrac{dg(x)}{dx}}{\left(g(x)\right)^2},$$

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{\left(g(x)\right)^2}$$

(**Quotientenregel**)

(iv) $\dfrac{d}{dx}\left(f(g(x))\right) = \left.\dfrac{df(y)}{dy}\right|_{y=g(x)} \cdot \dfrac{dg(x)}{dx}$,

$\left(f(g(x))\right)' = f'(g(x)) \cdot g'(x)$

(**Kettenregel**)

(v) Hat f die Umkehrfunktion $\overset{-1}{f}$ und ist $f'(x_0) \neq 0$ für $x_0 \in X$, so ist für $y_0 = f(x_0)$:

$$\left.\frac{d}{dy}\overset{-1}{f}(y)\right|_{y=y_0} = \frac{1}{\left.\dfrac{df(x)}{dx}\right|_{x=x_0}},$$

$$\left[\overset{-1}{f}\left(f(x_0)\right)\right]' = \frac{1}{f'(x_0)}.$$

Für einige grundlegende Beispiele soll die jeweilige Ableitung in einem Punkt x_0 des Definitionsbereichs berechnet werden. Dazu wird entweder der Quotient $\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}$ und dann der Grenzübergang $\Delta x \to 0$ vollzogen oder es werden die Regeln des Satzes 5.5-2 mit bereits bekannten Ableitungen verwendet.

Beispiel:

Für die durch $f(x)=x^n$ mit $n \in \mathbf{N}$ definierte Funktion ist

$$\begin{aligned}
\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x} &= \frac{(x_0+\Delta x)^n - x_0^n}{\Delta x} \\
&= \frac{\sum_{i=0}^{n}\binom{n}{i}\cdot x_0^{n-i}\cdot \Delta x^i - x_0^n}{\Delta x} \\
&= \frac{x_0^n + \Delta x\cdot\sum_{i=1}^{n}\binom{n}{i}\cdot x_0^{n-i}\cdot \Delta x^{i-1} - x_0^n}{\Delta x} \\
&= \sum_{i=1}^{n}\binom{n}{i}\cdot x_0^{n-i}\cdot \Delta x^{i-1} \\
&= \underbrace{n\cdot x_0^{n-1}}_{i=1} + \Delta x\cdot\sum_{i=2}^{n}\binom{n}{i}\cdot x_0^{n-i}\cdot \Delta x^{i-2}, \quad \text{also}
\end{aligned}$$

$\lim\limits_{\Delta x\to 0}\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x} = n\cdot x_0^{n-1}$ bzw.

$\left(x^n\right)' = n\cdot x^{n-1}$.

Beispiel:

Für die durch $\overset{-1}{f}(x)=\sqrt[n]{x}=x^{1/n}$ mit $n \in \mathbf{N}$ definierte Funktion ist nach Satz 5.6-2 (v) (da $\overset{-1}{f}(x)$ die Umkehrfunktion zu der Funktion $f(x)=x^n$ des vorherigen Beispiels ist) mit $y_0 = f(x_0)$, d.h. $x_0 = \overset{-1}{f}(y_0)=\sqrt[n]{y_0}$,

$$\left.\frac{d}{dy}\overset{-1}{f}(y)\right|_{y=y_0} = \frac{1}{\left.\frac{df(x)}{dx}\right|_{x=x_0}} = \frac{1}{n\cdot x_0^{n-1}} = \frac{1}{n\cdot\left(\sqrt[n]{y_0}\right)^{n-1}} = \frac{1}{n\cdot\left(y_0^{1/n}\right)^{n-1}} = \frac{1}{n\cdot y_0^{1-1/n}} = \frac{1}{n}\cdot y_0^{1/n-1}, \text{ bzw.}$$

$$\left(x^{1/n}\right)' = \frac{1}{n}\cdot x^{\frac{1}{n}-1}.$$

Beispiel:

Für die durch $g(x) = x^q$ mit $q \in \mathbf{Q}$ und $q > 0$, etwa $q = \frac{n}{m}$ mit $n \in \mathbf{N}$ und $m \in \mathbf{N}_{>0}$, definierte Funktion ist gemäß Kettenregel (Satz 5.6-2 (iv)):

$$\left(x^{\frac{n}{m}}\right)' = \frac{d}{dx}\left(x^{\frac{n}{m}}\right) = \frac{d}{dx}\left(x^n\right)^{\frac{1}{m}} = \frac{1}{m}\cdot\left(x^n\right)^{\frac{1}{m}-1}\cdot n\cdot x^{n-1} = \frac{n}{m}\cdot x^{\frac{n}{m}-1}.$$

Beispiel:

Die Berechnung der Ableitung der Exponentialfunktion $\exp(x) = e^x$ erfolgt wieder direkt über die Definition der Ableitung. Dazu zunächst eine Vorbemerkung: Gemäß Satz 5.1-12 ist für einen kleinen Wert $|\Delta x|$:

$$\exp(\Delta x) = \sum_{i=0}^{1}\frac{\Delta x^i}{i!} + R_2(\Delta x) = 1 + \Delta x + R_2(\Delta x) \text{ mit } |R_2(\Delta x)| \le \frac{2\cdot|\Delta x|^2}{2!} = |\Delta x|^2.$$

Damit ist

$$\frac{\exp(x_0+\Delta x)-\exp(x_0)}{\Delta x} = \frac{\exp(x_0)\cdot\exp(\Delta x)-\exp(x_0)}{\Delta x} = \exp(x_0)\cdot\frac{\exp(\Delta x)-1}{\Delta x},$$

und der Grenzübergang ergibt

$$\begin{aligned}\lim_{\Delta x\to 0}\frac{\exp(x_0+\Delta x)-\exp(x_0)}{\Delta x} &= \lim_{\Delta x\to 0}\exp(x_0)\cdot\frac{\exp(\Delta x)-1}{\Delta x}\\ &= \exp(x_0)\cdot\lim_{\Delta x\to 0}\frac{\exp(\Delta x)-1}{\Delta x}\\ &= \exp(x_0)\cdot\lim_{\Delta x\to 0}\frac{(\Delta x+R_2(\Delta x))}{\Delta x}\\ &= \exp(x_0)\cdot\left(1+\lim_{\Delta x\to 0}\frac{R_2(\Delta x)}{\Delta x}\right).\end{aligned}$$

Wegen $\left|\frac{R_2(\Delta x)}{\Delta x}\right| \le |\Delta x|$ folgt damit

$$\left(\exp(x)\right)' = \left(e^x\right)' = e^x = \exp(x)\,.$$

Wegen $a^x = e^{\ln(a)\cdot x}$ ist mit der Kettenregel (Satz 5.6-2 (iv)):

$$\left(\exp_a(x)\right)' = \left(a^x\right)' = \left(e^{\ln(a)\cdot x}\right)' = e^{\ln(a)\cdot x}\cdot\ln(a) = \ln(a)\cdot a^x\,.$$

Beispiel:

Die Ableitung des natürlichen Logarithmus als Umkehrfunktion der Exponentialfunktion wird wieder mit Hilfe von Satz 5.6-2 (v) berechnet:

Es sei $y_0 = \exp(x_0)$, d.h. $x_0 = \ln(y_0)$.

$$\frac{d}{dy}\overset{-1}{\exp}(y)\bigg|_{y=y_0} = \frac{1}{\left.\frac{d\exp(x)}{dx}\right|_{x=x_0}} = \frac{1}{\exp(x_0)} = \frac{1}{y_0}\,,\ \text{also}$$

$$\left(\ln(x)\right)' = \frac{1}{x}\,.$$

Die Ableitung der Logarithmusfunktion zu einer Basis $a > 1$ lautet nun

$$\left(\log_a(x)\right)' = \left(\frac{\ln(x)}{\ln(a)}\right)' = \frac{1}{\ln(a)\cdot x}\,.$$

Beispiel:

Es kann nun auch die Ableitung der durch $h(x) = x^r$ mit $r \in \mathbf{R}$ bestimmten Funktion ermittelt mit Hilfe der Kettenregel (Satz 5.6-2 (iv)) werden:

$$\left(x^r\right)' = \left(e^{\ln(x)\cdot r}\right)' = e^{\ln(x)\cdot r} \cdot r \cdot \frac{1}{x} = r \cdot x^r \cdot \frac{1}{x} = r \cdot x^{r-1}.$$

Beispiel:

Die durch $k(x) = \left(x^2+1\right)^x$ gegebene Funktion hat die Eigenschaft, dass x sowohl in der „Basis“ als auch im Exponenten vorkommt. In diesem Fall wendet man den Trick des Logarithmierens mit Satz 5.5-5 (iii) an:

$$\ln\left(k(x)\right) = x \cdot \ln\left(x^2+1\right);$$

die Ableitung der linken Seite lautet:

$$\left(\ln\left(k(x)\right)\right)' = \frac{\left(k(x)\right)'}{k(x)},$$

die Ableitung der rechten Seite lautet:

$$\left(x \cdot \ln\left(x^2+1\right)\right)' = 1 \cdot \ln\left(x^2+1\right) + x \cdot \frac{2 \cdot x}{x^2+1};$$

beide Seiten werden gleichgesetzt und die entstandene Gleichung nach $\left(k(x)\right)'$ aufgelöst:

$$\frac{\left(k(x)\right)'}{k(x)} = \ln\left(x^2+1\right) + \frac{2 \cdot x^2}{x^2+1},$$

$$\left(k(x)\right)' = \left(x^2+1\right)^x \cdot \left(\ln\left(x^2+1\right) + \frac{2 \cdot x^2}{x^2+1}\right).$$

Die folgende Tabelle fasst die Ergebnisse der Beispiele zusammen.

$f(x)$	$f'(x)$
$x^r, r \in \mathbf{R}$	$r \cdot x^{r-1}$
$c = const.$	0
$\sum_{i=0}^{n} a_i \cdot x^i$	$\sum_{i=0}^{n} i \cdot a_i \cdot x^{i-1}$
$1/x^n$	$-n/x^{n+1}$
$\sqrt{h(x)}$	$\frac{h'(x)}{2\sqrt{h(x)}}$
$\ln(x) = \log_e(x)$	$1/x$
$\ln(h(x))$	$\frac{h'(x)}{h(x)}$
$\log_a(x)$	$\frac{1}{x \cdot \ln(a)}$
e^x	e^x
$a^x, a > 0, a = const.$	$a^x \cdot \ln(a)$
$e^{h(x)}$	$h'(x) \cdot e^{h(x)}$
$a^{h(x)}, a > 0, a = const.$	$h'(x) \cdot a^{h(x)} \cdot \ln(a)$

Die Funktion $f: X \to \mathbf{R}$ sei differenzierbar (und damit auch stetig). Dann ist $f': X \to \mathbf{R}$ ebenfalls eine Funktion, die aber nicht unbedingt differenzierbar oder stetig sein muss. Ist sie jedoch differenzierbar, so kann man

$$\frac{df'(x)}{dx}$$

bilden und nennt dieses die 2. Ableitung von *f*.

Allgemein werden **Ableitungen höherer Ordnung** wie folgt definiert:

Es ist

$$f^{(0)}(x) = f(x);$$

ist die $(n-1)$-te Ableitung der Funktion $f: X \to \mathbf{R}$ im Intervall $I \subseteq X$ differenzierbar, so ist die *n*-te Ableitung von *f* gegeben durch

$$f^{(n)}(x) = \frac{d}{dx} f^{(n-1)}(x).$$

Existieren für *f* alle Ableitungen bis zur *n*-ten Ableitung, so heißt *f* ***n*-mal differenzierbar**.

Beispiel:

Für die durch $f(x) = x \cdot e^x$ definierte Funktion lauten die ersten beiden Ableitungen:

$$f'(x) = 1 \cdot e^x + x \cdot e^x = (1+x) \cdot e^x,$$
$$f''(x) = 1 \cdot e^x + (1+x) \cdot e^x = (2+x) \cdot e^x.$$

Zu vermuten ist, dass die *n*-te Ableitung $f^{(n)}(x) = (n+x) \cdot e^x$ lautet. Für $n = 0,1,2$ stimmt dieses, und die Vermutung gelte für $n \geq 2$. Die $(n+1)$-te Ableitung ist dann

$$f^{(n+1)}(x) = \left((n+x) \cdot e^x\right)' = 1 \cdot e^x + (n+x) \cdot e^x = (n+1+x) \cdot e^x, \text{ d.h.}$$

die Vermutung gilt für jedes $n \in \mathbf{N}$.

Beispiel:

Für das Polynom $p(x) = x^m$ lauten alle Ableitungen:

$$p^{(n)}(x) = \begin{cases} m \cdot (m-1) \cdot \ldots \cdot (m-n+1) \cdot x^{m-n} & \text{für } n \leq m \\ 0 & \text{für } n > m. \end{cases}$$

Ableitungen höherer Ordnung werden insbesondere zur Untersuchung des Kurvenverlaufs von Graphen zu reellen Funktionen (**Kurvendiskussion**) eingesetzt. Diesem Thema ist der Rest des Kapitels gewidmet.

Die Funktion $f: X \to \mathbf{R}$ hat an der Stelle $x_0 \in X$ ein **(lokales) Maximum**, wenn es eine ε-Umgebung $U(x_0, \varepsilon) = \{ x \mid |x - x_0| < \varepsilon \} = \{ x \mid x_0 - \varepsilon < x < x_0 + \varepsilon \}$ von x_0 gibt, so dass für alle $x \in U(x_0, \varepsilon)$ mit $x \neq x_0$ gilt: $f(x) < f(x_0)$.

Die Funktion $f: X \to \mathbf{R}$ hat an der Stelle $x_0 \in X$ ein **(lokales) Minimum**, wenn es eine ε-Umgebung $U(x_0, \varepsilon) = \{ x \mid |x - x_0| < \varepsilon \} = \{ x \mid x_0 - \varepsilon < x < x_0 + \varepsilon \}$ von x_0 gibt, so dass für alle $x \in U(x_0, \varepsilon)$ mit $x \neq x_0$ gilt: $f(x) > f(x_0)$.

Unter einem **(lokalen) Extremwert** versteht man ein lokales Maximum oder ein lokales Minimum.

Die Funktion $f: X \to \mathbf{R}$ hat an der Stelle x_W einen **Wendepunkt**, wenn es ε-Umgebung $U(x_W, \varepsilon) = \{ x \mid |x - x_W| < \varepsilon \} = \{ x \mid x_W - \varepsilon < x < x_W + \varepsilon \}$ von x_W gibt, so dass f für jedes $x \in U(x_W, \varepsilon)$ mit $x_W - \varepsilon < x < x_W$ streng konvex und für jedes $x \in U(x_W, \varepsilon)$ mit $x_W < x < x_W + \varepsilon$ streng konkav ist bzw. für jedes $x \in U(x_W, \varepsilon)$ mit $x_W - \varepsilon < x < x_W$ streng konkav und für jedes $x \in U(x_W, \varepsilon)$ mit $x_W < x < x_W + \varepsilon$ streng konvex ist.

Satz 5.6-3:

Die Funktion $f: X \to \mathbf{R}$ sei an der Stelle $x_0 \in X$ mindestens *n*-mal differenzierbar.

Ist

$f^{(k)}(x_0) = 0$ für $k = 1, \ldots, n-1$ und
$f^{(n)}(x_0) \neq 0$,
und ist *n gerade*,

so hat *f* an der Stelle x_0 einen Extremwert, und zwar ein (lokales) Maximum, wenn $f^{(n)}(x_0) < 0$ ist bzw. ein (lokales) Minimum, wenn $f^{(n)}(x_0) > 0$ ist.

Ist

$f^{(k)}(x_0) = 0$ für $k = 2, \ldots, n-1$ und
$f^{(n)}(x_0) \neq 0$
und ist *n ungerade*,

so hat *f* an der Stelle einen Wendepunkt; die Krümmung wechselt von konvex nach konkav, wenn $f^{(n)}(x_0) < 0$ ist; sie wechselt von konkav nach konvex, wenn $f^{(n)}(x_0) > 0$ ist. Gilt zusätzlich $f'(x_0) = 0$, so liegt ein Wendepunkt mit waagerechter Tangente (**Sattelpunkt**) vor.

Das folgende Beispiel untersucht den Kurvenverlauf des Graphen zum Polynom, das durch

$$p(x) = 0{,}2 \cdot x^5 - x^3 + 1$$

definiert wird. Die Ableitungen lauten:

$$p'(x) = x^4 - 3 \cdot x^2,$$
$$p''(x) = 4 \cdot x^3 - 6 \cdot x,$$
$$p'''(x) = 12 \cdot x^2 - 6,$$
$$p^{(4)}(x) = 24 \cdot x,$$
$$p^{(5)}(x) = 24,$$

$p^{(6)}(x) = 0$.

Die erste Ableitung $p'(x)$ hat die Nullstellen $x_{0,1} = 0$, $x_{0,2} = \sqrt{3}$, $x_{0,3} = -\sqrt{3}$. Diese Werte werden in die höheren Ableitungen eingesetzt:

$p''(x_{0,1}) = p''(0) = 0$, $p'''(x_{0,1}) = p'''(0) = -6$, also liegt hier ein Wendeüunkt mit waagerechter Tangente (Sattelpunkt) vor.

$p''(x_{0,2}) = p''(\sqrt{3}) = 6 \cdot \sqrt{3} > 0$, also liegt hier ein lokales Minimum vor.

$p''(x_{0,3}) = p''(-\sqrt{3}) = -6 \cdot \sqrt{3} < 0$, also liegt hier ein lokales Maximum vor.

Die Nullstellen der zweiten Ableitung lauten $x_{0,4} = x_{0,1} = 0$, $x_{0,5} = \sqrt{3/2}$, $x_{0,6} = -\sqrt{3/2}$. Die Werte $x_{0,5}$ und $x_{0,6}$ in die dritte Ableitung eingesetzt ergeben $p'''(x_{0,5}) = p'''(x_{0,6}) = 12$. Es liegen also an diesen Werten Wendepunkte vor.

Satz 5.6-4:

Ist die Funktion $f: X \to \mathbf{R}$ im Intervall $I \subseteq X$ differenzierbar, so sind folgende Aussagen (a) und (b) gleichbedeutend:

(a) f ist in I monoton fallend (bzw. steigend).
und
(b) Für jedes $x \in I$ gilt $f'(x) \leq 0$ (bzw. $f'(x) \geq 0$).

Satz 5.6-5:

Ist die Funktion $f: X \to \mathbf{R}$ im Intervall $I \subseteq X$ zweimal differenzierbar, so sind folgende Aussagen (a) bis (d) gleichbedeutend:

(a) f ist in I konvex (bzw. konkav).
und
(b) Für Werte $x_1 \in I$ und $x_2 \in I$ mit $x_1 < x_2$ gilt $\frac{f(x_2 - f(x_1)}{x_2 - x_1} \geq f'(x_1)$ (bzw. $\frac{f(x_2 - f(x_1)}{x_2 - x_1} \leq f'(x_1)$).
und
(c) Die Ableitung $f'(x)$ ist in I monoton steigend (bzw. monoton fallend).
und
(d) Für jedes $x \in I$ ist $f''(x) \geq 0$ (bzw. $f''(x) \leq 0$).

Beispiel:

Zu untersuchen ist das Krümmungsverhalten des durch

$$p(x) = -\frac{1}{10} \cdot x^5 + x^3$$

definierten Polynoms. Seine zweite Ableitung lautet

$$p''(x) = -2 \cdot x^3 + 6 \cdot x = -2 \cdot x \cdot \left(x^2 - 3\right).$$

Es ist

$p''(x) \leq 0$ genau dann wenn $(x \geq 0) \wedge \left(x^2 - 3 \geq 0\right)$ oder $(x \leq 0) \wedge \left(x^2 - 3 \leq 0\right)$ gilt. Im ersten Fall gilt $(x \geq 0) \wedge \left(\left(x \geq \sqrt{3}\right) \vee \left(x \leq -\sqrt{3}\right)\right)$, also $x \geq \sqrt{3}$. Im zweiten Fall ist $(x \leq 0) \wedge \left(\left(-\sqrt{3} \leq x \leq \sqrt{3}\right)\right)$, also $-\sqrt{3} \leq x \leq 0$. Für $x \geq \sqrt{3}$ oder für $-\sqrt{3} \leq x \leq 0$ ist also p konkav, für alle übrigen Bereiche konvex.

5.7 Die Regel von de l'Hospital

Häufig sind Grenzwerte der Form

$$\lim_{x \to x_0} f(x)$$

zu berechnen, wobei

$$f(x) = \frac{g(x)}{h(x)} \text{ und } \lim_{x \to x_0} g(x) = \lim_{x \to x_0} h(x) = 0$$

gelten. In diesem Fall ist der folgende Satz von Bedeutung (**Regel von de l'Hospital**, 1661-1704):

Satz 5.7-1:

Gegeben seien die Funktionen $g: X \to \mathbf{R}$ und $h: X \to \mathbf{R}$, $X \subseteq \mathbf{R}$. Dabei seien g und h $(n + 1)$-mal differenzierbar und ihre $(n + 1)$-te Ableitungen stetig. Für $x_0 \in X$ gelte

$$(*) \quad g(x_0) = g'(x_0) = \ \ldots \ = g^{(n)}(x_0) = 0,$$
$$h(x_0) = h'(x_0) = \ \ldots \ = h^{(n)}(x_0) = 0 \text{ und } h^{(n+1)}(x_0) \neq 0.$$

Dann gilt:

$$\lim_{x \to x_0} \frac{g(x)}{h(x)} = \lim_{x \to x_0} \frac{g'(x)}{h'(x)} = \ \ldots \ = \lim_{x \to x_0} \frac{g^{(n)}(x)}{h^{(n)}(x)} = \lim_{x \to x_0} \frac{g^{(n+1)}(x)}{h^{(n+1)}(x)} = \frac{g^{(n+1)}(x_0)}{h^{(n+1)}(x_0)} \quad .$$

Es werden also Zähler- und Nennerfunktion getrennt abgeleitet.

Beispiele:

$$\lim_{x \to 1} \frac{x^n - 1}{x - 1} = n$$

$$\lim_{x \to 0} \frac{e^x - 1}{x} = 1$$

Der hier zitierte Satz steht für den „Fall 0/0“. Der Satz gilt auch, wenn eventuell $x_0 \notin X$ (als einzelner Wert) ist und die Bedingung $(*)$ durch

$$(**)\quad g(x_0) = g'(x_0) = \ \ldots\ = g^{(n)}(x_0) = \infty,$$
$$h(x_0) = h'(x_0) = \ \ldots\ = h^{(n)}(x_0) = \infty \quad \text{und} \quad h^{(n+1)}(x) \neq 0 \text{ für jedes } x \in X.$$

ersetzt wird („Fall ∞ / ∞“).

Die Regel von de l'Hospital ermöglicht auch die Berechnung unbestimmter Ausdrücke der Art $\infty - \infty,\ 0 \cdot \infty,\ 1^{\infty},\ \infty^0,\ 0^0$. Diese werden zunächst so umgeformt, dass der „Fall 0/0“ oder der „Fall ∞ / ∞“ entsteht. Die folgende Tabelle gibt die Umformungen auf den „Fall 0/0“ an:

Typ	Funktion	Umformung	„Fall 0/0“
$\infty - \infty$	$g(x) - h(x)$	-	$\dfrac{1/h(x) - 1/g(x)}{1/h(x) \cdot 1/g(x)}$
$\infty - \infty$	$g(x) - h(x)$	Exponentieren	$\dfrac{1/e^{h(x)}}{1/e^{g(x)}}$
$0 \cdot \infty$	$g(x) \cdot h(x)$	-	$\dfrac{g(x)}{1/h(x)}$
1^{∞}	$g(x)^{h(x)}$	Logarithmieren	$\dfrac{\ln(g(x))}{1/h(x)}$
∞^0	$g(x)^{h(x)}$	Logarithmieren	$\dfrac{h(x)}{1/\ln(g(x))}$
0^0	$g(x)^{h(x)}$	Logarithmieren	$\dfrac{h(x)}{1/\ln(g(x))}$

Beispiel:

Es soll $\lim\limits_{x \to 0} (1+x)^{1/x}$ bestimmt werden. Dieser Grenzwert ist vom Typ „1^{∞}“ mit den Funktionen $g(x) = 1 + x$ und $h(x) = 1/x$. Logarithmieren der gesamten Funktion ergibt

$$\ln\left(g(x)^{h(x)}\right) = h(x)\cdot\ln\left(g(x)\right) = \frac{\ln(g(x))}{1/h(x)} = \frac{\ln(1+x)}{x}.$$

Jetzt liegt der „Fall 0/0“ vor:

$$\lim_{x\to 0}\frac{\ln(1+x)}{x} = \lim_{x\to 0}\frac{1/1+x}{1} = 1;$$

Die Logarithmierung wird durch Exponentiation wieder rückgängig gemacht, also

$$\lim_{x\to 0}(1+x)^{1/x} = e^1 = e.$$

Beispiel:

In Satz 5.5-6 (ii) wird formuliert, dass für jedes $a \in \mathbf{R}$ mit $a > 1$ und jedes Polynom $p(x)$

$$\lim_{x\to\infty}\frac{|p(x)|}{a^x} = 0$$

gilt. Mit Hilfe der Regel von de l'Hospital für den „Fall ∞/∞“ lässt sich dieses verifizieren:

Es sei $p(x)$ ein Polynom vom Grade *n*, d.h. $p(x) = \sum_{i=0}^{n} a_i \cdot x^i$ mit $a_n \neq 0$. Es erfolgt eine Beschränkung auf den Fall $p(x) \geq 0$, so dass in der Limesbetrachtung auf die Betragsstriche verzichtet werden kann. Dann ist

$$\lim_{x\to\infty}\frac{p(x)}{a^x} = \lim_{x\to\infty}\frac{(p(x))^{(n)}}{(a^x)^{(n)}} = \lim_{x\to\infty}\frac{n!\cdot a_n}{a^x\cdot(\ln(a))^n} = 0.$$

Beispiel:

In Satz 5.5-6 (ii) wird formuliert, dass für jedes $a \in \mathbf{R}$ mit $a > 1$ und $m \in \mathbf{N}$

$$\lim_{x\to\infty}\frac{(\log_a(x))^m}{x} = 0$$

gilt. Dieser Grenzwert ist ebenfalls ein Beispiel für den „Fall ∞/∞“; er wird verifiziert durch

$$\lim_{x\to\infty}\frac{(\log_a(x))^m}{x} = \lim_{x\to\infty}\frac{m\cdot(\log_a(x))^{m-1}\Big/ x\cdot\ln(a)}{1} = \lim_{x\to\infty}\frac{m\cdot(\log_a(x))^{m-1}}{x\cdot\ln(a)}$$

$$= \lim_{x\to\infty}\frac{m\cdot(m-1)\cdot(\log_a(x))^{m-2}\Big/ x\cdot\ln(a)}{\ln(a)} = \lim_{x\to\infty}\frac{m\cdot(m-1)\cdot(\log_a(x))^{m-2}}{x\cdot(\ln(a))^2}$$

$$= \ldots = \lim_{x\to\infty}\frac{m\cdot(m-1)\cdot\ldots\cdot(m-(m-2))\cdot(\log_a(x))^{m-(m-1)}}{x\cdot(\ln(a))^{m-1}}$$

$$= \lim_{x\to\infty}\frac{m!\cdot(\log_a(x))}{x\cdot(\ln(a))^{m-1}} = \lim_{x\to\infty}\frac{m!}{x\cdot(\ln(a))^m} = 0 \quad .$$

5.8 Das Newton-Verfahren

Bei der Lösung von Gleichungen kommt es häufig vor, dass eine explizite Auflösung nach der unbekannten Größe nicht möglich ist. Man ist dann an einer numerischen Lösung interessiert. Ähnliche Probleme ergeben sich bei der numerischen Bestimmung von Nullstellen von Funktionen.

