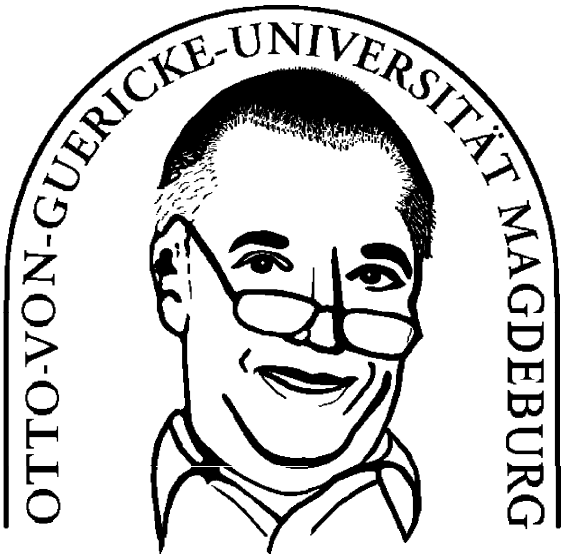


Mathematische Grundlagen der Informatik

Algebra, Graphen, Analysis, Stochastik, Numerik

H. Hollatz

Letzte Änderung am 3. September 2002

hh (http://horst.hollatz.de; horst@hollatz.de).

Anmerkung.
Die hier dargelegte Vorlesung halte ich seit dem Jahre 1985, dem Beginn des Studienganges Informatik an der damaligen Otto-von-Guericke-Hochschule in Magdeburg. Auf Drängen mehrerer Kollegen aus der Informatik (aber ohne ihre gerätetechnische Hilfe) habe ich im Sturmherbst des Jahres 1989 mit der Niederschrift begonnen. Im Jahre 1993 entschied ich mich unter dem Druck der neuen Medien, das Manuskript vollständig zu überarbeiten und ihm die hiermit vorliegende Form zu geben. Das Manuskript enthält den vollständigen, nicht erweiterten, um die eigenen Meinungen und Erlebnisse gekürzten Text der 4stündigen Vorlesung und besteht aus den folgenden Teilen:

1. Algebra
2. lineare Algebra,
3. Graphentheorie,
4. Analysis,
5. Stochastik,
6. Numerik.

Die Teile werden innerhalb von 3 Semestern in der angegebenen Reihenfolge gelesen. Bis zum Jahre 1999 war dies eine 4-semestrige Vorlesung und enthielt außerdem Einführungen in die mathematische Logik und die lineare Optimierung. Seit dem Jahre 2000 gibt es Programme zu Algorithmen. Die C-Programme können über meine www-Seite bezogen werden.
Das Manuskript ist kein Ersatz für die Vorlesung, was der geneigte Interessent beim Lesen bald merken wird. Neben der Vermittlung von grundlegendem, mathematischem Wissen besteht das Ziel der Einführung nicht vorrangig im Beschreiben und Üben mathematischer Techniken, sondern im Erlernen der mathematischen Denk- und Ausdrucksweise, treu dem Grundsatz: Nicht das Ziel ist das Leben, sondern der Weg, auf dem, wenn man das Ziel verfehlt. Das Wesen der Mathematik besteht nicht in ihren Resultaten, sondern in den Methoden, mit denen sie erreicht werden.
Bei 28 Vorlesungen pro Semester ist es nur mit großer Konzentration möglich, den Stoff zu vermitteln. Dies wird erschwert bzw. unmöglich gemacht, falls durch Feiertage entstehende Ausfalltage während eines Semesters nicht nachgeholt werden.
Von Studenten und Kollegen habe ich vielfache Unterstützung erhalten, wofür ich mich herzlich bedanke. Besonders möchte ich Frau Diana Kritsche hervorheben, die die Zeichnungen angefertigt und das gesamte Manuskript kritisch korrigierend studiert hat; fruchtbare Streitgespräche mit ihr haben zu Veränderungen der Darstellung geführt, die auch als Verbesserungen anzusehen sind. Die Übungsaufgaben wurden wesentlich von Frau Ute Förster zusammengestellt; sie konnte dabei Sammlungen meiner anderen Übungsleiter verwenden, wie z. B. von Dr. Peter Sigler, Dr. Norbert Schimark und Dr. Michael Schaper. Ich danke ihnen allen.

Mir ist bewußt, daß die Übungsaufgaben nicht den heutigen inhaltlichen Anforderungen, wohl aber der aktuellen Übungsliteratur entsprechen. Schwerpunkte der Übungen sollten insbesondere die warum-macht-man-das-so-Frage, das mathematische Modellieren, das Finden von Algorithmen für Aufgaben aus Mathematik und Informatik, das Entscheiden von Effizienzkriterien bei Algorithmen, das Finden von Ursachen, die zum Versagen von Algorithmen führen, sein. In diesem Sinn werden die Übungsaufgaben laufend überarbeitet. Desgleichen ermöglicht die vorliegende Netzversion, Korrekturen, Verbesserungen und Aktualisierungen in kurzen Zeitabständen einzuarbeiten. [1]

[1] Die hier dargelegte Vorlesung halte ich seit dem Jahre 1985, dem Beginn des Studienganges Informatik an der damaligen Otto-von-Guericke-Hochschule in Magdeburg. Auf Drängen mehrerer Kollegen aus der Informatik (aber ohne ihre gerätetechnische Hilfe) habe ich im Sturmherbst des Jahres 1989 mit der Niederschrift begonnen. Im Jahre 1993 entschied ich mich unter dem Druck der neuen Medien, das Manuskript vollständig zu überarbeiten und ihm die hiermit vorliegende Form zu geben. Das Manuskript enthält den vollständigen, nicht erweiterten, um die eigenen Meinungen und Erlebnisse gekürzten Text der 4-stündigen Vorlesung und besteht aus den Teilen Algebra, lineare Algebra, Graphentheorie, Analysis, Stochastik, Numerik. Die Teile werden innerhalb von 3 Semestern in der angegebenen Reihenfolge gelesen. Bis zum Jahre 1999 war dies eine 4-semestrige Vorlesung und enthielt außerdem Einführungen in die mathematische Logik und die lineare Optimierung. Seit dem Jahre 2000 gibt es eine Programme zu Algorithmen. Die C++-Programme können über meine www-Seite bezogen werden. Das Manuskript ist kein Ersatz für die Vorlesung, was der geneigte Interessent beim Lesen bald merken wird. Neben der Vermittlung von grundlegendem, mathematischem Wissen besteht das Ziel der Einführung nicht vorrangig im Beschreiben und Üben mathematischer Techniken, sondern im Erlernen der mathematischen Denk- und Ausdrucksweise, treu dem Grundsatz: Nicht das Ziel ist das Leben, sondern der Weg, auch dann, wenn man das Ziel verfehlt. Das Wesen der Mathematik besteht nicht in ihren Resultaten, sondern in den Methoden, mit denen sie erreicht wurden.
Bei 28 Vorlesungen pro Semester ist es nur mit großer Konzentration möglich, den Stoffumfang zu schaffen. Dies wird erschwert bzw. unmöglich gemacht, falls durch Feiertage entstehende Ausfalltage während eines Semesters nicht nachgeholt werden.
Von Studenten und Kollegen habe ich vielfache Unterstützung erhalten, wofür ich mich herzlich bedanke. Besonders möchte ich Frau Bianca Truthe hervorheben, die die Zeichnungen angefertigt und das gesamte Manuskript kritisch korrigierend studiert hat; fruchtbare Streitgespräche mit ihr haben zu Veränderungen der Darstellung geführt, die auch als Verbesserungen anzusehen sind. Die Übungsaufgaben wurden wesentlich von Frau Ute Förster zusammengestellt; sie konnte dabei Sammlungen meiner anderen Übungsleiter verwenden, wie z. B. von Dr. Peter Szyler, Dr. Norbert Schieweck und Dr. Michael Schaper. Ich danke ihnen allen.
Mir ist bewußt, daß die Übungsaufgaben nicht den heutigen inhaltlichen Anforderungen, wohl aber der aktuellen Übungsliteratur entsprechen. Schwerpunkte der Übungen sollten insbesondere die warum-macht-man-das-so-Frage, das mathematische Modellieren, das Finden von Algorithmen für Aufgaben aus Mathematik und Informatik, das Entscheiden von Effizienzkriterien bei Algorithmen, das Finden von Ursachen, die zum Versagen von Algorithmen führen, sein. In diesem Sinne werden die Übungsaufgaben laufend überarbeitet. Desgleichen ermöglicht die vorliegende Netzversion, Korrekturen, Verbesserungen und Aktualisierungen in kurzen Zeitabständen einzuarbeiten.

Inhaltsverzeichnis

Kapitel 1

Algebra

1.1. Mengen

Die Mathematik zeichnet sich im Gebäude der Wissenschaften durch viele Besonderheiten aus; wohl keine andere Wissenschaft ist fähig, innerhalb ihrer Grenzen, mit ihren spezifischen Untersuchungsmethoden, ihre eigenen wissenschaftlichen Grundlagen zu diskutieren. Mathematische Begriffe und Ergebnisse sind so exakt und unmißverständlich, daß jede andere Wissenschaft sich gezwungen sieht, mathematische Methoden anzuwenden, falls sie eine ähnliche Präzision in ihren Resultaten anstrebt. Es scheint fast so, als ob der Mathematisierungsgrad einer Einzelwissenschaft ein Maßstab für ihre Seriosität ist.
Auffällig wird wohl für jeden der hierarchische Begriffsaufbau in der Mathematik sein. Mögliche Sachverhalte, die auf viele Objekte zutreffen, werden in Begriffe gefaßt. Um einen Begriffsinhalt festzulegen, braucht man andere Begriffe, deren Inhalte bereits festgelegt sind. Ein hierarchischer Begriffsaufbau muß verständlicherweise eine oder mehrere Wurzeln haben; das sind Begriffe, deren Inhalte nicht bzw. nicht durch die Mathematik festzulegen sind. Zu diesen atomaren Begriffen gehört der Mengenbegriff. Nach Georg Cantor (berühmter Mathematiker des 19. Jahrhunderts) versteht man unter einer **Menge** die

> *Zusammenfassung von wohlbestimmten und wohlunterschiedenen Objekten aus der Anschauung oder dem Denken, die man Elemente der Menge nennt, zu einem Ganzen.*

Dies ist keine Definition für eine Menge; vielmehr soll hier eine begriffliche Vorstellung von dem gegeben werden, was wir meinen, wenn wir von einer Menge sprechen. Insbesondere dürfen wir nicht in den Fehler verfallen, eine Menge als körperlich gegeben anzusehen. Cantor spricht von der Zusammenfassung von Objekten, also von dem Ergebnis einer gedanklichen Tätigkeit. Der Mengenbegriff setzt daher voraus, daß es jemanden gibt, der zusammenfaßt. Ob Objekte wohlunterschieden sind, hängt wesentlich von dem ab, der Mengen bildet. Der Unterschied zwischen Objekten der Anschauung wird oft durch Eigenschaften der Objekte bestimmt. Eigenschaften wiederum sind meist durch Wörter ausgedrückt. Die Wortbildung innerhalb natürlicher, lebender Sprachen ist ein nicht endender Prozeß. Mit diesen Schwierigkeiten haben viele Mathematiker lange gerungen, ehe sie sich entschieden haben, den Mengenbegriff als nicht definierbar anzuerkennen.
Oft wird es sehr schwierig und aufwendig sein, von einem Objekt zu entscheiden, ob es zu einer Menge gehört oder nicht. Mengen sind gedankliche Konstrukte des Menschen, die es ihm ermöglichen, mit anderen über konkrete Dinge zu sprechen. Wenn wir z. B. den Begriff 'Stuhl' benutzen, so meinen wir damit ein beliebiges Element aus der Menge aller Stühle bzw. aus der Menge aller Stühle in einem Raum. Wenn wir in der Mathematik von der Existenz einer Menge sprechen, meinen wir stets die Existenz als gedankliche Konstruktion; ihre reale Existenz dagegen muß man bezweifeln bzw. schärfer: Mengen gibt es real nicht.
Die begriffliche Vorstellung einer Menge läßt sehr beliebige Mengenbildungen zu; so genügt es insbesondere, eine oder mehrere Eigenschaften anzugeben, durch die die Elemente der betreffenden Menge charakterisiert werden sollen, um sie so von anderen Objekten zu unterscheiden. Den dafür verwendbaren Eigenschaften sind keine Bedingungen auferlegt, so daß man sehr merkwürdige Eigenschaften zur Mengenbildung heranziehen darf. Der englische Philosoph B. Russel hat zur Bildung einer Menge (wir wollen sie mit A bezeichnen) eine zugelassene Eigenschaft angegeben. Die Elemente der Menge A seien durch die folgende Eigenschaft charakterisiert:

> *Ein Objekt ist genau dann Element der Menge A, wenn es sich nicht selbst als Element enthält.*

Wir fragen nun danach, ob die Menge A selbst Element dieser Menge ist. Sollte sie es sein, wäre sie ein Objekt, das sich selbst als Element enthält und könnte daher nicht zur Menge gehören. Gehört sie aber nicht zur Menge, so ist sie ein Objekt, das sich nicht selbst als Element enthält und müßte daher zur Menge gehören. Kurzum: Wie wir es auch drehen, es entsteht ein logischer Widerspruch, der gelöst werden kann, indem man solche Konstruktionen ausschließt. Dies wird noch dadurch unterstützt, daß solche Mengenbildungen in keinen wirklichen mathematischen Anwendungen vorkommen. Praktische Mengenbildungen gehen von einem stufenförmigen Aufbau aus:

Grundbereich - Mengen - Mengensystem - Mengenfamilie - ...,

wobei Teilmengen von Mengen einer Stufe Elemente der nächsten sind und niemals der gleichen. Für die Bildung von Mengensystemen dürfen als Objekte nur Mengen genommen werden. In irgendeiner Menge von Ländern der Erde hat z. B. ein einzelner Mensch nichts zu suchen. Wenn man über Tierarten spricht, d. h. über Teilmengen der Menge aller Tiere, so macht man keine Aussagen über einzelne Tiere, sondern stets über jene Teilmengen, die man zur Mengenbildung zugelassen hat. Oder anders: Eine Aussage über die Menschheit ist etwas prinzipiell anderes als eine Aussage über jeden einzelnen Menschen oder jedes Mitglied einer Gruppe von Menschen. Eine Aussage über die Menschheit wird nicht dadurch fragwürdig oder gar falsch, daß sie möglicherweise in ihrer Wirkung auf einzelne Menschen unannehmbar ist. Die beiden Aussagen 'Die Medizin ist ein Segen für Menschen' und 'Die Medizin schadet der Menschheit' sind zwei Aussagen, die sich nicht ausschließen; sie sprechen über Objekte aus unterschiedlichen Grundbereichen (wohl bestehen Beziehungen, Relationen zwischen ihnen). Durch den stufenförmigen Aufbau werden Widersprüche vermieden. Die Mengenerklärung von CANTOR interpretieren wir nun so, daß man zu einer gegebenen Aussage H alle jene Objekte x eines Grundbereiches E, auf die H zutrifft, zu einer Menge zusammenfassen darf, d. h. wir postulieren das Mengenbildungsprinzip.

Axiom 1 (Mengenbildungsprinzip). *Es gibt eine Menge X, die genau die Objekte x enthält, auf die H zutrifft.*

In Zweifelsfällen, bei denen aus dem Zusammenhang nicht klar hervorgeht, woher die Objekte zu nehmen sind, ist anzugeben, innerhalb welches Grundbereiches E die Betrachtungen verlaufen. In der Informatik nennt man einen solchen Grundbereich gewöhnlich auch **Universum**.
Die entstehenden Mengen sind neue, aber abstrakte Objekte. So ist etwa eine Menge von Zahlen etwas anderes als eine Zahl, selbst dann, wenn sie nur ein Element enthält. Es ist daher nicht sinnvoll zu fragen, ob eine Menge von Elementen des Universums Element einer anderen Menge von Elementen des gleichen Universums ist. Jedoch bildet die Gesamtheit aller über einem Grundbereich herstellbaren Mengen ein neues Universum, auf das wiederum das Mengenbildungsprinzip angewendet werden darf. Die dabei entstehenden Mengen nennt man **Mengensysteme** oder **Mengen zweiter Stufe**.
Für die Mathematik ist es zweckmäßig, auch die sog. **leere Menge** zuzulassen; also eine Menge, auf deren Elemente keine Aussage zutrifft und die daher auch kein Element enthält; sie wird mit $\emptyset$ bezeichnet. Die Zugehörigkeit eines Objektes x zu einer Menge X schreibt man in der Form „$x \in X$“ und spricht: „x ist Element von X“. Falls x nicht Element von X ist, schreibt man $x \notin X$. Mengen kann man durch Auflisten ihrer Elemente oder durch eine Eigenschaft, die allen Elementen der Menge gemeinsam ist, darstellen:

$$X = \{\, x, y, z, \dots \} \text{ oder } X = \{\, x \mid x \text{ hat die Eigenschaft } H \,\}.$$

Für die letztere Form schreibt man auch $X = \{\, x \mid H(x) \,\}$. Hier sei auf eine wichtige Tatsache hingewiesen: Mengen sind unsortiert und unnumeriert; jedes Element des Universums tritt in höchstens einem Exemplar in einer Menge auf.
Bei der Angabe von Eigenschaften benutzen wir gelegentlich abkürzende Zeichen, die der mathematischen Logik entlehnt sind:

$\forall$	für alle ... bzw. für jedes ... oder zu jedem ...
$\exists$	es gibt (mindestens) ein ...
$\Longleftrightarrow$	genau dann, wenn ...
$\Longrightarrow$	daraus folgt, daß ...
$\wedge$	und
$\vee$	oder (nicht ausschließend)
$\neg$	Verneinung (Negation)

Die eine Menge definierende Eigenschaft ist nicht eindeutig bestimmt; also können verschiedene Eigenschaften die gleiche Menge definieren. Daher benötigen wir ein Grundprinzip, das uns sagt, wann zwei Mengen aus dem gleichen Universum übereinstimmen, gleich sind. Die Gleichheit von Mengen wird festgelegt durch das

Axiom 2 (Extensionalitätsprinzip). *Zwei Mengen sind genau dann gleich, wenn sie die gleichen Elemente enthalten.*

Das Extensionalitätsprinzip können wir auch so ausdrücken: Zwei Mengen X und Y sind genau dann gleich, wenn für jedes x gilt: $x \in X$ genau dann, wenn $x \in Y$; in Zeichen:

$$X = Y \iff \forall x (x \in X \iff x \in Y).$$

Damit ist eine wichtige inhaltliche Vorstellung fixiert, die an den Mengenbegriff gebunden sein soll: Unabhängig davon, durch welche Aussage eine Menge ursprünglich definiert wurde, ist sie durch die in ihr enthaltenen

Elemente eindeutig bestimmt. Zu jeder Aussage H über Objekte eines gegebenen Universums gibt es genau eine Menge X, die alle und nur die Objekte x als Elemente enthält, auf die die Aussage H zutrifft. Denn aus

$$\forall x(x \in X \iff H(x)) \wedge \forall x(x \in Y \iff H(x))$$

folgt $\forall x(x \in X \iff x \in Y)$ und nach dem Extensionalitätsprinzip ist $X = Y$.
In der Mathematik hat man auch Mengen zu bilden, für deren Elemente sich nicht eine gemeinsame Eigenschaft angeben läßt. Solche Mengenbildungen verwenden das

Axiom 3 (Auswahlprinzip). *Zu jedem nichtleeren Mengensystem mit paarweise elementfremden Mengen gibt es eine Menge, die mit jeder Menge des Systems genau ein Element gemeinsam hat.*

Eine nach dem Auswahlprinzip gebildete Menge nennt man **Auswahlmenge**. Das Auswahlprinzip besagt, daß man aus jeder Menge eines Mengensystems mit paarweise elementfremden Mengen genau ein Element auswählen und die ausgewählten Elemente zu einer neuen Menge zusammenfassen darf. Die Auswahl kann nach sehr verschiedenen Vorschriften erfolgen, so daß die Mengenbildung nach dem Auswahlprinzip nicht eindeutig ist. So kann man z. B. aus der Menge aller geraden Zahlen und der Menge aller ungeraden Zahlen beliebige Mengen mit genau zwei Elementen bilden, von denen das eine eine gerade und das andere eine ungerade Zahl ist. Aus der Schule sind bereits wichtige Beispiele für Zahlenmengen bekannt:

$\mathbb{N}$: Menge der natürlichen Zahlen ohne 0,
$\mathbb{N}_0$: Menge der natürlichen Zahlen mit 0,
$\mathbb{Z}$: Menge der ganzen Zahlen,
$\mathbb{Q}$: Menge der rationalen Zahlen,
$\mathbb{R}$: Menge der reellen Zahlen.

Die üblichen Rechenregeln in diesen Zahlenmengen nehmen wir als bekannt an. Speziell wissen wir auch, daß jede Menge von natürlichen Zahlen ein kleinstes Element enthält.
Weiter sei an das **Prinzip der vollständigen Induktion** erinnert. Dazu sei H eine von einer natürlichen Zahl n abhängende Aussage. Das Induktionsprinzip lautet dann:

- *Es gibt eine natürliche Zahl n_0 mit: $H(n_0)$ ist eine wahre Aussage.*
- *Für alle $n \geqq n_0$ gilt: Aus $H(n)$ folgt $H(n+1)$.*
- *Dann gilt die Aussage H für alle $n \geqq n_0$.*

In formalisierter Form lautet dieses Prinzip:

$$\exists n_0((H(n_0) \wedge \forall n((n \geqq n_0) \wedge H(n) \Longrightarrow H(n+1))) \Longrightarrow \forall n(n \geqq n_0 \Longrightarrow H(n))).$$

Die erste genannte Eigenschaft nennt man **Induktionsanfang**, die zweite heißt **Induktionsschluß**; die Voraussetzung darin nennt man **Induktionsannahme**. Den Induktionsschluß kann man gleichwertig durch die folgende Formulierung ersetzen:

- Für alle $n \geqq n_0$ gilt: Aus $H(k)$ mit $k \leqq n$ folgt $H(n+1)$.

Das Prinzip der vollständigen Induktion dient zum Beweisen von Aussagen und zur induktiven Definition bzw. Konstruktion von Objekten unterschiedlichster Art: Man gibt erste Objekte - die atomaren Elemente - an und verkündet ein Verfahren, mit dem man aus schon vorhandenen Objekten neue gewinnen kann. Zu jedem Objekt gehört dann eine natürliche Zahl n, so daß man durch n-malige Anwendung des Verfahrens das Objekt aus den atomaren Elementen gewinnen kann; das Objekt ist damit n-stufig aus den atomaren Elementen ableitbar. Damit ist jedem Objekt eine natürliche Zahl zugeordnet, über die man Aussagen mit vollständiger Induktion beweisen kann. Diese Vorgehensweise wird in der mathematischen Logik und theoretischen Informatik sehr oft angewendet.
Wir wollen zwei Beispiele für die Anwendung der vollständigen Induktion betrachten. Gegeben sei eine Schokoladentafel, die in einzelne Riegel gebrochen werden soll, wobei über einen Bruch nicht gebrochen werden darf. Wie groß ist die minimale Anzahl von Brüchen? Wir behaupten dazu folgendes:

> Wenn die Tafel n Riegel hat, dann muß unabhängig von dem benutzten Bruchverfahren stets $(n-1)$-mal gebrochen werden.