Gegeben sei eine Funktion $f: X \to \mathbf{R}$, die auf einem Intervall $I = [a,b]$, $I \subseteq X$ mindestens zweimal differenzierbar mit stetiger 2. Ableitung ist. Außerdem seien folgende Bedingungen 1. bis 4. erfüllt:

1. $f(a)\cdot f(b) < 0$, d.h. *f* hat im Intervall *I* eine Nullstelle (das ergibt sich aus der Stetigkeit von *f* und der Tatsache, dass *f* im Intervall *I* das Vorzeichen wechselt, sie Satz 5.2-3(i)).

2. $f'(x) \neq 0$ für jedes $x \in I$, d.h. die Nullstelle ist eindeutig (da in *I* kein Extremwert von *f* existiert).

3. $f''(x) \leq 0$ oder $f''(x) \geq 0$ für jedes $x \in I$, d.h. *f* ist entweder konkav oder konvex auf *I*.

4. Bezeichnet *c* denjenigen Endpunkt von $[a,b]$, für den $|f'(x)|$ kleiner ist als am anderen Endpunkt, so gilt

$$\left|\frac{f(c)}{f'(c)}\right| \leq b-a\,,$$

d.h. die Tangente an den Graph von *f* in demjenigen Endpunkt des Intervalls *I*, für den $|f'(x)|$ am kleinsten ist, schneidet die *x*-Achse im Intervall *I*.

Gesucht wird eine Lösung der Gleichung

$f(x) = 0$ mit $x \in I$.

Bei diesen Voraussetzungen über *f* approximiert das folgende Verfahren die gesucht Lösung $x_0 \in I$ (für die dann $f(x_0) = 0$ gilt):

Man wählt einen beliebigen Punkt $a_0 \in I$.
Man berechnet für *n* = 0, 1, 2, ...

$$a_{n+1} = a_n - \frac{f(a_n)}{f'(a_n)},$$

bis sich aufeinanderfolgende Werte von a_{n+1} und a_n nur noch „wenig“ unterscheiden (weniger als eine vorgegebene Genauigkeitsschranke).

Die so definierte Folge $(a_n)_{n\in\mathbf{N}}$ approximiert die gesuchte Lösung $x_0 \in I$ der Gleichung $f(x) = 0$.

Beispiele:

Zur Bestimmung der Quadratwurzel $\sqrt{c}$ einer reellen Zahl *c* > 0 wählt man

$f(x) = x^2 - c$.

Die Folge $(a_n)_{n\in\mathbf{N}}$ lautet hierbei:

$$a_0 = c/2,$$
$$a_{n+1} = \frac{1}{2}\cdot\left(a_n + \frac{c}{a_n}\right),\ n \in \mathbf{N}.$$

Zur Bestimmung der beliebigen Wurzel $\sqrt[k]{c} = c^{1/k}$ mit $k \in \mathbf{N}_{>0}$ einer reellen Zahl *c* > 0 wählt man

$f(x) = x^k - c$.

Die Folge $(a_n)_{n \in \mathbf{N}}$ lautet hierbei:

$a_0 > 0$ beliebig,
$a_{n+1} = \left(1 - \frac{1}{k}\right) \cdot a_n + \frac{1}{k} \cdot c \cdot a_n^{1\text{-k}}$, $n \in \mathbf{N}$.

Zur Berechnung des inversen Werts $1/c$ einer reellen Zahl $c > 0$ sind keine Divisionen erforderlich: Man wählt

$f(x) = \frac{1}{x} - c$.

Die Folge $(a_n)_{n \in \mathbf{N}}$ lautet hierbei:

$a_0 =$ beliebig mit $0 < a_0 < 2c^{-1}$ (Schätzung),
$a_{n+1} = a_n \cdot (2 - c \cdot a_n)$, $n \in \mathbf{N}$.

Das Newton-Verfahren ist robust gegen Rundungsfehler. Ein Iterationsschritt im Verfahren, d.h. die Berechnung eines weiteren Werts a_{n+1}, verwendet nur den Wert a_n und nicht vorherige Werte, etwa a_{n-1}, a_{n-2}, ..., a_0. Der Wert a_{n+1} hängt also nur von a_n ab. Derartige „einstellige“ Iterationsverfahren haben den Vorteil, dass sich Rundungsfehler nicht akkumulieren.

Außerdem zeigt das Newton-Verfahren ein gutes Konvergenzverhalten („quadratische Konvergenz“), d.h. nach wenigen Iterationsschritten bekommt man bereits eine gute Näherung an die gesuchte Lösung.

Beispiel:

Es wird eine Nullstelle der durch $f(x) = x^3 - x^2 + 2 \cdot x + 5$ gegebenen Funktion gesucht. Es ist $f'(x) = 3 \cdot x^2 - 2 \cdot x + 2$ und $f''(x) = 6 \cdot x - 2$. Als „Suchintervall“ für eine Nullstelle kann $I = [-2, -1]$ genommen werden. Hierfür sind alle obigen Bedingungen 1. bis 4. erfüllt. Als Startwert der Iteration wird $a_0 = -1{,}5$ gewählt. Die folgende Tabelle zeigt das Ergebnis nach

dem Newton-Verfahren nach 6 Iterationsschritten (ermittelt mit einem Tabellenkalkulationsprogramm). Zusätzlich ist das Ergebnis angegeben, das das Tabellenkalkulationsprogramm mit der eingebauten „Berechne-für"-Funktion bei 100 Iterationen liefert. Offensichtlich ist hier das Newton-Verfahren bei weitem überlegen.

n	a_n	$f(a_n)$	$f'(a_n)$
0	− 1,5	− 3,625	11,75
1	− 1,1914893617021	− 0,49411980004431	8,6419194205523
2	− 1,1343122722156	− 0,014768016090808	8,1286175371279
3	− 1,1324954791357	$-1{,}4526940122457 \cdot 10^{-05}$	8,1126289890596
4	− 1,1324936884782	$-1{,}4100025400726 \cdot 10^{-11}$	8,1126132402848
5	− 1,1324936884764	$-7{,}6501305290577 \cdot 10^{-16}$	8,1126132402696
6	− 1,1324936884764	$-7{,}6501305290577 \cdot 10^{-16}$	8,1126132402696

Ergebnis der eingebauten „Berechne-für"-Funktion bei 100 Iterationen:
$x_0 = -1{,}1324940415747 \qquad f(x_0) = -2{,}8645504939842 \cdot 10^{-06}$

5.9 Taylorpolynome

Im folgenden sei $f: X \to \mathbf{R}$ mit $X \subseteq \mathbf{R}$ eine „genügend oft" differenzierbare Funktion. Der Wert $x_0 \in X$ sei ein festgewählter Punkt. Der Funktionsverlauf von *f* soll durch eine Folge $(T_n(x; x_0; f))_{n\in\mathbf{N}}$ „einfacherer" Funktionen angenähert werden, die mit *f* im Punkt $(x_0, f(x_0))$ übereinstimmen und folgenden Bedingungen genügen:

$T_n(x; x_0; f)$ für $n \geq 0$ ist dasjenige Polynom *n*-ten Grades, das mit *f* an der Stelle x_0 übereinstimmt und dessen sämtliche Ableitungen bis zur *n*-ten Ableitung mit den entsprechenden Ableitungen von *f* bei x_0 übereinstimmen. Zur Vereinfachung der Rechnung wird dabei nicht der Ansatz $T_n(x; x_0; f) = \sum_{i=0}^{n} b_i \cdot x^i$ gewählt, sondern

$T_n(x; x_0; f) = a_n \cdot (x - x_0)^n + a_{n-1} \cdot (x - x_0)^{n-1} + \ldots + a_1 \cdot (x - x_0) + a_0$ mit
$T_n(x_0; x_0; f) = f(x_0),$
$T_n'(x_0; x_0; f) = f'(x_0),$
...
$T_n^{(n)}(x_0; x_0; f) = f^{(n)}(x_0).$

Zur Berechnung der Koeffizienten $a_0, ..., a_n$ wird $T_n(x; x_0; f)$ nacheinander nach x abgeleitet und x_0 eingesetzt:

$$T_n^{(0)}(x; x_0; f)\big|_{x=x_0} = T_n(x_0; x_0; f) = a_0 = f(x_0),$$

$$T_n'(x; x_0; f)\big|_{x=x_0} = n \cdot a_n \cdot (x-x_0)^{n-1} + (n-1) \cdot a_{n-1} \cdot (x-x_0)^{n-2} + ... + a_1\big|_{x=x_0} = a_1 = f'(x_0),$$

$$T_n''(x; x_0; f)\big|_{x=x_0} = n \cdot (n-1) \cdot a_n \cdot (x-x_0)^{n-2} + (n-1) \cdot (n-2) \cdot a_{n-1} \cdot (x-x_0)^{n-3} + ... + 2 \cdot a_2\big|_{x=x_0}$$

$$= 2 \cdot a_2 = f''(x_0), \text{ also } a_2 = \frac{1}{2} \cdot f''(x_0),$$

...

$$T_n^{(n-1)}(x; x_0; f)\big|_{x=x_0} = n \cdot (n-1) \cdot ... \cdot 2 \cdot a_n \cdot (x-x_0) + (n-1)! \cdot a_{n-1}\big|_{x=x_0} = (n-1)! \cdot a_{n-1}$$

$$= f^{(n-1)}(x_0), \text{ also } a_{n-1} = \frac{1}{(n-1)!} \cdot f^{(n-1)}(x_0),$$

$$T_n^{(n)}(x; x_0; f)\big|_{x=x_0} = n \cdot (n-1) \cdot ... \cdot 2 \cdot 1 \cdot a_n\big|_{x=x_0} = n! \cdot a_n$$

$$= f^{(n)}(x_0), \text{ also } a_n = \frac{1}{n!} \cdot f^{(n)}(x_0).$$

Insgesamt ergibt sich also

$$T_n(x; x_0; f) = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!} \cdot (x-x_0)^i .$$

Die Polynome $T_n(x; x_0; f)$ für $n = 0$ und $n = 1$ lauten:

$n = 0$: $T_0(x; x_0; f) = f(x_0)$, ist also die konstante Funktion mit Wert $f(x_0)$.

$n = 1$: $T_1(x; x_0; f) = f(x_0) + f'(x_0) \cdot (x-x_0)$, d.h. T_1 ist die Tangente an den Graphen von f im Punkt $(x_0, f(x_0))$.

Selbstverständlich ist $f(x)$ in der Regel nicht gleich $T_n(x; x_0; f)$, sondern es gilt

$$f(x) = T_n(x; x_0; f) + R_n(x; x_0; f)$$

mit einer als Restglied bezeichneten Funktion $R_n(x; x_0; f)$. Mit Sätzen der Differentialrechnung, die in den bisherigen Kapiteln nicht zitiert wurden (siehe daher angegebene Literatur), lässt sich das Restglied funktional ausdrücken. Das Ergebnis ist in folgendem Satz zusammengefasst.

Satz 5.7-2:

Die Funktion $f: X \to \mathbf{R}$ sei in einer ε-Umgebung $U(x_0, \varepsilon) = \{ x \mid |x - x_0| < \varepsilon \} \subseteq X$ von $x_0 \in X$ $(n+1)$-mal differenzierbar. Dann gilt für alle $x \in U(x_0, \varepsilon)$:

$$f(x) = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!} \cdot (x - x_0)^i + R_n(\mathrm{x}; x_0; f).$$

Die Summe

$$T_n(x; x_0; f) = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!} \cdot (x - x_0)^i$$

heißt **Taylorpolynom *n*-ter Ordnung von *f* an der Stelle** x_0; $R_n(x; x_0; f)$ heißt **Restglied des Taylorpolynoms *n*-ter Ordnung von *f* an der Stelle** x_0. Die Darstellung von $f(x)$ in der Form $f(x) = T_n(x; x_0; f) + R_n(x; x_0; f)$ nennt man **Taylorentwicklung von *f*** an der Stelle x_0**.**

Für das Restglied gilt:

$$R_n(x; x_0; f) = \frac{1}{(n+1)!} f^{(n+1)}(z) \cdot (x - x_0)^{n+1}.$$

Dabei ist *z* ein Wert mit $x_0 < z < x$, falls $x_0 < x$ ist, bzw. mit $x < z < x_0$, falls $x < x_0$ ist. Für $x = x_0$ ist $R_n(x) = 0$.

Bemerkung: In der mathematischen Literatur werden noch andere Darstellungen des Restglieds angegeben.

Beispiel:

Es soll die Taylorentwicklung für die Funktion

$$f(x) = e^x$$

an der Stelle $x_0 = 0$ berechnet werden. Dabei sollen nur die Ableitungen von $f(x)$ verwendet werden. Da für alle Ableitungen $f^{(i)}(x) = e^x$ gilt und $e^0 = 1$ ist, ergibt sich für das *n*-te Taylorpolynom von $f(x) = e^x$ an der Stelle $x_0 = 0$:

$$T_n\left(x; 0; e^x\right) = \sum_{i=0}^{n} \frac{1}{i!} \cdot x^i = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \ldots + \frac{x^n}{n!}.$$

Mit Satz 5.7-2 ist

$$e^x = \sum_{i=0}^{n} \frac{1}{i!} \cdot x^i + R_n\left(x; 0; e^x\right) = \sum_{i=0}^{n} \frac{1}{i!} \cdot x^i + \frac{1}{(n+1)!} \cdot e^z \cdot x^{n+1} \text{ mit } 0 < z < x \text{ für } x > 0 \text{ und}$$

$$x < z < 0 \text{ für } x < 0.$$

Nun gilt:

$$\lim_{n\to\infty} R_n\left(x; 0; e^x\right) = \lim_{n\to\infty} \frac{1}{(n+1)!} \cdot e^z \cdot x^{n+1} = 0; \text{ denn}$$

für $x > 0$ ist $0 \le R_n\left(x; 0; e^x\right) < \frac{1}{(n+1)!} \cdot e^x \cdot x^{n+1}$, da $0 < z < x$ ist, und damit

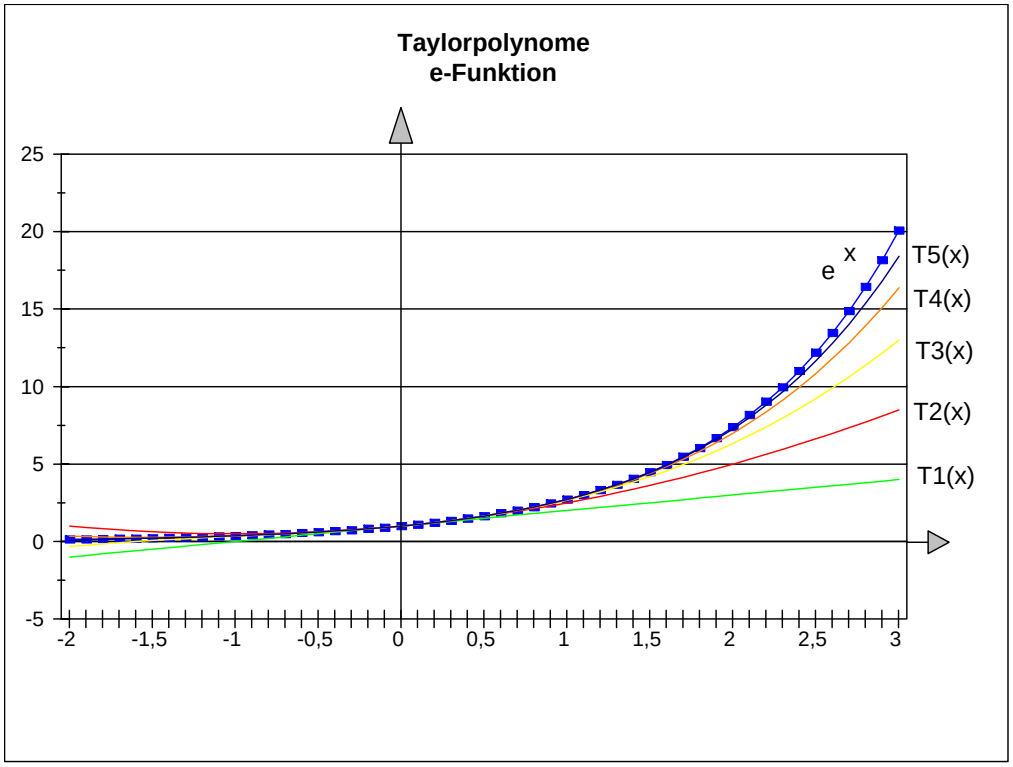
$$0 \le \lim_{n\to\infty} R_n\left(x; 0; e^x\right) \le \lim_{n\to\infty}\left(\frac{1}{(n+1)!} \cdot e^x \cdot x^{n+1}\right) = 0 \quad \text{(siehe Satz 5.1-5 (iii)).}$$

für $x < 0$ wird $\left|R_n\left(x; 0; e^x\right)\right| = \left|\frac{1}{(n+1)!} \cdot e^z \cdot x^{n+1}\right| = \frac{1}{(n+1)!} \cdot e^z \cdot \left|x^{n+1}\right| < \frac{\left|x^{n+1}\right|}{(n+1)!}$ betrachtet (die letzte Ungleichung ergibt sich aus $e^z < e^0 = 1$): hier ergibt sich mit demselben Argument $\lim_{n\to\infty}\left|R_n\left(x; 0; e^x\right)\right| = 0$, und mit Satz 5.1-2 (v) folgt $\lim_{n\to\infty} R_n\left(x; 0; e^x\right) = 0$.

Insgesamt ist

$$e^x = \sum_{i=0}^{\infty} \frac{1}{i!} \cdot x^i \quad \text{für } x \in \mathbf{R}.$$

Dieses ist ein Ergebnis, das nicht überrascht; denn so wurde die Exponentialfunktion ja definiert. Zu beachten ist aber, dass bei der Taylorentwicklung allein von der Tatsache Gebrauch gemacht wurde, dass $f^{(i)}(x) = f(x) = e^x$ und $e^0 = 1$ ist. Die folgende Abbildung zeigt den Verlauf der Exponentialfunktion und ihrer Taylorpolynome an der Stelle $x_0 = 0$ für $0 \le n \le 5$:



Um auch ein „numerisches Gefühl“ für die Qualität der Approximation der Exponentialfunktion durch ihr *n*-tes Taylorpolynom zu bekommen, werden in folgender Tabelle Werte des dritten Taylorpolynoms zur Exponentialfunktion an der Stelle $x_0 = 0$ dicht an der Entwicklungsstelle und weit entfernt von der Entwicklungsstelle mit numerisch ermittelten Werten von e^x verglichen. Man sieht dabei, dass für Werte, die dicht bei $x_0 = 0$ liegen, bereits $1 + x$ häufig eine befriedigende Annäherung an e^x liefert. Für große Werte von *x* ist ein *n*-tes Taylorpolynom mit einem großen Wert von *n* zu nehmen.

x	$T_3\left(x; 0; e^x\right) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$	e^x (letzte Stelle gerundet)
1	$2,\overline{6}...$	2,71828182845904523536028747135 27
1/2	$1,64583\overline{3}...$	1,64872127070012814684865078781 42
1/10	$1,10516\overline{6}...$	1,10517091807564762481170782649 02
1/100	$1,01005016\overline{6}...$	1,01005016708416805754216545690 29
20	$1554,\overline{3}...$	485 165 195,4097902779691068305415 4

Insbesondere gilt

$$e = \sum_{i=0}^{\infty} \frac{1}{i!} = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \ \ldots \ + \frac{1}{n!} + \sum_{i=n+1}^{\infty} \frac{1}{i!}.$$

Wegen $a^x = e^{\ln(a)\cdot x}$ ist

$$a^x = \sum_{i=0}^{\infty} \frac{\left(x \cdot \ln(a)\right)^i}{i!} \text{ für } x \in \mathbf{R}.$$

Beispiel:

Es soll nun die Taylorentwicklung der natürlichen Logarithmusfunktion $\ln(x)$ hergeleitet werden. Die Wahl $x_0 = 0$ der Entwicklungsstelle ist hierbei nicht möglich, da $\ln(x)\big|_{x=x_0}$ nicht definiert ist. Andererseits ist die Taylorentwicklung an der Entwicklungsstelle $x_0 = 0$ besonders einfach. Es wird daher zunächst die durch

$$f(x) = \begin{cases} \mathbf{R}_{>-1} & \to \ \mathbf{R} \\ x & \to \ \ln(1+x) \end{cases}$$

definierte Funktion in eine Taylorreihe an der Stelle $x_0 = 0$ entwickelt:

Es ist

$$f'(x) = \frac{1}{1+x} = (1+x)^{-1}, \quad f''(x) = -(1+x)^{-2}, \quad f'''(x) = 2\cdot(1+x)^{-3},$$
$$f^{(i)}(x) = (-1)^{i-1}\cdot(i-1)!\cdot(1+x)^{-i} \text{ und damit}$$

$$T_n(x;0;\ln(1+x)) = \frac{\ln(1+0)}{0!} + \sum_{i=1}^{n} \frac{(-1)^{i-1}\cdot(i-1)!\cdot(1+0)^{-i}}{i!}\cdot x^i$$
$$= \sum_{i=1}^{n} \frac{(-1)^{i-1}}{i}\cdot x^i = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + \frac{(-1)^{n-1}}{n}\cdot x^n$$

Mit Satz 5.7-2 ist

$$\ln(1+x) = \sum_{i=1}^{n} \frac{(-1)^{i-1}}{i}\cdot x^i + R_n(x;0;\ln(1+x)) = \sum_{i=1}^{n} \frac{(-1)^{i-1}}{i}\cdot x^i + \frac{(-1)^n}{(n+1)}\cdot(1+z)^{-(n+1)}\cdot x^{n+1}$$
$$\text{mit } 0 < z < x \text{ für } x > 0 \text{ und}$$
$$x < z < 0 \text{ für } -1 < x < 0.$$

Das Restglied

$$R_n(x;0;\ln(1+x)) = \frac{(-1)^n}{(n+1)} \cdot (1+z)^{-(n+1)} \cdot x^{n+1}$$

lässt sich betragsmäßig abschätzen:

Für $0 < z < x$ gilt (wegen $1+z>1$ bzw. $\frac{1}{1+z} < 1$) $\left|R_n(x;0;\ln(1+x))\right| < \frac{x^{n+1}}{n+1}$. Für $0 \le x \le 1$ folgt daher $\lim_{n\to\infty}\left|R_n(x;0;\ln(1+x))\right| = 0$.

Für $-1/2 < x < z < 0$ ist (wegen $0 < 1-|x| = 1-(-x) = 1+x < 1+z < 1$ bzw. $0 < \frac{1}{1+z} < \frac{1}{1-|x|}$)

$\left|R_n(x;0;\ln(1+x))\right| = \frac{|x|^{n+1}}{(n+1)} \cdot \left|(1+z)\right|^{-(n+1)} < \frac{1}{n+1} \cdot \left(\frac{|x|}{1-|x|}\right)^{n+1}$. Wegen $|x| < 1/2$ ist $\frac{|x|}{1-|x|} < 1$, so dass auch in diesem Fall $\lim_{n\to\infty}\left|R_n(x;0;\ln(1+x))\right| = 0$ gilt.

Für $-1 < x \le -1/2$ und $x < z < 0$ verwendet man eine andere, mathematisch anspruchsvollere Darstellung des Restglieds und kann dann auch hier zeigen, dass $\lim_{n\to\infty}\left|R_n(x;0;\ln(1+x))\right| = 0$ gilt.

Insgesamt ist

$$\ln(1+x) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot x^i$$
$$= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \ldots + \frac{(-1)^{n+1}}{n} x^n + \sum_{i=n+1}^{\infty} \frac{(-1)^{i+1}}{i} x^i \quad \text{für } x \in \mathbf{R} \text{ mit } -1 < x \le 1.$$

Die Taylorentwicklung für $f(x) = \ln(x)$ für $x > 0$ erhält man aus der Substitution $z = x-1$, $x = z+1$:

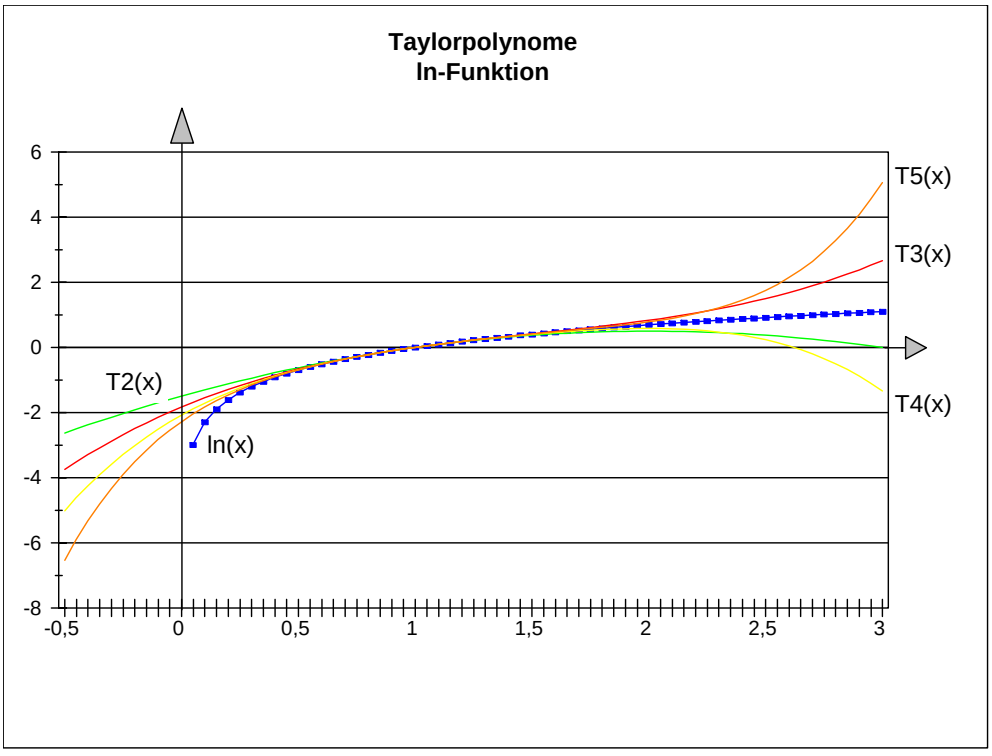
$$\ln(x) = \ln(z+1) = \sum_{i=1}^{n} \frac{(-1)^{i-1}}{i} \cdot (x-1)^i + R_n(x-1;0;\ln(1+z)) \text{ und}$$

$$\ln(x) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot (x-1)^i$$
$$= x - 1 - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \ldots + \frac{(-1)^{n+1}}{n}(x-1)^n + \sum_{i=n+1}^{\infty} \frac{(-1)^{i+1}}{i}(x-1)^i$$
für $x \in \mathbf{R}$ mit $0 < x \leq 2$.

Daraus ergibt sich das Ergebnis aus Satz 5.1-9 (iii)

$$\ln(2) = \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i}$$
$$= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \ldots + \frac{(-1)^{n+1}}{n} + \sum_{i=n+1}^{\infty} \frac{(-1)^{i+1}}{i} \approx 0{,}6931471.$$

Die folgende Abbildung zeigt den Verlauf der natürlichen Logarithmusfunktion und ihrer Taylorpolynome an der Stelle $x_0 = 1$ für $0 \leq n \leq 5$; man sieht sehr schön die Approximation der Taylorpolynome für $x \in \mathbf{R}$ mit $0 < x \leq 2$.



Es soll nun die Größe von *n* berechnet werden, die ausreicht, damit $T_n(x; 0; \ln(1+x))$ eine Approximation an $\ln(1+x)$ liefert, die einer vorgegebenen Genauigkeit genügt. Hierbei sei $0 \leq x \leq 1$. Soll also $T_n(x; 0; \ln(1+x))$ in der Dezimalentwicklung bis zur *m*-ten Nachkomma-

stelle genau sein, so kann man folgendermaßen vorgehen: Die Dezimalentwicklung von $\ln(1+x)$ bzw. von $T_n(x;0;\ln(1+x))$, die bis zur *m*-ten Nachkommastelle identisch sind, seien

$\ln(1+x) = [d_k d_{k-1} \ldots d_1 d_0, d_{-1} d_{-2} \ldots d_{-m} d_{-m-1} d_{-m-2} \ldots]_{10}$ bzw.

$T_n(x;0;\ln(1+x)) = [d_k d_{k-1} \ldots d_1 d_0, d_{-1} d_{-2} \ldots d_{-m} c_{-m-1} c_{-m-2} \ldots]_{10}$.

Dann ist

$$\begin{aligned} R_n(x;0;\ln(1+x)) &= \ln(1+x) - T_n(x;0;\ln(1+x)) \\ &= 0,\underbrace{0 \ldots 0}_{m-\text{mal}} + (d_{-m-1} - c_{-m-1}) \cdot 10^{-(m+1)} + (d_{-m-2} - c_{-m-2}) \cdot 10^{-(m+2)} + \ldots \\ &= \sum_{i=m+1}^{\infty} (d_{-i} - c_{-i}) \cdot 10^{-i}, \end{aligned}$$

$(-9) \cdot \sum_{i=m+1}^{\infty} 10^{-i} \le R_n(x;0;\ln(1+x)) \le 9 \cdot \sum_{i=m+1}^{\infty} 10^{-i}$, d.h. wegen $\sum_{i=m+1}^{\infty} 10^{-i} = \frac{1}{9} \cdot 10^{-m}$:

$-10^{-m} \le R_n(x;0;\ln(1+x)) \le 10^{-m}$ bzw. $0 \le |R_n(x;0;\ln(1+x))| \le 10^{-m}$.

Im angenommenen Fall $0 \le x \le 1$ ist $0 \le |R_n(x;0;\ln(1+x))| \le \frac{x^{n+1}}{n+1}$. Die Anforderung an *n* lautet also

$$\frac{x^{n+1}}{n+1} \le 10^{-m}.$$

Soll beispielsweise $\ln(1{,}5)$ auf 7 Nachkommastellen genau durch das *n*-te Taylorpolynom angegeben werden, so wird *n* so bestimmt, dass $\frac{1/2^{n+1}}{n+1} \le 10^{-7}$, d.h. $(n+1) \cdot 2^{n+1} \ge 10^7$ ist; $(n+1) \cdot 2^{n+1}$ muss also mindestens 8 Dezimalstellen aufweisen. Die folgende Tabelle listet einige Werte für $(n+1) \cdot 2^{n+1}$ auf:

n	$(n+1)\cdot 2^{n+1}$
1	8
2	24
...	...
10	22 528
11	49 152
...	...
17	4 718 592
18	9 961 472
19	10 971 520

In diesem Beispiel ist also $n \geq 19$ zu wählen.

Bemerkung:

ln(1,5) =
0,40546510810816438197801311546435
(letzte Stelle gerundet)

Zur Berechnung von ln(x) für $x \in \mathbf{R}$ mit $0 < x \leq 2$ kann man also ein n-tes Taylorpolynom mit genügend großem n als Approximation an ln(x) nehmen. Im allgemeinen (auch für $x > 2$) funktioniert dieser Ansatz nicht, da das Restglied nicht gegen 0 konvergiert. Nun gilt aber für $|h| < 1$:

$$\ln(1+h) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot h^i \quad \text{und} \quad \ln(1-h) = \ln(1+(-h)) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot (-h)^i = \sum_{i=1}^{\infty} \frac{-1}{i} \cdot h^i \quad \text{und damit}$$

$$\ln\left(\frac{1+h}{1-h}\right) = \ln(1+h) - \ln(1-h) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \cdot h^i - \sum_{i=1}^{\infty} \frac{-1}{i} \cdot h^i = 2 \cdot \sum_{i=0}^{\infty} \frac{1}{2 \cdot i + 1} \cdot h^{2 \cdot i + 1}$$

$$= 2 \cdot \left(h + \frac{h^3}{3} + \frac{h^5}{5} + \ldots + \frac{h^{2 \cdot n+1}}{2 \cdot n + 1} \right) + 2 \cdot \sum_{i=n+1}^{\infty} \frac{h^{2 \cdot i+1}}{2 \cdot i + 1}.$$

Für $x \in \mathbf{R}$ mit $x > 0$ ist $-1 < \frac{x-1}{x+1} < 1$. Setzt man $h = \frac{x-1}{x+1}$, so ist $|h| < 1$ und $x = \frac{1+h}{1-h}$. Damit ergibt sich

$$\ln(x) = 2 \cdot \left(h + \frac{h^3}{3} + \frac{h^5}{5} + \ldots + \frac{h^{2 \cdot n+1}}{2 \cdot n + 1} \right) + 2 \cdot \sum_{i=n+1}^{\infty} \frac{h^{2 \cdot i+1}}{2 \cdot i + 1} \Bigg|_{h=\frac{x-1}{x+1}} \quad \text{für } x \in \mathbf{R}.$$

Beispiel:

Für $m \in \mathbf{N}$ ist

$$(1 \pm x)^m = \sum_{i=0}^{m} (\pm\ 1)^i \frac{m \cdot (m-1) \cdot \ \ldots\ \cdot (m-i+1)}{i!} x^i = \sum_{i=0}^{m} (\pm\ 1)^i \binom{m}{i} x^i \quad \text{für } x \in \mathbf{R}.$$

Diese Formel ist ein Spezialfall der allgemeineren Formel

$$(a \pm b)^m = \sum_{i=0}^{m} (\pm\ 1)^i \frac{m \cdot (m-1) \cdot \ \ldots\ \cdot (m-i+1)}{i!} a^i b^{m-i} = \sum_{i=0}^{m} (\pm\ 1)^i \binom{m}{i} a^i b^{m-i}$$

für $a \in \mathbf{R},\ b \in \mathbf{R}$.

Für $m \in \mathbf{R} \setminus \mathbf{N}$ mit $m > 0$ ist

$$(1 \pm x)^m = \sum_{i=0}^{\infty} (\pm\ 1)^i \frac{m \cdot (m-1) \cdot \ \ldots\ \cdot (m-i+1)}{i!} x^i$$

$$= 1 \pm mx + \frac{m(m-1)}{2} x^2 \pm \frac{m(m-1)(m-2)}{6} x^3 + \sum_{i=4}^{\infty} (\pm\ 1)^i \frac{m(m-1) \cdot \ldots \cdot (m-i+1)}{i!} x^i$$

für $x \in \mathbf{R}$ mit $-1 \le x \le 1$.

Beispiel:

Zur Berechnung eines Wertes der Form $\frac{1{,}000001}{0{,}999999^2}$ mit großer Genauigkeit kann die Taylorentwicklung einer „geeigneten“ Funktion herangezogen werden. Mit

$$f : \begin{cases} \mathbf{R}_{\neq 1} & \to & \mathbf{R} \\ x & \to & \dfrac{1+x}{(1-x)^2} \end{cases}$$

ist $\frac{1{,}000001}{0{,}999999^2} = f\left(10^{-6}\right)$. Das n-te Taylorpolynom der Funktion f wird ermittelt; dazu werden die einige Ableitungen ermittelt, um daraus auf die Form der i-ten Ableitung zu schließen:

$$f^{(0)}(x) = f(x) = \frac{1+x}{(1-x)^2},$$

$$f'(x)=\frac{1\cdot(1-x)^2-(1+x)\cdot(-2)\cdot(1-x)}{(1-x)^4}=\frac{3+x}{(1-x)^3},$$

$$f''(x)=\frac{(1-x)^3-(3+x)\cdot(-3)\cdot(1-x)^2}{(1-x)^6}=\frac{10+2\cdot x}{(1-x)^4}=\frac{2\cdot(5+x)}{(1-x)^4},$$

$$f'''(x)=\frac{2\cdot(1-x)^4-2\cdot(5+x)\cdot(-4)\cdot(1-x)^3}{(1-x)^8}=\frac{42+6\cdot x}{(1-x)^5}=\frac{6\cdot(7+x)}{(1-x)^5};$$

die *i*-te Ableitung lautet also (das kann man durch vollständige Induktion beweisen):

$$f^{(i)}(x)=\frac{i!\cdot(2\cdot i+1+x)}{(1-x)^{i+2}}.$$

Damit gilt für das *n*-te Taylorpolynom an der Stelle $x_0=0$:

$$T_n(x;0;f(x))=\sum_{i=0}^{n}\frac{i!\cdot(2\cdot i+1)}{i!}\cdot x^i=\sum_{i=0}^{n}(2\cdot i+1)\cdot x^i=1+3\cdot x+5\cdot x^2+7\cdot x^3+\ldots+(2\cdot n+1)\cdot x^n.$$

Für $-1<x<1$ konvergiert das Restglied $R_n(x;0;f(x))=(2\cdot n+3+z)\cdot x^{n+1}$ gegen 0, so dass gilt:

$$\frac{1+x}{(1-x)^2}=\sum_{i=0}^{\infty}(2\cdot i+1)\cdot x^i \quad \text{für } -1<x<1.$$

Damit ist $\frac{1{,}000001}{0{,}999999^2}=\left.\frac{1+x}{(1-x)^2}\right|_{x=10^{-6}}=1{,}000003000005000007000009000011000013 0...$

Beispiel:

Über die auf ganz **R** stetigen und differenzierbaren Funktionen *g* und *h* seien folgende Eigenschaften bekannt:

$g(0)=0,\qquad h(0)=1,$
$g'(x)=h(x),\quad h'(x)=-g(x).$

Dann lautet wegen

$g(0)=0,$
$g'(0)=h(0)=1,$

$g''(0) = h'(0) = -g(0) = 0$,
$g'''(0) = -g'(0) = -h(0) = -1$,
$g^{(4)}(0) = -h'(0) = g(0) = 0$, d.h.

$$g^{(i)}(0) = \begin{cases} 0 & \text{für } (i \bmod 2) = 0 \\ 1 & \text{für } (i \bmod 4) = 1 \\ -1 & \text{für } (i \bmod 4) = 3 \end{cases}$$

das $(2 \cdot n + 1)$– te Taylorpolynom von g an der Stelle $x_0 = 0$:

$$T_{2 \cdot n+1}(x; 0; g(x)) = \sum_{i=0}^{n} \frac{(-1)^i}{(2 \cdot i + 1)!} \cdot x^{2 \cdot i + 1}.$$

Das Restglied konvergiert für jedes $x \in \mathbf{R}$ gegen 0 (die Begründung erfolgt wie bei der Exponentialfunktion), so dass

$$g(x) = \sum_{i=0}^{\infty} \frac{(-1)^i}{(2 \cdot i + 1)!} \cdot x^{2 \cdot i + 1}$$

gilt. Im vorliegenden Fall darf man die Bildung der unendlichen Reihe und die Ableitungsoperation miteinander vertauschen, so dass

$$h(x) = g'(x) = \sum_{i=0}^{\infty} \frac{(-1)^i \cdot (2 \cdot i + 1)}{(2 \cdot i + 1)!} \cdot x^{2 \cdot i} = \sum_{i=0}^{\infty} \frac{(-1)^i}{(2 \cdot i)!} \cdot x^{2 \cdot i}$$

gilt. Die Ausgangsbedingungen $g(0) = 0$, $h(0) = 1$, $g'(x) = h(x)$ und $h'(x) = -g(x)$ legen die Funktionen g und h eindeutig fest. Die aus der Schule bekannte Sinus- und Kosinusfunktionen besitzen diese Eigenschaften: $g(x) = \sin(x)$, $h(x) = \cos(x)$. Es ist daher

$$\sin(x) = \sum_{i=0}^{\infty} \frac{(-1)^i}{(2 \cdot i + 1)!} \cdot x^{2 \cdot i + 1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots + \frac{(-1)^n}{(2 \cdot n + 1)!} \cdot x^{2 \cdot n + 1} + \sum_{i=n+1}^{\infty} \frac{(-1)^i}{(2 \cdot i + 1)!} \cdot x^{2 \cdot i + 1}$$

für $x \in \mathbf{R}$.

$$\cos(x) = \sum_{i=0}^{\infty} \frac{(-1)^i}{(2 \cdot i)!} \cdot x^{2 \cdot i} = 1 - \frac{x^2}{2} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots + \frac{(-1)^n}{(2 \cdot n)!} \cdot x^{2 \cdot n} + \sum_{i=n+1}^{\infty} \frac{(-1)^i}{(2 \cdot i)!} \cdot x^{2 \cdot i}$$

für $x \in \mathbf{R}$.

Vergleicht man diese Taylorreihenentwicklung mit der Exponentialfunktion, so fällt die große Ähnlichkeit auf. Es gelten folgende Zusammenhänge zwischen der Exponential-, der Sinus- und der Kosinusfunktion:

Erweitert man die Definition der Exponentialfunktion von den reellen Zahlen auf die komplexen Zahlen (siehe Kapitel 1.4), so erhält man die komplexwertige Exponentialfunktion

$$\exp: \begin{cases} \mathbf{C} & \to & \mathbf{C} \\ z & \to & \sum_{i=0}^{\infty} \frac{z^i}{i!} \end{cases} .$$

Die unendliche Reihe in dieser Definition wird wieder über endliche Partialsummen definiert, wobei hierbei alle Operationen in den komplexen Zahlen ausgeführt werden. Die bei den Konvergenzbetrachtungen auftretenden Betragswerte sind dann wegen $|z| = |a+bi| = \sqrt{a^2+b^2}$ reelle Zahlen, so dass Konvergenzbetrachtungen von den komplexen Zahlen auf die reellen Zahlen übertragen werden und in **R** stattfinden. Es lässt sich zeigen, dass auch hier wieder $\sum_{i=0}^{\infty} \frac{z^i}{i!}$ absolut konvergiert, so dass der Grenzwert $\exp(z) = \sum_{i=0}^{\infty} \frac{z^i}{i!}$ mit $z \in \mathbf{C}$ in der Tat existiert. Zur Abkürzung schreibt man anstelle von $\exp(z)$ wieder e^z und meint damit den Grenzwert $\sum_{i=0}^{\infty} \frac{z^i}{i!}$.

Nun gilt für die imaginäre Zahl i: $i^2 = -1$. Für $x \in \mathbf{R}$ ergibt sich:

$$\begin{aligned} \frac{e^{i\cdot x} - e^{-i\cdot x}}{2} &= \frac{1}{2} \cdot \left(\sum_{k=0}^{\infty} \frac{(i\cdot x)^k}{k!} - \sum_{k=0}^{\infty} \frac{(-i\cdot x)^k}{k!} \right) \\ &= \frac{1}{2} \cdot \sum_{k=0}^{\infty} \frac{(i\cdot x)^k - (-i\cdot x)^k}{k!} \\ &= \frac{1}{2} \cdot \sum_{k=0}^{\infty} \frac{2\cdot (i\cdot x)^{2\cdot k+1}}{(2\cdot k+1)!} = i \cdot \sum_{k=0}^{\infty} \frac{i^{2\cdot k} \cdot x^{2\cdot k+1}}{(2\cdot k+1)!} = i \cdot \sum_{k=0}^{\infty} \frac{(-1)^k \cdot x^{2\cdot k+1}}{(2\cdot k+1)!} = i \cdot \sin(x) \quad , \end{aligned}$$

$$\begin{aligned} \frac{e^{i\cdot x} + e^{-i\cdot x}}{2} &= \frac{1}{2} \cdot \left(\sum_{k=0}^{\infty} \frac{(i\cdot x)^k}{k!} + \sum_{k=0}^{\infty} \frac{(-i\cdot x)^k}{k!} \right) \\ &= \frac{1}{2} \cdot \sum_{k=0}^{\infty} \frac{(i\cdot x)^k + (-i\cdot x)^k}{i!} \\ &= \frac{1}{2} \cdot \sum_{k=0}^{\infty} \frac{2\cdot (i\cdot x)^{2\cdot k}}{(2\cdot k)!} = \sum_{k=0}^{\infty} \frac{(-1)^k \cdot x^{2\cdot k}}{(2\cdot k)!} = \cos(x) \quad . \end{aligned}$$

Damit folgt

$$e^{i\cdot x} = \frac{e^{i\cdot x} + e^{-i\cdot x}}{2} + \frac{e^{i\cdot x} - e^{-i\cdot x}}{2} = \cos(x) + i\cdot\sin(x).$$

Für $x \in \mathbf{R}$ ist also $\cos(x)$ der Realteil und $\sin(x)$ der Imaginärteil von $e^{i\cdot x}$.

Die in Satz 5.5-1 für reelle Zahlen x und y definierte Funktionalgleichung $\exp(x+y) = \exp(x)\cdot\exp(y)$ lässt sich auch für die komplexwertige Exponentialfunktion nachweisen, so dass insgesamt für $z \in \mathbf{C}$, etwa $z = a + b\cdot i$ mit $a \in \mathbf{R}$ und $b \in \mathbf{R}$, gilt:

$$\exp(z) = \exp(a + i\cdot b) = \exp(a)\cdot\exp(i\cdot b) = \exp(a)\cdot(\cos(b) + i\cdot\sin(b)).$$

5.10 Fibonacci-Zahlen

In Kapitel 2.1 werden die Fibonacci-Zahlen als Funktion definiert:

$$fib: \begin{cases} \mathbf{N} & \to & \mathbf{N} \\ n & \to & \begin{cases} n & \text{für } n = 0 \text{ und } n = 1 \\ fib(n-1) + fib(n-2) & \text{für } n \geq 2. \end{cases} \end{cases}$$

Die Fibonacci-Zahlen spielen in vielen Teilen der Mathematik und der Informatik (z.B. bei der Laufzeitberechnung des Zugriffs auf Daten, die in Form höhenbalancierter Bäume gespeichert sind) eine wichtige Rolle.

Zur Vereinfachung der Darstellung wird $F_n = fib(n)$ gesetzt, d.h. $(F_n)_{n\in\mathbf{N}}$ ist die Folge der Fibonacci-Zahlen. Die ersten elf Fibonacci-Zahlen lauten:

n	0	1	2	3	4	5	6	7	8	9	10
F_n	0	1	1	2	3	5	8	13	21	34	55

Gemäß obiger Definition ist

$$F_n = \begin{cases} n & \text{für } n = 0 \text{ oder } n = 1 \\ F_{n-2} + F_{n-1} & \text{für } n \geq 2. \end{cases}$$

Zur Berechnung der n-ten Fibonacci-Zahlen kann man z.B. folgende PASCAL-Funktion einsetzen, die wohl elegant ist, aber schlechtes Laufzeitverhalten zeigt:

```
FUNCTION fib_1 (n : INTEGER) : INTEGER;

BEGIN { fib_1 }
  IF n < 0 THEN Exit;

  CASE OF n
  0    : fib_1 := 0;
  1    : fib_1 := 1;
  ELSE fib_1 := fib_1(n-2) + fib_1(n-1);
  END;
END   { fib_1 };
```

Optimales Laufzeitverhalten zeigt folgende PASCAL-Funktion, die zur Berechnung der *n*-ten Fibonacci-Zahl F_n nacheinander alle Fibonacci-Zahlen F_i mit $0 \le i < n$ berechnet:

```
FUNCTION fib_2 (n : INTEGER) : INTEGER;

VAR f_n1, f_n2, f_n : INTEGER;
    idx             : INTEGER;

BEGIN { fib_2 }
   IF n < 0 THEN Exit;

   CASE OF n
   0    : fib_2 := 0;
   1    : fib_2 := 1;
   ELSE BEGIN
          f_n2 := 0;
          f_n1 := 1;
          FOR idx := n DOWNTO 2 DO
            BEGIN
              f_n  := f_n2 + f_n1;
              f_n2 := f_n1;
              f_n1 := f_n;
            END;
          fib_2 := f_n;
        END;
   END { CASE };
END   { fib_2 };
```

Die Fibonacci-Zahlen sind rekursiv definiert, und es ist wünschenswert, den Wert der *n*-ten Fibonacci-Zahl direkt in Abhängigkeit von *n* zu erhalten. Hier hilft eine spezielle mathematische Methode, die Methode der erzeugenden Funktionen, weiter, die hier nur auf das vorliegende Beispiel angewandt werden soll.

Zunächst fasst man die beiden Fälle der definierenden Gleichung der Fibonacci-Zahlen zu einer Gleichung zusammen. Dazu definiert man

$$F_{-1} = F_{-2} = 0$$

und erhält

$F_n = F_{n-1} + F_{n-2} + a_n$ für jedes $n \in \mathbf{N}$; hierbei ist $a_1 = 1$ und $a_n = 0$ für $n \neq 1$.