Beweis. Den Beweis dieser Aussage führen wir durch vollständige Induktion über die Anzahl n der Riegel. Für $n = 1$ ist die Behauptung offenbar richtig, denn eine Tafel mit genau einem Riegel muß nicht mehr gebrochen werden. Hiermit ist der Induktionsanfang in diesem Falle bereits abgeschlossen; die atomaren Elemente (im Sinne der induktiven Definition von Objekten) sind alle Schokoladentafeln mit genau einem Riegel. Für den Induktionsschluß haben wir zu zeigen:

Wenn alle Tafeln mit m Riegeln, $m \leqq n$, genau $(m-1)$-mal gebrochen werden müssen, so wird jede Tafel mit $n+1$ Riegeln genau n-mal gebrochen.

Nehmen wir also eine beliebige Tafel mit $n+1$ Riegeln und versuchen, eine raffinierte Bruchmethode anzuwenden. Egal, wie diese Methode auch arbeitet: Nach dem ersten Bruch entstehen stets zwei kleinere Tafeln, von denen die eine etwa m und die andere dann $n+1-m$ Riegel hat; wichtig ist für uns, daß jede der beiden höchstens n Riegel hat. Nun können wir auf beide die Induktionsannahme anwenden: Die eine Tafel wird mit $m-1$ und die zweite mit $n+1-m-1 = n-m$ Brüchen zerlegt, was zusammen mit dem Anfangsbruch gerade $(m-1)+(n-m)+1 = n$ Brüche liefert; dies war aber unsere Induktionsbehauptung. Nach dem Induktionsprinzip gilt damit die eingangs aufgestellte Behauptung □

Als zweites Beispiel betrachten wir die Frage, wie man den größten gemeinsamen Teiler $\mathrm{ggT}(m,n)$ zweier natürlicher Zahlen m, n mit $m \geqq n$ ermitteln kann. Wir suchen also eine natürliche Zahl d, die einerseits m und n teilt und andererseits die Eigenschaft hat, daß jeder Teiler von d auch die beiden Zahlen m und n teilt. Setzen wir $m' = m-n$, so ist jeder Teiler von m und n auch ein Teiler von m'; umgekehrt ist jeder Teiler von m' auch ein Teiler von m und n. Deshalb gilt

$$\mathrm{ggT}(m,n) = \mathrm{ggT}(m',n), \quad m' = m-n.$$

Damit ist die Suche nach dem größten gemeinsamen Teiler auf einen einfacheren Fall reduziert, sofern nicht $m' = 0$ ausfällt. In diesem Falle ist aber $m = n$ und daher $\mathrm{ggT}(m,n) = m$. Aus diesen Überlegungen ergibt sich ein Weg, wie man den größten gemeinsamen Teiler ermitteln kann: Wir teilen m durch n mit Rest und setzen $m = an + m'$ mit $0 \leqq m' < n$. Danach wiederholen wir diesen Prozeß mit n und m'. Es sei z. B. $n_0 = 5725, n_1 = 135$. Dann folgt

$$\begin{aligned} 5725 &= 42 \cdot 135 + 55, \\ 135 &= 2 \cdot 55 + 25, \\ 55 &= 2 \cdot 25 + 5, \\ 25 &= 5 \cdot 5 + 0, \end{aligned}$$

also $\mathrm{ggT}(5725,135) = \mathrm{ggT}(135,55) = \mathrm{ggT}(55,25) = \mathrm{ggT}(25,5) = 5$. Diese Methode ist als **euklidischer Algorithmus** bekannt.

Satz 1 (Euklidischer Algorithmus). *Zu je zwei natürlichen Zahlen $n_0 \geqq n_1 > 0$ gibt es Zahlen $n_2, \ldots, n_{k+1}$ mit*

$$n_j = a_j n_{j+1} + n_{j+2}, \quad 0 \leqq n_{j+1} < n_{j+1},\ j = 0, \ldots, k-1, n_{k+1} = 0.$$

Außerdem gilt $n_k = \mathrm{ggT}(n_0, n_1)$.

Beweis. Wir führen den Beweis durch vollständige Induktion über n_0. Für den Induktionsanfang sei $n_0 = 1$. Dann ist $n_1 = 1$ und somit

$$n_0 = 1 \cdot n_1 + 0,$$

was uns sagt, daß die Zahl 1 der größte gemeinsame Teiler ist. Betrachten wir nun eine beliebige natürliche Zahl n_0 und nehmen als Induktionsvoraussetzung an, daß die Behauptung für $n_0 - 1$ gilt. Wir teilen n_0 durch n_1:

$$n_0 = a_0 \cdot n_1 + n_2, \quad 0 \leqq n_2 < n_1.$$

Im Falle $n_2 = 0$ gilt die Behauptung. Andernfalls ist $n_1 < n_0$ und wir dürfen die Induktionsvoraussetzung auf n_1 anwenden, womit wir die gesuchte Folge $n_1, n_2, \ldots, n_k, n_{k+1}$ mit $n_{k+1} = 0$ gefunden haben. Aus der Gleichung $n_j = a_j n_{j+1} + n_{j+2}$ ergibt sich

$$\mathrm{ggT}(n_j, n_{j+1}) = \mathrm{ggT}(n_{j+1}, n_{j+2})$$

und daraus durch Induktion über j:

$$\mathrm{ggT}(n_0, n_1) = \mathrm{ggT}(n_k, n_{k+1}) = \mathrm{ggT}(n_k, 0) = n_k.$$

□

Das folgende Programm GGT liefert mittels euklischem Algorithmus den größten gemeinsamen Teiler zweier natürlicher Zahlen.

```
//=========================================================================
//   Grösster gemeinsamer Teiler zweier natürlicher Zahlen
//=========================================================================
```

```
#define uint unsigned int
uint ggt(uint m, uint n){ uint r; while(r=m%n) m=n, n=r; return(n);}
```

Wichtige Grundbegriffe der Mengenlehre und damit der Mathematik sind u. a. folgende.
Unter einer **Teilmenge** oder **Untermenge** X einer Menge Y - symbolisch durch $X \subseteqq Y$ ausgedrückt - versteht man eine Menge, deren sämtliche Elemente auch in Y liegen:

$$X \subseteqq Y \iff \forall x(x \in X \Longrightarrow x \in Y).$$

Y heißt **Obermenge** von X. Die leere Menge ist natürlich Teilmenge jeder Menge. Eine Teilmenge heißt **echte Teilmenge**, wenn es in der Obermenge mindestens ein Element gibt, das nicht in der Untermenge liegt; in Zeichen $X \subset Y$. In diesem Falle heißt Y auch **echte Obermenge** von X. Die Menge aller Teilmengen einer Menge X heißt **Potenzmenge** und wird mit $\mathcal{P}(X)$ bezeichnet. Die Mathematik studiert vor allem Beziehungen zwischen Objekten, die von ihr selbst 'erfunden' worden sind. Dieses Erfinden geschieht aber nicht im luftleeren Raum, sondern ist auch an die Brauchbarkeit, Anwendbarkeit in anderen Gebieten oder sogar im Leben gebunden. Die Teilmengenbeziehung hat folgende Eigenschaften, die man wohl mühelos einsieht:

1. Jede Menge ist Untermenge von sich selbst, d. h. für jede Menge X gilt: $X \subseteqq X$.
2. Ist eine Menge Untermenge einer anderen und diese Untermenge einer dritten, so ist die erste auch Untermenge der dritten, d. h. für alle Mengen X, Y, Z gilt: Aus $X \subseteqq Y$ und $Y \subseteqq Z$ folgt $X \subseteqq Z$.
3. Wenn eine Menge Untermenge einer anderen ist und umgekehrt, so stimmen beide überein, d. h. für alle Mengen X, Y gilt: Aus $X \subseteqq Y$ und $Y \subseteqq X$ folgt $X = Y$.

Viele Beziehungen zwischen Objekten haben solche Eigenschaften. Wir bezeichnen eine Beziehung abstrakt mit R; gewöhnlich wird eine Beziehung auch mit einem Namen belegt. Betrachten wir Objekte x einer Menge X, so nennen wir eine Beziehung R **reflexiv**, wenn xRx für alle $x \in X$ gilt (x steht zu sich in der Beziehung R). Die Beziehung R heißt **transitiv**, wenn aus xRy und yRz stets folgt, daß auch xRz gilt. Schließlich heißt die Beziehung R **antisymmetrisch**, wenn aus xRy und yRx stets $x = y$ folgt. Formal stellt sich das so dar:

reflexiv : $\forall x(xRx)$.

transitiv : $\forall x \forall y \forall z(xRy \wedge yRz \Longrightarrow xRz)$.

antisymmetrisch : $\forall x \forall y(xRy \wedge yRx \Longrightarrow x = y)$.

Eine Beziehung, für die man Reflexivität, Antisymmetrie und Transitivität nachweisen kann, heißt **Halbordnung**. So ist z. B. die '$\leqq$'-Beziehung für natürliche Zahlen eine Halbordnung.
Aus gegebenen Mengen kann man auf sehr verschiedene Weise neue bilden. Dieses stellen wir uns mittels Mengenoperationen vor; einige der wichtigsten Mengenoperationen sollen nun eingeführt werden. Dazu seien X, Y beliebige Mengen von Objekten eines Universums. Wir wählen als definierende Aussage '$x \in X$ und $x \in Y$'. Nach dem Mengenbildungsprinzip gibt es dann eine Menge Z, die alle und nur die Objekte als Elemente enthält, die sowohl zu X als auch zu Y gehören:

$$x \in Z \iff x \in X \text{ und } x \in Y.$$

Nach dem Extensionalitätsprinzip ist die Menge Z eindeutig bestimmt; man nennt sie den **Durchschnitt** der beiden Mengen; in Zeichen:

$$Z = X \cap Y = \{\, x \mid x \in X \wedge x \in Y \,\}.$$

Die Aussage '$x \in X$ oder $x \in Y$' liefert die **Vereinigung** der beiden Mengen:

$$X \cup Y = \{\, x \mid x \in X \vee x \in Y \,\}.$$

Das Wörtchen 'oder' ist in der Mathematik stets nichtausschließend gemeint (entsprechend auch das Symbol '$\vee$'). Die Vereinigung enthält also genau jene Elemente, die in wenigstens einer der beiden Mengen X, Y enthalten sind einschließlich aller, die zu beiden Mengen gehören. Das ausschließende Oder (entweder oder) liefert die **symmetrische Differenz**:

$$X \triangle Y = \{\, x \mid \text{ entweder } x \in X \text{ oder } x \in Y \,\}.$$

Schließlich definieren wir die **Mengendifferenz** $X \setminus Y$ als Menge all jener Elemente aus der Menge X, die nicht zu Y gehören:

$$X \setminus Y = \{\, x \mid x \in X \wedge x \notin Y \,\}.$$

Sollte bei der Mengendifferenz die Menge Y eine Teilmenge von X sein, so nennt man $X \setminus Y$ das **Komplement** bzw. die **Komplementmenge** von Y in X und bezeichnet sie mit $\mathcal{C}_X(Y)$:

$$\mathcal{C}_X(Y) = \{\, x \mid x \in X \wedge x \notin Y \,\}.$$

Meist liegt bei der Komplementbildung die Menge X, bezüglich derer das Komplement zu bilden ist, dadurch fest, daß man das Universum wählt; dann schreibt man für das Komplement von Y einfach $\overline{Y}$. Wählt man z. B. als Universum die Menge der natürlichen Zahlen und setzt

$$X = \{\, n \mid n \text{ ist gerade} \,\}, Y = \{\, n \mid n \text{ ist Vielfaches von } 5 \,\},$$

so erhält man

$$\begin{aligned} X \setminus Y &= \{\, n \mid n \text{ ist gerade, aber nicht Vielfaches von } 5 \,\}, \\ \overline{X} &= \{\, n \mid n \text{ ist ungerade} \,\}, \\ X \cap Y &= \{\, n \mid n \text{ hat die Endziffer } 0 \,\}, \\ X \cup Y &= \{\, n \mid n \text{ ist gerade oder hat die Endziffer } 5 \,\}. \end{aligned}$$

Für die Mengenoperationen der Vereinigung, des Durchschnitts und der Komplementbildung gelten insbesondere die folgenden Regeln. Dazu sei E ein beliebiges Universum; X, Y, Z seien beliebige Untermengen von E. Wir stellen 9 Rechenregeln fest.

Satz 2. *Für beliebige Untermengen X, Y, Z eines Universums E gelten die folgenden Aussagen.*
1. **(Kommutativität)** *Durchschnitt und Vereinigung sind kommutativ:*

$$X \cap Y = Y \cap X, \quad X \cup Y = Y \cup X.$$

2. **(Assoziativität)** *Durchschnitt und Vereinigung sind assoziativ:*

$$(X \cap Y) \cap Z = X \cap (Y \cap Z), \quad (X \cup Y) \cup Z = X \cup (Y \cup Z).$$

3. **(Distributivität)** *Durchschnitt und Vereinigung sind distributiv:*

$$(X \cup Y) \cap Z = (X \cap Z) \cup (Y \cap Z),$$

$$(X \cap Y) \cup Z = (X \cup Z) \cap (Y \cup Z).$$

4. **(Idempotenz)** *Die Operationen Durchschnitt und Vereinigung sind idempotent:*

$$X \cap X = X, \quad X \cup X = X.$$

5. **(Absorption)** *Durchschnitt und Vereinigung sind absorptiv:*

$$(X \cup Y) \cap X = X, \quad (X \cap Y) \cup X = X.$$

6. **(Null und Eins)** *Leere Menge und Universum wirken als Null bzw. Eins:*

$$X \cap \emptyset = \emptyset, \quad X \cup \emptyset = X, \quad X \cap E = X, \quad X \cup E = E.$$

7. **(Komplementregeln)***:*

$$X \cap \overline{X} = \emptyset, \quad X \cup \overline{X} = E, \quad \overline{\overline{X}} = X.$$

8. **(Modulregel)***: Für alle Untermengen $X \subseteq Y$ und jede Menge Z gilt*

$$X \cup (Y \cap Z) = (X \cup Z) \cap Y.$$

9. **(de Morgansche Regeln)***:*

$$\overline{X \cap Y} = \overline{X} \cup \overline{Y}, \quad \overline{X \cup Y} = \overline{X} \cap \overline{Y}.$$

Man nennt eine Menge mit drei Operationen, die die Eigenschaften 1.-7. haben, eine **boolesche Algebra**. Daher können wir sagen: Die Potenzmenge $\mathcal{P}(E)$ eines beliebigen Universums E bildet mit den Operationen Durchschnitt, Vereinigung und Komplement eine boolesche Algebra. Zur Übung sollten die genannten Regeln vom Leser bewiesen werden. Dazu ein Hinweis. Was ist eigentlich zu beweisen? Auf beiden Seiten der obigen Gleichungen stehen Mengen. Also muß man in allen Fällen beweisen, daß Mengen gleich sind. Nach dem Extensionalitätsprinzip sind zwei Mengen genau dann gleich, wenn sie die gleichen Elemente enthalten. Mit der sog.

Tabellenmethode kann man die obigen Regeln leicht beweisen. Eine vollständige Charakterisierung der Mengen $X \cap Y$, $X \cup Y$, $X \triangle Y$, $X \setminus Y$ liefert die folgende Tabelle:

X	Y	$X \cap Y$	$X \cup Y$	$X \triangle Y$	$X \setminus Y$
1	1	1	1	0	0
1	0	0	1	1	1
0	1	0	1	1	0
0	0	0	0	0	0

.

Die vier Zeilen entsprechen den vier möglichen Fällen, daß nämlich ein gegebenes Objekt x entweder der im Kopf angegebenen Menge angehört (durch 1 angedeutet) oder nicht (durch 0 angedeutet). So ist z. B. die dritte Zeile wie folgt zu lesen: Wenn $x \notin X$ und $x \in Y$, so ist $x \notin X \cap Y$, $x \in X \cup Y$, $x \in X \triangle Y$ und $x \notin X \setminus Y$. Gleichzeitig beschreibt die Tabelle den genauen Gebrauch der logischen Konjunktion ('und'), der Alternative ('oder') und der Antivalenz ('entweder ... oder'). Die Tabellenmethode zum Beweisen der genannten Regeln besteht nun darin, daß man für die in den Regeln auftretenden Mengen entsprechende Tabellen aufstellt. Treten zwei gleiche Spalten auf, so stimmen die betreffenden Mengen überein, andernfalls nicht. Beispielhaft stellen wir die Beweistabelle für die erste Distributivregel auf (mit $U = (X \cup Y) \cap Z$ und $V = (X \cap Z) \cup (Y \cap Z)$):

X	Y	Z	$X \cup Y$	U	$X \cap Z$	$Y \cap Z$	V
0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	1	0	1	0	0	0	0
0	1	1	1	1	0	1	1
1	0	0	1	0	0	0	0
1	0	1	1	1	1	0	1
1	1	0	1	0	0	0	0
1	1	1	1	1	1	1	1

.

In der Tabelle stimmen die Spalten für die Mengen $(X \cup Y) \cap Z$ und $(X \cap Z) \cup (Y \cap Z)$ überein; also sind beide Mengen gleich. Für besonders interessierte sei erwähnt, daß die Tabellenmethode durch das Haubersche Theorem gerechtfertigt ist, nach welchem gilt: Wenn die Voraussetzungen gegebener Sätze alle möglichen Fälle erschöpfen und die Behauptungen sich gegenseitig ausschließen, dann gelten auch die Umkehrungen der Sätze.
Wir wollen noch ein in der Informatik wichtiges Beispiel für eine Boolesche Algebra angeben.
Unter einem **Schalter** verstehen wir eine Vorrichtung, die genau einen von zwei Zuständen annehmen kann; den einen nennen wir 'leitend' (oder auch 'geschlossen') und den anderen 'nicht leitend' ('offen'). Für den offenen Zustand benutzen wir das Symbol '0', für den geschlossenen das Symbol '1'. Diese Größen nennen wir **Schaltwerte**. Aus Schaltern bauen wir nun **Schaltkreise** auf. Es ist klar, daß dieses Modell verschiedenste konkrete Realisierungen zuläßt. Die Schaltkreise definieren wir induktiv:

- Ein Schalter ist ein Schaltkreis mit genau einem Schaltwert 0 oder 1.
- Sind x und y Schaltkreise, so auch $x \wedge y$ und $x \vee y$, sowie $\neg x$ mit Schaltwerten gemäß folgender Tabelle:

x	y	$x \wedge y$	$x \vee y$	$\neg x$
0	0	0	0	1
0	1	0	1	
1	0	0	1	0
1	1	1	1	

.

- Weitere Schaltkreise gibt es nicht.

Einiges zur Erläuterung: Sind x und y Schalter, so ist $x \wedge y$ offenbar eine Serienschaltung, da sie genau dann leitet, wenn beide Schalter geschlossen sind. Der Schaltkreis $x \vee y$ ist eine Parallelschaltung, da sie genau dann nicht leitet, wenn beide Schalter offen sind; $\neg x$ hat den zu x entgegengesetzten Zustand. Nach dieser induktiven Definition gibt es zu jedem Schaltkreis eine natürliche Zahl n, die angibt, daß man durch n-malige Anwendung der Operationen $\wedge, \vee, \neg$ den Schaltkreis aus Schaltern aufbauen kann; man sagt, daß der Schaltkreis n-stufig aus den Schaltern aufbaubar ist. Ferner entnimmt man der Definition von Schaltkreisen, daß das Aufbauen von Schaltkreisen als Ausführen von Operationen $\wedge, \vee, \neg$ auf den Schaltwerten 0, 1 gedeutet werden kann. Es ist leicht mit Hilfe der obigen Tabelle zu erkennen, daß die Menge {0,1} mit den durch die Tabelle definierten Operationen $\wedge$,$\vee$,$\neg$ eine Boolesche Algebra bildet.
Für die folgenden Abschnitte benötigen wir insbesondere noch den Begriff der **Produktmenge** (**Kreuzmenge**, **kartesisches Produkt**) $X \times Y$ von zwei Mengen X, Y:

$$X \times Y = \{\, (x, y) \mid x \in X \wedge y \in Y \,\}.$$

Ein Element der Produktmenge heißt **geordnetes Paar**. Zwei geordnete Paare (x, y), (u, v) sind genau dann gleich, wenn $x = u$ und $y = v$ gilt:

$$(x, y) = (u, v) \iff x = u \text{ und } y = v.$$

So ist z. B. $\mathbb{R} \times \mathbb{R}$ die Menge aller Paare von reellen Zahlen. Ein **geordnetes n-Tupel** $(x_1, \ldots, x_n)$ von Objekten $(n \geqq 2)$ führen wir induktiv ein:

$$(x_1, \ldots, x_n) = ((x_1, \ldots, x_{n-1}), x_n).$$

Das Objekt x_i aus einem n-Tupel heißt i-te **Komponente** (i-tes **Glied**) des n-Tupels. Aus der Gleichheit von geordneten Paaren folgt, daß zwei n-Tupel genau dann gleich sind, wenn sie komponentenweise übereinstimmen. Als Produktmenge $X_1 \times X_2 \times \cdots \times X_n$ von n Mengen $X_1, \ldots, X_n$ definiert man:

$$X_1 \times X_2 \times \cdots \times X_n = \{\, (x_1, \ldots, x_n) \mid x_i \in X_i, i = 1, \ldots, n \,\}.$$

Im Falle $X_1 = X_2 = \cdots = X_n = X$ schreibt man einfach X^n; speziell ist also $\mathbb{R}^n$ die Menge aller n-Tupel von reellen Zahlen. Wenn wir von Paaren bzw. n-Tupeln sprechen, meinen wir stets geordnete Paare bzw. geordnete n-Tupel.