Beide Seiten werden mit x^n multipliziert und alle Werte aufaddiert; hier ist nicht gesagt, welchen Wert *x* annimmt, und auch Fragen der Konvergenz spielen zunächst keine Rolle. Man erhält:

$$\begin{aligned}\sum_{n=0}^{\infty} F_n \cdot x^n &= \sum_{n=0}^{\infty} F_{n-1} \cdot x^n + \sum_{n=0}^{\infty} F_{n-2} \cdot x^n + \sum_{n=0}^{\infty} a_n \cdot x^n \\ &= \sum_{n=0}^{\infty} F_n \cdot x^{n+1} + \sum_{n=0}^{\infty} F_n \cdot x^{n+2} + x \qquad \text{wegen } F_{-1} = F_{-2} = 0 \\ &= x \cdot \sum_{n=0}^{\infty} F_n \cdot x^n + x^2 \cdot \sum_{n=0}^{\infty} F_n \cdot x^n + x.\end{aligned}$$

Setzt man $F(x) = \sum_{n=0}^{\infty} F_n \cdot x^n$, so erhält man die Gleichung

$$F(x) = x \cdot F(x) + x^2 \cdot F(x) + x,$$

die man nach $F(x)$ auflöst:

$$F(x) = \frac{x}{1 - x - x^2}.$$

Diese Funktion erfüllt für $x_0 = 0$ die Voraussetzungen von Satz 5.7-2, so dass man versuchen könnte, die Taylorentwicklung an der Stelle $x_0 = 0$ herzuleiten. Wieder unter der Annahme, dass Konvergenz vorliegt, ergäbe sich dann:

$$F(x) = \sum_{i=0}^{\infty} \frac{F^{(i)}(0)}{i!} \cdot x^i = \sum_{n=0}^{\infty} F_n \cdot x^n .$$

Ein Koeffizientenvergleich lieferte $F_n = \frac{F^{(n)}(0)}{n!}$. Dieser Weg ist jedoch mühsam, da die Ableiungen $F^{(i)}(x)$ schwierig zu bestimmen sind. Daher wird ein anderer Weg beschritten:

Es werden Zahlen A, B, α und β mit Hilfe der Partialbruchzerlegung bestimmt, für die gilt:

$$F(x) = \frac{x}{1-x-x^2} = \frac{A}{1-\alpha \cdot x} + \frac{B}{1-\beta \cdot x}.$$

$$\begin{aligned}\frac{A}{1-\alpha \cdot x} + \frac{B}{1-\beta \cdot x} &= \frac{A \cdot (1-\beta \cdot x) + B \cdot (1-\alpha \cdot x)}{(1-\alpha \cdot x) \cdot (1-\beta \cdot x)} \\ &= \frac{A - A \cdot \beta \cdot x + B - B \cdot \alpha \cdot x}{(1-\alpha \cdot x) \cdot (1-\beta \cdot x)} \\ &= \frac{x}{1-x-x^2}.\end{aligned}$$

Der Koeffizientenvergleich ergibt:

$A + B - (A \cdot \beta + B \cdot \alpha) \cdot x = x$ und
$(1-\alpha \cdot x) \cdot (1-\beta \cdot x) = 1-x-x^2$.

Mit $x = 0$ folgt aus der ersten Gleichung

$A + B = 0$ bzw. $A = -B$.

Zur Bestimmung von α und β werden die Nullstellen von $1-x-x^2$ bestimmt. Diese lauten (siehe Kapitel 5.3):

$$x_{01} = -\frac{1}{2} \cdot (1+\sqrt{5}) \text{ und } x_{02} = -\frac{1}{2} \cdot (1-\sqrt{5}).$$

Damit ist

$$\begin{aligned}1-x-x^2 &= -(x-x_{01}) \cdot (x-x_{02}) \\ &= -(x_{01}-x) \cdot (x_{02}-x) \\ &= -x_{01} \cdot x_{02} \cdot \left(1-\frac{x}{x_{01}}\right) \cdot \left(1-\frac{x}{x_{02}}\right) \\ &= \left(1-\frac{x}{x_{01}}\right) \cdot \left(1-\frac{x}{x_{02}}\right) \qquad \text{wegen } -x_{01} \cdot x_{02} = 1 \\ &= (1-\alpha \cdot x) \cdot (1-\beta \cdot x)\end{aligned}$$

und folglich

$$\alpha = \frac{1}{x_{01}} = -2 \cdot \frac{1}{1+\sqrt{5}} = -2 \cdot \frac{1-\sqrt{5}}{\left(1+\sqrt{5}\right)\cdot\left(1-\sqrt{5}\right)} = \frac{1}{2}\cdot\left(1-\sqrt{5}\right)$$
$$\approx -0{,}618034$$

und

$$\beta = \frac{1}{x_{02}} = -2 \cdot \frac{1}{1-\sqrt{5}} = \frac{1}{2}\cdot\left(1+\sqrt{5}\right)$$
$$\approx 1{,}618034.$$

Diese Werte für α und β werden in die Gleichung $A+B-\left(A\cdot\beta+B\cdot\alpha\right)\cdot x = x$ eingesetzt, wobei $A+B=0$ bzw. $A=-B$ bereits bekannt ist, und der Wert $x=1$ genommen wird. Man erhält

$$1 = -A\cdot\beta + A\cdot\alpha = A\cdot\left(\alpha-\beta\right) = A\cdot\left(-\sqrt{5}\right) \text{ bzw.}$$

$$A = -\frac{1}{\sqrt{5}} \text{ und } B = \frac{1}{\sqrt{5}}.$$

Damit ist

$$\begin{aligned} F(x) &= \frac{x}{1-x-x^2} \\ &= \frac{A}{1-\alpha\cdot x} + \frac{B}{1-\beta\cdot x} \\ &= \frac{1}{\sqrt{5}}\cdot\left(\frac{1}{1-\beta\cdot x} - \frac{1}{1-\alpha\cdot x}\right). \end{aligned}$$

Nun ist nach Satz 5.1-9 (i) für $\left|\alpha\cdot x\right| < 1$ bzw. $\left|\beta\cdot x\right| < 1$:

$$\frac{1}{1-\alpha\cdot x} = \sum_{i=0}^{\infty}\left(\alpha\cdot x\right)^i \text{ bzw. } \frac{1}{1-\beta\cdot x} = \sum_{i=0}^{\infty}\left(\beta\cdot x\right)^i.$$

und insgesamt

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{5}}\cdot\left(\frac{1}{1-\beta\cdot x} - \frac{1}{1-\alpha\cdot x}\right) \\ &= \sum_{n=0}^{\infty}\left(\frac{1}{\sqrt{5}}\cdot\beta^n - \frac{1}{\sqrt{5}}\cdot\alpha^n\right)\cdot x^n \\ &= \sum_{n=0}^{\infty} F_n\cdot x^n. \end{aligned}$$

Bemerkung: Diese Gleichung gilt wegen der Konvergenzanforderungen $|\alpha \cdot x| < 1$ und $|\beta \cdot x| < 1$ für jedes $x \in \mathbf{R}$ mit $|x| < \min\{|1/\alpha|, \|1/\beta\|\} = |1/\beta| < 0{,}618034$, eine Tatsache, die hier nicht von Belang ist.

Der Koeffizientenvergleich liefert:

Die durch

$$F_n = \begin{cases} n & \text{für } n = 0 \text{ oder } n = 1 \\ F_{n-2} + F_{n-1} & \text{für } n \geq 2 \end{cases}$$

definierten Fibonacci-Zahlen erfüllen die Gleichung

$$F_n = \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right) \text{ für jedes } n \in \mathbf{N}.$$

5.11 Erzeugende Funktionen

Im vorliegenden Kapitel wird die in Kapitel 5.10 erwähnte Methode der erzeugenden Funktionen beschrieben und im Kapitel 5.12 auf weitere interessante Fragestellungen der Informatik angewendet.

Es sei $(g_n)_{n \in \mathbf{N}}$ eine Folge reeller Zahlen. Die **erzeugende Funktion** G der Folge $(g_n)_{n \in \mathbf{N}}$ wird durch $G(z) = \sum_{n=0}^{\infty} g_n \cdot z^n$ definiert. Hierbei ist z eine „formale“ (komplexe oder reelle) Variable.

Bemerkung: Haben zwei Folgen dieselbe erzeugende Funktion, so sind die beiden Folgen gleich.

Die erzeugende Funktion fasst die gesamte Information über die Folge $(g_n)_{n \in \mathbf{N}}$ in einem einzigen arithmetischen Ausdruck zusammen. Die z-Potenzen separieren dabei die einzelnen Folgenglieder. Natürlich kann man nach Konvergenzbedingungen in Abhängigkeit von z fragen. Im vorliegenden Zusammenhang spielen diese aber eine untergeordnete Rolle.

In Kapitel 5.10 wird die erzeugende Funktion der Fibonacci-Folge $(F_n)_{n\in\mathbf{N}}$ berechnet zu $F(z)=\sum_{n=0}^{\infty} F_n\cdot z^n=\frac{z}{1-z-z^2}$. Hierbei ist zur Berechnung die Tatsache, dass diese Reihe nur für $|z|<2\cdot 1/\left(1+\sqrt{5}\right)\approx 0{,}618034$ konvergiert, ohne Belang. Die erzeugende Funktion der Fibonacci-Folge wird in eine Potenzreihe umgewandelt mit dem Ergebnis $F(z)=\frac{z}{1-z-z^2}=\sum_{n=0}^{\infty}\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^n-\left(\frac{1-\sqrt{5}}{2}\right)^n\right)\cdot z^n$. Daraus lässt sich die *n*-te Fibonacci-Zahl in geschlossener Form ablesen: $F_n=\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^n-\left(\frac{1-\sqrt{5}}{2}\right)^n\right)$.

Satz 5.11-1:

Es seien $(f_n)_{n\in\mathbf{N}}$ bzw. $(g_n)_{n\in\mathbf{N}}$ zwei Folgen mit den erzeugenden Funktionen $F(z)$ bzw. $G(z)$ und α und β Konstanten.

(i) Die erzeugende Funktion der Folge $(\alpha\cdot f_n+\beta\cdot g_n)_{n\in\mathbf{N}}$ lautet

$$\sum_{n=0}^{\infty}(\alpha\cdot f_n+\beta\cdot g_n)\cdot z^n=\alpha\cdot F(z)+\beta\cdot G(z).$$

(ii) Die erzeugende Funktion der Folge, die man erhält, indem man $(g_n)_{n\in\mathbf{N}}$ um *m* Plätze nach rechts verschiebt, also der Folge $\left(\underbrace{0,...,0}_{m},g_0,g_1,...\right)$, lautet

$$\sum_{n=m}^{\infty} g_{n-m}\cdot z^n=\sum_{n=0}^{\infty} g_n\cdot z^{n+m}=z^m\cdot\sum_{n=0}^{\infty} g_n\cdot z^n=z^m\cdot G(z).$$

(iii) Die erzeugende Funktion der Folge, die man erhält, indem man $(g_n)_{n\in\mathbf{N}}$ um *m* Plätze nach links verschiebt, also der Folge $(g_m,g_{m+1},g_{m+2},...)$, lautet

$$\sum_{n=0}^{\infty} g_{n+m}\cdot z^n=\frac{G(z)-g_0-g_1\cdot z-g_2\cdot z^2-...-g_{m-1}\cdot z^{m-1}}{z^m}.$$

(iv) Die erzeugende Funktion der Folge $(\alpha^n\cdot g_n)_{n\in\mathbf{N}}$ lautet

$$\sum_{n=0}^{\infty}\alpha^n\cdot g_n\cdot z^n=\sum_{n=0}^{\infty} g_n\cdot(\alpha\cdot z)^n=G(\alpha\cdot z).$$

../..

(v) Die erzeugende Funktion der Folge $\big((n+1)\cdot g_{n+1}\big)_{n\in\mathbf{N}} = \big(g_1, 2\cdot g_2, 3\cdot g_3, ...\big)$ lautet $\sum_{n=0}^{\infty}(n+1)\cdot g_{n+1}\cdot z^n = G'(z)$.

(vi) Die erzeugende Funktion der Folge $\big(n\cdot g_n\big)_{n\in\mathbf{N}} = \big(0, g_1, 2\cdot g_2, 3\cdot g_3, ...\big)$ lautet $\sum_{n=0}^{\infty} n\cdot g_n\cdot z^n = z\cdot G'(z)$.

(vii) Die erzeugende Funktion der Folge $\left(\frac{1}{n}\cdot g_{n-1}\right)_{n\in\mathbf{N}} = \left(0, g_0, \frac{1}{2}\cdot g_1, \frac{1}{3}\cdot g_2, ...\right)$, wobei das 0-te Folgenglied den Wert 0 hat, lautet $\sum_{n=1}^{\infty}\frac{1}{n}\cdot g_{n-1}\cdot z^n = \int_0^z G(t)dt$.

(viii) Die Faltung der Folgen $\big(f_n\big)_{n\in\mathbf{N}}$ und $\big(g_n\big)_{n\in\mathbf{N}}$, d.h. die Folge $\left(\sum_{k=0}^{n} f_k\cdot g_{n-k}\right)_{n\in\mathbf{N}}$, hat die erzeugende Funktion $\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} f_k\cdot g_{n-k}\right)\cdot z^n = F(z)\cdot G(z)$.

(ix) Die Folge, deren Glieder die *n*-ten Partialsummen der Folge $\big(g_n\big)_{n\in\mathbf{N}}$ bilden, d.h. die Folge $\big(g_0, g_0+g_1, g_0+g_1+g_2, ...\big) = \left(\sum_{k=0}^{n} g_k\right)_{n\in\mathbf{N}}$ hat die erzeugende Funktion $\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} g_k\right)\cdot z^n = \frac{1}{1-z}\cdot G(z)$.

(x) Die erzeugende Funktion der Folge, die aus $\big(g_n\big)_{n\in\mathbf{N}}$ entsteht, indem man die Folgenglieder mit ungeradem Index durch 0 ersetzt, d.h. der Folge $\big(g_0, 0, g_2, 0, g_4, 0, ...\big)$, lautet $\sum_{n=0}^{\infty} g_{2\cdot n}\cdot z^{2\cdot n} = \frac{G(z)+G(-z)}{2}$.

Die erzeugende Funktion der Folge, die aus $\big(g_n\big)_{n\in\mathbf{N}}$ entsteht, indem man die Folgenglieder mit geradem Index durch 0 ersetzt, d.h. der Folge $\big(0, g_1, 0, g_3, 0, g_5, 0, ...\big)$, lautet $\sum_{n=0}^{\infty} g_{2\cdot n+1}\cdot z^{2\cdot n+1} = \frac{G(z)-G(-z)}{2}$.

Die Aussagen (i) – (iv) sind unmittelbar einsichtig.

Aussage (v) ergibt sich aus

$$G'(z) = \left(\sum_{n=0}^{\infty} g_n \cdot z^n \right)' = g_1 + 2 \cdot g_2 \cdot z + 3 \cdot g_3 \cdot z^2 + \ldots$$
$$= \sum_{n=0}^{\infty} (n+1) \cdot g_{n+1} \cdot z^n \quad .$$

Aussage (vi) folgt aus (ii), indem man die Folge aus (v) um einen Platz nach rechts verschiebt.

Aussage (vii) ergibt sich wie folgt (hier wird benutzt, dass unter der Voraussetzung der Konvergenz die Operationen der Integral- und Summenbildung vertauschbar sind; hinzu kommt ein wenig elementares Schulwissen):
$\int_0^z G(t)dt = \int_0^z \left(\sum_{n=0}^{\infty} g_n \cdot t^n \right) dt = \sum_{n=0}^{\infty} g_n \cdot \int_0^z t^n dt = \sum_{n=0}^{\infty} g_n \cdot \frac{z^{n+1}}{n+1} = \sum_{n=1}^{\infty} \frac{1}{n} \cdot g_{n-1} \cdot z^n + 0 \cdot z^0$. Dazu gehört die Folge $\left(0, g_0, \frac{1}{2} \cdot g_1, \frac{1}{3} \cdot g_2, \ldots \right)$.

Für Aussage (viii) ist
$$F(z) \cdot G(z) = \left(\sum_{n=0}^{\infty} f_n \cdot z^n \right) \cdot \left(\sum_{n=0}^{\infty} g_n \cdot z^n \right)$$
$$= f_0 \cdot g_0 + (f_0 \cdot g_1 + f_1 \cdot g_0) \cdot z + (f_0 \cdot g_2 + f_1 \cdot g_1 + f_2 \cdot g_0) \cdot z^2 + \ldots$$
$$= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} f_k \cdot g_{n-k} \right) \cdot z^n$$
zu beachten.

Die erzeugende Funktion der konstanten Folge $(1,1,1,\ldots) = (1)_{n \in \mathbf{N}}$ lautet $F(z) = \sum_{n=0}^{\infty} z^n = \frac{1}{1-z}$.

Daher ist mit (viii) $\frac{1}{1-z} \cdot G(z)$ die erzeugende Funktion der Faltung von $(1,1,1,\ldots) = (1)_{n \in \mathbf{N}}$ und $(g_n)_{n \in \mathbf{N}}$. Das n-te Folgenglied der Faltung lautet $\sum_{k=0}^{n} 1 \cdot g_{n-k} = \sum_{k=0}^{n} g_k$. Das zeigt die Aussage (ix).

Aussage (x) rechnet man direkt nach:
$G(z) + G(-z) = \sum_{n=0}^{\infty} g_n \cdot \left(1 + (-1)^n\right) \cdot z^n = \sum_{\substack{n=0 \\ n \text{ ist gerade}}}^{\infty} g_n \cdot 2 \cdot z^n = 2 \cdot \sum_{n=0}^{\infty} g_{2 \cdot n} \cdot z^{2 \cdot n}$. Entsprechend ist

$G(z) - G(-z) = \sum_{n=0}^{\infty} g_n \cdot \left(1 - (-1)^n\right) \cdot z^n = \sum_{\substack{n=0 \\ n \text{ ist ungerade}}}^{\infty} g_n \cdot 2 \cdot z^n = 2 \cdot \sum_{n=0}^{\infty} g_{2 \cdot n+1} \cdot z^{2 \cdot n+1}$.

Die folgende Zusammenstellung zeigt einige wichtige Beispiele von Folgen mit ihren erzeugenden Funktionen. Auch hier sollen Fragen des Konvergenzverhaltens bezüglich der formalen Variablen z unberücksichtigt bleiben.

Folge $(g_n)_{n\in\mathbf{N}}$	erzeugende Funktion	geschlossene Form
(i) $g_0 = 1$, $g_n = 0$ für $n \geq 1$: $(1, 0, 0, 0, \ldots)$	$1 \cdot z^0$	$G(z) = 1$
(ii) $g_m = 1$, $g_n = 0$ für $n \neq m$, $m \in \mathbf{N}$ fest: $\left(0, \ldots, 0, \underbrace{1}_{\text{Position } m}, 0, 0, 0, \ldots\right)$	$1 \cdot z^m$	$G(z) = z^m$
(iii) $g_n = 1$ für $n \in \mathbf{N}$: $(1, 1, 1, 1, \ldots)$	$\sum_{n=0}^{\infty} z^n$	$G(z) = \frac{1}{1-z}$
(iv) $g_{2\cdot i} = 1$, $g_{2\cdot i+1} = -1$ für $i \in \mathbf{N}$: $(1, -1, 1, -1, \ldots)$	$\sum_{n=0}^{\infty} (-1)^n \cdot z^n$	$G(z) = \frac{1}{1+z}$
(v) $g_{2\cdot i} = 1$, $g_{2\cdot i+1} = 0$ für $i \in \mathbf{N}$: $(1, 0, 1, 0, 1, \ldots)$	$\sum_{n=0}^{\infty} z^{2\cdot n}$	$G(z) = \frac{1}{1-z^2}$
(vi) $g_{m\cdot n} = 1$ für $n \in \mathbf{N}$, $g_i = 0$ sonst: $(1, 0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots)$	$\sum_{n=0}^{\infty} z^{m\cdot n}$	$G(z) = \frac{1}{1-z^m}$
(vii) $g_n = n+1$ für $n \in \mathbf{N}$: $(1, 2, 3, 4, \ldots)$	$\sum_{n=0}^{\infty} (n+1) \cdot z^n$	$G(z) = \frac{1}{(1-z)^2}$
(viii) $g_n = c^n$ für $n \in \mathbf{N}$, $c \in \mathbf{R}$ fest: $(1, c, c^2, c^3, \ldots)$	$\sum_{n=0}^{\infty} c^n \cdot z^n$	$G(z) = \frac{1}{1-c\cdot z}$
(ix) $g_n = \binom{m}{n}$ für $n \in \mathbf{N}$, $m \in \mathbf{N}$ fest: $\left(1, m, \binom{m}{2}, \binom{m}{3}, \ldots, m, 1, 0, 0, \ldots\right)$	$\sum_{n=0}^{\infty} \binom{m}{n} \cdot z^n$	$G(z) = (1+z)^m$
(x) $g_n = \binom{m+n-1}{n}$ für $n \in \mathbf{N}$, $m \in \mathbf{N}$, $m \geq 1$ fest: $\left(1, m, \binom{m+1}{2}, \binom{m+2}{3}, \ldots\right)$	$\sum_{n=0}^{\infty} \binom{m+n-1}{n} \cdot z^n$	$G(z) = \frac{1}{(1-z)^m}$

../..

(xi) $g_n = \binom{m+n}{n}$ für $n \in \mathbf{N}$, $m \in \mathbf{N}$ fest: $\left(1, \binom{m+1}{2}, \binom{m+2}{3}, \binom{m+3}{3}, \ldots\right)$	$\sum_{n=0}^{\infty} \binom{m+n}{n} \cdot z^n$	$G(z) = \frac{1}{(1-z)^{m+1}}$
(xii) $g_0 = 0$, $g_n = 1/n$ für $n \geq 1$: $(0, 1, 1/2, 1/3, 1/4 \ldots)$	$\sum_{n=1}^{\infty} 1/n \cdot z^n$	$G(z) = \ln\left(\frac{1}{1-z}\right)$
(xiii) $g_0 = 0$, $g_n = (-1)^{n+1} \cdot 1/n$ für $n \geq 1$: $(0, 1, -1/2, 1/3, -1/4 \ldots)$	$\sum_{n=1}^{\infty} (-1)^{n+1} \cdot 1/n \cdot z^n$	$G(z) = \ln(1+z)$
(xiv) $g_0 = 0$, $g_n = 1/(n!)$ für $n \geq 1$: $(1, 1, 1/2, 1/6, 1/24, 1/120, \ldots)$	$\sum_{n=0}^{\infty} 1/(n!) \cdot z^n$	$G(z) = e^z$

Die Berechnung der geschlossenen Form der erzeugenden Funktion in den Beispielen (i) – (vi) und (viii) ist klar bzw. ergibt sich aus Satz 5.1-9. Beispiel (vii) kann ebenfalls mit Satz 5.1-9 oder auch mit Satz 5.11-1 (v) begründet werden: Setzt man dort $g_n = 1$, so folgt mit Beispiel (iii):

$$\sum_{n=0}^{\infty} (n+1) \cdot z^n = \sum_{n=0}^{\infty} (n+1) \cdot 1 \cdot z^n = \left(\frac{1}{1-z}\right)' = \frac{1}{(1-z)^2}.$$

Beispiel (ix) ist die binomische Formel.

Beispiel (x) lässt sich durch Induktion über *m* zeigen:

Für $m = 1$ ist $\sum_{n=0}^{\infty} \binom{1+n-1}{n} \cdot z^n = \sum_{n=0}^{\infty} z^n = 1/(1-z) = 1/(1-z)^1$. Die Behauptung gelte für $m \geq 1$.

Dann ist $\sum_{n=0}^{\infty} \binom{m+1+n-1}{n} \cdot z^n = \sum_{n=0}^{\infty} \binom{m+1}{n} \cdot z^n$. Nach Satz 4.1-3 (v) (dort für *i* den Wert *n* und für *n* den Ausdruck $m+n-1$ einsetzen) ist $\binom{m+n}{n} = \sum_{k=0}^{n} \binom{m-1+k}{k}$. Setzt man $g_n = \binom{m+n-1}{n}$, so ist $\sum_{n=0}^{\infty} \binom{m+1}{n} \cdot z^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} \binom{m-1+k}{k} \right) \cdot z^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} g_k \right) \cdot z^n$. Nach Induktionsvoraussetzung lautet die erzeugende Funktion der Folge $(g_n)_{n \in \mathbf{N}}$:

$G(z) = \sum_{n=0}^{\infty} \binom{m+n-1}{n} \cdot z^n = \frac{1}{(1-z)^m}$. Mit Satz 5.11-1 (ix) folgt

$$\sum_{n=0}^{\infty} \binom{m+1}{n} \cdot z^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} g_k \right) \cdot z^n = \frac{1}{1-z} \cdot G(z) = \frac{1}{(1-z)^{m+1}}.$$

Beispiel (xi) ergibt sich aus (x) wie im Induktionsschritt zu (x).

Die Folge $(0,1,1/2,1/3,1/4...)$ in Beispiel (xii) hat die erzeugende Funktion $\sum_{n=1}^{\infty} 1/n \cdot z^n = \sum_{n=1}^{\infty} 1/n \cdot 1 \cdot z^n = \sum_{n=1}^{\infty} 1/n \cdot g_{n-1} \cdot z^n$ mit der konstanten Folge $(g_n)_{n\in\mathbf{N}} = (1)_{n\in\mathbf{N}}$. Die Folge $(g_n)_{n\in\mathbf{N}}$ hat nach (iii) die erzeugende Funktion $G(z) = \frac{1}{1-z}$. Mit Satz 5.11-1 (vii) folgt

$$\sum_{n=1}^{\infty} 1/n \cdot z^n = \sum_{n=1}^{\infty} \frac{1}{n} \cdot g_{n-1} \cdot z^n = \int_0^z G(t)dt = \int_0^z \frac{1}{1-t} dt = \int_1^{1-z} \left(-\frac{1}{x}\right) dx = -\ln(1-z) = \ln\left(\frac{1}{1-z}\right).$$

Die Folge $(0,1,-1/2,1/3,-1/4...)$ in Beispiel (xiii) hat die erzeugende Funktion

$$\sum_{n=1}^{\infty} (-1)^{n+1} \cdot 1/n \cdot z^n = -\sum_{n=1}^{\infty} 1/n \cdot (-z)^n = -\ln\left(\frac{1}{1+z}\right) = \ln(1+z).$$

Beispiel (xiv) ist die Taylorentwicklung der Exponentialfunktion e^z.

Erzeugende Funktionen stellen ein sehr gutes Werkzeug bei der **Auflösung von rekursiv definierten Folgen** zur Verfügung.

Dabei geht man wie folgt vor:

Gegeben sei die Folge $(g_n)_{n\in\mathbf{N}}$ über ein rekursives Gleichungssystem. Gesucht ist eine geschlossene Darstellung von g_n in Abhängigkeit von *n*.