1.2. Relationen und Abbildungen

Objekte der Realität und des Denkens stehen in Beziehungen zueinander. Es scheint, daß die Beziehungen zwischen den Objekten für den Menschen wichtiger sind als die Objekte selbst. So wissen wir z. B. aus vergleichenden Geschichtsanalysen, daß es kein absolutes Schönheitsidol für die Menschen gibt. Wir bezeichnen manche Autos als schön, weil es auch andere gibt, die wir häßlich nennen. Die Beziehungen der Eskimos zu ihrer Natur werden von ihnen in Begriffen erfaßt, die man in keine Weltsprache übersetzen kann; sie haben über 100 verschiedene Wörter für das sie umgebende Weiß. Für den Menschen wichtige Beziehungen sind einerseits in vielen, leicht differenzierten Begriffen und andererseits durch viele Synonyma in seiner Sprache repräsentiert. Denken wir nur an die Liebe oder den Tod. Welche Beziehungen durch wieviele Begriffe beschrieben werden, hat tiefe soziale, materielle Wurzeln; in ihnen widerspiegeln sich ethnische Besonderheiten eines Volkes. Die Mathematik als eine über den Völkern stehende Wissenschaft versucht nun, typische Eigenschaften von Beziehungen zu modellieren und zu untersuchen. Natürlich modelliert die Mathematik keine persönlichen Beziehungen zwischen Menschen, sondern solche, die (möglicherweise) gefühlsunabhängig sind. So gibt es z. B. Beziehungen zwischen Studenten und Hochschulen, Straßen und Wegen, die Städte verbinden, Punkten im Raum, die auf einer Geraden liegen usw. Mathematisch läßt sich das dadurch erfassen, daß man die in Beziehung stehenden Objekte, die oft aus verschiedenen Universen stammen dürfen, zu Paaren zusammenfaßt.
Gegeben seien zwei Mengen X, Y. Eine Untermenge R der Produktmenge $X \times Y$ nennt man **binäre Relation**: $R \subseteqq X \times Y$. Im Falle $X = Y$ spricht man von einer binären Relation über X. Zu einer binären Relation gehört die Schreibweise: $xRy \iff (x, y) \in R$.
In einer Relation müssen nicht alle Elemente der betreffenden Mengen erfaßt sein. Eine Relation kann man sich in folgender Weise gebildet denken: Es sei eine Aussage $H(x, y)$ für die Objekte x eines Universums E_1 und die Objekte y eines Universums E_2 gegeben. Nach dem Mengenbildungsprinzip gibt es dann eine Menge R, die genau alle Paare (x, y) aus dem Universum $E_1 \times E_2$ enthält, auf die die Aussage $H(x, y)$ zutrifft. Nach dem Extensionalitätsprinzip ist die Menge R eindeutig bestimmt:

$$R = \{\, (x, y) \mid H(x, y) \,\}.$$

Diese Menge R wird nun als Relation über $E_1 \times E_2$ aufgefaßt. In diesem Sinne ist $H(x, y)$ eine definierende Aussage für die Relation R. Ausdrücklich sei darauf hingewiesen, daß im Relationsbegriff nicht der Mengenbildungsprozeß, sondern nur das Ergebnis einer Mengenbildung erfaßt ist. Aus den obigen Beispielen gewinnt man folgende Relationen:

$$R_1 = \{\, (P, G) \mid \text{der Punkt } P \text{ liegt auf der Geraden } G \,\},$$
$$R_2 = \{\, (S, T) \mid S, T \text{ sind Studenten der gleichen Hochschule} \,\},$$
$$R_3 = \{\, (k, l) \mid k, l \text{ sind Wegstrecken mit } k < l \,\}.$$

Für die Informatik wichtig sind Darstellungen von binären Relationen über endlichen Mengen. Die geeignete Darstellung einer Relation dient einerseits der guten Veranschaulichung; andererseits soll sie das Ausführen von Operationen mit Relationen unterstützen. Es seien also X, Y Mengen mit endlich vielen Elementen; X habe n, Y habe m Elemente und R sei eine beliebige Relation $R \subseteqq X \times Y$. Eine erste Darstellungsmöglichkeit für R ist die **Tabelle**. Wir erhalten eine Tabelle der Relation R, indem wir jeder Zeile der Tabelle genau ein Element aus X und jeder Spalte genau ein Element aus Y zuordnen; an den Schnittpunkt der Zeile zu $x \in X$ mit der

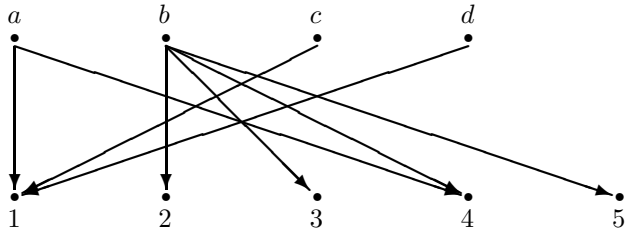
Spalte zu $y \in Y$ schreiben wir die Zahl 1, falls $(x, y) \in R$, andernfalls die Zahl 0. Die so entstandene Tabelle beschreibt die Relation R vollständig. Für

$$X = \{\, a, b, c, d \,\},$$
$$Y = \{\, 1, 2, 3, 4, 5 \,\},$$
$$R = \{\, (a,1), (c,1), (d,1), (b,2), (b,3), (b,5), (a,4)(b,4) \,\}$$

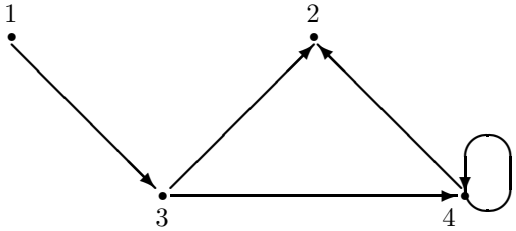
ergibt sich die folgende Tabelle:

	1	2	3	4	5
a	1	0	0	1	0
b	0	1	1	1	1
c	1	0	0	0	0
d	1	0	0	0	0

Wesentliche Nachteile dieser Darstellung sind folgende. Durch die Tabellenform der Relation wird neben den in Relation stehenden Paaren stets eine Anordnung der Elemente mitgeliefert, obwohl diese Anordnung nichts mit der Relation zu tun hat. Damit wird es sehr aufwendig, zwei Relationen auf Übereinstimmung zu prüfen. Die Prüfungsszeit läßt sich reduzieren, wenn man die Elemente beider Relationen nach den gleichen Prinzipien ordnet. Dies erfordet jedoch das Neuordnen nach jeder Änderung. Beim Hinzufügen von Elementen zu einer Relation müssen Duplikate entfernt werden. Relationen zwischen Objekten verschiedener Universen sind in großen Datenbanken abgelegt, wobei die Objekte durch Wörter repräsentiert sind. Dabei richtet sich der zu verwendende Speicherplatz für ein Element nach dem schlechtesten Fall, d. h. nach jenem Wort, das den größten Speicherplatz benötigt. Dieser Umstand bedingt, daß selbst Relationen mit relativ wenig Elementen viel Speicherplatz verschwenden können. Gegenwärtig verringern sich die Operationszeiten von Rechnern in einem viel größeren Maße als die Zugriffszeiten auf externe Speichermedien. Darum ist es eine ständige Forschungsaufgabe, die Speicher- und Zugriffsmechanismen zu den Elementen einer Relation in Datenbanken so zu optimieren, daß die zu lösenden Aufgaben möglichst schnell bearbeitet werden.
Eine andere Darstellungsform für R ist das **Pfeildiagramm** (**gerichteter Graph**). Hier werden die Elemente von X und Y durch Punkte in der Ebene repräsentiert und zwei Punkte durch eine gerichtete Strecke (Pfeil) verbunden, falls das zugeordnete Paar zu R gehört. Die obige Relation könnte dann wie folgt aussehen:



Bei einer binären Relation R über einer Menge X läßt sich die graphische Darstellung noch vereinfachen: Jedem Element $x \in X$, das als eine Komponente in einem Paar aus R auftritt, wird ein Punkt (Knoten) in der Ebene zugeordnet. Sodann zeichnet man einen Pfeil von x nach y, falls x und y in der Relation R stehen, d. h. falls $(x, y) \in R$ gilt. Dabei entsteht ein gerichteter Graph, der im Falle $(x, x) \in R$ auch Schlingen enthält:



Diese Darstellungsform ist besonders für die optische Veranschaulichung von Zusammenhängen gut geeignet. Bei der rechnerinternen Abspeicherung kann man verkettete Listen verwenden. Die oben genannten Nachteile bleiben aber prinzipiell bestehen. Zusätzlich verschärft sich hier das Problem, zwei Relationen auf Gleichheit zu prüfen.
Binäre Relationen lassen sich klassifizieren. Eine binäre Relation R über X heißt **reflexiv**, falls xRx für alle $x \in X$ gilt; sollte für kein $x \in X$ xRx gelten, heißt R **irreflexiv**. Wir nennen eine binäre Relation R **symmetrisch**, wenn für alle $x, y \in X$ aus xRy stets yRx geschlossen werden kann. Sollte aus xRy und yRx stets $x = y$ folgen, nennen wir R **antisymmetrisch**. Im Falle, daß aus xRy mit $x \neq y$ stets folgt, daß yRx nicht gilt, heißt die binäre Relation **asymmetrisch**. Eine binäre Relation R soll **transitiv** heißen, wenn für alle $x, y, z \in X$ aus xRy und yRz stets folgt, daß auch xRz gilt. Schließlich heißt eine binäre Relation R **connex**, wenn für alle $x, y \in X$ gilt: xRy oder yRx oder $x = y$. Eine reflexive, symmetrische und transitive binäre Relation nennt man **Äquivalenzrelation** auf X, während eine reflexive, antisymmetrische, transitive binäre Relation **Halbordnung** heißt. Eine irreflexive, transitive und connexe binäre Relation soll **Ordnung** heißen. So ist z. B. die obige Studentenmenge eine Äquivalenzrelation auf der Menge aller Studenten, die $<$-Relation ist asymmetrisch und eine Ordnung auf der Menge aller natürlichen Zahlen; die $\subseteqq$-Relation ist eine Halbordnung auf der Potenzmenge einer Menge.
Durch Äquivalenzrelationen werden Mengen in Untermengen zerlegt. Ist etwa R eine Äquivalenzrelation auf der Menge X, so sei

$$[x]_R = \{\, y \in X \mid yRx \,\},$$

also die Menge aller jener Elemente aus der Menge X, die zum Element x in der Relation R stehen. Diese Menge heißt **Äquivalenzklasse** von x bezüglich R, oder kurz R-Klasse von x bzw. **Restklasse** von x. Jedes Element aus der Menge X erzeugt mittels der Äquivalenzrelation R eine Restklasse. Offenbar gehört stets das Element x zu seiner Restklasse. Es gilt nun

Satz 3. *Es sei eine Äquivalenzrelation auf der Menge X gegeben. Dann gehört jedes Element $x \in X$ zu genau einer Äquivalenzklasse und jedes Element einer fixierten Äquivalenzklasse erzeugt diese (und nur diese) mit der Äquivalenzrelation.*

Beweis. Wir zeigen, daß im Falle zRx stets $[x]_R = [z]_R$ gelten muß. Für jedes $y \in [x]_R$ gilt yRx; aus zRx folgt wegen der Symmetrie auch xRz, also zusammen yRx und xRz, was uns mit der Transitivität yRz liefert, was aber gleichbedeutend mit $y \in [z]_R$ ist. Folglich gilt $[x]_R \subseteqq [z]_R$. Indem wir nun $y \in [z]_R$ annehmen, schließen wir ganz analog $[z]_R \subseteqq [x]_R$, was zusammen $[x]_R = [z]_R$ bedeutet. □
Diese Überlegung zeigt uns, daß verschiedene Äquivalenzklassen elementfremd sind. Jede Äquivalenzklasse ist durch die Angabe eines ihrer Elemente, eines **Repräsentanten** festgelegt und besteht aus genau allen zu diesem Repräsentanten äquivalenten Elementen. Die Menge aller Äquivalenzklassen bezeichnen wir mit X/R; sie enthält nur elementfremde Teilmengen von X als Elemente und heißt **Zerlegung** von X oder auch **Restsystem** nach der Relation R. Jeder Äquivalenzrelation ist also eine Zerlegung zugeordnet. Ist umgekehrt eine Zerlegung für die Menge X gegeben, so entspricht ihr eine wohlbestimmte Äquivalenzrelation R durch: xRy soll bedeuten, daß die Elemente x und y in der gleichen Teilmenge der Zerlegung liegen. Wir fassen beides zusammen zum

Satz 4. *Jede Zerlegung von X induziert eine Äquivalenzrelation auf X und umgekehrt.*

Beispiele.
1. Es sei folgende Relation auf $\mathbb{N}_0$ gegeben:

kRn genau dann, wenn $k - n$ Vielfaches von 3 ist.

R ist eine Äquivalenzrelation auf $\mathbb{N}_0$, denn die induzierte Zerlegung lautet:

$$[0]_R = \{\, 0, 3, 6, 9, \ldots \,\}, [1]_R = \{\, 1, 4, 7, \ldots\}, [2]_R = \{2, 5, 8, \ldots \,\}.$$

Eine R-Klasse enthält genau jene natürlichen Zahlen, die bei Division durch die Zahl 3 den gleichen Rest lassen.
2. Die Zerlegung von $\mathbb{N}$ in die Teilmengen

$$\{\, 1, \ldots, 9 \,\}, \{\, 10, \ldots, 99 \,\}, \{\, 100, \ldots, 999 \,\}, \ldots$$

induziert folgende Äquivalenzrelation S auf $\mathbb{N}$:

kSn genau dann, wenn k und n die gleiche Ziffernanzahl zur Basis 10 haben.

Allgemein versteht man unter einer n-stelligen Relation R über gegebenen Mengen

$$A_1, \ldots, A_n$$

eine Teilmenge der betreffenden Produktmenge:

$$R \subseteq A_1 \times \cdots \times A_n.$$

In diesem Sinne sind also binäre Relationen 2-stellig. Die Mengen A_i heißen **Faktoren** der Relation. Jeder Faktor A_i wird durch einen Namen identifiziert. Für uns soll der Index i der Name des Faktors A_i sein. In der Informatik kommen Relationen insbesondere im Zusammenhang mit Datenbanken vor. Eine Datenbank mit n Spalten kann als n-stellige Relation aufgefaßt werden. Für den Augenblick nehmen wir an, daß die beteiligten Faktoren einer n-stelligen Relation nur endlich viele Elemente enthalten. Dann kann man sich jede n-stellige Relation in Listenform niedergeschrieben denken: In jeder Zeile steht ein Element (n-Tupel) aus der Relation; in der i-ten Spalte stehen nur Elemente aus dem i-ten Faktor A_i. Die Elemente der Relation dürfen in einer beliebigen Reihenfolge in der Liste auftreten. Anfragen an Datenbanken werden mittels einiger elementarer Operationen über Relationen realisiert. Vor allem sind dies natürlich die elementaren Mengenoperationen Durchschnitt, Vereinigung und Komplement. Mit diesen Operationen ist es nicht möglich, die Stelligkeit von Relationen zu ändern, d. h. Spalten zu streichen bzw. hinzuzufügen. Darum soll der Operationspool um wenige neue Operationen erweitert werden. Als begleitendes Beispiel wählen wir 4 Faktoren:
A_1: Menge von Waren, A_2: Menge von Herstellern,
A_3: Menge von Transportmitteln, A_4: Menge von Verkaufsstellen.
Ein definierender Satz für eine 4-stellige Relation R könnte dann lauten:
$(a_1, a_2, a_3, a_4) \in R$ genau dann, wenn die Ware $a_1 \in A_1$ von Hersteller $a_2 \in A_2$ mit dem Transportmittel $a_3 \in A_3$ in die Verkaufsstelle $a_4 \in A_4$ gebracht wird.
Es sei also $R_4 \subseteq A_1 \times A_2 \times A_3 \times A_4$.
1^0 **Projektion**:
Es sei $R \subseteq A_1 \times \cdots \times A_n$ und $L = \{ l_1, l_2, \ldots, l_s \}$ eine Teilmenge aus $\{ 1, 2, \ldots, n \}$. Aus den n Spalten der zu R gehörenden Liste wählen wir s Spalten mit den Nummern $l_1, \ldots, l_s$ aus und streichen alle anderen; bei zwei gleichen Zeilen streichen wir eine von beiden. Die so entstandene Liste repräsentiert eine Relation $R[A_i, i \in L]$ über $A_{l_1} \times \cdots \times A_{l_s}$, die **Projektion** von R auf $A_{l_1} \times \cdots \times A_{l_s}$. Formal können wir die Elemente aus $R[A_i, i \in L]$ wie folgt charakterisieren: $(a_1, \ldots, a_s) \in R[A_i, i \in L]$ genau dann, wenn ein n-Tupel $(b_1, \ldots, b_n) \in R$ existiert mit $a_i = b_{l_i}, i = 1, \ldots, s$.
Bilden wir etwa für R_4 die Projektion $R_4[A_1, A_2, A_4]$, so enthält die Projektion nur 3-Tupel (a_1, a_2, a_4) und ein definierender Satz für die neue Relation könnte sein: Die Ware a_1 wird von Hersteller a_2 in der Verkaufsstelle a_4 angeboten.
2^0 **Verbund** (**Join**):
Es seien R eine n-stellige Relation über $A_1 \times \cdots \times A_n$, S eine m-stellige Relation über $B_1 \times \cdots \times B_m$, wobei für ein gewisses Indexpaar $(i, j), 1 \leqq i \leqq n, 1 \leqq j \leqq m$ gelte $A_i = B_j$. Wir stellen uns beide Relationen wieder in Listenform vor. Dann bedeutet unsere Voraussetzung, daß die beiden Relationen eine Spalte haben mögen, in denen nur Elemente aus der gleichen Menge stehen dürfen; sie haben also einen gemeinsamen Faktor. Aus den beiden Listen bilden wir nun eine neue mit $n + m - 1$ Spalten $A_1, \ldots, A_n, B_1, \ldots, B_{j-1}, B_{j+1}, \ldots, B_m$ und folgenden Zeilen: Ist $(a_1, \ldots, a_n)$ eine Zeile aus R, $(b_1, \ldots, b_m)$ eine Zeile aus S und gilt $a_i = b_j$, so wird die Zeile $(a_1, \ldots, a_n, b_1, \ldots, b_{j-1}, b_{j+1}, \ldots, b_m)$ in die neue Liste aufgenommen. Diese Liste repräsentiert eine $(n + m - 1)$-stellige Relation $R[A_i]S$ über $A_1 \times \cdots \times A_n \times B_1 \times \cdots \times B_{j-1} \times B_{j+1} \times \cdots \times B_m$ und ist wie folgt definiert:

$(a_1, \ldots, a_n, b_1, \ldots, b_{j-1}, b_{j+1}, \ldots, b_m) \in R[A_i]S$
genau dann, wenn
$(a_1, \ldots, a_n) \in R, (b_1, \ldots, b_m) \in S, a_i = b_j.$

Die so gebildete Relation nennt man **Verbund** oder **Join**.
Mit den obigen Beispielmengen $A_1, \ldots, A_4$ seien
R : Relation über A_1, A_4: Die Ware a_1 wird in a_4 verkauft,
S : Relation über A_1, A_2, A_3: Die Ware a_1 wird von a_2 mit a_3 geliefert. Dann ist $R[A_1]S$ eine Relation über A_1, A_2, A_3, A_4,die genau jene Zeilen (a_1, a_2, a_3, a_4) enthält, wo $(a_1, a_4) \in R$ und $(a_1, a_2, a_3) \in S$ gilt.
Man kann natürlich auch einen Verbund über mehr als eine Spalte machen. Dazu läßt sich die obige Vorgehensweise sofort verwenden, indem man die für den Verbund vorgesehenen Spalten zu einer Superspalte zusammenfaßt.
Allgemein kann man sich einen Join über l gemeinsame Spalten zweier Relationen R und S wie folgt vorstellen. Mit jeder Zeile x von R und jeder Zeile y von S mache man folgende Operation: Wenn x und y in den ausgewählten gemeinsamen l Spalten übereinstimmende Werte haben, werden beide Zeilen aneinander geheftet, einer der gemeinsamen Spaltensätze gestrichen und das so entstandene $(n + m - l)$-Tupel in den Join aufgenommen. Die Join-Operation ist eine der wichtigsten für Datenbanken, da sie es gestattet, zwei Datenbanken zu verschmelzen.
3^0 **Division**:
Es sei R eine Relation über $A_1, \ldots, A_n$, L eine Teilmenge von $\{ 1, \ldots, n \}$, etwa $L = \{ 1, \ldots, m \}$, K enthalte die restlichen Indices:

$$K = \{ 1, \ldots, n \} \setminus L = \{ m+1, \ldots, n \},$$

und S sei eine Relation über $A_1, \ldots, A_m$. Ein $(n-m)$-Tupel liegt genau dann in der Division R/S, wenn es sich durch ein m-Tupel aus S zu einem n-Tupel aus R machen läßt:

$(a_{m+1}, \ldots, a_n) \in R/S$ genau dann, wenn
es ein $(a_1, \ldots, a_m) \in S$ gibt mit $(a_1, \ldots, a_m, a_{m+1}, \ldots, a_n) \in R$.

Algorithmisch kann man sich die Division R/S wie folgt erzeugt denken: Zunächst muß jede Spalte von S auch Spalte von R sein. Mit jeder Zeile von S mache man folgende Operation: Man suche aus R alle Zeilen heraus, die in allen mit S gemeinsamen Spalten überstimmende Werte haben. Aus jeder solchen Zeile werden die mit S gemeinsamen Spalten gestrichen; der Rest bildet eine Zeile der Division, sofern sie noch nicht in der Relation vorkommt.
Formal kann man schreiben:

$$(R[A_i, i \in L]S)[A_i, i \in K].$$

Als Beispiel nehmen wir zu R_4 noch eine Relation S über A_1, A_2 hinzu:

$(a_1, a_2) \in S$ genau dann, wenn a_1 durch a_2 hergestellt wird.

Dann lautet ein definierender Satz für die Division R_4/S : Das Transportmitttel a_3 beliefert die Verkaufsstelle a_4. Falls man die Herkunft der Relation betonen möchte, könnte der Satz auch so lauten: Es gibt für eine Ware a_1 einen Hersteller a_2, so daß $(a_1, a_2) \in S$ gilt und mit dem Transportmittel a_3 die Ware nach a_4 geliefert wird. Insbesondere zeigt sich, daß die Operationen Projektion und Join sehr starke Werkzeuge für die Realisierung von Anfragen bei Datenbanken sind. Dazu ein Beispiel. Es sei eine Relation R gegeben, die alle Studenten des Landes mit den Informationen Name, Vorname, Geburtsjahr, Geburtsmonat, Geburtstag, Geburtsort, Wohnort, Straße, Universität, Fachrichtung erfaßt. Die Anfrage soll lauten: Gib mir bitte eine Namensliste aller Studenten, die Egon heißen, im Januar oder im Mai geboren wurden und in Sachsen-Anhalt studieren. Die Anfrage kann man z. B. so bearbeiten: Wir bilden eine 1-stellige Relation S über die Spalte Vorname; S soll nur ein Element enthalten: $S = \{ (Egon) \}$. Mit dem Join $X = R$[Vorname]S erhalten wir eine Liste, die alle Studenten enthält, die den Vornamen Egon tragen. Nun bilden wir eine 1-stellige Relation T über Geburtsmonat, die nur die beiden Elemente (Januar), (Mai) enthält. Ein Join $Y = X$[Geburtsmonat]T enthält genau alle Studenten, die Egon heißen und im Januar oder im Mai geboren sind. Weiter bilden wir eine 1-stellige Relation U über Universität, in der nur die Elemente (Magdeburg) und (Halle) liegen. Der Join $Z = Y$[Universität]U schränkt die letzte Liste auf die beiden Universitäten Magdeburg und Halle ein. Führen wir abschließend eine Projektion auf die Spalte Name aus, erhalten wir die gewünschte Liste. Formal kann man die Aufgabe auch schneller lösen: Wir bilden eine Relation S über Vorname, Geburtsmonat, Universität mit den Elementen:

(Egon,Januar,Halle), (Egon,Januar,Magdeburg),
(Egon,Mai,Halle), (Egon,Mai,Magdeburg).