1. Schritt: Man schreibt eine einzige Gleichung auf, die g_n mit Hilfe anderer Folgenglieder definiert. Diese Gleichung sollte für alle *n* gelten; eventuell muss man Folgenglieder mit negativen Indizes formal anfügen, für die dann $g_{-1} = g_{-2} = ... = 0$ gesetzt wird.

2. Schritt: Beide Seiten der Gleichung aus dem 1. Schritt werden nacheinander mit z^n multipliziert und jeweils aufsummiert, so dass auf der linken Seite der Gleichung die erzeugende Funktion $G(z) = \sum_{n=0}^{\infty} g_n \cdot z^n$ entsteht. Die rechte Seite der Gleichung wird so umgeformt, dass sie ebenfalls einen arithmetischen Ausdruck in $G(z)$ darstellt.

3. Schritt: Die Gleichung aus dem 2. Schritt wird nach $G(z)$ aufgelöst.

4. Schritt: $G(z)$ wird in eine formale Potenzreihe entwickelt. Der Koeffizient von z^n ist g_n in geschlossener Form.

Bemerkung: Der komplexeste Schritt ist im allgemeinen der 4. Schritt.

Zur Illustration wird das Verfahren an der Darstellung der Fibonacci-Zahlen in geschlossener Form gezeigt. Die Einzelheiten sind in Kapitel 5.10 bereits ausgeführt.

Die Folge $(F_n)_{n\in\mathbf{N}}$ der Fibonacci-Zahlen ist rekursiv definiert durch

$$F_n = \begin{cases} n & \text{für } n = 0 \text{ oder } n = 1 \\ F_{n-1} + F_{n-2} & \text{für } n \geq 2 \end{cases} .$$

Die Ergebnisse der einzelnen Schritte lauten:

1. Schritt: Mit $F_{-1} = F_{-2} = 0$ ist $F_n = F_{n-1} + F_{n-2} + a_n$ für jedes $n \in \mathbf{N}$; hierbei ist $a_1 = 1$ und $a_n = 0$ für $n \neq 1$.

2. Schritt:

$$\begin{aligned} F(z) &= \sum_{n=0}^{\infty} F_n \cdot z^n = \sum_{n=0}^{\infty} F_{n-1} \cdot z^n + \sum_{n=0}^{\infty} F_{n-2} \cdot z^n + \sum_{n=0}^{\infty} a_n \cdot z^n \\ &= \sum_{n=0}^{\infty} F_n \cdot z^{n+1} + \sum_{n=0}^{\infty} F_n \cdot z^{n+2} + z \\ &= z \cdot \sum_{n=0}^{\infty} F_n \cdot z^n + z^2 \cdot \sum_{n=0}^{\infty} F_n \cdot z^n + z \\ &= z \cdot F(z) + z^2 F(z) + z \end{aligned}$$

3. Schritt: $F(z) = \dfrac{z}{1 - z - z^2}$

4. Schritt:

$$\begin{aligned} F(z) &= \sum_{n=0}^{\infty} \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right) \cdot z^n \\ &= \sum_{n=0}^{\infty} F_n \cdot z^n . \end{aligned}$$

5.12 Anzahlbetrachtungen in Binärbäumen

Eine der wichtigsten Datenstrukturen, die in der Informatik vorkommen, sind Binärbäume. Dazu einige einführende Definitionen:

Ein **gerichteter Graph** $G = (V, E)$ besteht aus einer endlichen Menge $V = \{v_1, \ldots, v_n\}$ von **Knoten** (vertices) und einer endlichen Menge $E = \{e_1, \ldots, e_k\} \subseteq V \times V$ von **Kanten** (edges).

Die Kante $e = (v_i, v_j)$ läuft von v_i nach v_j (verbindet v_i mit v_j). Der Knoten v_i heißt **Anfangsknoten** der Kante $e = (v_i, v_j)$, der Knoten v_j **Endknoten** von $e = (v_i, v_j)$. Zu einem Knoten $v \in V$ heißt $pred(v) = \{v' \mid (v', v) \in E\}$ die **Menge der direkten Vorgänger** von v, $succ(v) = \{v' \mid (v, v') \in E\}$ die **Menge der direkten Nachfolger** von v.

Ein **Binärbaum** $B_n = (V, E)$ mit n Knoten wird durch folgende Eigenschaften 1. – 4. charakterisiert:

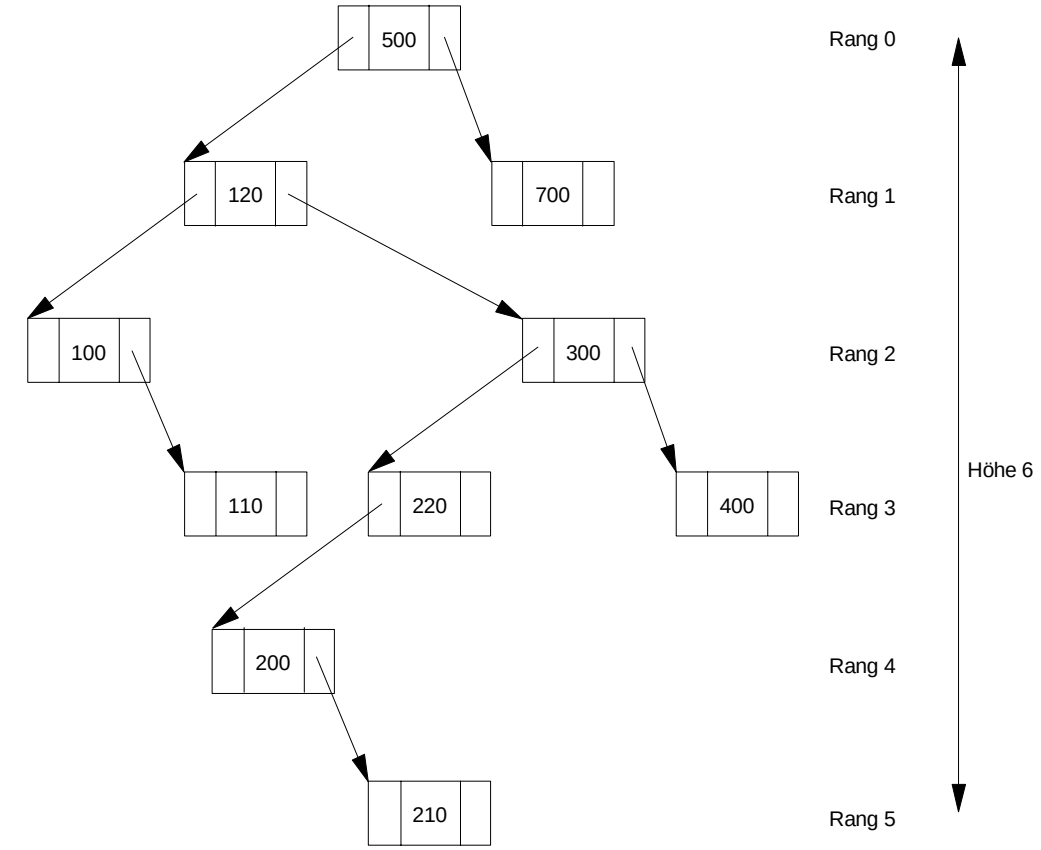
1. Entweder ist $n \geq 1$ und $|V| = n \geq 1$ und $|E| = n - 1$,
   oder es ist $n = 0$ und $V = E = \varnothing$ (**leerer Baum**)

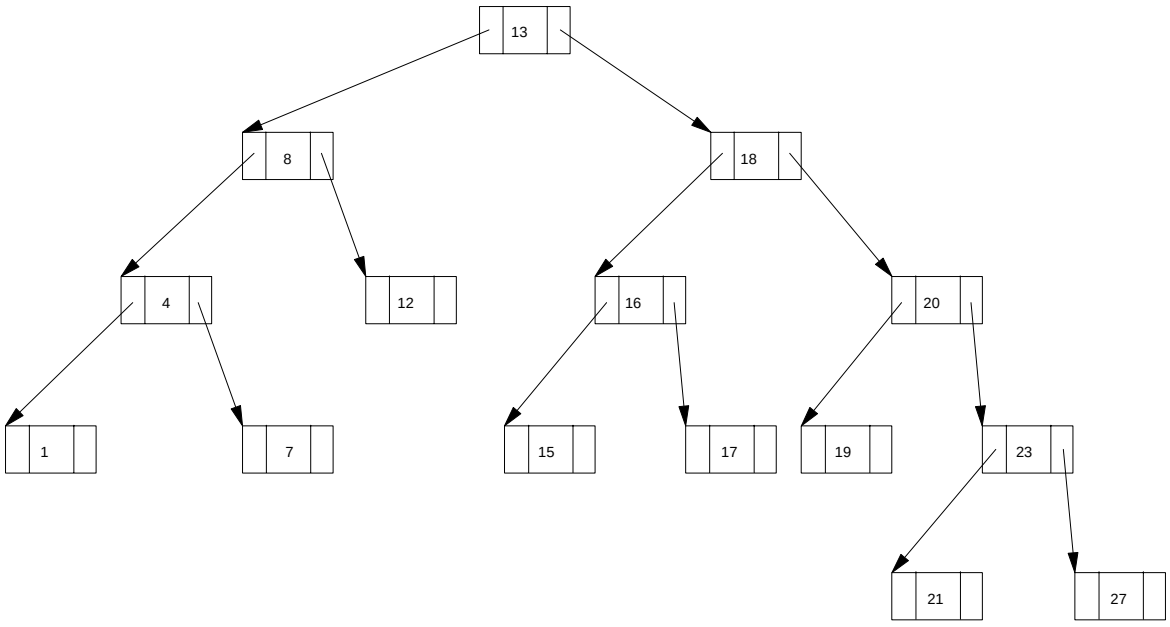
2. Bei $n \geq 1$ gibt es genau einen Knoten $r \in V$, dessen Menge direkter Vorgänger leer ist; dieser Knoten heißt **Wurzel** von B_n.

3. Bei $n \geq 1$ besteht die Menge der direkten Vorgänger eines jeden Knotens, der nicht die Wurzel ist, aus genau einem Element.

4. Bei $n \geq 1$ besteht die Menge der direkten Nachfolger eines jeden Knotens aus einem Element oder zwei Elementen oder ist leer. Ein Knoten, dessen Menge der direkten Nachfolger leer ist, heißt **Blatt**.

Beispiele:





In einem Binärbaum $B = (V, E)$ gibt es für jeden Knoten $v \in V$ genau einen **Pfad** von der Wurzel r zu v, d.h. es gibt eine Folge $\left((a_0, a_1), (a_1, a_2), \ldots, (a_{m-1}, a_m)\right)$ mit $r = a_0$, $v = a_m$ und $(a_{i-1}, a_i) \in E$ für $i = 1, ..., m$. Der Wert m gibt die **Länge des Pfads** an. Um den Knoten v

von der Wurzel aus über die Kanten des Pfads zu erreichen, werden *m* Kanten durchlaufen. Diese Länge wird auch als **Rang des Knotens** *v* bezeichnet.

Der Rang eines Knotens lässt sich auch folgendermaßen definieren:

1. Die Wurzel hat den Rang 0.
2. Ist *v* ein Knoten im Baum mit Rang $r-1$ und *w* ein direkter Nachfolger von *v*, so hat *w* den Rang *r*.

Unter der **Höhe eines Binärbaums** versteht man den maximal vorkommenden Rang eines Blattes + 1.

Der zweite Binärbaum zeichnet sich dadurch aus, dass sich die Höhen der Teilbäume, die von einem Knoten ausgehen, höchstens um 1 unterscheiden. Bäume mit dieser Eigenschaft heißen **AVL-Bäume**.

In einem Binärbaum bilden alle Knoten mit demselben Rang ein **Niveau des Baums**. Das Niveau 0 eines Binärbaums enthält genau einen Knoten, nämlich die Wurzel. Das Niveau 1 enthält mindestens 1 und höchstens 2 Knoten. Das Niveau *j* enthält höchstens doppelt so viele Knoten wie das Niveau $j-1$. Daher gilt:

Satz 5.12-1:

(i) Das Niveau $j \geq 0$ eines Binärbaums enthält mindestens einen und höchstens 2^j Knoten. Die Anzahl der Knoten vom Niveau 0 bis zum Niveau *j* (einschließlich) beträgt mindestens $j+1$ Knoten und höchstens $\sum_{i=0}^{j} 2^j = 2^{j+1}-1$ Knoten.

(ii) Ein Binärbaum hat maximale Höhe, wenn jedes Niveau genau einen Knoten enthält. Er hat minimale Höhe, wenn jedes Niveau eine maximale Anzahl von Knoten enthält. Also gilt für die Höhe $h(B_n)$ eines Binärbaums mit *n* Knoten:

$$\lfloor \log_2(n) \rfloor + 1 = \lceil \log_2(n+1) \rceil \leq h(B_n) \leq n .$$

(iii) Für die Höhe $h(B_n)$ eines AVL-Baums mit *n* Knoten gilt

$$\lceil \log_2(n+1) \rceil \leq h(B_n) < 1{,}4404201 \cdot \log_2(n+2).$$

Aussage (i) ergibt sich durch vollständige Induktion.

Aussage (ii) ergibt sich aus folgenden Überlegungen: Die obere Abschätzung $h(B_n) \le n$ ist offensichtlich. Für die untere Abschätzung betrachtet man einen Binärbaum mit *n* Knoten und minimaler Höhe *h*. Jedes Niveau, bis eventuell das höchste Niveau *m*, ist vollständig gefüllt. Es ist $h = m+1$. Bis zum Niveau $m-1$ enthält der Binärbaum gemäß (i) insgesamt $2^{m-1+1}-1 = 2^m - 1$ viele Knoten, auf Niveau *m* sind es mindestens einer und höchstens 2^m. Daraus folgt: $2^m - 1 + 1 \le n \le 2^m - 1 + 2^m$, also $2^m \le n \le 2^{m+1} - 1$ und damit $m < \log_2(n+1) \le m+1$, d.h. $\lceil \log_2(n+1) \rceil = m+1 = h$. Für einen beliebigen Binärbaum mit *n* Knoten gilt daher $\lceil \log_2(n+1) \rceil \le h(B_n)$.

Aussage (iii) zeigt, dass die Höhe eines AVL-Baums durch eine logarithmischen Größenordnung, gemessen in der Anzahl der Knoten bewegt. Aussage (iii) lässt sich mit Hilfe der Methode der erzeugenden Funktion (siehe Ende des Kapitels 5.11) nachweisen. Dazu werden die dort beschriebenen 4 Schritte durchgeführt.

Da ein Binärbaum, dessen Niveaus, bis eventuell das höchste Niveau *m*, vollständig gefüllt sind, ein AVL-Baum ist, folgt die untere Abschätzung $\lceil \log_2(n+1) \rceil \le h(B_n)$.

Es sei ein AVL-Baum mit Höhe $h+1$ gegeben, der eine minimale Knotenanzahl enthält. Dann sind unter der Wurzel zwei Binärbäume mit Höhen *h* und $h-1$ mit jeweils minimaler Knotenanzahl. Es sei K_h die minimale Knotenanzahl eines AVL-Baums bei Höhe *h*. Dann gilt:

$K_0 = 0$, $K_1 = 1$, $K_h = K_{h-1} + K_{h-2} + 1$ für $h \ge 2$. Die erzeugende Funktion der Folge $(K_h)_{h \in \mathbf{N}}$ sei $K(z)$.

1. Schritt: $K_h = K_{h-1} + K_{h-2} + 1 + a_h$ für $h \ge 0$ mit $a_0 = -1$, $a_i = 0$ für $i \ge 1$.

2. Schritt:

$$\begin{aligned} K(z) &= \sum_{h=0}^{\infty} K_h \cdot z^h = \sum_{h=1}^{\infty} K_{h-1} \cdot z^h + \sum_{h=2}^{\infty} K_{h-2} \cdot z^h + \sum_{h=0}^{\infty} z^h - 1 \\ &= \sum_{h=0}^{\infty} K_h \cdot z^{h+1} + \sum_{h=0}^{\infty} K_h \cdot z^{h+2} + \sum_{h=0}^{\infty} z^h - 1 \\ &= z \cdot K(z) + z^2 \cdot K(z) + \frac{1}{1-z} - 1 \quad . \end{aligned}$$

3. Schritt: $K(z) = \dfrac{z}{(1-z)\cdot(1-z-z^2)} = F(z) \cdot \dfrac{1}{1-z}$; hierbei ist $F(z)$ die erzeugende Funktion der Folge der Fibonacci-Zahlen. Die Folge $(K_h)_{h \in \mathbf{N}}$ ist also nach Satz 5.11-1 (viii)

die Faltung der Folge der Fibonacci-Zahlen $(F_n)_{n\in\mathbf{N}}$ mit der konstanten Folge $(1,1,1,...)$.

4. Schritt: Dieser erübrigt sich, da die Lösung im 3. Schritt bereits ermittelt wurde: $K_h = \sum_{k=0}^{n} F_k \cdot 1 = \sum_{k=0}^{n} F_k = F_{n+2} - 1$. Die letzte Gleichung zeigt man beispielsweise durch vollständige Induktion (Übungsaufgabe 5.30 (b)).

Für einen AVL-Baum mit Höhe *h* und *n* Knoten ist
$n \geq$ minimale Knotenanzahl bei Höhe $h = F_{h+2} - 1$ bzw. $F_{h+2} \leq n+1$. In Kapitel 5.10 wird $F_n = \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right)$ für jedes $n \in \mathbf{N}$ ermittelt. Da F_n eine natürliche Zahl ist, folgt $F_n = \left\lfloor \frac{1}{\sqrt{5}} \cdot \left(\frac{1+\sqrt{5}}{2} \right)^n + \frac{1}{2} \right\rfloor$. Eingesetzt in die obige Ungleichung ergibt sich $F_{h+2} = \frac{1}{\sqrt{5}} \cdot \left(\frac{1+\sqrt{5}}{2} \right)^{h+2} + \frac{1}{2} \leq n+2$. Löst man diese Ungleichung nach *h* auf, folgt unter Verwendung von Satz 5.5-5: $h < 1{,}4404201 \cdot \log_2(n+2) - 0{,}327724 < 1{,}4404201 \cdot \log_2(n+2)$.

Satz 5.12-2:

(i) Die Anzahl strukturell verschiedener Binärbäume mit *n* Knoten mit $n \geq 0$ beträgt

$$\frac{1}{n+1} \cdot \binom{2 \cdot n}{n} = \frac{4^n}{(n+1) \cdot \sqrt{\pi \cdot n}} + C$$ mit einer reellen Konstanten $C > 0$.

(ii) Die *mittlere* Anzahl von Knoten, die von der Wurzel aus bis zur Erreichung eines beliebigen Knotens eines Binärbaums mit *n* Knoten (gemittelt über alle *n* Knoten) besucht werden, d.h. der *mittlere „Abstand" eines Knotens von der Wurzel* in einem Binärbaum mit *n* Knoten, ist $C' \cdot \sqrt{\pi n} + C''$ mit reellen Konstanten $C' > 0$ und $C'' > 0$. Im günstigsten Fall (wenn also alle Niveaus voll besetzt sind) ist der größte Abstand eines Knotens von der Wurzel in einem Binärbaum mit *n* Knoten gleich $\lfloor \log_2(n) \rfloor + 1 = \lceil \log_2(n+1) \rceil \approx \log_2(n)$, im ungünstigsten Fall ist dieser Abstand gleich *n*.

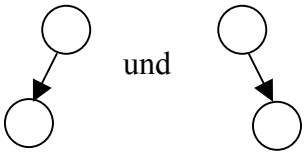
Beide Aussagen mit Hilfe des am Ende von Kapitel 5.11 beschriebenen Vefahrens bewiesen werden.

Es sei b_n für $n \in \mathbf{N}$ die Anzahl strukturell verschiedener Binärbäume mit *n* Knoten. $B(z)$ sei die erzeugende Funktion der Folge $(b_n)_{n\in\mathbf{N}}$. Für kleine Werte von *n* kann man b_n direkt angeben:

$b_0 = 1$: der einzige Binärbaum ohne Knoten ist der leere Baum;

$b_1 = 1$: der einzige Binärbaum mit genau 1 Knoten ist der Baum, der nur aus der Wurzel besteht;

$b_2 = 2$: die beiden Binärbäume mit genau 2 Knoten sind:



Für $n \geq 1$ besteht ein Binärbaum mit *n* Knoten aus der Wurzel und zwei Binärbäumen mit zusammen $n-1$ Knoten, die an den beiden Nachfolgerpositionen der Wurzel beginnen. An der linken Nachfolgerposition befindet sich ein Binärbaum mit *k* vielen Knoten, an der rechten Nachfolgerposition ein Binärbaum mit $n-1-k$ vielen Knoten. Daher gilt

$$b_n = \sum_{k=0}^{n-1} b_k \cdot b_{n-1-k} \quad \text{für } n \geq 1.$$

1. Schritt: Alle Gleichungen werden in einer einzigen Gleichung zusammengefasst:

$$b_n = \sum_{k=0}^{n-1} b_k \cdot b_{n-1-k} + a_n \quad \text{für } n \in \mathbf{N}\text{; hierbei ist } a_0 = 1 \text{ und } a_n = 0 \text{ für } n \geq 1.$$

2. Schritt: Beide Seiten werden mit z^n multipliziert und alle Werte aufaddiert. Man erhält:

$$
\begin{aligned}
B(z) &= \sum_{n=0}^{\infty} b_n \cdot z^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n-1} b_k \cdot b_{n-1-k} + a_n \right) \cdot z^n \\
&= \sum_{n=1}^{\infty} \left(\sum_{k=0}^{n-1} b_k \cdot b_{n-1-k} \right) \cdot z^n + 1 \\
&= z \cdot \sum_{n=1}^{\infty} \left(\sum_{k=0}^{n-1} b_k \cdot b_{n-1-k} \right) \cdot z^{n-1} + 1 \\
&= z \cdot \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} b_k \cdot b_{n-k} \right) \cdot z^n + 1 \\
&= z \cdot \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} b_k \cdot b_{n-k} \cdot z^k \cdot z^{n-k} \right) + 1 \\
&= z \cdot \left(\sum_{n=0}^{\infty} b_n \cdot z^n \right) \cdot \left(\sum_{n=0}^{\infty} b_n \cdot z^n \right) + 1 \qquad \text{nach Satz 5.1-10} \\
&= z \cdot \left(B(z) \right)^2 + 1 \quad .
\end{aligned}
$$

3. Schritt: Es gilt $\left(B(z)\right)^2 - \frac{1}{z} \cdot B(z) + \frac{1}{z} = 0$ bzw.

$$
\begin{aligned}
B(z) &= \frac{1}{2 \cdot z} \pm \sqrt{\frac{1}{4 \cdot z^2} - \frac{1}{z}} \\
&= \frac{1 \pm \sqrt{1 - 4 \cdot z}}{2 \cdot z} \quad .
\end{aligned}
$$

Falls $B(z) = \frac{1 + \sqrt{1 - 4 \cdot z}}{2 \cdot z}$ gilt, so ergibt sich der Widerspruch $B(0) = b_0 = \infty$, also ist $B(z) = \frac{1 - \sqrt{1 - 4 \cdot z}}{2 \cdot z}$.