Sodann liefert

$$(R[\text{Vorname,Geburtsmonat,Universität}]S)[\text{Name}]$$

die Antwort.
Eine sehr wichtige Operation kann hier noch nicht angegeben werden, da entsprechende Hilfsmittel zu ihrer Beschreibung fehlen. Es handelt sich um die Selektion, eine Operation auf Relationen, die aus einer Relation Elemente auswählt und zu einer neuen Relation zusammenfaßt. Im Logik-Kapitel werden wir über die dazu nötigen Hilfsmittel verfügen.
Für Datenbankanwendungen erwähnen wir den Begriff **Schlüssel**. Eine Menge $\{ A_i \mid i \in L \}$ von Faktoren der Produktmenge $A_1 \times \cdots \times A_n$ heißt **Schlüssel**, wenn für jede Relation R aus einer Menge $\mathcal{R}$ von Relationen über $A_1, \ldots, A_n$ gilt: $|R| = |R[\{ A_i \mid i \in L \}]|$. Ein Schlüssel in einer Relation R dient zum eindeutigen Identifizieren eines n-Tupels in R. Meist bildet ein Faktor einen Schlüssel. Wenn man dann R auf die Schlüsselspalte projiziert, haben R und die Projektion gleichviele Elemente. Die Daten in einer Schlüsselspalte sind Identifikatoren für alle n-Tupel, die in den betrachteten Relationen auftreten. Oft wird noch zwischen Primär- und Sekundärschlüssel unterschieden. Ein Sekundärschlüssel bezieht sich auf eine Untermenge der betrachteten Relationenmenge.
Jede binäre Relation R auf X, Y hat einen sog. Vorbereich $D(R)$ und einen Nachbereich $W(R)$ gemäß:

$$D(R) = \{ x \in X \mid \text{es gibt ein } y \in Y : (x, y) \in R \},$$

$$W(R) = \{ y \in Y \mid \text{es gibt ein } x \in X : (x, y) \in R \}.$$

Der Vorbereich enthält alle jene Elemente aus X, die als erste Komponente eines Paares aus R auftreten; der Nachbereich enthält alle jene Elemente aus Y, die als zweite Komponente eines Paares aus R auftreten. Dabei kann natürlich ein Element aus dem Vorbereich in mehreren Paaren aus der Relation vorkommen. Wenn dies

aber nicht der Fall ist, sprechen wir von einer Abbildung. Eine binäre Relation f über X, Y (d. h. $f \subseteqq X \times Y$) heißt **Abbildung**, wenn es zu jedem $x \in D(f)$ genau ein $y \in W(f)$ gibt, das mit x in der Relation f steht. Im Zusammenhang mit Abbildungen nennt man den Vorbereich auch **Definitionsbereich** oder **Argumentbereich**. Den Nachbereich einer Abbildung nennt man **Wertebereich** oder auch **Bildbereich**. In einem Paar $(x, y) \in f$ spricht man bei x von dem Argument und bei y vom Wert bzw. Bild von x bei der Abbildung f und bezeichnet es mit $f(x)$. Anstelle von Abbildung sagt man oft auch **Funktion**. Beide Namen haben hier den gleichen Inhalt. Allgemein spricht man von einer Abbildung f aus X in Y. Im Falle $D(f) = X$ liegt eine Abbildung von X in Y vor. Die Abbildung heißt **surjektiv**, falls $W(f) = Y$ gilt; bei $D(f) = X, W(f) = Y$ ist f eine Abbildung von X auf Y. Die übliche Schreibweise

$$f : X \longmapsto Y$$

meint stets, daß f eine Abbildung von X in Y sein soll. Die Abbildung f heißt **injektiv**, wenn aus $f(x) = f(y)$ stets folgt, daß auch $x = y$ gilt, d. h. wenn verschiedene Argumente auch verschiedene Werte haben. Unter dem Zeichen $f(U)$ mit $U \subseteqq X$ versteht man das Bild der Elemente aus U bei der Abbildung f:

$$f(U) = \{\, f(x) \mid x \in D(f) \cap U \,\}.$$

Entsprechend ist $f^{-1}(V)$ mit $V \subseteqq Y$ das Urbild der Elemente aus V bei der Abbildung f:

$$f^{-1}(V) = \{\, x \in D(f) \mid f(x) \in V \,\}.$$

Einfache Abbildungen sind die konstante Abbildung

$$f_c : X \longmapsto Y \text{ mit } f_c(x) = c \ \forall x \in X$$

und die identische Abbildung

$$id_X : X \longmapsto X \text{ mit } id_X(x) = x \ \forall x \in X.$$

Eine injektive und surjektive Abbildung f heißt **bijektiv**, umkehrbar eindeutig oder eineindeutig. Bei einer bijektiven Abbildung tritt jedes $y \in Y$ als Bild $f(x)$ von genau einem Element $x \in X$ auf; also kann man die Abbildung f umkehren:

$$f(x) \longmapsto x \ \forall x \in X.$$

Die so definierte Abbildung von Y auf X heißt zu f **invers**, in Zeichen f^{-1}. Sind Argument- und Wertebereich einer Abbildung f endlich, kann man f in Listenform darstellen, wobei eine Spalte den Namen 'Argumente' und die andere den Namen 'Werte' haben könnte. Aus einer solchen Liste kann man bei einer bijektiven Abbildung sofort die inverse gewinnen, indem man die Inhalte der beiden Spalten vertauscht.
Analog zur Komposition von Relationen kann man Abbildungen unter gewissen Umständen verknüpfen (verketten, nacheinander ausführen). Es seien dazu

$$f : X \longmapsto Y,\ g : Y \longmapsto Z.$$

Unter $g \circ f$ versteht man die Nacheinanderausführung der beiden Abbildungen:

$$g \circ f : X \longmapsto Z \text{ mit } x \longmapsto g(f(x)).$$

Für beliebige Abbildungen f, g, h mit

$$f : X \longmapsto Y,\ g : Y \longmapsto Z,\ h : Z \longmapsto U$$

sieht man leicht, daß die Verkettung assoziativ ist:

$$(h \circ g) \circ f = h \circ (g \circ f).$$

Sind f und g injektiv (surjektiv, bijektiv) so auch $g \circ f$. Existiert die zu f inverse Abbildung, so ist

$$f \circ f^{-1} = id_Y,\ f^{-1} \circ f = id_X.$$

Es sei $F(X)$ die Menge aller Abbildungen der Menge X in sich. Eine Teilmenge davon ist die Menge $S(X)$ aller bijektiven Abbildungen von X in sich. Die Verkettung ist dann eine Operation auf der Menge $F(X)$, und in der Menge $S(X)$ gibt es zu jeder Abbildung eine inverse. Ist $f : X \longmapsto Y$ eine Abbildung, so ist die Relation F mit

$$xFy \iff f(x) = f(y)$$

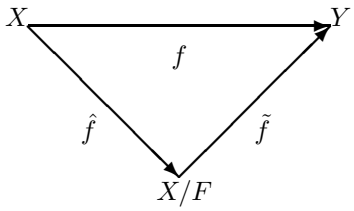
eine Äquivalenzrelation auf X; man nennt sie durch die Abbildung f **induziert**. Ist umgekehrt F eine Äquivalenzrelation auf X, so stellt

$$\hat{f} : X \longmapsto X/F \text{ mit } x \longmapsto [x]_F$$

offenbar eine Abbildung dar, die die Äquivalenzrelation F induziert; daher nennen wir die Abbildung $\hat{f}$ von der Äquivalenzrelation F induziert. Zusammen können wir somit sagen

Satz 5. *Jede Abbildung induziert eine Äquivalenzrelation und umgekehrt.*

Schematisch wird der Sachverhalt durch folgendes Bild veranschaulicht:



Dieses Schema gilt bei Vorgabe einer beliebigen surjektiven Abbildung f. Die bijektive Abbildung $\tilde{f} : X/F \longmapsto Y$ ist definiert als $\tilde{f}([x]_F) = f(x)$. In gewisser Weise kann man das Diagramm auch umkehren. Eine Zerlegung der Menge X induziert eine Äquivalenzrelation auf X, durch die das Restsystem X/F erklärt ist; durch die Äquivalenzrelation F wird eine Abbildung $\hat{f}$ von X auf X/F induziert. Die Verkettung dieser Abbildung mit einer beliebigen bijektiven Abbildung $\tilde{f}$ von X/F auf eine Menge Y liefert eine Abbildung f von X auf Y, und jede Abbildung von X auf Y entsteht in dieser Weise. Bis auf bijektive Abbildungen sind damit durch alle möglichen Zerlegungen von X auch alle möglichen, auf X definierbaren Abbildungen charakterisiert.
Es soll hier das Schaltkreisbeispiel fortgeführt werden. Zwei Schalter sollen **äquivalent** heißen, wenn sie die gleichen Schaltwerte haben. Offenbar ist dies eine Äquivalenzrelation auf der Menge aller Schalter. Da es uns nur auf die Schaltwerte von Schaltern und nicht auf ihre technische Ausführung ankommt, nennen wir die Äquivalenzklassen wieder Schalter. Wir betrachten sinnvoll nur den Fall, daß endlich viele Schalter $X_1, \ldots, X_n$ verfügbar sind. Jeder Stellung der Schalter entspricht ein n-Tupel $(x_1, \ldots, x_n)$ mit $x_i \in \{ 0, 1 \}, i = 1, \ldots, n$ und umgekehrt. Da der Schaltwert eines Schaltkreises bereits durch die Schaltwerte der Schalter bestimmt ist, definiert jeder Schaltkreis $S(X_1, \ldots, X_n)$ genau eine **Boolesche Funktion** (**Schaltfunktion**) f_S auf der Booleschen Algebra $B = \{ 0, 1 \}$ mit den Operationen $\wedge, \vee, \neg$:

$$f_S : B^n \longmapsto B.$$

Zwei verschiedene Schaltkreise können durchaus die gleiche Schaltfunktion definieren. Zwei Schaltkreise heißen daher **äquivalent**, wenn sie die gleiche Schaltfunktion realisieren. Dies ist auch eine Äquivalenzrelation, jetzt aber auf der Menge aller Schaltkreise über den Schaltern $X_1, \ldots, X_n$. Mit dieser Abstraktion haben wir die Untersuchung von Schaltkreisen auf das Studium Boolescher Funktionen reduziert und damit einer mathematischen Behandlung zugängig gemacht. Wesentliche Probleme des Schaltkreisentwurfes sind die Analyse einer gegebenen und die Synthese einer gesuchten Schaltung bei Einhaltung gewisser technischer Bedingungen.
In der Mathematik ist oft eine Indexschreibweise für Abbildungen üblich. Ist etwa f eine Abbildung der Menge I in die Menge Y, so schreibt man für das Bild des Elementes $i \in I$ einfach y_i. Ist f surjektiv, so gilt

$$Y = \{ y_i \mid i \in I \}.$$

Der Argumentbereich I heißt **Indexmenge** für die Elemente von Y, falls alle Elemente aus Y als Bilder auftreten; bei nichtinjektivem f können mehrere y_i gleich sein. Häufig verwendet man die Indexschreibweise für endliche Mengen. Um den Begriff einer endlichen Menge korrekt einzuführen, definieren wir zunächst die Gleichmächtigkeit. Zwei Mengen X, Y heißen **gleichmächtig**, wenn es eine bijektive Abbildung von X auf Y gibt. Leicht überlegt man sich, daß dies eine Äquivalenzrelation G ist. In der Äquivalenzklasse $[X]_G$ liegen alle und nur die zu X gleichmächtigen Mengen. Eine Menge X heißt **endlich** , wenn es eine natürliche Zahl n gibt, so daß $\{ 1, \ldots, n \} \in [X]_G$ gilt; die Zahl n heißt dann **Ordnung** oder **Mächtigkeit** der Menge X und wird mit $|X|$ bezeichnet. Die leere Menge hat die Ordnung 0. Sollte keine solche natürliche Zahl existieren, heißt die Menge **unendlich**. Bei den unendlichen Mengen unterscheiden wir zwischen abzählbar und überabzählbar. Den endlichen Mengen ist damit eine Zahl, ihre Mächtigkeit, zugeordnet. Dies kann man auch für unendliche Mengen durchführen, indem man festsetzt: Jede Klasse gleichmächtiger Mengen definiert eine sog. transfinite Zahl, die wir Mächtigkeit einer Menge dieser Klasse nennen und mit $|X|$ bezeichnen, wobei X eine beliebige Menge der betrachteten Klasse sein soll. Die Mächtigkeiten lassen sich ordnen durch folgende Betrachtung: Wir sagen $|X| \leqq |Y|$, wenn X gleichmächtig zu einer Untermenge von Y ist und $|X| < |Y|$, falls $|X| \leqq |Y|$ und beide Mengen nicht gleichmächtig sind. Damit ist die Mächtigkeit einer Menge X mit der Mächtigkeit einer Menge Y im Sinne der üblichen $\leqq$-Relation und der $<$-Relation vergleichbar.
Ist eine unendliche Menge gleichmächtig zur Menge der natürlichen Zahlen, so heißt sie **abzählbar**, andernfalls **überabzählbar**.
Über unendliche Mengen wollen wir einige Aussagen beweisen. Zunächst gilt

Satz 6. *Die Vereinigung von abzählbar vielen endlichen Mengen ist abzählbar.*

Beweis. Zum Abzählen von abzählbar vielen endlichen Mengen $X_1, \ldots, X_n, \ldots$ geben wir einen Algorithmus an. Es sei $m_i = |X_i|, i = 1, \ldots, n, \ldots$. Der Algorithmus lautet:

$k := 0$;
for $i = 1$ to n **do**
die Elemente von X_i werden mit den Zahlen $k+1, k+2, \ldots, k+m_i$ indiziert;
$k := k + m_i$
end for

Damit ist schon alles bewiesen. □

Satz 7. *Die rationalen Zahlen sind abzählbar.*

Beweis. Es reicht sicherlich aus, die positiven rationalen Zahlen abzuzählen. Jede positive rationale Zahl ist Quotient zweier natürlicher Zahlen. Wir bilden nun die Mengen

$$X_i = \left\{ \frac{p}{q} \;\middle|\; p, q \in \mathbb{N}, p + q = i + 1 \right\},\ i = 1, 2, 3, \ldots.$$

Jede dieser Mengen ist endlich: $|X_i| = i$ und die Vereinigung aller ist die Menge aller positiven rationalen Zahlen. Indem wir den obigen Algorithmus anwenden, stellen wir fest, daß die Vereingung aller dieser Mengen abzählbar ist. □

Satz 8. *Die Vereinigung von abzählbar vielen abzählbaren Mengen ist abzählbar.*

Beweis. Es seien

$$X_i = \{\, x_{1i}, x_{2i}, \ldots, x_{ni}, \ldots \},\ i = 1, \ldots, m, \ldots$$

als abzählbare Mengen gegeben. Offenbar sind die Mengen

$$Y_k = \{\, (i, j) \mid i, j \in \mathbb{N}, i + j = k + 1 \},\ k = 1, 2, \ldots$$

endlich. Damit bilden wir aus den Mengen X_i neue Mengen

$$Z_k = \{\, z \mid z = x_{ij} \in X_j \text{ oder } z = x_{ji} \in X_i, (i, j) \in Y_k \},\ k = 1, 2, \ldots.$$

Die Mengen Z_k sind endlich: $|Z_k| = k$, und die Vereinigung aller Mengen X_i stimmt mit der Vereinigung aller Mengen Z_k überein; diese ist aber mit dem obigen Algorithmus als abzählbar nachgewiesen. □

Satz 9. *Die reellen Zahlen sind überabzählbar.*

Beweis. Um dies zu beweisen, widerlegen wir die Annahme, daß die reellen Zahlen im Intervall $(0, 1)$ abzählbar sind. Bekanntlich läßt sich jede reelle Zahl aus dem Intervall $(0, 1)$ als unendlicher Dezimalbruch

$$0, z_1 z_2 \ldots z_n \ldots$$

schreiben, wobei die Größen z_i Ziffern zwischen 0 und 9 darstellen. Auf Grund der Abzählung können wir alle reellen Zahlen aus (0,1) in einer unendlichen Liste aufführen:

$$\begin{array}{l}
0, z_{11} z_{12} z_{13} z_{14} \ldots z_{1i} \ldots \\
0, z_{21} z_{22} z_{23} z_{24} \ldots z_{2i} \ldots \\
0, z_{31} z_{32} z_{33} z_{34} \ldots z_{3i} \ldots \\
\ldots\ldots\ldots \\
0, z_{i1} z_{i2} z_{i3} z_{i4} \ldots z_{ii} \ldots \\
\ldots\ldots\ldots..
\end{array}$$

Wir konstruieren nun eine Zahl $0, z_1 z_2 \ldots z_i \ldots$, die in dieser Aufzählung nicht vorkommt. In dieser Zahl wählen wir als Ziffer z_1 eine beliebige, aber von z_{11} verschiedene, für z_2 wählen wir eine von z_{22} verschiedene usw., für z_i wählen wir eine von z_{ii} verschiedene Ziffer. Die so entstehende reelle Zahl ist sicher von jeder in der obigen Liste verschieden, denn sie hat an der i-ten Stelle eine Ziffer, die von jener Ziffer verschieden ist, welche in der i-ten Zahl an der i-ten Stelle steht. Auf Grund der angenommenen Abzählbarkeit dürfte es aber eine solche Zahl nicht geben. Dieser Widerspruch löst sich, indem wir davon ausgehen, daß die reellen Zahlen überabzählbar sind. □

An dieser Stelle sei auf einen wichtigen Umstand hingewiesen: Überlicherweise verwendet man in der Mathematik das Prinzip des indirekten Beweises. Um die Richtigkeit einer Aussage zu beweisen, nimmt man an, sie sei falsch und leitet daraus einen Widerspruch zu den Voraussetzungen her. Auf diese Weise hat man lediglich bewiesen, daß die Negation der betrachteten Aussage im Widerspruch zu den Voraussetzungen steht. Diese Tatsache reicht uns, um die Richtigkeit der betrachteten Aussage zu postulieren. Hinter dieser Konstruktion stehen das

Prinzip der Zweiwertigkeit von Aussagen „Jede Aussage ist entweder wahr oder falsch“ und das Prinzip vom ausgeschlossenen Widersruch „Es gibt keine Aussage, die sowohl wahr als auch falsch ist“. Es gibt Mathematiker, die den indirekten Beweis ablehnen und nur solche mathematischen Aussagen akzeptieren, die auf direktem Wege beweisbar sind. Wir werden hier den indirekten Beweis als gültige Beweismethode verwenden.
Falls der Vorbereich X einer Abbildung $*$ eine Produktmenge aus n Komponenten darstellt:

$$X = X_1 \times \cdots \times X_n,$$

spricht man von einer n-stelligen **Operation** :

$$* : X_1 \times X_2 \times \cdots \times X_n \longmapsto Y.$$

Das einem n-Tupel $(x_1, \ldots, x_n) \in X_1 \times \cdots \times X_n$ zugeordnete Element $*(x_1, \ldots, x_n)$ heißt **Resultat** der Operation $*$ für die Operanden $x_1, \ldots, x_n$. Je nachdem, in welchem mathematischen Umfeld man sich bewegt, sind auch andere Namen für den gleichen Begriff gebräuchlich. So ist z. B. in der Analysis der Begriff einer Funktion von n Veränderlichen gleichwertig zur n-stelligen Operation. Bedingt durch die Entwicklung der Informatik werden beschränkt und unbeschränkt ausführbare Operationen betrachtet. Eine Operation ist nur beschränkt ausführbar, wenn nicht alle Elemente der Grundmenge als Operanden zugelassen sind. Im Rahmen dieser Einführung werden unbeschränkt ausführbare Operationen betrachtet. Im Falle einer n-stelligen Operation der Form

$$* : X^n \longmapsto X$$

spricht man von einer n-stelligen Operation auf X. In unserer Definition liegt das Resultat einer Operation automatisch in der Menge X; man sagt, daß die Menge bezüglich der Operation $*$ **abgeschlossen** ist. In der Informatik werden aber auch Operationen betrachtet, die aus der Menge herausführen. So ist z. B. das Produkt zweier Gleitpunktzahlen auf einem Rechner i. a. keine im Rechner darstellbare Gleitpunktzahl; also ist die Menge der Gleitpunktzahlen auf einem Rechner bezüglich der arithmetischen Operationen nicht abgeschlossen. Besonders wichtig sind die binären (zweistelligen) Operationen auf einer Menge X, die wir einfach Operationen nennen:

$$* \text{ heißt Operation auf } X \Longleftrightarrow * : X^2 \longmapsto X.$$

Hier wird das Resultat $*(x, y)$ wie gewöhnlich mit $x * y$ bezeichnet. In diesen Begriff ordnen sich viele bekannte Operationen ein: Addition und Multiplikation von reellen Zahlen, Mengen-Operationen. Die für eine Operation gewählte Bezeichnung (das die Operation symbolisierende Zeichen) ist generisch gemeint, d. h. seine wirkliche Bedeutung hängt von den Operanden ab. So sind die Addition von natürlichen Zahlen und die Addition von reellen Zahlen verschiedene Operationen, beide werden aber mit dem Symbol '+' beschrieben. Generische Funktionen sind typisch für den mathematischen Formalismus. Sofern nicht ausdrücklich etwas anderes vereinbart wird, soll im folgenden unter einer Operation stets eine binäre gemeint sein. In der Algebra betrachtet man Operationen, die verschiedene Eigenschaften haben. Die für uns wichtigen sollen kurz zusammengestellt werden:

$*$ ist **kommutativ** $\Longleftrightarrow x * y = y * x \; \forall x, y \in X$,
$*$ ist **assoziativ** $\Longleftrightarrow x * (y * z) = (x * y) * z \; \forall x, y, z \in X$,
$*$ ist **links-distributiv** bzgl. $\circ$
$\Longleftrightarrow x * (y \circ z) = (x * y) \circ (x * z) \; \forall x, y, z \in X$,
$*$ ist **rechts-distributiv** bzgl. $\circ$
$\Longleftrightarrow (x \circ y) * z = (x * z) \circ (y * z) \; \forall x, y, z \in X$,
$*$ ist **distributiv** bzgl. $\circ$
$*$ ist links- und rechts-distributiv bzgl. $\circ$
$*$ ist **idempotent** $\Longleftrightarrow x * x = x \; \forall x \in X$.

Es ist leicht, sich diese abstrakten Eigenschaften an bekannten Operationen zu veranschaulichen.
Bei einer endlichen Menge wird eine Operation auch oft mittels einer **Operationstafel** beschrieben, z. B.

$*$	a	b	c	d
a	b	c	d	a
b	c	d	a	b
c	d	a	b	c
d	a	b	c	d

.

Mit solcher Tabelle kann man die obigen abstrakten Eigenschaften für eine konkrete Operation studieren.

1.3. Algebraische Strukturen

1.3.1. Homomorphie

Eine **allgemeine Algebra** oder **algebraische Struktur** bzw. einfach **Struktur** ist ein 4-Tupel

$$\mathcal{S} = (S; Kons; Oper; Rela)$$

mit folgender Bedeutung: Die erste Komponente stellt eine beliebige, nichtleere Menge dar, die man auch **Trägermenge** oder **Universum** nennt. Die Voraussetzung, daß S nichtleer sein soll, verhindert, daß man sich mathematische Objekte ausdenkt, für die es keine Realisierung gibt. Die Komponente $Kons$ enthält ausgewählte Elemente, sog. Konstanten aus der Trägermenge S, deren Existenz gesichert sein muß bzw. die zum Formulieren von Eigenschaften und Regeln dienen, deren Gültigkeit von der Struktur verlangt wird. Man nennt sie oft **Alphabet** und ihre Elemente **Atome**, weil sich oft alle Elemente der Trägermenge aus ihnen erzeugen lassen. Danach folgt eine Menge $Oper$ von Operationen über S, wobei jede eine Stelligkeit besitzt; diese Operationen sind charakteristisch für die Struktur und a priori definiert. Die letzte Komponente stellt eine Menge von Relationen auf der Trägermenge dar; jede Relation aus $Rela$ hat eine gewisse Stelligkeit. Ausdrücklich sei bemerkt, daß die Fälle $Kons = \emptyset, Rela = \emptyset, Oper = \emptyset$ eingeschlossen sind. Sind die Mengen endlich, werden ihre Elemente (eventuell mit den Stelligkeiten) im Tupel explizit aufgeführt.
Wir kennen Beispiele für algebraische Strukturen:

- die Boolesche Algebra $(\mathcal{P}(M); \cap, \cup, ^-)$ der Potenzmenge einer Menge M mit den Operationen Durchschnitt, Vereinigung und Komplement,
- die Boolesche Algebra $(\{0,1\}; \wedge, \vee, \neg)$ bei Schaltkreisen.

Meist ist bei algebraischen Strukturen mindestens eine Operation gegeben. Es gibt jedoch auch sehr wichtige Strukturen ohne Operation: Die Graphen. Ein Graph besteht aus einer endlichen Trägermenge S, deren Elemente man Knoten nennt und endlich vielen symmetrischen oder asymmetrischen binären Relationen auf S. Wir werden Graphen im Kap. 3 genauer studieren.
In vielen Anwendungen ist die Trägermenge durch das Alphabet und die Operationen definiert: Man definiert die Objekte, die beim Ausführen der Operationen als Resultat auftreten dürfen; die Resultate dürfen wieder Operanden sein usw. Die Gesamtheit aller dieser auftretenden Objekte bildet dann die Trägermenge der Struktur. Eine solche Struktur nennt man **frei**. So definiert man z. B. eine Sprache durch ein Alphabet $\{x_1, \ldots, x_l\}$ und die Operation des Aneinanderreihens. Die Trägermenge ist dann die Menge aller durch Aneinanderreihen aus den Buchstaben $x_1, \ldots, x_l$ gebildeten Objekte (Wörter). Bei einer freien Struktur darf die Trägermenge weggelassen werden; jedoch muß ein Alphabet angegeben sein.
Durch Erweiterung auf mehrere Trägermengen und Einführung beschränkt ausführbarer Operationen ordnet sich hier praktisch jedes, innerhalb der Informatik betrachtete formale System ein. Um den einführenden Charakter unserer Darlegungen zu betonen, beschränken wir uns auf zweistellige Relationen, Operationen und lassen den Stelligkeitsindex weg. Die folgenden Betrachtungen gelten sinngemäß auch im Falle von Operationen und Relationen mit beliebigen (endlichen) Stelligkeiten. In den Anwendungen sind die Operations- bzw. Relationsmenge durchaus nicht endlich; bei einem der wichtigsten Beispiele, den Vektorräumen, liegt eine unendliche Relationenmenge vor, wie wir noch sehen werden. Die folgenden Begriffe im Zusammenhang mit einer allgemeinen Algebra sind so fundamental, daß sie die gesamte heutige Mathematik durchziehen.
Eine Struktur

$$\mathcal{S}' = (S'; Kons'; Oper'; Rela')$$

heißt **Substruktur** (**Unterstruktur**, **Teilstruktur**) einer Struktur

$$\mathcal{S} = (S; Kons; Oper; Rela),$$

wenn $S' \subseteq S, Kons' \subseteq Kons, Oper' \subseteq Oper, Rela' \subseteq Rela$ gilt. Bei einer Substruktur muß also jede Relation über der Trägermenge S' die Einschränkung einer entsprechenden Relation über der Trägermenge S sein. Die Trägermenge muß abgeschlossen bezüglich der auf S definierten Operationen sein. Z. B. ist $(\mathbb{N}_0; 0; +; <)$ eine Substruktur von $(\mathbb{Z}; 0; +; <)$, wobei in der ersten Struktur die $<$-Relation über den natürlichen Zahlen und mit '+' die Addition von natürlichen Zahlen gemeint sind; entsprechend bei der zweiten Struktur im Bereich der ganzen Zahlen.
Den Abbildungsbegriff übertragen wir sinngemäß auf Strukturen. Gegeben seien zwei Strukturen

$$\mathcal{S} = (S; Kons; Oper; Rela), \quad \mathcal{S}' = (S'; Kons'; Oper'; Rela').$$

Ein Abbildung $f = (f_1, f_2, f_3)$ mit

$$f_1 : S \longmapsto S', \quad f_2 : Oper \longmapsto Oper', \quad f_3 : Rela \longmapsto Rela'$$

heißt **Strukturabbildung**, wenn die beiden Abbildungen f_2, f_3 bijektiv sind. Eine Strukturabbildung $f = (f_1, f_2, f_3)$ heißt surjektiv (injektiv, bijektiv), wenn f_1 eine surjektive (injektive, bijektive) Abbildung ist. Wir lassen meist den Index für die einzelnen Komponenten einer Strukturabbildung f weg, da jeweils aus dem Argument sofort ersichtlich ist, um welche Komponente es sich handelt. Dies ist eine auch in der Informatik übliche Schreibweise; sie tritt z. B. bei generischen Funktionen und Operationen auf: Erst zur Übersetzungszeit wird mittels der aktuellen Parameter bzw. Operanden entschieden, welche Funktion bzw. Operation auszuführen ist.
Der nächste Begriff ist die **Isomorphie** von Strukturen. Eine Struktur $\mathcal{S}$ heißt **isomorph** zu einer Struktur $\mathcal{S}'$ wenn eine bijektive Strukturabbildung f existiert mit folgenden Eigenschaften

Relationstreue bei Isomorphie:
$f(x)f(R)f(y) \Longleftrightarrow xRy \quad \forall x, y \in S,\ \forall R \in Rela,$
Operationstreue:
$f(x * y) = f(x)f(*)f(y) \quad \forall x, y \in S,\ \forall * \in Oper.$

Die bijektive Strukturabbildung f heißt **Isomorphismus** von der Struktur $\mathcal{S}$ auf die Struktur $\mathcal{S}'$.
Bei der Isomorphie von Strukturen übertragen sich also die Grundelemente, die Relationen und Operationen; isomorphe Strukturen sind mit den verwendeten Methoden nicht zu unterscheiden. Man überzeugt sich, daß die Isomorphie eine Äquivalenzrelation ist. Mit einem Isomorphismus f von einer Struktur $\mathcal{S}$ auf eine Struktur $\mathcal{S}'$ wird jede Eigenschaft der Elemente der Trägermenge S von $\mathcal{S}$ bezüglich der Grundelemente, -relationen und -operationen in eine analoge Eigenschaft der Elemente der Trägermenge S' von $\mathcal{S}'$ übersetzt; mit der inversen Abbildung f^{-1} geschieht die Übersetzung in umgekehrter Richtung; f^{-1} ist ein Isomorphismus von $\mathcal{S}'$ auf $\mathcal{S}$. In der Algebra werden Strukturen nur bis auf Isomorphie untersucht.
Eine Abschwächung der Isomorphie ist die Homomorphie. Eine Struktur $\mathcal{S}'$ (wie oben) heißt **homomorph** zur Struktur $\mathcal{S}$, wenn die beim Isomorphismus postulierte Strukturabbildung bezüglich der Mengen S, S' lediglich surjektiv, die Abbildung relationstreu und operationstreu ist. Diese Strukturabbildung nennt man dann **Homomorphismus** von $\mathcal{S}$ auf $\mathcal{S}'$. Die Homomorphie von Strukturen ist reflexiv und transitiv, aber im allgemeinen nicht symmetrisch. Bei einem Homomorphismus dürfen mehrere Elemente das gleiche Bild haben. Aus diesem Grunde muß die Relationstreue neu gefaßt werden. Dazu übertragen wir zunächst in natürlicher Weise die Relationen auf der Struktur $\mathcal{S}$ auf Relationen zwischen Mengen. Für beliebige Mengen $X, Y \subseteq S$ und eine beliebige (binäre) Relation R setzen wir fest, daß $X\ R\ Y$ gilt, falls es zu jedem $x \in X$ ein $y \in Y$ mit $(x, y) \in R$ und zu jedem $y \in Y$ ein $x \in X$ mit $(x, y) \in R$ gibt. Damit lautet die Relationstreue eines Homomorphismus (bei binären Relationen):

Relationstreue bei Homomorphie:
$f(x)f(R)f(y) \Longleftrightarrow f^{-1}f(x)Rf^{-1}f(y) \quad \forall x, y \in S,\ \forall R \in Rela,$

Mit dem Homomorphismus-Begriff versucht die Mathematik, die folgenden bekannten Sachverhalte zu modellieren: Bei der Übertragung von Informationen können keine Informationen gewonnen werden; es gehen höchstens welche verloren. Man kann nur solche Informationen aus einer Struktur separieren, die auch in ihr enthalten sind. So produziert ein Rechner niemals Informationen; er bereitet lediglich die eingegebene Information so auf, daß der Mensch bzw. vom Menschen geschaffene Geräte mit ihr als Eingabe effektiv umgehen können. Bei jeder Informationsverarbeitung oder -umwandlung kann höchstens Informationsverlust eintreten. Ganz natürlich ergibt sich hier die Frage, was eigentlich Information ist. Man spricht in der Informatik von Informationsverarbeitung und gar von informationsverarbeiteten Maschinen. In Wahrheit gibt es aber keine solchen Geräte: Maschinen können nur Träger von Informationen verarbeiten. So kann man z. B. Texte von einer Sprache in eine andere übersetzen mit dem Ziel, daß sich der Informationsgehalt nicht ändert. Dies bedeutet: Beide Texte sollen die gleiche Information beinhalten. Aus Bitfolgen werden durch einen Rechner neue erzeugt. Die menschliche Vorstellung, daß es sich dabei um Informationsverarbeitung handelt, bedeutet jedoch nur, daß die Bitfolgen nach gewissen Regeln aufgebaut sind und ihnen damit ein „Sinn“ gegeben wird. Auch die Transformation in neue Bitfolgen erfolgt nach vorgegebenen Regeln (Algorithmen) so, daß das Ergebnis einer menschlichen Sinngebung zugängig ist. Mathematisch kann man daher sagen: Information ist eine binäre Relation zwischen Objekten, bei der mindestens ein Objekt ein Lebewesen ist. Sollten alle Lebewesen einer Art aussterben, so gehen auch alle ausschließlich mit ihnen verbundenen Informationen verloren. In einer natürlichen Sprache geschriebene Texte sind nur dann Träger von Information, wenn es Menschen gibt.
Wir wollen nun den grundlegenden Homomorphiesatz formulieren und beweisen. Dazu seien

$$\mathcal{S} = (S; Kons; Oper; Rela), \quad \mathcal{S}' = (S'; Kons'; Oper'; Rela')$$

Strukturen und f ein Homomorphismus von $\mathcal{S}$ auf $\mathcal{S}'$. Zusätzlich verwenden wir die durch f auf S induzierte Äquivalenzrelation F und das Restsystem S/F; dabei sind die Äquivalenzklassen $[x]_F$ gerade das Urbild $f^{-1}f(x)$ der Strukturabbildung f. Der Homomorphismus f induziert in natürlicher Weise eine Struktur

$$\mathcal{S}/F = (S/F; Kons/F; \overline{Oper}; \overline{Rela}).$$

Zu jeder Relation $R \in Rela$ definiert man $\overline{R} \in \overline{Rela}$ wie folgt:

$$[x]_F \overline{R} [y]_F \Longleftrightarrow f^{-1}f(x) R f^{-1}f(y) \quad \forall [x]_F, [y]_F \in S/F.$$

Für diese Definition muß man zeigen, daß sie unabhängig von der speziellen Auswahl der Repräsentanten aus den Klassen $[x]_F, [y]_F$ ist; d. h., daß wir die gleichen Relationen definieren, wenn wir beliebig andere Elemente aus den Klassen wählen. Es seien also $u \in [x]_F, v \in [y]_F$ beliebig. Wegen der Äquivalenz von u und x gilt $f(u) = f(x)$, entsprechend gilt $f(v) = f(y)$. Aus der Relationstreue von f erhalten wir für $R \in Rela$ aus xRy, daß auch $f(x)f(R)f(y)$ gilt und damit $f(u)f(R)f(v)$, woraus $f^{-1}f(u)Rf^{-1}f(v)$ folgt, was aber gleichbedeutend mit $f^{-1}f(x)Rf^{-1}f(y)$ ist. Damit ist die Definition der Relation $\overline{R}$ repräsentantenunabhängig. Analog erklären wir auf X/F die Operation $\overline{*} \in \overline{Oper}$:

$$[x]_F \overline{*} [y]_F = [x * y]_F.$$

Wie eben zeigt man, daß diese Definition repräsentantenunabhängig ist. Schließlich setzen wir noch die atomaren Elemente für $y \in Kons$:

$$\overline{y} = \{\, x \mid x \in X \text{ und } xFy \,\}.$$

Offenbar gilt $\overline{y} = [y]_F$, $\forall y \in Kons$.
Die Struktur $\mathcal{S}/F$ nennt man die durch den Homomorphismus f erzeugte **Faktorstruktur** (**Restklassenstruktur**, **Reststruktur**). Jede Äquivalenzrelation auf der Trägermenge einer Struktur $\mathcal{S}$ induziert damit eine Strukturabbildung $\hat{f}$ von $\mathcal{S}$ auf die Reststruktur $\mathcal{S}/F$. Die Strukturabbildung $\hat{f}$ ist ein Homomorphismus von $\mathcal{S}$ auf die Reststruktur $\mathcal{S}/F$, was unmittelbar aus der Definition der Operationen und Relationen in $\mathcal{S}/F$ folgt. Zwischen $\mathcal{S}/F$ und $\mathcal{S}'$ gibt es eine natürliche bijektive Strukturabbildung $\tilde{f}$ gemäß:

$$\tilde{f}([x]_F) = f(x), \quad \tilde{f}(\overline{R}) = f(R), \ \tilde{f}(\overline{*}) = f(*).$$

Die von f induzierte natürliche Strukturabbildung $\tilde{f}$ ordnet jeder Restklasse das allen seinen Elementen gemeinsame Bild in S' zu; jeder Relation $\overline{R}$ wird das Bild jener Relation R zugeordnet, mit deren Hilfe sie definiert worden ist; analog bei den Operationen. Wir zeigen, daß $\tilde{f}$ ein Isomorphismus von $\mathcal{S}/F$ auf $\mathcal{S}'$ ist. Da die Abbildung $\tilde{f}$ ohnehin bijektiv ist, genügt es zu zeigen, daß $\tilde{f}$ relations- und operationstreu ist. Für den Nachweis der Relationstreue von $\tilde{f}$ nutzen wir die Definition von $\overline{R}, \overline{R} \in \overline{Rela}$, die Relationstreue von f, $\tilde{f}([x]_F) = f(x)$ und erhalten:

$$[x]_F \overline{R} [y]_F \Longleftrightarrow f^{-1}f(x) R f^{-1}f(y) \Longleftrightarrow f(x)f(R)f(y) \Longleftrightarrow \tilde{f}([x]_F)\tilde{f}(\overline{R})\tilde{f}([y]_F).$$

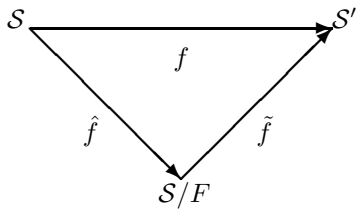
Für den Nachweis, daß $\tilde{f}$ operationstreu ist, verwenden wir die Operationstreue von f und erhalten für $* \in Oper$:

$$\begin{aligned} \tilde{f}([x]_F)\tilde{f}(\overline{*})\tilde{f}([y]_F) &= f(x)f(*)f(y) \\ &= f(x * y) \\ &= \tilde{f}([x * y]_F) \\ &= \tilde{f}([x]_F \overline{*} [y]_F), \end{aligned}$$

womit gezeigt ist, daß die Abbildung $\tilde{f}$ operationstreu ist. Sie erfüllt zusammen alle Eigenschaften eines Isomorphismus. Nun können wir unsere Überlegungen zu dem folgenden Homomorphiesatz zusammenfassen.

Satz 10 (Homomorphiesatz). *Es seien f ein Homomorphismus, der die Struktur $\mathcal{S}$ mit der Trägermenge S auf die Struktur $\mathcal{S}'$ mit der Trägermenge S' abbildet, F die durch f auf S induzierte Äquivalenzrelation, $\hat{f}$ die durch F induzierte Strukturabbildung von $\mathcal{S}$ auf die Reststruktur $\mathcal{S}/F$ und $\tilde{f}$ die natürliche bijektive Strukturabbildung von $\mathcal{S}/F$ auf $\mathcal{S}'$. Dann ist $\hat{f}$ ein Homomorphismus von $\mathcal{S}$ auf die von f erzeugte Reststruktur $\mathcal{S}/F$ und $\tilde{f}$ ein Isomorphismus von $\mathcal{S}/F$ auf $\mathcal{S}'$.*

Nach dem Homomorphiesatz kann jeder Homomorphismus f von $\mathcal{S}$ auf $\mathcal{S}'$ als Verkettung $\tilde{f} \circ \hat{f}$ des durch f induzierten Homomorphismus $\hat{f}$ von $\mathcal{S}$ auf die Reststruktur $\mathcal{S}/F$ und eines Isomorphismus $\tilde{f}$ von $\mathcal{S}/F$ auf $\mathcal{S}'$ dargestellt werden. Diesen Tatbestand zeigt das folgende Diagramm:



Der Inhalt des Homomorphiesatzes läßt sich bei Vorgabe einer geeigneten Äquivalenzrelation F (anstelle eines Homomorphismus f) umkehren, da zwischen beiden als Abbildungen eine umkehrbar eindeutige Beziehung besteht. Wir haben aus den Äquivalenzrelationen nur jene auszusondern, die den Homomorphismen umkehrbar eindeutig zugeordnet sind. Wie oben sei $\mathcal{S}$ eine Struktur mit der Trägermenge S. Eine Äquivalenzrelation F auf S heißt **Kongruenzrelation** auf $\mathcal{S}$, wenn sie mit den auf $\mathcal{S}$ definierten Relationen und Operationen in folgendem Sinne verträglich ist:

relationsverträglich:
$xRu \Longleftrightarrow yRv \quad \forall R \in Rela,\ \forall (x,y),(u,v) \in F,$
operationsverträglich:
$(x * u, y * v) \in F \quad \forall * \in Oper,\ \forall (x,y),(u,v) \in F.$

Die Relationsverträglichkeit besagt: Die ersten Komponenten von zwei Paaren aus einer beliebigen Äquivalenzklasse von F stehen genau dann in der Relation R, wenn auch die beiden zweiten Komponenten der betreffenden Paare in der Relation R stehen. Die Operationsverträglichkeit verlangt: Jede Äquivalenzklasse ist abgeschlossen hinsichtlich der komponentenweisen Ausführung aller Operationen aus der Operationenmenge.
In methodisch gleicher Weise wie oben zeigt man nun

Satz 11. *Es sei F eine Kongruenzrelation auf einer Struktur $\mathcal{S}$ mit der Trägermenge S und $\hat{f}$ die von F induzierte Strukturabbildung von $\mathcal{S}$ auf die Reststruktur $\mathcal{S}/F$. Dann ist $\hat{f}$ ein Homomorphismus von $\mathcal{S}$ auf $\mathcal{S}/F$ und jede Verkettung dieses Homomorphismus mit einem Isomorphismus von $\mathcal{S}/F$ auf eine beliebige Struktur $\mathcal{S}'$ liefert einen Homomorphismus von $\mathcal{S}$ auf $\mathcal{S}'$.*

Durch die Kongruenzrelationen auf einer Struktur $\mathcal{S}$ sind daher alle möglichen Homomorphismen von dieser Struktur auf irgendeine andere Struktur bis auf Isomorphie charakterisiert. Und umgekehrt: Kennt man alle Homomorphismen von $\mathcal{S}$ auf eine beliebige andere Struktur, so kennt man auch alle Kongruenzrelationen auf $\mathcal{S}$. In diesem Sinne kann man algebraisch die Homomorphismen und die Kongruenzrelationen als zwei verschiedene Darstellungen des gleichen Sachverhaltes bezüglich Strukturen auffassen.
In den folgenden Abschnitten werden wir etwas über klassische algebraische Strukturen erfahren. Im Zusammenhang mit algebraischen Strukturen werden wir weiterhin von Abbildungen sprechen, obwohl stets Strukturabbildungen gemeint sind. Eine Struktur $\mathcal{S} = (S; Kons; Oper; Rela)$ wird meist durch $S(Kons, Oper, Rela)$ bezeichnet, wobei nur die nichtleeren Mengen und bei endlichen Mengen die Elemente aufgeführt sind. Spezielle Strukturen erhalten auch spezielle Bezeichnungen. So bezeichnen wir mit $\mathbb{N}$ die Struktur der natürlichen Zahlen mit den gebräuchlichen Operationen, entsprechend gelten die Bezeichnungen $\mathbb{N}_0, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$ für die natürlichen Zahlen mit Null, die ganzen Zahlen, die rationalen Zahlen und die reellen Zahlen. Falls eine Operation angegeben sein sollte, so betrachten wir nur die dadurch erzeugte Struktur. So bedeutet z. B. $\mathbb{N}_0(+)$ die algebraische Struktur mit den natürlichen Zahlen als Trägermenge und der Addition als einzige Operation.

1.3.2. Halbgruppen und Gruppen

Zur Motivation betrachten wir einen Zug auf einer Modelleisenbahn. Ein Fahrbefehl repräsentiert sich in der Angabe, um wieviele Haltestellen der Zug vorfahren soll. Falls n Haltestellen betrachtet werden, sind also die Zahlen $1, \ldots, n-1$ mögliche Fahrbefehle. Sollen mehrere Fahrbefehle nacheinander ausgeführt werden, so wird dies durch eine Folge von Zahlen $a_1, \ldots, a_l$ symbolisiert. Als natürliche Verknüpfung von Befehlsfolgen tritt hier die Aneinanderreihung auf. Ein anderes Beispiel ist die Menge aller Wörter über einem Alphabet $x_1, \ldots, x_n$. Hier wird man zunächst jeden Buchstaben x_i des Alphabets als Wort bezeichnen und als natürliche Verknüpfung die Aneinanderreihung von Wörtern ansehen: Sind w_1, w_2 Wörter, so ist auch $w = w_1 \circ w_2$ ein Wort, wobei das Zeichen $\circ$ die Aneinanderreihung symbolisieren soll. Alle Objekte, die man nicht auf diese Weise suksessiv aus dem Alphabet gewinnen kann, wird man nicht als Wörter bezeichnen. In beiden Fällen ist das Endergebnis mehrerer Aneinanderreihungen unabhängig davon, in welcher Reihenfolge sie ausgeführt wurden. Sind a, b, c Befehlsfolgen für den Zug, so liefern $(ab)c$ und $a(bc)$ das gleiche Resultat; ebenso bei den Wörtern: Sind u, v, w Wörter, so repräsentieren $(u \circ v) \circ w$ und $u \circ (v \circ w)$ das gleiche Wort. Diese Beispiele führen uns zum Begriff der Halbgruppe: Eine Menge H zusammen mit einer assoziativen, binären Operation $*$ heißt **Halbgruppe**. Für die auf einer Halbgruppe $H(*)$ definierte Operation $*$ müssen also zwei Grundbedingungen erfüllt sein:

1. Die Trägermenge H ist abgeschlossen bezüglich der Operation $*$; die Ausführung der Operation mit Elementen aus H liefert stets ein Element aus H: $\forall x, y \in H : x * y \in H$.
2. Die Operation $*$ ist assoziativ:

$$(x * y) * z = x * (y * z) \ \ \forall x, y, z \in H.$$

Hier sind einige Beispiele für Halbgruppen:

1. Die Menge aller obigen Befehlsfolgen mit der Operation des Aneinanderreihens. Ebenso die Menge aller Wörter über einem Alphabet mit der Operation des Aneinanderreihens.