4. Schritt: Mit der Regel von de l'Hospital gilt

$$
b_0 = B(0) = \left. \frac{1 - \sqrt{1 - 4 \cdot z}}{2 \cdot z} \right|_{z=0} = \left. \frac{-(1/2) \cdot (-4) \cdot \left(1 / \sqrt{1 - 4 \cdot z}\right)}{2} \right|_{z=0} = 1 \, .
$$

Aus der in Kapitel 5.10 hergeleiten Formel

$$
(1 \pm x)^m = \sum_{i=0}^{\infty} (\pm 1)^i \, \frac{m \cdot (m-1) \cdot \ldots \cdot (m-i+1)}{i!} x^i \, ,
$$

wobei $m = 1/2$ und anstelle von x der Wert $-4 \cdot z$ gesetzt wird, ergibt sich

$$
\begin{aligned}
1 - \sqrt{1 - 4 \cdot z} &= 1 - \sum_{i=0}^{\infty} \frac{(1/2) \cdot (-1/2) \cdot \ldots \cdot (1/2 - i + 1)}{i!} \cdot (-4 \cdot z)^i \\
&= 1 - \left(1 + (-4 \cdot z) \cdot \sum_{i=1}^{\infty} \frac{(1/2) \cdot (-1/2) \cdot \ldots \cdot (1/2 - i + 1)}{i!} \cdot (-4 \cdot z)^{i-1} \right) .
\end{aligned}
$$

Damit ist

$$B(z) = \frac{1-\sqrt{1-4\cdot z}}{2\cdot z}$$
$$= 2\cdot\sum_{i=1}^{\infty}\frac{(1/2)\cdot(-1/2)\cdot\ldots\cdot(1/2-i+1)}{i!}\cdot(-4\cdot z)^{i-1}.$$

Um diese etwas „unangenehm“ aussehende unendliche Reihe zu vereinfachen, wird folgende Nebenrechnung durchgeführt:

$$\frac{(1/2)\cdot(-1/2)\cdot\ldots\cdot(1/2-i+1)}{i!} = \frac{(1/2)\cdot(-1/2)\cdot(-3/2)\ldots\cdot\left(-\frac{2\cdot i-3}{2}\right)}{i!}$$
$$= \frac{\left(\frac{1}{2}\right)^{i}\cdot(-1)^{i-1}\cdot 1\cdot 3\cdot\ldots\cdot(2\cdot i-3)}{i!}\cdot\frac{2\cdot 4\cdot\ldots\cdot(2\cdot i-2)}{2\cdot 4\cdot\ldots\cdot(2\cdot i-2)}$$
$$= \frac{\left(\frac{1}{2}\right)^{i}\cdot(-1)^{i-1}\cdot(2\cdot i-2)!}{i!\cdot 2^{i-1}\cdot(i-1)!}$$
$$= \frac{1}{2}\cdot\left(-\frac{1}{4}\right)^{i-1}\cdot\frac{1}{i}\cdot\binom{2\cdot i-2}{i-1}.$$

Damit ist

$$B(z) = 2\cdot\sum_{i=1}^{\infty}\frac{(1/2)\cdot(-1/2)\cdot\ldots\cdot(1/2-i+1)}{i!}\cdot(-4\cdot z)^{i-1}.$$
$$= \sum_{i=1}^{\infty}\left(-\frac{1}{4}\right)^{i-1}\cdot\frac{1}{i}\cdot\binom{2\cdot i-2}{i-1}\cdot(-4\cdot z)^{i-1}$$
$$= \sum_{i=0}^{\infty}\left(-\frac{1}{4}\right)^{i}\cdot\frac{1}{i+1}\cdot\binom{2\cdot i}{i}\cdot(-4\cdot z)^{i}$$
$$= \sum_{i=0}^{\infty}\frac{1}{i+1}\cdot\binom{2\cdot i}{i}\cdot z^{i}$$
$$= \sum_{n=0}^{\infty}\frac{1}{n+1}\cdot\binom{2\cdot n}{n}\cdot z^{n}.$$

Der Koeffizientenvergleich in $B(z)=\sum_{n=0}^{\infty}b_n\cdot z^n=\sum_{n=0}^{\infty}\frac{1}{n+1}\cdot\binom{2\cdot n}{n}\cdot z^n$ ergibt

$$b_n = \frac{1}{n+1}\cdot\binom{2\cdot n}{n}.$$

Damit ist der erste Teil der Formel in Satz 5.12-2 (i) bewiesen. Für den zweiten Teil wird die Stirling’sche Formel (siehe angegebene Literatur)

$$n!\sim\sqrt{2\cdot\pi\cdot n}\cdot\left(\frac{n}{e}\right)^{n}$$

zusammen mit Satz 4.1-2 eingesetzt. Die Stirling'sche Formel ist eine sehr gute Approximation; der relative Fehler ist etwa $1/(12 \cdot n)$. Damit ergibt sich:

$$\begin{aligned}\frac{1}{n+1} \cdot \binom{2 \cdot n}{n} &\sim \frac{1}{n+1} \cdot \frac{(2 \cdot n)!}{n! \cdot n!} \\ &= \frac{\sqrt{4 \cdot \pi \cdot n} \cdot (2 \cdot n)^{2 \cdot n} \cdot e^{2 \cdot n}}{(n+1) \cdot e^{2 \cdot n} \cdot (2 \cdot \pi \cdot n) \cdot n^{2 \cdot n}} \\ &= \frac{4^n}{(n+1) \cdot \sqrt{\pi \cdot n}}.\end{aligned}$$

Für einen Knoten *v* in einem Baum *G* mit *n* Knoten bezeichne $r(v)$ den Rang des Knotens *v*, d.h. die Anzahl der Kanten, die von der Wurzel aus durchlaufen werden, um *v* zu erreichen. Um *v* von der Wurzel aus zu erreichen, werden $r(v)+1$ Knoten durchlaufen. Es sei $I(G) = \sum_v (r(v)+1)$, d.h. die Summe aller in *G* möglichen Pfadlängen, gemessen in der Anzahl besuchter Knoten.

Der linke Teilbaum unterhalb der Wurzel des Baum *G* sei G_l; der rechte Teilbaum unterhalb der Wurzel sei G_r. G_l oder G_r können auch leer sein. Für einen Knoten *v* in G_l sei $r_l(v)$ der Rang von *v* bezogen auf G_l. Entsprechend bezeichne $r_r(v)$ für einen Knoten *v* in G_r den Rang von *v* bezogen auf G_r. Dann ist für einen Knoten *v* in G_l $r_l(v) = r(v) - 1$. Es gilt daher:

$$\begin{aligned}I(G) &= \sum_{v \text{ in } G} (r(v)+1) \\ &= \sum_{v \text{ in } G_l} (r_l(v)+2) + \sum_{v \text{ in } G_r} (r_r(v)+2) + 1 \qquad \text{(der Rang der Wurzel ist 1)} \\ &= I(G_l) + \sum_{v \text{ in } G_l} 1 + I(G_r) + \sum_{v \text{ in } G_r} 1 + 1 \\ &= I(G_l) + I(G_r) + n \qquad \text{für } n > 0, \\ I(G) &= 0 \text{ für } n = 0.\end{aligned}$$

Es sei $I_n = \sum_{\substack{G \text{ ist ein Baum} \\ \text{mit } n \text{ Knoten}}} I(G)$, d.h. die Summe aller möglichen Pfadlängen in Binärbäumen mit *n* Knoten, gemessen in der Anzahl besuchter Knoten.

Mit Hilfe des am Ende von Kapitel 5.11 beschriebenen Vefahrens wird eine geschlossene Formel für I_n hergeleitet. Es bezeichne $I(z)$ die erzeugende Funktion der Folge $(I_n)_{n \in \mathbf{N}}$.

1. Schritt:

$$I_n = \sum_{\substack{G \text{ ist ein Baum} \\ \text{mit } n \text{ Knoten}}} I(G) \qquad \text{für } n \geq 1, \text{ Aufteilung nach linken Teilbäumen :}$$

$$= \sum_{i=0}^{n-1} \left(\underbrace{I(G_l)}_{\substack{G_l \text{ enthält} \\ i \text{ Knoten}}} + \underbrace{I(G_r)}_{\substack{G_r \text{ enthält} \\ n-i-1 \text{ Knoten}}} + n \right)$$

$$= \sum_{i=0}^{n-1} \left(I_i \cdot b_{n-i-1} + I_{n-i-1} \cdot b_i \right) + n \cdot b_n \quad .$$

Hierbei bezeichnet b_n für $n \in \mathbf{N}$ wieder die Anzahl strukturell verschiedener Binärbäume mit n Knoten.

2. Schritt:

$$I(z) = \sum_{n=1}^{\infty} I_n \cdot z^n$$

$$= \sum_{n=1}^{\infty} \left(\sum_{i=0}^{n-1} \left(I_i \cdot b_{n-i-1} + I_{n-i-1} \cdot b_i \right) + n \cdot b_n \right) \cdot z^n$$

$$= \sum_{n=1}^{\infty} \left(\sum_{i=0}^{n-1} \left(I_i \cdot b_{n-i-1} + I_{n-i-1} \cdot b_i \right) \cdot z^n + n \cdot b_n \cdot z^n \right)$$

$$= \sum_{n=1}^{\infty} \left(\sum_{i=0}^{n-1} \left(I_i \cdot b_{n-i-1} \right) \cdot z^n \right) + \sum_{n=1}^{\infty} \left(\sum_{i=0}^{n-1} \left(I_{n-i-1} \cdot b_i \right) \cdot z^n \right) + \sum_{n=1}^{\infty} n \cdot b_n \cdot z^n$$

Der Ausdruck $f_{n-1} = \sum_{i=0}^{n-1} I_i \cdot b_{n-i-1}$ ist das $(n-1)$-te Folgenglied der Faltung der Folgen $(I_n)_{n \in \mathbf{N}}$ und $(b_n)_{n \in \mathbf{N}}$. Dasselbe gilt für $\sum_{i=0}^{n-1} I_{n-i-1} \cdot b_i = f_{n-1}$. Daher ist

$$I(z) = \sum_{n=1}^{\infty} f_{n-1} \cdot z^n + \sum_{n=1}^{\infty} f_{n-1} \cdot z^n + \sum_{n=1}^{\infty} n \cdot b_n \cdot z^n \qquad \text{mit Satz 5.11-1}$$

$$= \sum_{n=0}^{\infty} f_n \cdot z^{n+1} + \sum_{n=0}^{\infty} f_n \cdot z^{n+1} + \sum_{n=1}^{\infty} n \cdot b_n \cdot z^n$$

$$= 2 \cdot z \cdot I(z) \cdot B(z) + z \cdot B'(z) \quad .$$

3. Schritt: $I(z) = \dfrac{1}{1 - 2 \cdot z \cdot B(z)} \cdot z \cdot B'(z)$.

Mit $B(z) = \dfrac{1 - \sqrt{1 - 4 \cdot z}}{2 \cdot z}$ und $B'(z) = \dfrac{1 - \sqrt{1 - 4 \cdot z} - 2 \cdot z}{2 \cdot z^2 \cdot \sqrt{1 - 4 \cdot z}}$ ist

$$I(z) = \frac{2 \cdot z - \sqrt{1 - 4 \cdot z} + 1 - 4 \cdot z}{2 \cdot z \cdot (1 - 4 \cdot z)}$$

$$= \frac{1}{1 - 4 \cdot z} - \frac{1}{2 \cdot z \cdot \sqrt{1 - 4 \cdot z}} + \frac{1}{2 \cdot z} \quad .$$

4. Schritt:

$$I(z) = \frac{1}{1-4\cdot z} - \frac{1}{2\cdot z\cdot\sqrt{1-4\cdot z}} + \frac{1}{2\cdot z}$$
$$= \sum_{n=0}^{\infty} 4^n \cdot z^n - \frac{1}{2\cdot z}\cdot\left(\frac{1}{\sqrt{1-4\cdot z}} - 1\right).$$

Das Beispiel $\sum_{n=0}^{\infty}\binom{m+n-1}{n}\cdot z^n = \frac{1}{(1-z)^m}$ aus Kapitel 5.10 ist auf Werte $m \in \mathbf{R}$ verallgemeinerbar. Damit ist

$$\frac{1}{\sqrt{1-4\cdot z}} = \sum_{n=0}^{\infty}\binom{n-1/2}{n}\cdot(4\cdot z)^n = \sum_{n=1}^{\infty}\binom{n-1/2}{n}\cdot(4\cdot z)^n + 1 \text{ und}$$

$$\frac{1}{2\cdot z}\cdot\left(\frac{1}{\sqrt{1-4\cdot z}} - 1\right) = \sum_{n=1}^{\infty}\binom{n-1/2}{n}\cdot 2\cdot(4\cdot z)^{n-1}$$
$$= 2\cdot\sum_{n=1}^{\infty}\binom{2\cdot n}{n}\cdot\frac{4^{n-1}}{4^n}\cdot z^{n-1}.$$

Hierbei wurde die Identität $\binom{n-1/2}{n} = \binom{2\cdot n}{n} \Big/ 2^{2\cdot n}$ verwendet. Mit

$\binom{2\cdot n}{n} = \frac{(2\cdot n-1)\cdot 2}{n}\cdot\binom{2\cdot(n-1)}{n-1}$ folgt weiter:

$$2\cdot\sum_{n=1}^{\infty}\binom{2\cdot n}{n}\cdot\frac{4^{n-1}}{4^n}\cdot z^{n-1} = \frac{1}{2}\cdot\sum_{n=1}^{\infty}\binom{2\cdot n}{n}\cdot z^{n-1}$$
$$= \sum_{n=1}^{\infty}\binom{2\cdot(n-1)}{n-1}\cdot\frac{2\cdot n-1}{n}\cdot z^{n-1}$$
$$= \sum_{n=0}^{\infty}\binom{2\cdot n}{n}\cdot\frac{2\cdot n+1}{n+1}\cdot z^n$$
$$= \sum_{n=0}^{\infty}(2\cdot n+1)\cdot b_n\cdot z^n.$$

Insgesamt ist damit

$$I(z) = \sum_{n=0}^{\infty}\left(4^n - (2\cdot n+1)\cdot b_n\right)\cdot z^n \text{ und } I_n = 4^n - (2\cdot n+1)\cdot b_n.$$

Der Ausdruck $I_n \big/ n\cdot b_n$ ist die mittlere Anzahl an Knoten, die in Binärbäumen mit n Knoten von der Wurzel aus zu einem Knoten durchlaufen werden. Mit dem gerade hergeleiteten Ergebnis folgt $I_n \big/ n\cdot b_n = \frac{4^n}{n\cdot b_n} - \frac{2\cdot n+1}{n}$. Mit der Stirlingschen Formel ist

$b_n = \frac{1}{n+1} \cdot \binom{2 \cdot n}{n} \sim \frac{4^n}{(n+1) \cdot \sqrt{\pi \cdot n}}$ und

$I_n \big/ n \cdot b_n \sim \sqrt{\pi \cdot n} \cdot \frac{n+1}{n} - 2 + \frac{1}{n} \leq C' \cdot \sqrt{\pi \cdot n} + C''$ mit reellen Konstanten $C' > 0$ und $C'' > 0$.

6 Ausgewählte Themen der Linearen Algebra

Das vorliegende Kapitel wählt aus einem Gebiet der Mathematik, der Linearen Algebra, spezielle Themen aus, ohne die grundlegenden Theorien explizit zu behandeln. Im wesentlichen geht hier darum, ein effizientes Verfahren zur Lösung linearer Gleichungssysteme, wie sie in vielen Anwendungen der Mathematik vorkommen, vorzustellen.

6.1 Matrizen und Vektoren

Ein rechteckiges Zahlenschema aus (reellen) Zahlen

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,j} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,j} & \dots & a_{2,n} \\ . & . & \dots & . & \dots & . \\ a_{i,1} & a_{i,2} & \dots & a_{i,j} & \dots & a_{i,n} \\ . & . & \dots & . & \dots & . \\ a_{m,1} & a_{m,2} & \dots & a_{m,j} & \dots & a_{m,n} \end{bmatrix} = \mathbf{A}_{(m,n)} = \left[a_{i,j}\right]_{i=1,\dots,m;\, j=1,\dots,n}$$

heißt **(reellwertige) Matrix vom Typ (*m, n*)**. Im Schnittpunkt der **Zeile** *i* und der **Spalte** *j* steht das **Matrixelement** $a_{i,j} \in \mathbf{R}$. Der erste Index gibt die Zeilennummer, der zweite Index die Spaltennummer an. Im folgenden werden Matrizen durch fett gedruckte Buchstaben bezeichnet.

Zwei Matrizen $\mathbf{A}_{(m,n)} = \left[a_{i,j}\right]$ und $\mathbf{B}_{(r,s)} = \left[b_{l,k}\right]$ sind gleich, wenn sie vom selben Typ sind, d.h. $m = r$ und $n = s$, und sie elementweise gleich sind, d.h. wenn $a_{i,j} = b_{i,j}$ für $i = 1, ..., m$ und $j = 1, ..., n$ gilt.

Eine Matrix vom Typ (*n*, *n*) heißt **quadratische Matrix**.

Eine Matrix, deren sämtliche Elemente 0 sind, heißt **Nullmatrix**; sie wird mit **0** bezeichnet.

Die quadratische Matrix

$$\mathbf{I}_{(n,n)} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 \\ . & . & \ldots & . & \ldots & . & & & & \\ 0 & 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 & 0 \\ . & . & \ldots & . & \ldots & . & & & & \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 1 \end{bmatrix}$$

vom Typ (*n*, *n*), die in der Diagonalen die Zahlen 1 und sonst nur Nullen enthält, heißt **Einheitsmatrix vom Typ (*n*, *n*)**. Es ist

$$\mathbf{I}_{(n,n)} = \left\lfloor \delta_{i,j} \right\rfloor \text{ mit } \delta_{i,j} = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{für } i \neq j \end{cases}.$$

Eine Matrix vom Typ (1, *n*) heißt **Zeilenvektor** der Länge *n*. Eine Matrix vom Typ (*m*, 1) heißt **Spaltenvektor** der Länge *m*. In beiden Fällen verzichtet man meist auf die doppelte Indizierung:

Ein Zeilenvektor wird geschrieben als

$$\vec{a} = \left[a_1 \quad a_2 \quad \ldots \quad a_j \quad \ldots \quad a_n \right].$$

Ein Spaltenvektor wird geschrieben als

$$\vec{b} = \begin{bmatrix} b_1 \\ . \\ . \\ . \\ b_i \\ . \\ . \\ . \\ b_m \end{bmatrix}.$$

Es sollen nun **Rechenoperationen auf Matrizen** definiert werden:

Es seien **A** und **B** zwei Matrizen *vom (gleichen) Typ* (*m*, *n*).

Die **Summe** von **A** und **B** ist die Matrix $\mathbf{C} = \left\lfloor c_{i,j} \right\rfloor = \mathbf{A} + \mathbf{B}$ mit $c_{i,j} = a_{i,j} + b_{i,j}$.

Die **Differenz** von **A** und **B** ist die Matrix $\mathbf{C} = \lfloor c_{i,j} \rfloor = \mathbf{A} - \mathbf{B}$ mit $c_{i,j} = a_{i,j} - b_{i,j}$.

Die Summe (Differenz) zweier Matrizen vom Typ (*m*, *n*) ist wieder vom Typ (*m*, *n*). Man erhält sie also, indem man die Elemente an den sich entsprechenden Positionen addiert (subtrahiert).

Es sei $k \in \mathbf{R}$. Das **Skalarprodukt** der Matrix **A** mit (dem Skalar) *k* ist die Matrix $\mathbf{D} = \lfloor d_{i,j} \rfloor = k \cdot \mathbf{A} = \mathbf{A} \cdot k$ mit $d_{i,j} = k \cdot a_{i,j} = a_{i,j} \cdot k$.

Das Skalarprodukt einer Matrix **A** vom Typ (*m*, *n*) mit einer reellen Zahl ist wieder eine Matrix vom Typ (*m*, *n*). Bei der Bildung des Skalarprodukts einer Matrix mit einer Zahl werden also alle Matrixelemente mit dieser Zahl multipliziert.

Es seien **A**, **B** und **C** Matrizen gleichen Typs, $k \in \mathbf{R}$ und $h \in \mathbf{R}$. Dann gelten folgende Regeln:

$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$,
$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$,
$\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$,
Mit $-\mathbf{A} = (-1) \cdot \mathbf{A}$ ist $\mathbf{A} - \mathbf{A} = \mathbf{0}$,

$k \cdot (\mathbf{A} + \mathbf{B}) = k \cdot \mathbf{A} + k \cdot \mathbf{B}$,
$(k + h) \cdot \mathbf{A} = k \cdot \mathbf{A} + h \cdot \mathbf{A}$,
$(k \cdot h) \cdot \mathbf{A} = k \cdot (h \cdot \mathbf{A})$,
$1 \cdot \mathbf{A} = \mathbf{A}$.

Die Menge der Matrizen vom (gleichen) Typ (*m*, *n*) mit der definierten Addition von Matrizen und der Multiplikation von reellen Zahlen mit Matrizen bildet einen **Vektorraum** über **R**.

Erläuterung:

Eine algebraische Struktur $(V, \oplus, K, \cdot)$ heißt **Vektorraum über *K***, wenn gilt:

(i) $(V, \oplus)$ ist eine kommutative Gruppe

(ii) *K* ist ein Körper („Skalarkörper“)

(iii) die Abbildung $\cdot : \begin{cases} K \times V & \to & V \\ (k, \mathbf{v}) & \to & k \cdot \mathbf{v} \end{cases}$ genügt den folgenden Regeln:

für jedes $k \in K$, für jedes $l \in K$, für jedes $\mathbf{v} \in V$ und jedes $\mathbf{w} \in V$ gilt
$k \cdot (l \cdot \mathbf{v}) = (k \cdot l) \cdot \mathbf{v}$,
$1 \cdot \mathbf{v} = \mathbf{v}$,
$k \cdot (\mathbf{v} \oplus \mathbf{w}) = (k \cdot \mathbf{v}) \oplus (k \cdot \mathbf{w})$,
$(k + l) \cdot \mathbf{v} = (k \cdot \mathbf{v}) \oplus (l \cdot \mathbf{v})$.

Das **Produkt** der beiden Matrizen $\mathbf{A}_{(m,n)}$ und $\mathbf{B}_{(n,k)}$ ist nur dann definiert, wenn der erste Faktor $\mathbf{A}_{(m,n)}$ genauso viele Spalten wie der zweite Faktor $\mathbf{B}_{(n,k)}$ Zeilen hat. Das Produkt ist eine Matrix $\mathbf{C}_{(m,k)} = [c_{r,s}] = \mathbf{A} \cdot \mathbf{B}$ vom Typ (m, k) mit

$$c_{r,s} = \sum_{i=1}^{n} a_{r,i} \cdot b_{i,s} \text{ für } r = 1, ..., m, \; s = 1, ..., n .$$

Es gilt:

$$\mathbf{A} \cdot (\mathbf{B} \pm \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} \pm \mathbf{A} \cdot \mathbf{C},$$
$$(\mathbf{A} \pm \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot \mathbf{C} \pm \mathbf{B} \cdot \mathbf{C}.$$

Im allgemeinen ist $\mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}$.

Eine Matrix

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & ... & a_{1,j} & ... & a_{1,n} \\ a_{2,1} & a_{2,2} & ... & a_{2,j} & ... & a_{2,n} \\ . & . & ... & . & ... & . \\ a_{i,1} & a_{i,2} & ... & a_{i,j} & ... & a_{i,n} \\ . & . & ... & . & ... & . \\ a_{m,1} & a_{m,2} & ... & a_{m,j} & ... & a_{m,n} \end{bmatrix} = \mathbf{A}_{(m,n)} = [a_{i,j}]_{i=1,...,m;\, j=1,...,n}$$

kann als Menge $\{\vec{a}_1, \vec{a}_2, ..., \vec{a}_i, ..., \vec{a}_m\}$ ihrer Zeilenvektoren mit

$$\vec{a}_i = [a_{i,1} \quad a_{i,2} \quad ... \quad a_{i,j} \quad ... \quad a_{i,n}] \text{ für } i = 1, ..., m$$

bzw. als Menge $\{\vec{b}_1, \vec{b}_2, ..., \vec{b}_j, ..., \vec{b}_n\}$ ihrer Spaltenvektoren mit

$$\vec{b}_j = \begin{bmatrix} a_{1,j} \\ a_{2,j} \\ . \\ . \\ a_{i,j} \\ . \\ . \\ . \\ a_{m,j} \end{bmatrix} \text{ für } j = 1, ..., n$$

aufgefasst werden.

Eine Menge $\{\vec{a}_1, ..., \vec{a}_r\}$ von Vektoren heißt **linear unabhängig**, wenn gilt:

Aus der Gleichung

$k_1 \cdot \vec{a}_1 + ... + k_r \cdot \vec{a}_r = \mathbf{0}$ mit $k_i \in \mathbf{R}$ für $i = 1, ..., r$ folgt $k_1 = ... = k_r = 0$.

Andernfalls heißt $\{\vec{a}_1, ..., \vec{a}_r\}$ **linear abhängig**.

Um zu überprüfen, ob eine Menge von Vektoren linear unabhängig ist, stellt man also die „Vektorgleichung“ $k_1 \cdot \vec{a}_1 + ... + k_r \cdot \vec{a}_r = \mathbf{0}$ auf, wobei die reellen Zahlen $k_1, ..., k_r$ zunächst „Unbekannte“ sind, und zeigt dann, dass diese Gleichung nur gültig sein kann, wenn alle Unbekannten $k_1, ..., k_r$ gleich 0 sind. Kann man andererseits die Gleichung $k_1 \cdot \vec{a}_1 + ... + k_r \cdot \vec{a}_r = \mathbf{0}$ aufstellen, wobei mindestens eine der Zahlen $k_1, ..., k_r$ von 0 verschieden ist, so sind die Vektoren linear abhängig.

Sind die Vektoren $\vec{a}_1, ..., \vec{a}_r$ jeweils Spaltenvektoren mit *m* Komponenten, so ist die Vektorgleichung $k_1 \cdot \vec{a}_1 + ... + k_r \cdot \vec{a}_r = \mathbf{0}$ ein **Gleichungssystem** mit *m* Zeilen.

Ein Vektor $\vec{a}$ ist eine **Linearkombination** der Vektoren $\vec{a}_1, ..., \vec{a}_n$, wenn es Zahlen $k_1 \in \mathbf{R}$, ..., $k_n \in \mathbf{R}$ gibt mit

$\vec{a} = k_1 \cdot \vec{a}_1 + ... + k_n \cdot \vec{a}_n$.

In diesem Fall gilt die Vektorgleichung

$k_1 \cdot \vec{a}_1 + ... + k_n \cdot \vec{a}_n - 1 \cdot \vec{a} = \mathbf{0}$, d.h.

die Menge $\{\vec{a}_1, ..., \vec{a}_r, \vec{a}\}$ ist nicht linear unabhängig. Ist umgekehrt die Menge $\{\vec{a}_1, ..., \vec{a}_r\}$ linear abhängig, so sind in der Vektorgleichung

$k_1 \cdot \vec{a}_1 + ... + k_r \cdot \vec{a}_r = \mathbf{0}$

nicht alle Skalare gleich 0, etwa $k_j \neq 0$. Es ist dann

$k_j \cdot \vec{a}_j = -k_1 \cdot \vec{a}_1 - ... - k_{j-1} \cdot \vec{a}_{j-1} - k_{j+1} \cdot \vec{a}_{j+1} - ... - k_r \cdot \vec{a}_r$, also

$$\vec{a}_j = \left(-\frac{k_1}{k_j}\right) \cdot \vec{a}_1 + ... + \left(-\frac{k_{j-1}}{k_j}\right) \cdot \vec{a}_{j-1} + \left(-\frac{k_{j+1}}{k_j}\right) \cdot \vec{a}_{j+1} + ... + \left(-\frac{k_r}{k_j}\right) \cdot \vec{a}_r .$$

Insgesamt ergibt sich damit

Satz 6.1-1:

Es sei $n \geq 2$.

Die Vektoren $\vec{a}_1, ..., \vec{a}_n$ sind genau dann linear abhängig, wenn sich wenigstens ein Vektor dieser Menge als Linearkombination der anderen Vektoren dieser Menge darstellen lässt.

Unter dem **Zeilenrang $r_Z(\mathbf{A})$ einer Matrix $\mathbf{A} = \mathbf{A}_{(m,n)}$** versteht man die Maximalzahl linear unabhängiger Zeilen (-vektoren). Unter dem **Spaltenrang $r_S(\mathbf{A})$ einer Matrix $\mathbf{A} = \mathbf{A}_{(m,n)}$** versteht man die Maximalzahl linear unabhängiger Spalten (-vektoren).

Offensichtlich gilt $r_Z(\mathbf{A}) \leq m$ und $r_S(\mathbf{A}) \leq n$.

Der Beweis des folgenden Satzes erfordert eine Reihe weiterführender Überlegungen und einen ziemlich trickreichen Umgang mit den beteiligten Indizes.

Satz 6.1-2:

Für jede Matrix **A** gilt:

$r_Z(\mathbf{A}) = r_S(\mathbf{A})$.

Wegen Satz 6.1-2 kann man **Rang $r(\mathbf{A})$ einer Matrix A** durch $r(\mathbf{A}) = r_Z(\mathbf{A}) = r_S(\mathbf{A})$ definieren. Ist $\mathbf{A} = \mathbf{A}_{(m,n)}$, d.h. **A** besitzt *m* Zeilen und *n* Spalten, dann ist $r(\mathbf{A}) \le \min\{n, m\}$.

Satz 6.1-3:

Gegeben sei die Matrix

$$\mathbf{A} = \mathbf{A}_{(m,n)} = \begin{bmatrix} \vec{a}_1 \\ . \\ . \\ . \\ \vec{a}_m \end{bmatrix} = \begin{bmatrix} \vec{b}_1 & \dots & \vec{b}_n \end{bmatrix}.$$

(Die Vektoren $\vec{a}_1, ..., \vec{a}_m$ sind die Zeilen der Matrix **A**, die Vektoren $\vec{b}_1, ..., \vec{b}_n$ sind die Spalten von **A**.)

Dann gilt:

Zeilenrang und Spaltenrang von **A** ändern sich nicht, wenn man die Matrix **A** einer der folgenden **elementaren Umformungen** unterwirft:

(z1) Zwei Zeilen von **A** werden vertauscht.

(z2) Eine Zeile $\vec{a}_i$ von **A** wird ersetzt durch $\vec{a}_i + k \cdot \vec{a}_j$, wobei $k \in \mathbf{R} \setminus \{0\}$, $1 \le i \le m$, $1 \le j \le m$ und $i \ne j$ gilt.
(Auf $\vec{a}_i$ wird ein Vielfaches einer anderen Zeile addiert.)

(z3) Eine Zeile $\vec{a}_i$ von **A** wird ersetzt durch $k \cdot \vec{a}_i$, wobei $k \in \mathbf{R} \setminus \{0\}$ und $1 \le i \le m$ gilt.
($\vec{a}_i$ wird um ein Vielfaches verändert.)

(s1) Zwei Spalten von **A** werden vertauscht.

(s2) Eine Spalte $\vec{b}_i$ von **A** wird ersetzt durch $\vec{b}_i + k \cdot \vec{b}_j$, wobei $k \in \mathbf{R} \setminus \{0\}$, $1 \le i \le n$, $1 \le j \le n$ und $i \ne j$ gilt.
(Auf $\vec{b}_i$ wird ein Vielfaches einer anderen Spalte addiert.)

(s3) Eine Spalte $\vec{b}_i$ von **A** wird ersetzt durch $k \cdot \vec{b}_i$, wobei $k \in \mathbf{R} \setminus \{0\}$ und $1 \le i \le n$ gilt.
($\vec{b}_i$ wird um ein Vielfaches verändert.)

6.2 Lineare Gleichungssysteme

Eine Menge von *m* Gleichungen in *n* Variablen der Form

$$
\begin{array}{ccccccccccccc}
a_{1,1}\cdot x_1 & + & a_{1,2}\cdot x_2 & + & \ldots & + & a_{1,j}\cdot x_j & + & \ldots & + & a_{1,n}\cdot x_n & = & b_1 \\
a_{2,1}\cdot x_1 & + & a_{2,2}\cdot x_2 & + & \ldots & + & a_{2,j}\cdot x_j & + & \ldots & + & a_{2,n}\cdot x_n & = & b_2 \\
\vdots & & & & & & & & & & & & \\
a_{i,1}\cdot x_1 & + & a_{i,2}\cdot x_2 & + & \ldots & + & a_{i,j}\cdot x_j & + & \ldots & + & a_{i,n}\cdot x_n & = & b_i \\
\vdots & & & & & & & & & & & & \\
a_{m,1}\cdot x_1 & + & a_{m,2}\cdot x_2 & + & \ldots & + & a_{m,j}\cdot x_j & + & \ldots & + & a_{m,n}\cdot x_n & = & b_m
\end{array}
$$

heißt **lineares Gleichungssystem (in den Variablen** $x_1, \ldots, x_n$**)**. Abgekürzt lässt es sich schreiben als

$$\mathbf{A}_{(m,n)} \cdot \vec{x}_{(n,1)} = \vec{b}_{(m,1)}.$$

Die Matrix $\mathbf{A} = \mathbf{A}_{(m,n)} = \begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,j} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,j} & \ldots & a_{2,n} \\ \vdots & & & & & \\ a_{i,1} & a_{i,2} & \ldots & a_{i,j} & \ldots & a_{i,n} \\ \vdots & & & & & \\ a_{m,1} & a_{m,2} & \ldots & a_{m,j} & \ldots & a_{m,n} \end{bmatrix}$ heißt **Koeffizientenmatrix**.

Die Elemente der Koeffizientenmatrix und des Vektors $\vec{b} = \vec{b}_{(m,1)}$ sind vorgegebene reelle Zahlen.

Jeder Vektor $\vec{x} = \vec{x}_{(n,1)} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}$ mit $\mathbf{A}_{(m,n)} \cdot \vec{x}_{(n,1)} = \vec{b}_{(m,1)}$ heißt **Lösung des linearen Gleichungssystems**.

Ein lineares Gleichungssystem $\mathbf{A}_{(m,n)} \cdot \vec{x}_{(n,1)} = \vec{b}_{(m,1)}$ heißt **homogen**, wenn $b_1 = b_2 = ... = b_m = 0$ ist. Andernfalls heißt es **inhomogen**.

Im linearen Gleichungssystem $\mathbf{A}_{(m,n)} \cdot \vec{x}_{(n,1)} = \vec{b}_{(m,1)}$ heißt die Matrix

$$\left[\mathbf{A}_{(m,n)} \middle| \vec{b}_{(m,1)}\right] = \left[\mathbf{A} \middle| \vec{b}\right] = \left[\begin{array}{ccccccc} a_{1,1} & a_{1,2} & ... & a_{1,j} & ... & a_{1,n} & \mid b_1 \\ a_{2,1} & a_{2,2} & ... & a_{2,j} & ... & a_{2,n} & \mid b_2 \\ . & & & & & & . \; . \\ . & & & & & & . \; . \\ . & & & & & & . \; . \\ a_{i,1} & a_{i,2} & ... & a_{i,j} & ... & a_{i,n} & \mid b_i \\ . & & & & & & . \; . \\ . & & & & & & . \; . \\ . & & & & & & . \; . \\ a_{m,1} & a_{m,2} & ... & a_{m,j} & ... & a_{m,n} & \mid b_m \end{array}\right]$$

die **erweiterte Koeffizientenmatrix**.

Es stellt sich die Frage nach der **Lösbarkeit eines linearen Gleichungssystems** (existiert überhaupt eine Lösung? Ist die Lösung eindeutig bestimmt?)

Gegeben sei das lineare Gleichungssystem $\mathbf{A}_{(m,n)} \cdot \vec{x}_{(n,1)} = \vec{b}_{(m,1)}$. Gesucht wird eine Lösung

$$\vec{x} = \vec{x}_{(n,1)} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}.$$

Im folgenden wird die Typangabe zur Vereinfachung der Schreibweise weggelassen, so dass das Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$ lautet.

Satz 6.2-1:

Das lineare Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$ ist genau dann lösbar, wenn der Rang der Koeffizientenmatrix gleich dem Rang der erweiterten Koeffizientenmatrix ist, d.h. wenn $r(\mathbf{A}) = r(\left[\mathbf{A} \middle| \vec{b}\right])$ gilt. Ist $r(\mathbf{A}) < r(\left[\mathbf{A} \middle| \vec{b}\right])$, so heißt das Gleichungssystem **inkonsistent** (und ist nicht lösbar).

Für das lineare Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$ mit einer *m*-zeiligen und *n*-spaltigen Koeffizientenmatrix $\mathbf{A} = \mathbf{A}_{(m,n)}$ gelte $r(\mathbf{A}) = r(\left[\mathbf{A} \middle| \vec{b}\right]) = r$, so dass das **Gleichungssystem lösbar** ist. Es ist $r \le m$ und $r \le n$.

Ist $r < m$ (= Anzahl der Zeilen bzw. Gleichungen), so ist das Gleichungssystem lösbar, aber $m - r$ Gleichungen sind „überflüssig“, genauer: **redundant**, da sie Linearkombinationen der übrigen Gleichungen sind.

Die Anzahl der Lösungen des Gleichungssystems hängt davon ab, wie sich der Rang *r* zu der Anzahl *n* der Variablen verhält:

Ist $r < n$ (= Anzahl der Spalten bzw. Variablen), so sind $n - r$ Spaltenvektoren Linearkombinationen der anderen Spaltenvektoren. Das System ist lösbar, jedoch mit $n - r$ **freien Variablen**, denen beliebige reelle Werte zugeordnet werden können. Es gibt also **unendlich viele Lösungen**. Die Werte, die den übrigen *r* Variablen zugeordnet werden, hängen von den zugeordneten Werten der freien Variablen ab.

Ist $r = n$ (= Anzahl der Spalten bzw. Variablen), so ist $n \le m$ ($m - n$ Gleichungen sind redundant). Das System ist **eindeutig lösbar**, d.h. es gibt genau eine Lösung.

Ist die Koeffizientenmatrix **A** eines linearen Gleichungssystems quadratisch, d.h. $n = m$, d.h. es gibt soviele Gleichungen wie Variablen, dann gilt:

Für $r(\mathbf{A}) < r(\left[\mathbf{A} \middle| \vec{b}\right]) \le n$ gibt es keine Lösung;

für $r(\mathbf{A}) = r(\left[\mathbf{A} \middle| \vec{b}\right]) = n$ gibt es genau eine Lösung;

für $r(\mathbf{A}) = r(\left[\mathbf{A} \middle| \vec{b}\right]) < n$ gibt es unendlich viele Lösungen.

In einem homogenen linearen Gleichungssystem ist immer $r(\mathbf{A}) = r(\left[\mathbf{A} \middle| \vec{b}\right])$. Es gibt dann wenigstens eine Lösung (nämlich die **triviale Lösung** $x_1 = x_2 = \ldots x_n = 0$). Ist zudem $r(\mathbf{A}) = n$, dann gibt es nur diese Lösung. Ist $r(\mathbf{A}) = r < n$, dann gibt es weitere Lösungen mit $n - r$ freien Variablen. Ist $m < n$, d.h. es gibt weniger Gleichungen als Unbekannte, dann ist auch $r(\mathbf{A}) < n$. Für $m = n$ gibt es nur dann mehr als die triviale Lösung, wenn $r(\mathbf{A}) < n$ ist.

Im folgenden wird eine **Methode zur Lösung eines linearen Gleichungssystems (Gaußscher Algorithmus)** und damit einhergehend eine **Methode zur Bestimmung des Rangs einer Matrix** vorgestellt.

Gegeben sei das lineare Gleichungssystem

$$
\begin{array}{ccccccccccccc}
a_{1,1}\cdot x_1 & + & a_{1,2}\cdot x_2 & + & \ldots & + & a_{1,j}\cdot x_j & + & \ldots & + & a_{1,n}\cdot x_n & = & b_1 \\
a_{2,1}\cdot x_1 & + & a_{2,2}\cdot x_2 & + & \ldots & + & a_{2,j}\cdot x_j & + & \ldots & + & a_{2,n}\cdot x_n & = & b_2 \\
\vdots & & & & & & & & & & & & \\
a_{i,1}\cdot x_1 & + & a_{i,2}\cdot x_2 & + & \ldots & + & a_{i,j}\cdot x_j & + & \ldots & + & a_{i,n}\cdot x_n & = & b_i \\
\vdots & & & & & & & & & & & & \\
a_{m,1}\cdot x_1 & + & a_{m,2}\cdot x_2 & + & \ldots & + & a_{m,j}\cdot x_j & + & \ldots & + & a_{m,n}\cdot x_n & = & b_m
\end{array}
$$

bzw.

$$\mathbf{A}\cdot\vec{x}=\vec{b}\,.$$

Hierbei sei mindestens einer der Werte $a_{i,1}$ in der ersten Spalte von 0 verschieden; denn sonst käme x_1 im Gleichungssystem gar nicht vor. Die erweiterte Koeffizientenmatrix sei wieder

$$
\left[\mathbf{A}_{(m,n)}\middle|\,\vec{b}_{(m,1)}\right]=\left[\mathbf{A}\middle|\,\vec{b}\right]=\begin{bmatrix}
a_{1,1} & a_{1,2} & \ldots & a_{1,j} & \ldots & a_{1,n} & | & b_1 \\
a_{2,1} & a_{2,2} & \ldots & a_{2,j} & \ldots & a_{2,n} & | & b_2 \\
\vdots & & & & & & & \vdots \\
a_{i,1} & a_{i,2} & \ldots & a_{i,j} & \ldots & a_{i,n} & | & b_i \\
\vdots & & & & & & & \vdots \\
a_{m,1} & a_{m,2} & \ldots & a_{m,j} & \ldots & a_{m,n} & | & b_m
\end{bmatrix}=\begin{bmatrix}
\vec{a}_1 & | & b_1 \\
\vec{a}_2 & | & b_2 \\
\vdots & & \vdots \\
\vec{a}_i & | & b_i \\
\vdots & & \vdots \\
\vec{a}_m & | & b_m
\end{bmatrix}.
$$

Sie wird durch elementare Umformungen in eine „Treppenmatrix“ (siehe unten) umgewandelt, aus der man dann die Lösung des Gleichungssystems ablesen kann. Bei diesem Umformungsvorgang wird schrittweise eine Folge von Matrizen $\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(r)}$ erzeugt, die alle den-

selben Rang wie $\left[\mathbf{A}\middle|\vec{b}\right]$ haben. Hierbei ist $\mathbf{A}^{(i)}$ das Ergebnis der Umformung von $\mathbf{A}^{(i-1)}$ nach dem *i*-ten Schritt ($i = 1, ..., r$).

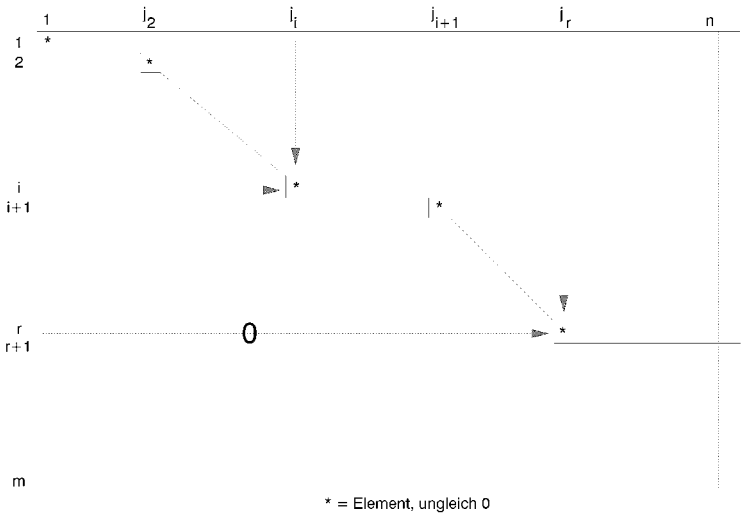
Um die Einträge von $\mathbf{A}^{(i)}$ von den Einträgen der übrigen Matrizen unterscheiden zu können, wird

$$\mathbf{A}^{(i)} = \left[\begin{array}{ccccccc} a_{1,1}^{(i)} & a_{1,2}^{(i)} & \dots & a_{1,j}^{(i)} & \dots & a_{1,n}^{(i)} & \left| \; b_1^{(i)} \right. \\ a_{2,1}^{(i)} & a_{2,2}^{(i)} & \dots & a_{2,j}^{(i)} & \dots & a_{2,n}^{(i)} & \left| \; b_2^{(i)} \right. \\ \cdot & & & & & & \cdot \;\; \cdot \\ \cdot & & & & & & \cdot \;\; \cdot \\ \cdot & & & & & & \cdot \;\; \cdot \\ a_{i,1}^{(i)} & a_{i,2}^{(i)} & \dots & a_{i,j}^{(i)} & \dots & a_{i,n}^{(i)} & \left| \; b_i^{(i)} \right. \\ \cdot & & & & & & \cdot \;\; \cdot \\ \cdot & & & & & & \cdot \;\; \cdot \\ \cdot & & & & & & \cdot \;\; \cdot \\ a_{m,1}^{(i)} & a_{m,2}^{(i)} & \dots & a_{m,j}^{(i)} & \dots & a_{m,n}^{(i)} & \left| \; b_m^{(i)} \right. \end{array}\right] = \left[\begin{array}{cc} \vec{a}_1^{(i)} & \left| \; b_1^{(i)} \right. \\ \vec{a}_2^{(i)} & \left| \; b_2^{(i)} \right. \\ \cdot \;\; \cdot & \cdot \\ \cdot \;\; \cdot & \cdot \\ \cdot \;\; \cdot & \cdot \\ \vec{a}_i^{(i)} & \left| \; b_i^{(i)} \right. \\ \cdot \;\; \cdot & \cdot \\ \cdot \;\; \cdot & \cdot \\ \cdot \;\; \cdot & \cdot \\ \vec{a}_m^{(i)} & \left| \; b_m^{(i)} \right. \end{array}\right]$$

gesetzt.

Zusätzlich wird eine Folge von Spaltennummern $j_1, ..., j_r$ erzeugt, deren Bedeutung aus dem Zusammenhang klar wird.

Die Matrix $\mathbf{A}^{(r)}$, die nach dem *r*-ten Umformungsvorgang entstanden ist, hat die Form



1. Schritt:

In der 1. Spalte von $\left[\mathbf{A}\middle|\,\vec{b}\right]$ wird von oben nach unten das erste von 0 verschiedene Element, d.h. ein Element der Form $a_{s,1} \neq 0$, gesucht.

Es wird $p := a_{s,1}$ gesetzt. Man nennt *p* das **Pivot-Element (im 1. Schritt)**.

Ist $s > 1$, so wird die erste Zeile $\left[\vec{a}_1\middle|\,b_1\right]$ von $\left[\mathbf{A}\middle|\,\vec{b}\right]$ mit der *s*-ten Zeile ausgetauscht; ist bereits $a_{1,1} \neq 0$ (d.h. $s = 1$), so findet kein Austausch statt. Die erste Zeile der durch den eventuellen Zeilenaustausch entstandenen Matrix werde wieder mit $\left[\vec{a}_1\middle|\,b_1\right]$ bezeichnet; entsprechend erhält die ursprünglich erste und nun an der *s*-ten Position stehende Zeile wieder die Bezeichnung $\left[\vec{a}_s\middle|\,b_s\right]$. Insbesondere ist mit dieser Numerierung $p = a_{1,1}$.

Für $k = 2, ..., m$ wird anschließend die Zeile $\left[\vec{a}_k\middle|\,b_k\right]$ durch

$$-a_{k,1} \cdot \left[\vec{a}_1\middle|\,b_1\right] + p \cdot \left[\vec{a}_k\middle|\,b_k\right]$$

ersetzt.

$\mathbf{A}^{(1)}$ ist die so aus $\left[\mathbf{A}\middle|\,\vec{b}\right]$ entstandene Matrix. Es wird $j_1 := 1$ gesetzt.

Ergebnis: Alle Zeilen von $\mathbf{A}^{(1)}$ ab Zeile 2 enthalten mindestens in der ersten Spalte den Wert 0; es gilt außerdem $a^{(1)}_{1,j_1} \neq 0$. Eventuell sind auch in der zweiten und einigen folgenden Spalten von Zeile 2 abwärts ausschließlich die Werte 0 entstanden.

Es wird $i = 2$ gesetzt und im *i*-ten Schritt fortgefahren.

i-ter Schritt für $1 < i \leq m$ *:*

Die Matrix $\mathbf{A}^{(i-1)}$ sei bereits bestimmt. Sie hat die Form

$$\mathbf{A}^{(i-1)} = \left[\begin{array}{cccccccccccc|c}
a_{1,1}^{(i-1)} & a_{1,2}^{(i-1)} & \ldots & . & . & . & . & . & . & . & \ldots & a_{1,n}^{(i-1)} & b_1^{(i-1)} \\
0 & 0 & \ldots & 0 & a_{2,j_2}^{(i-1)} & a_{2,j_2+1}^{(i-1)} & . & . & . & . & \ldots & a_{2,n}^{(i-1)} & b_2^{(i-1)} \\
. & . & \ldots & . & . & . & . & . & . & . & \ldots & . & . \\
0 & 0 & \ldots & . & . & . & . & 0 & a_{i-1,j_{i-1}}^{(i-1)} & a_{i-1,j_{i-1}+1}^{(i-1)} & \ldots & a_{i-1,n}^{(i-1)} & b_{i-1}^{(i-1)} \\
0 & 0 & \ldots & . & . & . & . & 0 & 0 & a_{i,j_{i-1}+1}^{(i-1)} & \ldots & a_{i,n}^{(i-1)} & b_i^{(i-1)} \\
0 & 0 & \ldots & . & . & . & . & 0 & 0 & a_{i+1,j_{i-1}+1}^{(i-1)} & \ldots & a_{i+1,n}^{(i-1)} & b_{i+1}^{(i-1)} \\
. & . & \ldots & . & . & . & . & . & . & . & \ldots & . & . \\
0 & 0 & \ldots & . & . & . & . & 0 & 0 & a_{m,j_{i-1}+1}^{(i-1)} & \ldots & a_{m,n}^{(i-1)} & b_m^{(i-1)}
\end{array}\right]$$

$$= \left[\begin{array}{c|c}
\vec{a}_1^{(i-1)} & b_1^{(i-1)} \\
\vec{a}_2^{(i-1)} & b_2^{(i-1)} \\
. & . \\
\vec{a}_{i-1}^{(i-1)} & b_{i-1}^{(i-1)} \\
\vec{a}_i^{(i-1)} & b_i^{(i-1)} \\
\vec{a}_{i+1}^{(i-1)} & b_{i+1}^{(i-1)} \\
. & . \\
\vec{a}_m^{(i-1)} & b_m^{(i-1)}
\end{array}\right]$$

Es gilt für jede Zeile k mit $1 \le k \le i-1$:

- alle Elemente bis zur Spalte $j_k - 1$ (einschließlich) sind gleich 0
- $a_{k,j_k}^{(i-1} \neq 0$
- alle Elemente in der Teilmatrix, die durch die Zeilen i und m und die Spalten 1 und j_{i-1} (einschließlich) begrenzt wird, sind gleich 0.

In der Teilmatrix

$$\left[\begin{array}{ccc|c}
a_{i,j_{i-1}+1}^{(i-1)} & \ldots & a_{i,n}^{(i-1)} & b_i^{(i-1)} \\
a_{i+1,j_{i-1}+1}^{(i-1)} & \ldots & a_{i+1,n}^{(i-1)} & b_{i+1}^{(i-1)} \\
. & \ldots & . & . \\
a_{m,j_{i-1}+1}^{(i-1)} & \ldots & a_{m,n}^{(i-1)} & b_m^{(i-1)}
\end{array}\right]$$

(das ist der untere rechte Teil) wird von links nach rechts gehend diejenige Spalte bestimmt, die zum ersten Mal Einträge enthält, die nicht sämtlich gleich 0 sind (hierbei wird die Zeilen- und Spaltennumerierung aus der Matrix übernommen):

Es wird also $j = j_{i-1} + 1$ gesetzt und die Bedingung

$$a_{i,j}^{(i-1)} = a_{i+1,j}^{(i-1)} = \ldots = a_{m,j}^{(i-1)} = 0$$

geprüft. Gilt diese Bedingung und ist $j < n$ (= Anzahl der Spalten von **A**), so wird *j* um 1 erhöht und die Bedingung erneut geprüft; gilt die Bedingung und ist bereits $j = n$, so ist das Verfahren beendet.

Im folgenden sei *j* der kleinste Wert, für den die Bedingung nicht gilt, d.h. die Teilmatrix

$$\left[\begin{array}{ccc|c} a_{i,j_{i-1}+1}^{(i-1)} & \ldots & a_{i,n}^{(i-1)} & b_i^{(i-1)} \\ a_{i+1,j_{i-1}+1}^{(i-1)} & \ldots & a_{i+1,n}^{(i-1)} & b_{i+1}^{(i-1)} \\ \cdot & \ldots & \cdot & \cdot \\ a_{m,j_{i-1}+1}^{(i-1)} & \ldots & a_{m,n}^{(i-1)} & b_m^{(i-1)} \end{array}\right]$$

enthält in der Spalte *j* ein Element, das ungleich 0 ist. Die kleinste Zeilennummer, für die das zutrifft, laute *s*, d.h.

$$\begin{aligned} & a_{i,j_{i-1}+1}^{(i-1)} = a_{i+1,j_{i-1}+1}^{(i-1)} = \ldots a_{m,j_{i-1}+1}^{(i-1)} \\ & = \ldots \\ & = a_{i,j-1}^{(i-1)} = a_{i+1,j-1}^{(i-1)} = \ldots a_{m,j-1}^{(i-1)} \\ & = a_{i,j}^{(i-1)} = a_{i+1,j}^{(i-1)} = \ldots a_{s-1,j}^{(i-1)} \\ & = 0 \\ & \text{und } a_{s,j}^{(i-1)} \neq 0. \end{aligned}$$

Es wird $p = a_{s,j}^{(i-1)}$ gesetzt. Der Wert *p* heißt **Pivot-Element (im *i*-ten Schritt)**.