2. $\mathbb{N}(+)$.
3. $\mathbb{N}(\cdot), \mathbb{Z} \setminus \{0\}(\cdot), \mathbb{Q} \setminus \{0\}(\cdot), \mathbb{R} \setminus \{0\}(\cdot)$.
4. $\mathbb{Z}(+), \mathbb{Q}(+), \mathbb{R}(+)$.
5. Die Menge aller Abbildungen einer Menge X in sich mit der Operation der Verkettung (Nacheinanderausführung). Diese Struktur wollen wir mit $\mathcal{F}(X)$ bezeichnen.
6. Die Menge aller bijektiven Abbildungen einer endlichen Menge X auf sich mit der Operation der Verkettung. Diese Struktur wird mit $\mathcal{S}(X)$ bezeichnet.

Diese Beispiele haben Gemeinsamkeiten und Unterschiede. In den Beispielen 1, 5 und 6 ist die betreffende Operation nicht kommutativ, bei den übrigen ist sie es. In dem Beispiel 2 ist das Resultat einer Operation stets von den Operanden verschieden. Bei einigen Beispielen gibt es Elemente, die bei der Verknüpfung mit einem anderen dieses ungeändert lassen, z. B. Addition mit 0 (Beispiele in 4.), Multiplikation mit 1 (Beispiele in 3.), Verkettung mit der identischen Abbildung. Schließlich gibt es in den Beispielen 4 und 6 stets Elemente, die eine Verknüpfung rückgängig machen können. Um diese Unterschiede zu modellieren, müssen wir weitere Begriffe bilden.
Wenn in einer Halbgruppe $H(*)$ ein Element $e \in H$ existiert mit

$$e * x = x * e = x \ \ \forall x \in H,$$

dann heißt $H(*)$ **Monoid** und e nennt man **neutrales Element** bzw. **Einselement**. So ist z. B. $\mathbb{N}_0(+)$ ein Monoid. Wenn in einem Monoid $G(*)$ zu jedem $x \in G$ ein $\overline{x} \in G$ existiert mit der Eigenschaft

$$x * \overline{x} = \overline{x} * x = e,$$

so heißt die Struktur $G(*)$ **Gruppe** und das Element $\overline{x}$ nennt man **invers** zu x. Der Leser möge sich überlegen, daß in einer Struktur bezüglich einer Operation höchstens ein Einselement und zu jedem Element höchstens ein inverses existieren kann. Die Beispiele 4 und 6 stellen offenbar Gruppen dar; in den Beipielen 3 gibt es Gruppen.
Ein Element $o \in H$, wobei $H(*)$ eine Halbgruppe sein möge, nennt man **Nullelement**, falls

$$x * o = o * x = o \ \forall x \in H$$

gilt. Hier sind Null- und Einselement wohl zu unterscheiden. Oft ist das Nullelement für eine Operation gerade das neutrale Element für eine andere.
Ein Element $a, a \neq o$ aus einer Halbgruppe $H(*)$ mit Nullelement o heißt **Nullteiler**, falls ein $b \in H, b \neq o$ existiert mit $a * b = o$ oder $b * a = o$. In einem solchen Falle ist auch b ein Nullteiler. Man überlege sich, daß es in einem Monoid zu einem Nullteiler kein inverses Element geben kann; folglich gibt es in einer Gruppe keinen Nullteiler.
Eine Halbgruppe $H(*)$ heißt **abelsch** (nach dem norwegischen Mathematiker N. H. Abel), falls die Operation kommutativ ist, also $x * y = y * x$ für alle $x, y \in H$ gilt. Oft schreiben wir für die Operation '$*$' das Pluszeichen '+', falls die betreffende Operation kommutativ ist. Eine additiv geschriebene abelsche Gruppe nennt man auch **Modul**. Als kleine Übung möge man beweisen, daß in einer Halbgruppe stets nur höchstens ein Nullelement, Einselement existieren. Ebenso gibt es auch nur höchstens ein inverses Element zu einem gegebenen. In den obigen Beispielen sind 4. und 6. Gruppen, 3. und 5. sind Monoide und 2. und 4. abelsche Halbgruppen. Zur Notation sei noch angemerkt: Die Operation '$*$' wird oft als Multiplikation mit dem Malzeichen geschrieben. Entsprechend werden bei multiplikativ geschriebener Operation das Einselement mit 1 und das inverse mit x^{-1} bezeichnet. Bei additiver Schreibweise sind das Einselement (= neutrales Element) durch 0 und das inverse durch $-x$ symbolisiert.
Die Verknüpfung von endlich vielen Elementen ist in Halbgruppen wegen der Assoziativität unabhängig von der Reihenfolge ihrer Ausführung; daher kann man eventuelle Klammern weglassen und einfach

$$x_1 \cdot x_2 \cdot \ldots \cdot x_n = \prod_{i=1}^{n} x_i$$

schreiben. Existiert zu jedem $x \in H$ ein inverses Element, so ist

$$(x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{-1} = x_n^{-1} \cdot x_{n-1}^{-1} \cdot \ldots \cdot x_1^{-1}$$

und bei additiver Schreibweise

$$x_1 + x_2 + \ldots + x_n = \sum_{k=1}^{n} x_k,$$

$$-(x_1 + x_2 + \ldots + x_n) = -x_n + (-x_{n-1}) + \ldots + (-x_1).$$

Im Falle $x_1 = x_2 = \ldots = x_n = x$ schreibt man anstelle von $x_1 \cdot \ldots \cdot x_n$ einfach x^n; existiert ein neutrales Element e, so setzt man $x^0 = e$ für alle $x \in H$. Existiert zu x ein inverses Element x^{-1}, so ist $(x^{-1})^n = x^{-n}$ und damit

$$x^m \cdot x^n = x^{n+m}, \qquad (x^m)^n = x^{n \cdot m}.$$

Bei additiver Schreibweise setzt man im Falle $x_1 = x_2 = \ldots = x_n$:

$$x_1 + x_2 + \ldots + x_n = n \cdot x;$$

existiert ein neutrales Element $e \in H$, so setzt man $0 \cdot x = e$. Für inverse Elemente erhält man

$$-(n \cdot x) = n \cdot (-x) = -n \cdot x.$$

Die Ordnung $|H(\cdot)|$ einer Halbgruppe $H(\cdot)$ ist die Mächtigkeit $|H|$ von H. Die Strukturen in den Beispielen 1.-4. haben unendliche Ordnungen, während 5. und 6. bei endlicher Menge X auch endliche Ordnungen haben. Es sei etwa $|X| = n$. Dann liefert die folgende Überlegung die Ordnung der Struktur $\mathcal{F}(X)$ aller Abbildungen von X in sich: Jede Abbildung aus $\mathcal{F}(X)$ ist durch Angabe der Bilder aller Elemente aus X eindeutig festgelegt; da für jedes der n Elemente auch n Möglichkeiten für das Bildelement existieren, gibt es also n^n Abbildungen von X in X: $|\mathcal{F}(X)| = n^n$. Bei den Abbildungen aus $\mathcal{S}(X)$ darf jedes Element genau einmal als Bild vorkommen. Beim ersten Element hat man n Möglichkeiten zur Auswahl, beim zweiten nur noch $n-1$, beim dritten noch $n-2$ usw. also gilt

$$|\mathcal{S}(X)| = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1 = n!.$$

Wir wollen nun eine spezielle endliche Gruppe etwas genauer betrachten.
Es sei M eine endliche Menge; ihre Elemente mögen sich in einem symbolischen Korb befinden. Wir nehmen nacheinander die Elemente aus dem Korb und versehen jedes mit einer fortlaufenden Nummer, um sie danach zurückzulegen. Wenn dadurch n Nummern vergeben wurden, können wir uns die Menge M indiziert vorstellen:

$$M = \{\, m_1, m_2, \ldots, m_n \,\}.$$

Nun schütteln wir den Korb und entnehmen die Elemente erneut nacheinander; jetzt wird sich die Reihenfolge geändert haben, und wir erhalten so eine neue Anordnung der Elemente:

$$M = \{\, m_{i_1}, m_{i_2}, \ldots, m_{i_n} \,\}.$$

Jede Reihenfolge der Elemente aus M wird durch eine Anordnung (Permutation) der Zahlen $1, 2, \ldots, n$ repräsentiert. Eine solche Anordnung beschreibt offenbar eine bijektive Abbildung von M auf sich und umgekehrt. Das Studium der bijektiven Abbildungen auf einer endlichen Menge M mit n Elementen ist also gleichbedeutend mit der Untersuchung aller Anordnungen der Zahlen $1, 2, \ldots, n$. Die Struktur $\mathbb{S}_n$ aller Anordnungen von n Elementen mit der Operation der Nacheinanderausführung heißt **symmetrische Gruppe** auf n Elemente. Diese soll nun genauer untersucht werden. Es sei

$$\pi \in \mathbb{S}_n : j \longrightarrow \pi(j), \qquad j = 1, \ldots, n.$$

Wir schreiben die Abbildung π in folgender Form:

$$\begin{pmatrix} 1 & 2 & 3 & \ldots & n \\ \pi(1) & \pi(2) & \pi(3) & \ldots & \pi(n) \end{pmatrix}.$$

Als Operation haben wir die Nacheinanderausführung (Verkettung):

$$\pi_3(j) = \pi_1(\pi_2(j)), j = 1, \ldots, n,$$

also z. B.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 1 & 5 & 2 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 3 & 4 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 1 & 5 & 2 \end{pmatrix}.$$

Zunächst stellen wir fest, daß jede bijektive Abbildung $\pi \in \mathbb{S}_n$ als Verkettung von elementfremden Zyklen dargestellt werden kann. Unter **Zyklus** versteht man dabei eine Anordnung der Form

$$\begin{pmatrix} t_1 & t_2 & t_3 & \ldots & t_{m-1} & t_m \\ t_2 & t_3 & t_4 & \ldots & t_m & t_1 \end{pmatrix},$$

kurz als $(t_1, t_2, \ldots, t_m)$ geschrieben. Die elementfremden Zyklen einer Anordnung π erhält man wie folgt. Man setze $t_1 = 1$; dazu wird $t_2 = \pi(t_1)$ bestimmt, danach $t_3 = \pi(t_2)$ usw.; da es nur n Elemente gibt, muß sich die Folge der t_i schließen, d. h. es gibt ein $m \leqq n$ und $\pi(t_m) = t_j$ mit $1 \leqq j < m$. Wäre nun $j > 1$, so hätte man t_j als Bild von t_m und t_{j-1}, was aber unmöglich ist; folglich gilt $j = 1$, d. h. $\pi(t_m) = t_1 = 1$. Das Verfahren setzt man mit einer noch nicht verwendeten Zahl, etwa der kleinsten, fort und erhält den nächsten, zum ersten elementfremden Zyklus usw. bis alle Elemente in jeweils einem Zyklus erfaßt sind.
Beispiel:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 6 & 3 & 7 & 8 & 1 & 4 & 5 \end{pmatrix} = (1 \quad 2 \quad 6) \cdot (4 \quad 7) \cdot (5 \quad 8) \ .$$

Wir wollen die dargestellte Methode zur Bestimmung aller elementfremden Zyklen einer Anordnung in algorithmischer Form aufbereiten. Vorgegeben sei also eine Anordnung π als Feld π mit n Elementen derart, daß π_i das Bild von i ist. Das Ergebnis wird auf einem Feld ϱ mit n Elementen abgelegt, wobei wir das letzte Element in einem Zyklus negativ eintragen. Der Algorithmus ZYKLEN leistet das Verlangte.

```
//=========================================================================
//     Bestimmung aller Zyklen einer Anordnung
//=========================================================================
void zyklen(int n,      // Länge der Anordnung
            int *pi,    // Feld, das die Anordnung enthält
            int **rho)  // Ausgabefeld (oder NULL)
{  int i, j, k, l, m, *r=*rho;
   if(!r) r=*rho=new int[n];
   for(j=0; j<n; r[j]=++j);
   j=0;
   while(j<n)
   { j++, k=r[j-1], m=k;
     while(k!=pi[m-1])
     { m=pi[m-1];
       for(i=j+1; i<n; i++)
         if(m==r[i-1]) j++, r[i-1]=r[j-1], r[j-1]=m, l=pi[j-1],
           pi[j-1]=pi[i-1], pi[i-1]=l,m=j;
     }
     r[j-1]=-r[j-1];
   }
}
```

Bei Eingabe des letzten Beispiels in diesen Algorithmus erhält man

$$\varrho = (\ 1 \quad 2 \quad -6 \quad 4 \quad -7 \quad -3 \quad 5 \quad -8\ ).$$

Die Anzahl der Elemente in einem Zyklus heißt **Länge** des Zyklus; einen Zyklus der Länge 2 nennt man **Transposition**. Bei Zyklen der Länge 1 bleibt das Element fest; in der Darstellung einer Anordnung durch elemenfremde Zyklen kann man diese auslassen. Bei der Verknüpfung von elementfremden Zyklen kommt es nicht auf die Reihenfolge der Verkettung an; im allgemeinen ist jedoch $\mathbb{S}_n (n > 2)$ nicht kommutativ, wie folgendes Beispiel zeigt:

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} = (1 \quad 2),$$

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} = (1 \quad 3).$$

Bei der Verkettung von Abbildungen in unserer hier gewählten Schreibweise ist zu beachten, daß die Operation von rechts nach links auszuführen ist.
In einer Anordnung spricht man von einer **Inversion**, wenn eine größere Zahl vor einer kleineren steht. Ist π eine Anordnung, so sei $f(\pi)$ die Anzahl der Inversionen von π. Unter dem **Signum** (Vorzeichen) einer Anordnung π versteht man die Größe

$$\operatorname{sgn}(\pi) = (-1)^{f(\pi)}.$$

So erhält man z. B. für

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix}, \qquad f(\pi) = 4, \qquad \operatorname{sgn}(\pi) = 1.$$

Für die Berechnung von $\operatorname{sgn}(\pi)$ beweisen wir die folgende Formel:

$$\operatorname{sgn}(\pi) = \prod_{\substack{(j,t) \\ j<t}} \frac{t - j}{\pi(t) - \pi(j)}.$$

Beweis. Die Abbildung π ist eineindeutig, also kommen alle Paare $(j,t), j < t$ in Zähler und Nenner jeweils genau einmal vor. Daher hat das Produkt den Betrag 1. Die Zählerfaktoren sind sämtlich positiv. Im Nenner ist genau dann ein Faktor negativ, wenn eine Inversion vorliegt. Also ist das Produkt bei einer geraden Anzahl von Inversionen gleich 1 und sonst gleich -1. □
Nach dieser Formel folgt im letzten Beispiel

$$\operatorname{sgn}\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} = \frac{2-1}{2-3}\cdot\frac{3-1}{4-3}\cdot\frac{3-2}{4-2}\cdot\frac{4-1}{1-3}\cdot\frac{4-2}{1-2}\cdot\frac{4-3}{1-4} = 1.$$

Auf Grund der Darstellung einer Anordnung π durch elementfremde Zyklen kann man leicht die zu π inverse Anordnung bestimmen, indem man zu jedem Zyklus den inversen ermittelt und alle miteinander verknüpft. Für einen Zyklus $z = (t_1 t_2 \cdots t_m)$ gilt $z^{-1} = (t_1 t_m t_{m-1} \cdots t_2)$, denn

$$\begin{pmatrix} t_1 & t_2 \cdots & t_m \\ t_2 & t_3 \cdots & t_1 \end{pmatrix}\begin{pmatrix} t_1 & t_2 & t_3 \cdots & t_m \\ t_m & t_1 & t_2 \cdots & t_{m-1} \end{pmatrix} = \begin{pmatrix} t_1 & t_2 \cdots & t_m \\ t_1 & t_2 \cdots & t_m \end{pmatrix}.$$

Insbesondere ist dadurch der Begriff 'symmetrische Gruppe' gerechtfertigt. Man überzeugt sich leicht von folgenden Rechenregeln:

- $\operatorname{sgn}(\varrho \cdot \pi) = \operatorname{sgn}(\varrho) \cdot \operatorname{sgn}(\pi)$
- Jeder m-elementige Zyklus kann als Verkettung von $m-1$ Transpositionen geschrieben werden:

$$(t_1 t_2 \cdots t_m) = (t_1 t_2) \cdot (t_2 t_3) \cdot \ldots \cdot (t_{m-1} t_m).$$

- Das Signum einer Transposition ist gleich -1.

Hieraus folgt eine neue Formel für das Signum. Hat die Anordnung π genau j elementfremde Zyklen mit den Längen $m_1, \ldots, m_j$, dann gilt

$$\operatorname{sgn}(\pi) = (-1)^{m_1-1+m_2-1+\ldots+m_j-1} = (-1)^{m_1+\ldots+m_j-j}.$$

Meist vereinfacht sich damit die Berechnung des Signums wesentlich:

$$\begin{aligned}\operatorname{sgn}\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 6 & 5 & 7 & 3 & 1 & 4 \end{pmatrix} &= \operatorname{sgn}\begin{pmatrix} ((1 & 2 & 6)\cdot(4 & 7)\cdot(3 & 5)) \end{pmatrix} \\ &= (-1)^{2+1+1} = 1.\end{aligned}$$

Durch das Signum lassen sich zwei Klassen von Anordnungen unterscheiden: Die geraden Anordnungen (bei ihnen gilt $\operatorname{sgn}(\pi) = 1$) und die ungeraden. Die geraden Anordnungen bilden eine Gruppe, die man **alternierende Gruppe** $\mathbb{A}_n$ auf n Elemente nennt. Allgemein nennt man eine beliebige Untergruppe einer symmetrischen Gruppe auch **Anordnungsgruppe** oder **Permutationsgruppe**.
Eine Substruktur $U(\cdot)$ einer Halbgruppe (Gruppe) $H(\cdot)$ heißt **Unterhalbgruppe** (**Untergruppe**). So ist die alternierende Gruppe auf n Elemente eine Untergruppe der symmetrischen Gruppe auf n Elemente. Oft betrachtet man auch Untergruppen von Halbgruppen. Wir erwähnen die folgenden Kriterien für Unterhalbgruppen bzw. Untergruppen:

1. Eine nichtleere Untermenge $U \subseteqq H$ ist genau dann Trägermenge einer Unterhalbgruppe von $H(\cdot)$, wenn sie abgeschlossen bezüglich der Operation ist.
2. Eine nichtleere Untermenge $U \subseteqq G$ ist genau dann Trägermenge einer Untergruppe der Gruppe $G(\cdot)$, wenn sie abgeschlossen ist bezüglich der Operation und bezüglich der Inversenbildung, d. h. wenn

$$u \cdot v^{-1} \in U \quad \forall u, v \in U$$

gilt.

Für endliche Untermengen gilt verschärfend

Satz 12. *Eine endliche, nichtleere Untermenge $U \subseteqq G$ ist genau dann Trägermenge einer Untergruppe der Gruppe $G(\cdot)$, wenn sie abgeschlossen bezüglich der in G erklärten Operation ist.*

Beweis. Dieses Kriterium ist offenbar bewiesen, wenn wir aus der Abgeschlossenheit bezüglich der in G erklärten Operation auf die Inversenbildung schließen können. Es sei also U abgeschlossen bezüglich der in G definierten Operation; mit $a \in U$ gilt dann sicher

$$a \cdot U = \{\, a \cdot u \mid u \in U \,\} \subseteqq U.$$

Angenommen, die Menge $a \cdot U$ ist echte Untermenge von U; dann müssen zwei Elemente $u, v \in U$ mit dem Element a das gleiche Produkt bilden:

$$a \cdot u = a \cdot v.$$

Indem wir diese Gleichung von links mit a^{-1} multiplizieren, erhalten wir

$$u = a^{-1} \cdot (a \cdot u) = a^{-1} \cdot (a \cdot v) = v,$$

also $a \cdot U = U$. Wegen $a \in U$ existiert ein $x \in U$ mit $a \cdot x = a$; dies liefert durch Linksmultiplikation mit a^{-1}:

$$e = a^{-1} \cdot a = a^{-1}a \cdot x = x \in U.$$

Somit muß auch das Einselement unter den Produkten aus $a \cdot U$ auftreten; also existiert ein $y \in U$ mit $a \cdot y = e$, d.h.

$$a^{-1} = a^{-1} \cdot e = a^{-1} \cdot a \cdot y = y \in U,$$

womit das Kriterium bewiesen ist. □

Aus diesen Kriterien schließen wir zunächst, daß der Durchschnitt von beliebig vielen Unterhalbgruppen wieder eine Unterhalbgruppe sein muß (falls er nicht leer ist). Entsprechendes gilt auch für Untergruppen. Wir merken an, daß man durch ein Gegenbeispiel zeigen kann: Für die Vereinigung gilt dies nicht.

Es sei nun eine beliebige Teilmenge $M \subseteqq H$ einer Halbgruppe $H(\cdot)$ gegeben. Wir bilden den Durchschnitt $\mathcal{U}(M)$ aller jener Unterhalbgruppen $U(\cdot)$ von $H(\cdot)$, die die Menge M als Teilmenge haben. Dieser Durchschnitt ist wieder eine Unterhalbgruppe und offenbar die kleinste von allen, in der die Menge M enthalten ist. Man nennt daher $\mathcal{U}(M)$ die von M erzeugte Unterhalbgruppe und bezeichnet sie einfach mit (M). Im Falle $H(\cdot) = (M)$ heißt die Menge M ein **Erzeugendensystem** der Struktur $H(\cdot)$. Analog lauten die Begriffsbildungen bei Gruppen. Im Falle einer Halbgruppe $H(\cdot)$ besteht (M) aus allen Elementen der Menge H, die sich als endliches Produkt von Elementen aus M darstellen lassen, d. h. aus allen Elementen der Form

$$u = m_1 \cdot m_2 \cdot \ldots \cdot m_l,\ m_i \in M,\ i = 1, \ldots, l,\ l \in \mathbb{N},$$

und im Falle einer Gruppe $G(\cdot)$ enthält (M) genau alle Produkte der Form

$$u = m_1^{i_1} \cdot m_2^{i_2} \cdot \ldots \cdot m_l^{i_l},\ i_j \in \{+1, -1\},\ m_j \in M,\ j = 1, \ldots, l,\ l \in \mathbb{N}.$$

Im Falle $M = \{m\}$ schreibt man einfach (m) anstelle von $(\{m\})$ und spricht von der durch m erzeugten Unterhalbgruppe (Untergruppe); diese nennt man **zyklische Unterhalbgruppe** bzw. **zyklische Untergruppe**. Bei Halbgruppen besteht eine zyklische Unterhalbgruppe (m) aus allen Potenzen m^l mit $l \in \mathbb{N}$ und bei Gruppen aus allen Potenzen m^l mit $l \in \mathbb{Z}$. Speziell gilt offenbar:

$\mathbb{Z}(+) = (1)$ als Gruppe,

$\mathbb{Z}(+) = (\{1, -1\})$ als Halbgruppe,

$\mathbb{N}(+) = (1)$ als Halbgruppe.

Die durch ein Gruppenelement g erzeugte zyklische Untergruppe (g) enthält alle voneinander verschiedenen Potenzen $g^k (k \in \mathbb{Z})$, und unter der **Ordnung** $o(g)$ eines Gruppenelementes g versteht man die Mächtigkeit der von g erzeugten Untergruppe: $o(g) = |(g)|$.

Satz 13. *Für die Ordnung $o(g)$ eines Gruppenelementes g gilt entweder $o(g) = \infty$ und $g^n \neq e$ für alle $n \in \mathbb{N}$ oder $o(g) = k$ und k ist die kleinste natürliche Zahl mit $g^k = e$; in diesem Falle ist $g^m = e$ für alle $m = l \cdot k$.*

Beweis. Aus $g^n = e$ folgt, daß die Menge $\{ g^1, g^2, \ldots, g^n \}$ Trägermenge einer Untergruppe ist; also kann $g^n = e$ nur für $n \geqq o(g)$ eintreten. Es sei $o(g) = l$. Dann gibt es zwei verschiedene natürliche Zahlen i, j mit $l \geqq i > j \geqq 0$ und $g^i = g^j$, also $g^{i-j} = e$, woraus $i - j \leqq l = o(g)$ folgt. Nach der ersten Überlegung muß $i - j \geqq l = o(g)$ sein, was zusammen $i - j = l$ und damit $i = l, j = 0$ ergibt. Natürlich ist $g^{l \cdot z} = e^z = e$ für alle $z \in \mathbb{Z}$. Hat m die Form $m = l \cdot z + i$ $(1 \leqq i \leqq l - 1)$, so folgt $g^m = g^i \neq e$. □

Mit jeder Untergruppe $U(\cdot)$ einer Gruppe $G(\cdot)$ verbinden sich zwei wichtige Äquivalenzrelationen. Zwei Elemente $x, y \in G$ heißen **linksäquivalent** bezüglich der Untergruppe $U(\cdot)$, wenn $x^{-1} \cdot y \in U$ gilt. Wir verifizieren die Eigenschaften einer Äquivalenzrelation. Wegen $x^{-1} \cdot x = e \in U$ ist die Relation reflexiv. Aus $x^{-1} \cdot y \in U$ folgt $y^{-1} \cdot x = (x^{-1} \cdot y)^{-1} \in U$ und umgekehrt; also ist die Relation symmetrisch. Aus $x^{-1} \cdot y \in U$ und $y^{-1} \cdot z \in U$ folgt

$$U \ni (x^{-1} \cdot y) \cdot (y^{-1} \cdot z) = x^{-1} \cdot (y \cdot y^{-1}) \cdot z = x^{-1} \cdot z,$$

also ist die Linksäquivalenz transitiv. Somit liefert die Linksäquivalenz bezüglich einer beliebig fixierten Untergruppe eine Zerlegung der Gruppe $G(\cdot)$. Die Restklassen der Zerlegung nennt man **Linksnebenklassen** der Untergruppe $U(\cdot)$:

$$[x] = \{\, z \mid z = x \cdot y,\ y \in U \,\} = x \cdot U.$$

Man erhält die Linksnebenklasse $[x]$ des Elementes x, indem man alle Elemente aus U von links mit x multipliziert. Alle diese Produkte sind verschieden. Ist nun die Untergruppe $U(\cdot)$ endlich, dann haben alle Linksnebenklassen gleichviel Elemente: $|x \cdot U| = |U|\ \forall x \in G$. Ist überdies auch noch die Gruppe $G(\cdot)$ endlich, so gibt es auch nur endlich viele Linksnebenklassen. Da außerdem jedes Element in genau einer Linksnebenklasse liegt, folgt daraus der Satz von **LAGRANGE**

Satz 14. *In jeder endlichen Gruppe ist die Ordnung jeder Untergruppe ein Teiler der Gruppenordnung.*

Insbesondere ist die Ordnung eines Gruppenelementes ein Teiler der Gruppenordnung und somit

$$g^{|G|} = e\ \forall g \in G.$$

Hat $G(\cdot)$ also Primzahlordnung, so kann $G(\cdot)$ keine nichttrivialen, d. h. vom Einselement verschiedenen, Untergruppen haben. Folglich erhalten wir aus dem Satz von Lagrange

Satz 15. *Jede Gruppe von Primzahlordnung ist zyklisch.*

Analog definiert man die **Rechtsnebenklassen** von $U(\cdot)$ und erhält sie als Mengen der Form

$$U \cdot x = \{\, y \cdot x \mid y \in U \,\}.$$

Jedes Gruppenelement von $G(\cdot)$ liegt in genau einer Rechtsnebenklasse; im allgemeinen sind aber $x \cdot U$ und $U \cdot x$ verschiedene Mengen; sie stimmen im kommutativen Fall überein. In jedem Falle gilt aber

$$x \cdot U = y \cdot U \iff x^{-1} \cdot y \in U \iff U \cdot x^{-1} = U \cdot y^{-1},$$

was uns sagt, daß es gleichviel Links- und Rechtsnebenklassen gibt. Diese Anzahl ist also eine zweite charakteristische Größe für jede Untergruppe $U(\cdot)$ aus einer Gruppe $G(\cdot)$; man nennt sie **Index** der Untergruppe $U(\cdot)$. Zusammen mit dem Satz von Lagrange können wir daher den folgenden Satz aussprechen.

Satz 16. *Die Ordnung einer endlichen Gruppe ist gleich dem Produkt aus Ordnung und Index einer beliebigen Untergruppe.*

Nun werden wir die Begriffe Homomorphie und Isomorphie auf algebraische Strukturen mit einer Operation anwenden. Dazu seien $H(\cdot)$ und $M(*)$ zwei Strukturen und φ ein Homomorphismus von $H(\cdot)$ auf $M(*)$, also eine Abbildung von H auf M mit der Eigenschaft

$$\varphi(x \cdot y) = \varphi(x) * \varphi(y) \quad \forall x, y \in H.$$

Ist die Abbildung überdies noch bijektiv, so ist sie ein Isomorphismus. Wir können sofort den Homomorphiesatz für Gruppen aussprechen, da wir ihn für algebraische Strukturen bewiesen haben.

Satz 17. *Jeder auf einer Gruppe definierte Homomorphismus läßt sich als Verkettung eines Homomorphismus von der Gruppe auf die Faktorgruppe und eines Isomorphismus von der Faktorgruppe auf die Bildstruktur darstellen.*

Sämtliche Faktorgruppen einer Gruppe sind bis auf Isomorphie durch die Homomorphismen auf der Gruppe beschrieben.

Satz 18. *Es sei φ ein Homomorphismus von der Halbgruppe $H(\cdot)$ in die Halbgruppe $M(*)$. Dann gelten die folgende Aussagen.*

1. *Das homomorphe Bild $\varphi(H(\cdot))$ ist eine Unterhalbgruppe von $M(*)$, d. h. $\varphi(H)(*)$ ist eine Halbgruppe.*
2. *Das homomorphe Bild einer kommutativen Struktur ist kommutativ.*
3. *Ist e das neutrale Element in $H(\cdot)$, so ist $\varphi(e)$ das neutrale Element in der Bildstruktur.*
4. *Ist a^{-1} invers zu a, so ist $\varphi(a^{-1})$ invers zu $\varphi(a)$ in der Bildstruktur.*
5. *Das homomorphe Bild $\varphi(H(\cdot))$ einer Gruppe $H(\cdot)$ ist eine Untergruppe von $M(*)$.*
6. *Ein Homomorphismus bildet Unterhalbgruppen (Untergruppen) auf Unterhalbgruppen (Untergruppen) ab.*

7. *Das Urbild einer Unterhalbgruppe von $\varphi(H)(*)$ ist eine Unterhalbgruppe von $H(\cdot)$; analog für Gruppen.*

8. *Ist $H(\cdot)$ eine Gruppe, so gilt $o(\varphi(a)) \leqq o(a)$ für alle $a \in H$. Hat a eine endliche Ordnung, so ist $o(\varphi(a))$ ein Teiler von $o(a)$.*

Die letzte Aussage folgt etwa mittels Aussage 7 und dem Satz von Lagrange. Die Beweise der einzelnen Aussagen sind sehr einfach und sollten vom Leser selbst gefunden werden.
Die Umkehrabbildung eines Isomorphismus ist wieder ein Isomorphismus; gibt es also einen Isomorphismus von $H(\cdot)$ auf $M(*)$, so sagt man, daß beide Strukturen isomorph sind, sich algebraisch nicht unterscheiden. Das bedeutet jedoch nicht, daß sie in Wirklichkeit gleich sind. So sind z. B. die Strukturen $\left\{ 2^i \;\middle|\; i \in \mathbb{N} \right\}(\cdot)$ und $\mathbb{N}(+)$ isomorph; in der letzteren ist das Rechnen für den Menschen leichter als in der ersten. Ein Rechner arbeitet aber mit der ersten Struktur.
Für endliche Gruppen gilt der Satz von **CAYLEY**.

Satz 19. *Jede endliche Gruppe ist zu einer Anordnungsgruppe isomorph.*

Beweis. Es sei $G(\cdot)$ mit $G = \{ g_1, \ldots, g_n \}$ eine Gruppe. Für jeden Index i definieren wir eine Abbildung f_i von G auf sich gemäß: $f_i(g) = g_i \cdot g$. Jede Abbildung f_i ist durch eine Anordnung der Elemente von G charakterisiert. Die Menge $\mathcal{S}(G)$ aller Anordnungen der Elemente von G bildet mit der Verkettung als Operation eine Gruppe. Die Abbildung

$$\varphi: \quad G \longmapsto \mathcal{S}(G) \text{ mit } \varphi(g_i) = f_i$$

ist ein Homomorphismus: Mit $g_i \cdot g_j = g_k$ folgt nämlich

$$\varphi(g_i \cdot g_j) = \varphi(g_k) = f_k$$

und

$$\varphi(g_i) \circ \varphi(g_j) = f_i \circ f_j,$$

sowie

$$f_i \circ f_j(g) = f_i(f_j(g)) = f_i(g_j \cdot g) = g_i \cdot g_j \cdot g = g_k \cdot g = f_k(g),$$

also

$$\varphi(g_i \cdot g_j) = \varphi(g_i) \circ \varphi(g_j),$$

was gerade die Operationstreue bedeutet. Wegen $|G| = |\mathcal{S}(G)|$ ist φ sogar ein Isomorphismus. □
Der Satz von Cayley hebt die prinzipielle Bedeutung von Anordnungsgruppen hervor: Algebraisch gesehen genügt es, Anordnungsgruppen, also Untergruppen einer symmetrischen Gruppe $\mathbb{S}_n$ zu studieren, weil man damit bis auf Isomorphie bereits alle endlichen Gruppen erfaßt hat. Für theoretische Untersuchungen ist diese Vorgehensweise nicht zweckmäßig, wohl aber für konkrete Berechnungen, insbesondere auf einem Rechner. Nach dem Satz von Cayley ist es erlaubt, Gruppenelemente im Rechner durch Zahlen darzustellen, wodurch z. B. die Typverträglichkeit von Prozeduren, die mit Gruppenelementen operieren, gesichert ist.
Wir kehren zum motivierenden Beispiel vom Anfang dieses Abschnittes zurück. Es sei $\mathcal{B}(\circ)$ die Halbgruppe mit den Befehlsfolgen als Trägermenge B und der Aneinanderreihung als Operation. Bei der Bewegung des Zuges auf den Schienen hat man z. B. folgende Fragen: Welche Wegstrecke wurde nach Ausführung einer Befehlsfolge $a_1, a_2, \ldots, a_l$ zurückgelegt? Welches ist die relative Endposition? In Stationen gezählt ist die Wegstrecke gleich der Summe der Zahlen $|a_i|, i = 1, \ldots, l$. Bei der Berechnung der relativen Endposition muß man beachten, daß zwar der Zug nach Ausführung der Befehlsfolge um $\sum_{i=1}^{l} a_i$ Stationen verschoben ist, aber jede Verschiebung um n Stationen die Rückkehr zur Ausgangsstation bedeutet, also die Endposition durch den nichtnegativen Rest $r_n(\sum_{i=1}^{l} a_i)$ von $\sum_{i=1}^{l} a_i$ bei Division durch n gegeben ist. Unsere Fragen werden daher durch folgende Abbildungen beantwortet:

$$\varphi: \quad \mathcal{B} \longmapsto \mathbb{Z} \text{ mit } \varphi(a_1, \ldots, a_l) = \sum_{i=1}^{l} |a_i|,$$

$$\psi: \quad \mathcal{B} \longmapsto \mathbb{Z} \text{ mit } \psi(a_1, \ldots, a_l) = r_n(\sum_{i=1}^{l} a_i),$$

und ψ ist die Verkettung der beiden Abbildungen

$$\sigma: \quad \mathcal{B} \longmapsto \mathbb{Z} \text{ mit } \sigma(a_1, \ldots, a_l) = \sum_{i=1}^{l} a_i,$$

$$r_n: \quad \mathbb{Z} \longmapsto \{\, 0, 1, \ldots, n-1 \,\} \text{ mit } r_n(i) = k, \text{ falls } i = l \cdot n + k.$$

Durch Nachrechnen erkennt man sogleich, daß φ und σ Homomorphismen von $\mathcal{B}(\circ)$ in $\mathbb{Z}(+)$ sind. Auf der Menge $\mathbb{Z}_n = \{\, 0, 1, \ldots, n-1 \,\}$ der nichtnegativen Reste einer ganzen Zahl bei Division durch n führen wir eine Addition ein:

$$i \oplus j = r_n(i + j).$$

Man sieht gleich, daß $\mathbb{Z}_n(\oplus)$ einen Modul darstellt und r_n ein Homomorphismus vom Modul $\mathbb{Z}(+)$ auf $\mathbb{Z}_n(\oplus)$ ist. Offenbar ist $r_n(i) = r_n(j)$ genau dann, wenn $i - j$ Vielfaches von n ist. Die durch r_n auf $\mathbb{Z}$ induzierte Äquivalenzrelation R_n führt auf Äquivalenzklassen der Form

$$\overline{i} = \{\, j \in \mathbb{Z} \mid j = i + s \cdot n, s \in \mathbb{Z} \,\}.$$

Die natürliche Operation (Addition) in $\mathbb{Z}/R_n$ wird durch

$$\overline{i} \tilde{+} \overline{j} = \overline{i + j}$$

definiert. Nach dem Homomorphiesatz ist die Faktorgruppe $\mathbb{Z}/R_n(\tilde{+})$ isomorph zu $\mathbb{Z}_n(\oplus)$, die daher **additive Restklassengruppe** modulo n genannt und mit $\mathbb{Z}_n$ bezeichnet wird. Mit diesen zusätzlichen Gedanken überzeugt man sich leicht, daß die obige Abbildung ψ ein Homomorphismus von $\mathcal{B}(\circ)$ auf $\mathbb{Z}_n$ ist.
Nach den allgemeinen Überlegungen zu algebraischen Strukturen wird durch einen Homomorphismus f auf einer Struktur $H(\cdot)$ eine Kongruenzrelation F induziert, und diese ist - zusätzlich zu den Bedingungen für eine Äquivalenzrelation - durch die Bedingung

$$xFy \text{ und } uFv \Longrightarrow x \cdot uFy \cdot v$$

charakterisiert. Umgekehrt: Jede Äquivalenzrelation F auf $H(\cdot)$, die diese Bedingung erfüllt, induziert einen Homomorphismus f von $H(\cdot)$ auf die Faktorstruktur $H(\cdot)/F$. Kurzum: Um alle Faktorstrukturen zu erfassen, benötigt man alle Kongruenzrelationen der betrachteten Struktur. Ehe wir die Verhältnisse bei Gruppen klären, beweisen wir den folgenden Satz.

Satz 20. *Für jede Kongruenzklasse $[x]_F$ einer Kongruenzrelation F auf einer Gruppe $G(\cdot)$ gilt*

$$[x]_F = x \cdot [e]_F = [e]_F \cdot x.$$

Beweis. Es sei f der von der Kongruenzrelation F induzierte Homomorphismus von $G(\cdot)$ in $G'(*)$. Die folgende Schlußkette beweist die Behauptung:

$$\begin{aligned}
y \in [x]_F \iff f(y) = f(x) &= f(x) * f(z) \quad \forall z: \ f(z) = f(e) \\
&= f(x \cdot z) \quad \forall z \in [e]_F \\
\iff y \in x \cdot [e]_F &\iff f(y) = f(x \cdot z) \quad \forall z \in [e]_F \\
&= f(x) * f(z) \quad \forall z \in [e]_F \\
&= f(x) * f(e) \\
&= f(e) * f(x) \\
&= f(z) * f(x) \quad \forall z: \ f(z) = f(e) \\
&\iff y \in [e]_F \cdot x.
\end{aligned}$$

□

Dieser Satz sagt uns, daß alle Kongruenzklassen einer gegebenen Kongruenzrelation durch die Kongruenzklasse $[e]_F$ zum neutralen Element vollständig beschrieben sind; die Kongruenzklasse zum Gruppenelement $x \in G$ läßt sich in der Form $x \cdot [e]_F$ darstellen, und es gilt außerdem noch

$$x \cdot [e]_F = [e]_F \cdot x.$$

Jede solche Menge ist Trägermenge des Urbildes der trivialen Untergruppe $(f(e))$ in der Bildstruktur; also müssen dies Untergruppen sein. Sie heißen **Normalteiler**. Genauer: Eine Untergruppe $N(\cdot)$ von $G(\cdot)$ heißt **Normalteiler**, wenn

$$x \cdot N = N \cdot x \quad \forall x \in G$$

gilt. Die Normalteiler und die Kongruenzrelationen einer Gruppe entsprechen einander umkehrbar eindeutig: Jeder Kongruenz F ist der Normalteiler $N(\cdot)$ mit $N = [e]_F$ zugeordnet; verschiedenen Kongruenzen entsprechen verschiedene Normalteiler. Umgekehrt definiert ein Normalteiler N durch

$$xR_Ny \iff x \cdot N = y \cdot N$$

eine Kongruenzrelation R_N auf $G(\cdot)$ mit $[x]_{R_N} = x \cdot N = N \cdot x$. Um dies einzusehen, bemerken wir zunächst, daß R_N eine Äquivalenzrelation ist und daher nur die zusätzliche Bedingung für eine Kongruenzrelation nachzuweisen ist. Aus $x \cdot N = y \cdot N$ und $u \cdot N = v \cdot N$ folgt

$$x \cdot u \cdot N = x \cdot N \cdot u = x \cdot N \cdot N \cdot u = y \cdot N \cdot v = y \cdot v \cdot N,$$

also ist R_N eine Kongruenzrelation. Insbesondere haben wir damit den folgenden Satz bewiesen.

Satz 21. *Eine Untergruppe $N(\cdot)$ einer Gruppe $G(\cdot)$ ist genau dann Normalteiler, wenn eine Kongruenzrelation F auf $G(\cdot)$ existiert mit $N = [e]_F$.*

Der Zusammenhang zwischen den Normalteilern und den Kongruenzrelationen einer Gruppe $G(\cdot)$ erlaubt es, von den Faktorgruppen

$$G/N(\cdot) = G/R_N(\cdot)$$

nach ihren Normalteilern zu sprechen. Das volle Urbild vom neutralen Element aus der Bildgruppe ist gerade der einen Homomorphismus f definierende Normalteiler $N(\cdot)$ und heißt **Kern** - in Zeichen '$\ker(f)$' - des Homomorphismus:

$$N = \ker(f) = [e]_F = \{\, x \in G \mid f(x) = f(e) \,\}.$$

Der Kern eines Homomorphismus ist also gerade die Menge aller jener Elemente, die auf das neutrale Element in der Bildstruktur abgebildet werden.
Abschließend sei eine weitere Charakterisierung der Normalteiler gegeben.

Satz 22. *Eine Untergruppe $N(\cdot)$ einer Gruppe $G(\cdot)$ ist genau dann Normalteiler, wenn $x \cdot N \cdot x^{-1} \subseteqq N$ gilt für alle $x \in G$.*

Beweis. Es gilt offenbar $N = x^{-1} \cdot x \cdot N \cdot x^{-1} \cdot x$. Wenn also die Menge $x \cdot N \cdot x^{-1}$ eine Untermenge von N ist, so folgt

$$N \subseteqq x^{-1} \cdot N \cdot x$$

und damit $N = x \cdot N \cdot x^{-1}$. Also gilt die Schlußkette

$$\begin{aligned} x \cdot N = N \cdot x \quad \forall x \in G &\iff x \cdot N \cdot x^{-1} = N \quad \forall x \in G \\ &\iff x \cdot N \cdot x^{-1} \subseteqq N \quad \forall x \in G. \end{aligned}$$

□

1.3.3. Ringe und Körper

Viele elementare Beispiele für eine algebraische Struktur haben nicht nur eine binäre Operation, sondern zwei. So kann man z. B. Zahlen addieren, multiplizieren und Mengen schneiden und vereinigen. Wir nennen eine algebraische Struktur $R(+,\cdot)$ mit zwei Operationen, die wir als Addition '+' und als Multiplikation '$\cdot$' bezeichnen, **Ring**, wenn $R(+)$ ein Modul, $R(\cdot)$ eine Halbgruppe und die Multiplikation distributiv bezüglich der auf R definierten Addition ist, d. h. für alle $x, y, z \in R$ gilt:

$$x \cdot (y + z) = x \cdot y + x \cdot z, \qquad (y + z) \cdot x = y \cdot x + z \cdot x.$$

Aus den definierenden Eigenschaften kann man Rechenregeln ableiten, die uns vom Rechnen mit ganzen oder reellen Zahlen bestens bekannt sind. Wesentlich ist hier, daß wir zum Beweis dieser Rechenregeln nur die definierenden Eigenschaften der Struktur ausnutzen. Wir bezeichnen im Ring mit 0 das neutrale Element bezüglich der Addition. Mit der Distributivität schließt man

$$\begin{aligned} &0 \cdot a = (0 + 0) \cdot a = 0 \cdot a + 0 \cdot a \Longrightarrow 0 = 0 \cdot a, \\ &a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0 \Longrightarrow 0 = a \cdot 0, \\ &\quad 0 = a \cdot 0 = a \cdot (b + (-b)) = a \cdot b + a \cdot (-b) \Longrightarrow -(a \cdot b) = a \cdot (-b), \\ &\quad 0 = 0 \cdot b = (a + (-a)) \cdot b = a \cdot b + (-a) \cdot b \Longrightarrow -(a \cdot b) = (-a) \cdot b. \end{aligned}$$

Falls $a \neq 0$ und kein Nullteiler in $R(\cdot)$ ist, folgt die übliche Kürzungsregel

$$a \cdot x = a \cdot y \Longrightarrow a \cdot x + a \cdot (-y) = a \cdot (x + (-y)) = 0 \Longrightarrow x = y.$$

Einige Beispiele für Ringe.