Die *i*-te Zeile von $\mathbf{A}^{(i-1)}$ wird mit der *s*-ten Zeile ausgetauscht (und die Numerierungen der Zeilen wie im ersten Schritt angepaßt).

Es wird $j_i = j$ gesetzt.

Für $k = i+1, \ldots, m$ wird nun Zeile $\left[\vec{a}_k^{(i-1)} \middle| b_k^{(i-1)}\right]$ durch

$$-a_{k,j_i} \cdot \left[\vec{a}_i^{(i-1)} \middle| b_i^{(i-1)}\right] + p \cdot \left[\vec{a}_k^{(i-1)} \middle| b_k^{(i-1)}\right]$$

ersetzt.

Die so entstandene Matrix ist $\mathbf{A}^{(i)}$.

Es wird *i* um 1 erhöht und der *i*-te Schritt mit diesem neuen Wert für *i* wiederholt.

Nach dem *r*-ten Schritt hat die Matrix $\mathbf{A}^{(r)}$ die oben dargestellte Form

$$\mathbf{A}^{(r)} = \left[\begin{array}{ccccccccc|c}
a_{1,j_1}^{(r)} & \ldots & . & . & . & . & \ldots & a_{1,n}^{(r)} & | & b_1^{(r)} \\
0 & \ldots & 0 & a_{2,j_2}^{(r)} & . & . & \ldots & a_{2,n}^{(r)} & | & b_2^{(r)} \\
. & \ldots & . & . & . & . & \ldots & . & | & . \\
0 & \ldots & 0 & . & 0 & a_{r,j_r}^{(r)} & \ldots & a_{r,n}^{(r)} & | & b_r^{(r)} \\
0 & \ldots & 0 & . & 0 & . & \ldots & 0 & | & b_{r+1}^{(r)} \\
. & \ldots & . & . & . & . & \ldots & 0 & | & . \\
0 & \ldots & 0 & . & 0 & . & \ldots & 0 & | & b_m^{(r)}
\end{array}\right]$$

mit $1 = j_1 < j_2 < \ldots < j_r$ und $a_{i,j_i}^{(r)} \neq 0$ für $i = 1, \ldots, r$.

Ist $b_{r+1}^{(r)} = \ldots b_m^{(r)} = 0$, so ist $r(\mathbf{A}) = r\left(\left[\mathbf{A} \middle| \vec{b}\right]\right) = r\left(\mathbf{A}^{(r)}\right) = r$, und das Gleichungssystem ist lösbar. Andernfalls ist das Gleichungssystem nicht lösbar.

Im folgenden sei $r(\mathbf{A}) = r\left(\left[\mathbf{A} \middle| \vec{b}\right]\right) = r\left(\mathbf{A}^{(r)}\right) = r$.

Es gilt:

Das ursprüngliche Gleichungssystem $\mathbf{A} \cdot \vec{x} = \vec{b}$ und das *r*-zeilige Gleichungssystem $\mathbf{A}^{(r)} \cdot \vec{x} = \vec{b}^{(r)}$ haben dieselbe Lösung $\vec{x}$ (da nur elementare Umformungen durchgeführt wurden). In ausgeschriebener Form (ohne die Zeilen, die nur Nullen enthalten) lautet

$\mathbf{A}^{(r)} \cdot \vec{x} = \vec{b}^{(r)}$:

$$\begin{array}{ccccccccccc}
a_{1,1}^{(r)} \cdot x_1 & + & & & \ldots & & & + & a_{1,n}^{(r)} \cdot x_n & = & b_1^{(r)} \\
& & a_{2,j_2}^{(r)} \cdot x_2 & + & & \ldots & & + & a_{2,n}^{(r)} \cdot x_n & = & b_2^{(r)} \\
& & & & . \quad . \quad . \quad . \quad . \quad . & & & & . \quad . \quad . & & \\
& & & & a_{r,j_r}^{(r)} \cdot x_{j_r} & & & & a_{r,n}^{(r)} \cdot x_n & = & b_r^{(r)}
\end{array}$$

Die Zeilen dieser Matrix (es sind die ersten *r* Zeilen von $\mathbf{A}^{(r)}$) werden von ***unten nach oben*** bearbeitet, und dabei werden den Variablen $x_n, x_{n-1}, \ldots, x_1$ Werte zugeordnet:

(r) Bearbeitung der Zeile mit der Nummer r:

Den Variablen $x_{j_r+1}, x_{j_r+2}, ..., x_n$ werden beliebige Werte aus **R** zugewiesen (freie Variablen):

$$x_{j_r+1} := u_{j_r+1}, \quad x_{j_r+2} := u_{j_r+2}, \quad ... \quad , x_n := u_n .$$

x_{j_r} wird aus der letzten Gleichung berechnet:

$$x_{j_r} = \frac{1}{a_{r,j_r}^{(r)}} \cdot \left(b_r^{(r)} - \sum_{k=j_r+1}^{n} a_{r,k}^{(r)} \cdot x_k \right) = \frac{1}{a_{r,j_r}^{(r)}} \cdot \left(b_r^{(r)} - \sum_{k=j_r+1}^{n} a_{r,k}^{(r)} \cdot u_k \right).$$

(i) Bearbeitung der Zeile mit der Nummer i mit $1 \le i < r$:

Die Zeilen $i+1, ..., r$ seien bereits bearbeitet. Die Variablen, die bisher entweder als freie oder berechnete Variablen ermittelt wurden, seien in aufsteigender Numerierung $x_k, x_{k+1}, ..., x_n$.

Den Variablen $x_{j_i+1}, ..., x_{k-1}$ werden wieder beliebige Werte aus **R** zugewiesen (freie Variablen):

$$x_{j_i+1} = u_{j_i+1}, \quad ... \quad , x_{k-1} := u_{k-1} .$$

x_{j_i} wird berechnet zu:

$$x_{j_i} = \frac{1}{a_{i,j_i}^{(r)}} \cdot \left(b_i^{(r)} - \sum_{k=j_i+1}^{n} a_{r,k}^{(r)} \cdot x_k \right).$$

Beispiel:

Das Gleichungssystem

$$\begin{array}{rrrrrrrrrrrrrrrrrrrrrrrrr}
-4 & x_1 & + & 4 & x_2 & - & 8 & x_3 & - & 24 & x_4 & - & 44 & x_5 & + & 4 & x_6 & - & 56 & x_7 & - & 44 & x_8 & = & - & 24 \\
3 & x_1 & - & 3 & x_2 & + & 6 & x_3 & + & 18 & x_4 & + & 30 & x_5 & - & 9 & x_6 & + & 42 & x_7 & + & 24 & x_8 & = & & 15 \\
2 & x_1 & - & 2 & x_2 & + & 4 & x_3 & + & 10 & x_4 & + & 16 & x_5 & - & 4 & x_6 & + & 20 & x_7 & + & 12 & x_8 & = & & 8 \\
-2 & x_1 & + & 2 & x_2 & - & 4 & x_3 & - & 12 & x_4 & - & 18 & x_5 & + & 10 & x_6 & - & 28 & x_7 & - & 10 & x_8 & = & - & 8 \\
2 & x_1 & - & 2 & x_2 & + & 4 & x_3 & + & 10 & x_4 & + & 18 & x_5 & & & & + & 20 & x_7 & + & 18 & x_8 & = & & 10
\end{array}$$

hat die erweiterte Koeffizientenmatrix

$$\left[\begin{array}{rrrrrrrr|r} -4 & 4 & -8 & -24 & -44 & 4 & -56 & -44 & -24 \\ 3 & -3 & 6 & 18 & 30 & -9 & 42 & 24 & 15 \\ 2 & -2 & 4 & 10 & 16 & -4 & 20 & 12 & 8 \\ -2 & 2 & -4 & -12 & -18 & 10 & -28 & -10 & -8 \\ 2 & -2 & 4 & 10 & 18 & 0 & 20 & 18 & 10 \end{array}\right]$$

1. Schritt:

$p = a_{1,1} = -4 \neq 0$; die *k*-te Zeile für $k = 2, 3, 4, 5$ wird ersetzt durch $-a_{k,1} \cdot (1.\,\text{Zeile}) - 4 \cdot (k\text{-te Zeile})$. Das ergibt:

$$\left[\begin{array}{rrrrrrrr|r} -4 & 4 & -8 & -24 & -44 & 4 & -56 & -44 & -24 \\ 0 & 0 & 0 & 0 & 12 & 24 & 0 & 36 & 12 \\ 0 & 0 & 0 & 8 & 24 & 8 & 32 & 40 & 16 \\ 0 & 0 & 0 & 0 & -16 & -32 & 0 & -48 & -16 \\ 0 & 0 & 0 & 8 & 16 & -8 & 32 & 16 & 8 \end{array}\right]$$

Um die Größen der Zahlen zu reduzieren, werden die einzelnen Zeilen jeweils durch einen geeigneten Faktor dividiert, z.B. wird die 1. Zeile durch -4, die 2. Zeile durch 12, die 3. Zeile durch 8, die 4. Zeile durch -16 und die 5. Zeile durch 8 dividiert, und es entsteht:

$$\mathbf{A}^{(1)} = \left[\begin{array}{rrrrrrrr|r} 1 & -1 & 2 & 6 & 11 & -1 & 14 & 11 & 6 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 & 3 & 1 & 4 & 5 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 & 2 & -1 & 4 & 2 & 1 \end{array}\right]$$

$j_1 = 1$

i-ter Schritt für i = 2:

Es ist $j_2 = 4$, $s = 3$, $p = 1$. Die 2. Zeile von $\mathbf{A}^{(1)}$ wird mit der 3. Zeile ausgetauscht, und es ergibt sich

$$\left[\begin{array}{cccccccc|c} 1 & -1 & 2 & 6 & 11 & -1 & 14 & 11 & 6 \\ 0 & 0 & 0 & 1 & 3 & 1 & 4 & 5 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 & 2 & -1 & 4 & 2 & 1 \end{array}\right]$$

Für $k = 3, 4, 5$ wird die *k*-te Zeile ersetzt durch

$$-a_{k,4} \cdot (2.\,\text{Zeile}) + 1 \cdot (k\text{-te Zeile}).$$

Damit ergibt sich

$$\mathbf{A}^{(2)} = \left[\begin{array}{cccccccc|c} 1 & -1 & 2 & 6 & 11 & -1 & 14 & 11 & 6 \\ 0 & 0 & 0 & 1 & 3 & 1 & 4 & 5 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & -1 & -2 & 0 & -3 & -1 \end{array}\right]$$

i-ter Schritt für i = 3:

Es ist $j_3 = 5$, $s = 3$, $p = 1$.

Für $k = 4, 5$ wird die *k*-te Zeile ersetzt durch

$$-a_{k,5} \cdot (3.\,\text{Zeile}) + 1 \cdot (k\text{-te Zeile})$$

Damit ergibt sich

$$\mathbf{A}^{(3)} = \left[\begin{array}{cccccccc|c} 1 & -1 & 2 & 6 & 11 & -1 & 14 & 11 & 6 \\ 0 & 0 & 0 & 1 & 3 & 1 & 4 & 5 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

i-ter Schritt für i = 4:

Für $j = j_3 + 1 \;\; (= 6), ..., 8 \;\; (= n)$ gilt jeweils $a_{4,j}^{(3)} = a_{5,j}^{(3)} = 0$, so dass das Verfahren abbricht. Es ist $r = 3$.

Es wird daher gesetzt:

$$x_6 = u_6, \quad x_7 = u_7, \quad x_8 = u_8,$$

$$x_5 = \frac{1}{1} \cdot \left(1 - \left(2 \cdot x_6 + 0 \cdot x_7 + 3 \cdot x_8\right)\right) = 1 - 2 \cdot u_6 - 3 \cdot u_8,$$

$$x_4 = \frac{1}{1} \cdot \left(2 - \left(3 \cdot x_5 + 1 \cdot x_6 + 4 \cdot x_7 + 5 \cdot x_8\right)\right) = -1 + 5 \cdot u_6 - 4 \cdot u_7 + 4 \cdot u_8,$$

$$x_2 = u_2, \quad x_3 = u_3,$$

$$\begin{aligned} x_1 &= \frac{1}{1} \cdot \left(6 - \left(-1 \cdot x_2 + 2 \cdot x_3 + 6 \cdot x_4 + 11 \cdot x_5 - 1 \cdot x_6 + 14 \cdot x_7 + 11 \cdot x_8\right)\right) \\ &= 1 + u_2 - 2 \cdot u_3 - 7 \cdot u_6 + 10 \cdot u_7 - 2 \cdot u_8. \end{aligned}$$

Die (unendlich große) Lösungsmenge des Gleichungssystems ist also

$$L = \left\{ \vec{X}_{(8,1)} \;\middle|\; \vec{X} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + u_2 \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + u_3 \cdot \begin{bmatrix} -2 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + u_6 \cdot \begin{bmatrix} -7 \\ 0 \\ 0 \\ 5 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + u_7 \cdot \begin{bmatrix} 10 \\ 0 \\ 0 \\ -4 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + u_8 \cdot \begin{bmatrix} -2 \\ 0 \\ 0 \\ 4 \\ -3 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

mit beliebigen reellen Zahlen u_2, u_3, u_6, u_7 und u_8

6.3 Invertieren von Matrizen

Eine quadratische Matrix $\mathbf{A}_{(n,n)}$ vom Typ (n, n) heißt **regulär**, wenn $r\left(\mathbf{A}_{(n,n)}\right) = n$ ist. Sie heißt **singulär**, wenn $r\left(\mathbf{A}_{(n,n)}\right) < n$ gilt.

Es sei $\mathbf{A}_{(n,n)}$ eine *quadratische* Matrix vom Typ (n, n). Gibt es eine Matrix $\mathbf{B}_{(n,n)}$ vom Typ (n, n) mit $\mathbf{A}_{(n,n)} \cdot \mathbf{B}_{(n,n)} = \mathbf{I}_{(n,n)}$, dann heißt $\mathbf{B}_{(n,n)}$ die zu $\mathbf{A}_{(n,n)}$ **inverse Matrix** und wird mit $\mathbf{A}^{-1}_{(n,n)}$ bezeichnet.

Zur Erinnerung: Mit $\mathbf{I}_{(n,n)}$ wird die quadratische Matrix bezeichnet, die in der Diagonalen die Zahlen 1 und sonst nur Nullen enthält (Einheitsmatrix).

Satz 6.3-1:

A und **B** seien quadratische Matrizen, zu denen jeweils die inversen Matrizen $\mathbf{A}^{-1}$ und $\mathbf{B}^{-1}$ existieren. Dann gilt:

$\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$,
$\left(\mathbf{A} \cdot \mathbf{B}\right)^{-1} = \mathbf{B}^{-1} \cdot \mathbf{A}^{-1}$,
$\left(k \cdot \mathbf{A}\right)^{-1} = 1/k \cdot \mathbf{A}^{-1}$ für $k \in \mathbf{R}_{\neq 0}$

Satz 6.3-2:

Für eine quadratische Matrix **A** sind folgende Aussagen (a) und (b) äquivalent:

(a) **A** ist eine reguläre Matrix.

(b) Zu **A** existiert die inverse Matrix $\mathbf{A}^{-1}$.

Satz 6.3-3:

Die Lösung der Matrixgleichung $\mathbf{A} \cdot \mathbf{X} = \mathbf{B}$ mit einer regulären Matrix **A** lautet $\mathbf{X} = \mathbf{A}^{-1} \cdot \mathbf{B}$.

Die Berechnung der inversen Matrix zu einer gegebenen quadratischen regulären Matrix **A** heißt **Invertieren der Matrix A**.

Die quadratische Matrix $\mathbf{A}_{(n,n)} = \left\lfloor a_{i,j} \right\rfloor_{(n,n)}$ sei regulär. Man kann zeigen, dass man die Zeilen einer regulären Matrix so vertauschen kann, dass nach dem Austausch alle Elemente der Diagonalen von 0 verschieden sind. Daher kann man gleich für $a_{i,i} \neq 0$ für $i = 1, ..., n$ voraussetzen.

Die Matrix $\mathbf{X}_{(n,n)} = \left\lfloor x_{i,j} \right\rfloor_{(n,n)}$ sei in Spaltenschreibweise:

$\mathbf{X} = [\vec{x}_1, ..., \vec{x}_n]$.

Der Vektor $\vec{e}_i$ für $i = 1, ..., n$ sei der Spaltenvektor, der in der *i*-ten Zeile eine 1 und sonst nur Nullen hat.

Zur Invertierung der Matrix **A** sind simultan die *n* linearen Gleichungssysteme

$$\mathbf{A} \cdot \vec{x}_1 = \vec{e}_1, ..., \mathbf{A} \cdot \vec{x}_n = \vec{e}_n$$

zu lösen. Diese Gleichungssysteme kann man zu einem Gleichungssystem

$$\mathbf{A}_{(n,n)} \cdot \mathbf{X}_{(n,n)} = \mathbf{I}_{(n,n)}$$

zusammenfassen und mit einer Variante des Gaußschen Verfahrens lösen (anstelle des Vektors $\vec{b}_{(m,1)}$ steht jetzt die Einheitsmatrix $\mathbf{I}_{(n,n)}$):

Die zu **A** gehörende erweiterte Matrix hat die Form

$$[\mathbf{A} \mid \mathbf{I}] = \begin{bmatrix} a_{1,1} & a_{1,2} & ... & a_{1,n} & \mid & 1 & 0 & 0 & ... & 0 & 0 \\ a_{2,1} & a_{2,2} & ... & a_{2,n} & \mid & 0 & 1 & 0 & ... & 0 & 0 \\ . & . & ... & . & \mid & . & . & . & ... & . & . \\ a_{n,1} & a_{n,2} & ... & a_{n,n} & \mid & 0 & 0 & 0 & ... & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \vec{a}_1 & \mid & 1 & 0 & 0 & ... & 0 & 0 \\ \vec{a}_1 & \mid & 0 & 1 & 0 & ... & 0 & 0 \\ . & \mid & . & . & . & ... & . & . \\ \vec{a}_n & \mid & 0 & 0 & 0 & ... & 0 & 1 \end{bmatrix}$$

Sie wird schrittweise durch elementare Umformungen in eine „erweiterte Diagonalmatrix“ umgewandelt. Dabei wird eine Folge von Matrizen $\mathbf{A}^{(1)}, ..., \mathbf{A}^{(n)}$ erzeugt; $\mathbf{A}^{(i)}$ ist das Ergebnis nach dem *i*-ten Schritt:

1. Schritt:

Es ist $a_{1,1} \neq 0$. Die 1. Zeile der erweiterten Matrix wird durch $a_{1,1}$ geteilt. Anschließend wird für $k = 2, ..., n$ die mit $-a_{k,1}$ multiplizierte 1. Zeile zur *k*-ten Zeile addiert. Das Ergebnis ist $\mathbf{A}^{(1)}$. Das Element in der 1. Spalte und der 1. Zeile ist gleich 1; alle Elemente der 1. Spalte ab Zeile 2 sind gleich 0.

Anschließend wird $i = 2$ gesetzt.

i-ter Schritt für $1 < i \leq n$:

Die Matrix $\mathbf{A}^{(i-1)}$ sei bereits ermittelt. Sie hat die Form

$$\mathbf{A}^{(i-1)} = \left[\begin{array}{cccccc|ccc|ccc}
1 & 0 & 0 & \ldots & 0 & 0 & a_{1,i}^{(i-1)} & a_{1,i+1}^{(i-1)} & \ldots & a_{1,n}^{(i-1)} & u_{1,1}^{(i-1)} & \ldots & u_{1,n}^{(i-1)} \\
0 & 1 & 0 & \ldots & 0 & 0 & a_{2,i}^{(i-1)} & a_{2,i+1}^{(i-1)} & \ldots & a_{2,n}^{(i-1)} & u_{2,1}^{(i-1)} & \ldots & u_{2,n}^{(i-1)} \\
. & . & . & \ldots & . & . & . & . & \ldots & . & . & \ldots & . \\
0 & 0 & 0 & \ldots & 0 & 1 & a_{i-1,i}^{(i-1)} & a_{i-1,i+1}^{(i-1)} & \ldots & a_{i-1,n}^{(i-1)} & u_{i-1,1}^{(i-1)} & \ldots & u_{i-1,n}^{(i-1)} \\
0 & 0 & 0 & \ldots & 0 & 0 & a_{i,i}^{(i-1)} & a_{i,i+1}^{(i-1)} & \ldots & a_{i,n}^{(i-1)} & u_{i,1}^{(i-1)} & \ldots & u_{i,n}^{(i-1)} \\
0 & 0 & 0 & \ldots & 0 & 0 & a_{i+1,i}^{(i-1)} & a_{i+1,i+1}^{(i-1)} & \ldots & a_{i+1,n}^{(i-1)} & u_{i+1,1}^{(i-1)} & \ldots & u_{i+1,n}^{(i-1)} \\
. & . & . & \ldots & . & . & . & . & \ldots & . & . & \ldots & . \\
0 & 0 & 0 & \ldots & 0 & 0 & a_{n,i}^{(i-1)} & a_{n,i+1}^{(i-1)} & \ldots & a_{n,n}^{(i-1)} & u_{n,1}^{(i-1)} & \ldots & u_{n,n}^{(i-1)}
\end{array}\right]$$

$$= \left[\begin{array}{c|c}
\vec{a}_1^{(i-1)} & \vec{u}_1^{(i-1)} \\
\vec{a}_2^{(i-1)} & \vec{u}_2^{(i-1)} \\
. & . \\
\vec{a}_{i-1}^{(i-1)} & \vec{u}_{i-1}^{(i-1)} \\
\vec{a}_i^{(i-1)} & \vec{u}_i^{(i-1)} \\
\vec{a}_{i+1}^{(i-1)} & \vec{u}_{i+1}^{(i-1)} \\
. & . \\
\vec{a}_n^{(i-1)} & \vec{u}_n^{(i-1)}
\end{array}\right]$$

Man kann $a_{i,i}^{(i-1)} \neq 0$ annehmen; ansonsten gibt es wegen der Regularität von **A** ein $s \in \{i, i+1, ..., n\}$ mit $a_{s,i}^{(i-1)} \neq 0$, und die *s*-te Zeile wird mit der *i*-ten Zeile ausgetauscht.
Die *i*-te Zeile wird durch $a_{i,i}^{(i-1)}$ geteilt, so dass sie in der *i*-ten Spalte (im Diagonalelement) den Wert 1 hat. Anschließend wird für $k = 1, ..., i-1, i+1, ..., n$ die mit $-a_{k,i}^{(i-1)}$ multiplizierte *i*-te Zeile zur *k*-ten Zeile addiert. Zu beachten ist, dass hierbei sowohl Zeilen behandelt werden, die oberhalb der *i*-ten Zeile stehen ($k = 1, ..., i-1$), als auch Zeilen, die unterhalb der *i*-ten Zeile stehen ($k = i+1, ..., n$).

$\mathbf{A}^{(i)}$ ist die so entstandene Matrix.

Es wird *i* um1 erhöht und der *i*-te Schritt mit diesem neuen Wert für *i* wiederholt.

$\mathbf{A}^{(n)}$ hat die Form

$$\mathbf{A}^{(n)} = \left[\begin{array}{cccccccccc|ccc} 1 & 0 & 0 & ... & 0 & 0 & 0 & 0 & ... & 0 & u_{1,1}^{(n)} & ... & u_{1,n}^{(n)} \\ 0 & 1 & 0 & ... & 0 & 0 & 0 & 0 & ... & 0 & u_{2,1}^{(n)} & ... & u_{2,n}^{(n)} \\ . & . & . & ... & . & . & . & . & ... & . & . & ... & . \\ 0 & 0 & 0 & ... & 0 & 1 & 0 & 0 & ... & 0 & u_{i-1,1}^{(n)} & ... & u_{i-1,n}^{(n)} \\ 0 & 0 & 0 & ... & 0 & 0 & 1 & 0 & ... & 0 & u_{i,1}^{(n)} & ... & u_{i,n}^{(n)} \\ 0 & 0 & 0 & ... & 0 & 0 & 0 & 1 & ... & 0 & u_{i+1,1}^{(n)} & ... & u_{i+1,n}^{(n)} \\ . & . & . & ... & . & . & . & . & ... & . & . & ... & . \\ 0 & 0 & 0 & ... & 0 & 0 & 0 & 0 & ... & 1 & u_{n,1}^{(n)} & ... & u_{n,n}^{(n)} \end{array}\right]$$

$$= \left[\mathbf{I}_{(n,n)} \quad | \quad \mathbf{U}_{(n,n)}\right]$$

Es gilt $\mathbf{U}_{(n,n)} = \mathbf{A}^{-1}$.

Beispiel:

Bestimmung der zu $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$ inversen Matrix:

Die folgenden Matrizen sind die Ergebnisse nach den Schritten :

$$\mathbf{A}^{(1)} = \begin{bmatrix} 1 & 2 & 3 & | & 1 & 0 & 0 \\ 0 & -1 & -4 & | & -2 & 1 & 0 \\ 0 & 0 & -1 & | & -1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{A}^{(2)} = \begin{bmatrix} 1 & 0 & -5 & | & -3 & 2 & 0 \\ 0 & 1 & 4 & | & 2 & -1 & 0 \\ 0 & 0 & -1 & | & -1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{A}^{(3)} = \begin{bmatrix} 1 & 0 & 0 & | & 2 & 2 & -5 \\ 0 & 1 & 0 & | & -2 & -1 & 4 \\ 0 & 0 & 1 & | & 1 & 0 & -1 \end{bmatrix}$$

Es gilt $\mathbf{A}^{-1} = \begin{bmatrix} 2 & 2 & -5 \\ -2 & -1 & 4 \\ 1 & 0 & -1 \end{bmatrix}$.