1. $\mathbb{Z}(+,\cdot)$,
2. $\mathbb{Q}(+,\cdot), \mathbb{R}(+,\cdot)$,
3. jeder Modul $G(+)$ mit der Nullmultiplikation auf G: $a \cdot b = 0$ für alle $a, b \in G$ (**Nullring** auf $G(+)$),
4. $\mathbb{Z}_n(+,\cdot)$ (Restklassenring modulo n).

In den Beispielen sind $\mathbb{Q}\setminus\{0\}(\cdot)$ und $\mathbb{R}\setminus\{0\}(\cdot)$ sogar kommutative Gruppen. Daher spezifizieren wir genauer. Ein Ring $R(+,\cdot)$ heißt kommutativ, wenn $R(\cdot)$ kommutativ ist. Sollte die Struktur $R(\cdot)$ ein Monoid sein, so nennt man den Ring $R(+,\cdot)$ **Ring mit Einselement**. Schließlich heißt ein Ring $R(+,\cdot)$ **Körper**, wenn die Struktur $R \setminus \{0\}(\cdot)$ eine kommutative Gruppe darstellt. Damit sind $\mathbb{Q}(+,\cdot)$ und $\mathbb{R}(+,\cdot)$ Körper. Die Begriffe Unterring und Unterkörper werden sinngemäß zu den entsprechenden Begriffen bei Gruppen und Halbgruppen gebildet. Eine nichtleere Untermenge $U \subseteq R$ der Trägermenge eines Ringes $R(+,\cdot)$ ist Trägermenge eines **Unterring**es von $R(+,\cdot)$, wenn $U(+,\cdot)$ ein Ring ist, wenn also $U(+)$ ein Modul und $U(\cdot)$ eine Halbgruppe darstellen (Die Distributivgesetze gelten dann automatisch!). Analog ist eine nichtleere Untermenge $U \subseteq K$ der Trägermenge eines Körpers $K(+,\cdot)$ Trägermenge eines **Unterkörper**, falls $U(+,\cdot)$ ein Körper ist. Wann ein Unterring bzw. Unterkörper vorliegt, sagen uns die folgenden beiden Kriterien.

Satz 23. *Eine nichtleere Untermenge $U \subseteq R$ eines Ringes $R(+,\cdot)$ ist genau dann Trägermenge eines Unterringes, wenn die Menge U abgeschlossen bezüglich der beiden Operationen '+','·' und der additiven Inversenbildung ist, d.h.*

$$u - v \in U, \quad u \cdot v \in U \quad \forall u, v \in U.$$

Eine nichtleere Untermenge $U \subseteq K$ der Trägermenge eines Körpers $K(+,\cdot)$ ist genau dann Trägermenge eines Unterkörpers, wenn

$$u - v \in U \quad \forall u, v \in U \text{ und } u^{-1} \cdot v \in U \quad \forall u, v \in U \setminus \{0\}.$$

Das Kriterium für einen Unterkörper sagt aus, daß die Trägermenge U eines Unterkörpers abgeschlossen bezüglich '+', der additiven Inversenbildung und $U \setminus \{0\}$ abgeschlossen bezüglich '·' und der multiplikativen Inversenbildung ist. Analog zu Gruppen definiert man den von einer Untermenge X erzeugten Unterring (Unterkörper) als den Durchschnitt aller jener Unterringe (Unterkörper), deren Trägermengen die Menge X enthalten. Dem allgemeinen Homomorphiebegriff folgend liegt bei einer Abbildung φ eines Ringes $R(+,\cdot)$ in einen Ring $S(*,\circ)$ ein **Ringhomomorphismus** vor, wenn

$$\varphi(x+y) = \varphi(x) * \varphi(y) \text{ und } \varphi(x \cdot y) = \varphi(x) \circ \varphi(y) \quad \forall x, y \in R$$

gilt. Ist überdies die Strukturabbildung φ sogar bijektiv, spricht man von einem **Ringisomorphismus**. Als Beispiel sei erwähnt, daß

$$\varphi: \quad \mathbb{Z} \longmapsto \mathbb{Z}_n \text{ mit } i \longmapsto r_n(i),$$

wobei $r_n(i)$ den nichtnegativen Rest von i bei Division durch n bedeutet, einen Ringhomomorphismus darstellt. Im Zusammenhang mit Ringhomomorphismen formulieren wir 6 Eigenschaften.

Satz 24. *Es sei φ ein Ringhomomorphismus von $R(+,\cdot)$ in $S(*,\circ)$. Dann gelten die folgenden Aussagen.*

1. *Der Homomorphismus φ überführt Ringe in Ringe, d. h. $\varphi(R)$ ist ein Unterring von S.*
2. *Das homomorphe Bild eines kommutativen Ringes ist wieder kommutativ.*
3. *Wenn e das Einselement im Urbildring darstellt, dann ist $\varphi(e)$ das Einselement im Bildring.*
4. *Das homomorphe Bild eines Unterringes ist ein Unterring in der Bildstruktur.*
5. *Das volle Urbild eines Unterringes aus dem Bildring ist ein Unterring im Urbildring.*

Die Kongruenzrelationen auf einem Ring $R(+,\cdot)$ sind gerade jene Äquivalenzrelationen auf R, die sowohl Kongruenzen auf $R(+)$ als auch auf $R(\cdot)$ sind. Nun ist $R(+)$ ein Modul, also sind alle Untergruppen auch Normalteiler; daher entspricht jeder Untergruppe von $R(+)$ umkehrbar eindeutig eine Kongruenz, wobei einer Kongruenzrelation S die Untergruppe $[0]_S$ entspricht. Ist nun S außerdem auch noch Kongruenz auf $R(\cdot)$, so gilt

$$r \cdot [0]_S \subseteq [0]_S \text{ und } [0]_S \cdot r \subseteq [0]_S \quad \forall r \in R,$$

denn für alle $x \in [0]_S$ ist

$$\varphi(r \cdot x) = \varphi(r) \cdot \varphi(x) = \varphi(r) \cdot \varphi(0) = \varphi(0).$$

Diese Untergruppen heißen Ideale. Eine Untergruppe $I(+)$ von $R(+)$, für die $r \cdot I \subseteq I$ und $I \cdot r \subseteq I$ für alle $r \in R$ gilt, heißt **Ideal**. Zwischen den Idealen, Kongruenzrelationen und den homomorphen Bildern eines Ringes besteht folgender Zusammenhang.

Satz 25. *Zwischen den Idealen I und den Kongruenzrelationen S eines Ringes $R(+,\cdot)$ besteht eine eineindeutige Beziehung:*

$$S \longmapsto I = [0]_S,$$

$$I \longmapsto S_I \text{ mit } xS_Iy \iff x + I = y + I.$$

Beweis. Für die letzte Beziehung ist noch zu zeigen, daß S_I eine Kongruenzrelation auf $R(\cdot)$ ist. Aus $x+I = y+I$ und $u + I = v + I$ folgt

$$\begin{aligned}(y + I) \cdot (v + I) = (x + I) \cdot (u + I) &= x \cdot u + x \cdot I + I \cdot u + I \cdot I \\ &\subseteqq x \cdot u + I \\ &\Longrightarrow y \cdot v \in x \cdot u + I \Longrightarrow y \cdot v + I = x \cdot u + I.\end{aligned}$$

□

Satz 26. *Das homomorphe Bild $\varphi(R)$ eines Ringes $R(+,\cdot)$ ist isomorph zum Restklassenring R/I von R nach dem Ideal $I = [0]_{S_I} = \ker(\varphi)$.*

Dieser Satz folgt sofort aus dem Homomorphiesatz für algebraische Strukturen. Es sei noch angemerkt, daß die obigen Eigenschaften von Normalteilern in Gruppen sinngemäß auch für Ideale gelten.
Abschließend sollen die endlichen Restklassenringe $\mathbb{Z}_m(+,\cdot)$ etwas näher betrachtet werden, da es unter ihnen Körper gibt, die in der Codierungstheorie angewendet werden. Zwei Elemente x, y einer Restklasse unterscheiden sich nur um ein ganzzahliges Vielfaches von m:

$$x = \lambda \cdot m + l, \quad y = \mu \cdot m + l \Longrightarrow x - y = \sigma \cdot m.$$

Daher haben alle Elemente einer Restklasse die gleichen Teiler mit m. Eine Restklasse nennt man **prime Restklasse** modulo m, wenn ihre Elemente zu m teilerfremd sind. Sind $[x]$ und $[y]$ prime Restklassen modulo m, so ist auch $[x] \cdot [y] = [x \cdot y]$ eine prime Restklasse modulo m. Aus der Schule wissen wir, daß es zu zwei teilerfremden Zahlen x, m stets ganze Zahlen u, v gibt mit $x \cdot u + m \cdot v = 1$, wobei u teilerfremd zu m ist. Die letzte Gleichung bedeutet, daß es zu jeder primen Restklasse $[x]$ mod m eine prime Restklasse $[y]$ mod m gibt mit $[x] \cdot [y] = [1]$. Somit ist gezeigt, daß die primen Restklassen mod m mit der Restklassenmultiplikation eine kommutative Gruppe bilden. Im Falle, daß m eine Primzahl ist, sind $[1], [2], \ldots, [m-1]$ sämtlich prime Restklassen und daher $\mathbb{Z}_m(+,\cdot)$ für jede Primzahl m ein endlicher Körper. Der erste von ihnen ist $\mathbb{Z}_2(+,\cdot)$ mit nur zwei Elementen.

1.4. Übungen

1. Man zeige die Gültigkeit des Distributivgesetzes und der Assoziativgesetze im Bereich der komplexen Zahlen.

2. Man gebe zu den folgenden komplexen Zahlen jeweils die alternativen Darstellungen an (arithmetische bzw. trigonometrische Darstellung):

$$-2 + 3i, \quad 8 - 6i, \quad 5\left(\cos\frac{\pi}{6} + i\sin\frac{\pi}{6}\right),$$

$$-5 - \sqrt{3}i, \quad 2\left(\cos\frac{5\pi}{4} + i\sin\frac{5\pi}{4}\right).$$

3. Für

$$z_1 = \frac{3}{2}\sqrt{2} - \frac{3}{2}\sqrt{2}i, \quad z_2 = 1 + i,$$

$$z_3 = -4 - 5i, \quad z_4 = -\frac{1}{2} + \frac{1}{2}\sqrt{3}i$$

berechne man:

$$\frac{z_1(z_2 + z_3)}{z_4}, \quad z_1^2 z_4, \quad \frac{z_3 z_4}{z_2},$$

$$\frac{z_1 - z_4^2}{z_2 + z_3} \quad (z_1 + z_2 + z_3 + z_4)^2.$$

Man gebe die Lösungen in der arithmetischen Darstellung an.

4. Man berechne die folgenden Wurzeln:

$$\sqrt[3]{1+i}, \quad \sqrt[5]{-1}, \quad \sqrt[4]{\sqrt{3}-i} \quad \text{und} \quad \sqrt[3]{-16+16i}$$

und stelle die Ergebnisse in der Gauss-schen Zahlenebene dar.

5. Man berechne:

$$(1-\sqrt{3})^{\frac{5}{2}}, \quad \left[(\sqrt{6}+\sqrt{2})+(\sqrt{6}-\sqrt{2})i\right]^{\frac{3}{2}}, \quad \text{und} \quad (-i)^{\frac{1}{4}}.$$

6. Man drücke $\sin x$, $\cos x$ und $\tan x$ durch $\tan \frac{x}{2}$ aus.

7. Man löse die goniometrischen Gleichungen

   (a)
   $$\sin x + \cos x = \frac{1}{\sin x},$$

   (b)
   $$\sin^4 x + \cos^4 x = \cos 4x.$$

8. Man ermittle jeweils alle reellen Zahlen x, die die folgenden Ungleichungen erfüllen.

   (a) $\sin^2 x + 2\sin x > 0$,
   (b) $|\sin 2x| \geqq \frac{1}{2}\sqrt{3}$,
   (c) $5\sin 2x + 2\cos x < 7$.

9. Man beweise die Gültigkeit von

$$\sum_{i=1}^{n} \frac{2^{(2^{i-1})}}{1-2^{(2^i)}} = \frac{2^{(2^n)}-2}{1-2^{(2^n)}}$$

für alle natürlichen Zahlen n.

10. Man beweise die Gültigkeit von

$$\underbrace{\sqrt{2+\sqrt{2+\cdots+\sqrt{2}}}}_{n-mal} = 2\cos\frac{\pi}{2^{n+1}}$$

für alle natürlichen Zahlen n.

11. Man gebe sämtliche Lösungen x der folgenden Gleichungen an:

    (a)
    $$\sqrt{p+x}+\sqrt{p-x} = x \qquad (p \text{ beliebige reelle Zahl}),$$

    (b)
    $$\left(\frac{x}{x+1}\right)^2 + \left(\frac{x+1}{x}\right)^2 = \frac{5}{2},$$

    (c)
    $$\sqrt{6x+1} - \sqrt{2x+1} = \sqrt{x}.$$

12. Man zeige die Gültigkeit der folgenden Beziehungen für alle natürlichen Zahlen n:

    (a) $\displaystyle\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$,
    (b) $\displaystyle\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$,
    (c) $\displaystyle\sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4}$.

13. Man zeige die Gültigkeit der folgenden Beziehung für alle natürlichen Zahlen n:

$$\sum_{k=1}^{n} k \cdot k! = (n+1)! - 1.$$

14. Man berechne
$$\sum_{k=1}^{n-1}(n-k)(n-k+1).$$

15. Man beweise:
Für alle von Null verschiedenen reellen Zahlen a und b und alle natürlichen Zahlen n gilt
$$(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}.$$

16. Man beweise oder widerlege folgende Aussagen:
    (a) Für alle $n \in \mathbb{N}$ ist $n^3 - n$ durch 6 teilbar.
    (b) Für alle $n \in \mathbb{N}$ ist $(n-1)^2 + n + 40$ eine Primzahl.

17. Es sei (a_k) eine arithmetische Zahlenfolge mit $a_k \neq 0$ für alle k. Man beweise, daß für alle $k \in \mathbb{N}$, $k \geqq 2$ gilt:
$$\frac{1}{a_1 \cdot a_2} + \frac{1}{a_2 \cdot a_3} + \cdots + \frac{1}{a_{k-1} \cdot a_k} = \frac{k-1}{a_1 \cdot a_k}.$$
    Hinweis: Eine Folge (a_k) ist eine arithmetische Zahlenfolge, wenn zwei beliebige, aufeinander folgende Zahlen eine feste Differenz haben.

18. Man beweise:
    (a) Für beliebige Mengen X, Y gilt $X \cup Y = X \cap Y$ genau dann, wenn $X = Y$.
    (b) Für beliebige Mengen X, Y, Z folgt aus $X \subseteqq Y, X \subseteqq Z$, daß auch $X \subseteqq Y \cap Z$ gilt.

    Was kann man zur Umkehrung von Aussage (b) sagen?

19. Man skizziere die folgenden Mengen:
    (a) $M_1 = \{\, (x,y) \mid y + 1 \geqq x \,\}$,
    (b) $M_2 = \{\, (x,y) \mid y = -x^2 \,\}$,
    (c) $M_3 = \{\, (x,y) \mid x^2 + y^2 \leqq 2 \,\}$,
    (d) $M_4 = \{\, (x,y) \mid \max(|x|,|y|) \leqq 2 \,\}$,
    (e) $M_1 \cup M_2, M_1 \cup M_2, M_1 \setminus M_3, M_3 \setminus M_2, M_3 \times M_4$.

20. Man zeige die Gültigkeit der folgenden Beziehungen für beliebige Mengen M_1, M_2 und M_3:
    (a) $M_1 \setminus (M_2 \cup M_3) = (M_1 \setminus M_2) \cap (M_1 \setminus M_3)$,
    (b) $M_1 \setminus (M_2 \cap M_3) = (M_1 \setminus M_2) \cup (M_1 \setminus M_3)$.

21. Man untersuche, ob für beliebige Mengen M_1, M_2 und M_3 die folgenden Beziehungen gelten:
    (a) $M_1 \cap (M_2 \setminus M_3) = (M_1 \cap M_2) \setminus (M_1 \cap M_3)$,
    (b) $M_1 \cup (M_2 \setminus M_3) = (M_1 \cup M_2) \setminus (M_1 \cup M_3)$.

22. Gilt für beliebige Mengen M_1 und M_2 die Beziehung
$$M_1 \cap M_2 = M_1 \setminus (M_1 \setminus M_2)?$$

23. Man untersuche die Eigenschaften der folgenden binären Relationen R auf der Menge X. Durch welche Relationen ist eine Ordnung, Halbordnung bzw. Äquivalenzrelation gegeben? Falls R eine Äquivalenzrelation ist, so charakterisiere man die Äquivalenzklassen.
    (a) $X = \mathbb{N}$, $\quad xRy \iff x|y \quad$ (x ist Teiler von y),
    (b) $X = \mathbb{N}$, $\quad xRy \iff 2|x^2 + y^2$,
    (c) $X = \{\,$ Menge der Geraden im Raum $\,\}$, $\quad xRy \iff x$ und y sind parallel,
    (d) $X = \{\,$ Menge der Geraden im Raum $\,\}$,
    $xRy \iff x$ und y besitzen mindestens einen gemeinsamen Punkt,

(e) $X = \mathbb{R}^2$, $(a_1, a_2), (b_1, b_2)$ fixiert, $(x_1, x_2)R(y_1, y_2)$ genau dann, wenn

$$\sqrt{(x_1 - a_1)^2 + (x_2 - a_2)^2} + \sqrt{(x_1 - b_1)^2 + (x_2 - b_2)^2}$$
$$= \sqrt{(y_1 - a_1)^2 + (y_2 - a_2)^2} + \sqrt{(y_1 - b_1)^2 + (y_2 - b_2)^2},$$

(f) $X = \mathcal{P}(M), \quad xRy \Longleftrightarrow x \cap y = \emptyset,$

(g) $X = \mathcal{P}(M), \quad xRy \Longleftrightarrow x \cup y = x,$

(h) $X = \mathbb{C}, \quad xRy \Longleftrightarrow x\bar{y} = \bar{x}y$, wobei $\bar{x}$ die konjugiert komplexe Zahl zu x bezeichnet.

24. Man untersuche die folgenden Relationen R über den jeweiligen Mengen X hinsichtlich ihrer Eigenschaften.

(a) $X = \mathbb{R}, \quad xRy \iff x \leqq y,$

(b) $X = \mathbb{N} \times \mathbb{N}, \quad (i, j)R(k, l) \iff i \cdot l = j \cdot k,$

(c) $X = \{\, 1, 2, 3 \,\}, \quad R = \{\, (1,1), (2,2), (3,3), (1,2), (2,3) \,\},$

(d) $X = \mathbb{N}, \quad mRn \iff m \cdot n$ ist gerade oder $m = n,$

(e) $X = \mathbb{N}, \quad mRn \iff \mathrm{ggT}(m, n) > 1$ (ggT - größter gemeinsamer Teiler).

Welche der Relationen bilden eine Äquivalenzrelation, eine Halbordnung oder eine Ordnung?

25. Man zeige, daß die Potenzmenge $\mathcal{P}(M)$ jeder endlichen Menge M mächtiger ist als M selbst.

26. Man suche Beispiele für Relationen, die

(a) reflexiv und symmetrisch, aber nicht transitiv,

(b) symmetrisch und antisymmetrisch zugleich sind.

27. Man zeige, daß durch die Relation S:

$$(a, b)S(c, d) \iff a + b = c + d$$

eine Äquivalenzrelation im $\mathbb{R}^2$ definiert wird.
Man veranschauliche sich die $[1, 1]_S$-Klasse sowie die Menge aller Äquivalenzklassen.

28. Man untersuche, ob die folgenden Relationen R über der Menge X Äquivalenzrelationen sind und beschreibe gegebenenfalls die Äquivalenzklassen.

(a) $X = \mathbb{N},$

$$mRn \Longleftrightarrow \sin\frac{m\pi}{2} \cdot \sin\frac{n\pi}{2} > 0 \text{ oder } \left|\sin\frac{m\pi}{2}\right| + \left|\sin\frac{n\pi}{2}\right| = 0,$$

(b) $X = \mathbb{R},$

$$xRy \Longleftrightarrow [x] = [y],$$

wobei $[x]$ die größte ganze Zahl z bezeichnet, die nicht größer als x ist,

(c) $X = \mathcal{P}(M),$

$$X_1 R X_2 \Longleftrightarrow X_1 = C_M(X_2).$$

29. Man untersuche folgende Abbildungen $f: \quad X \to Y$ auf ihre Eigenschaften:

(a) $X = [0, 1], Y = [-\frac{1}{8}, 1], f(x) = 2x^2 - x,$

(b) $X = [1, 2], Y = [1, 3], f(x) = |x|,$

(c) $X = [-1, 1], Y = [0, 1], f(x) = |x|.$

30. Man untersuche, ob folgende Teilmengen $f \subset \mathbb{R} \times \mathbb{R}$ Abbildungen von $\mathbb{R}$ in $\mathbb{R}$ sind.

(a)

$$f = \left\{\, (x, y) \in \mathbb{R}^2 \;\middle|\; y^2 = 9 - x^2 \,\right\},$$

(b)

$$f = \left\{\, (x, y) \in \mathbb{R}^2 \;\middle|\; (y + 3)^2 = 2\cos 5x \,\right\},$$

(c)

$$f = \left\{ (x, y) \in \mathbb{R}^2 \;\middle|\; y = \frac{x + 3}{x - 2} \right\},$$

(d)
$$f = \{\ (x,y) \in \mathbb{R}^2 \mid y = \sqrt{x}\ln x\ \},$$

(e)
$$f = \left\{\ (x,y) \in \mathbb{R}^2 \;\middle|\; y = \frac{x^3 - x + 2}{x^2 + 2}\ \right\},$$

(f)
$$f = \{\ (x,y) \in \mathbb{R}^2 \mid e^y = x^2 - x - 2\ \},$$

(g)
$$f = \{\ (x,y) \in \mathbb{R}^2 \mid e^y = x^4 - x + 7\ \}.$$

31. Gegeben seien die Mengen $X = \{\ 1,2,3,5,6,10,15,30\ \}$ und $Y = \{\ 2,3,5\ \}$. Man konstruiere eine bijektive Abbildung der Menge X auf die Potenzmenge $\mathcal{P}(Y)$, so daß für beliebige $m, n \in X$ gilt:
$$m|n \Longleftrightarrow f(m) \subseteq f(n).$$

32. Man zeige, daß es keine bijektive Abbildung einer Menge auf ihre Potenzmenge gibt.

33. Es seien f und g Abbildungen. Man finde Bedingungen, unter denen $f \circ g$ surjektiv bzw. injektiv bzw. bijektiv ist.

34. Man zeige, daß die Verknüpfung von Abbildungen assoziativ ist.

35. Eine Abbildung $f : X \to Y$ heißt linear, wenn gilt:
$$\forall x_1, x_2 \in X \quad \forall a, b \in \mathbb{R} : f(ax_1 + bx_2) = af(x_1) + bf(x_2).$$
Man untersuche, ob folgende Abbildungen linear sind.
   (a) $X = \mathbb{R}, \quad Y = \mathbb{R}, \quad f(x) = 3x + 4,$
   (b) $X = \mathbb{R}, \quad Y = \mathbb{R}, \quad f(x) = 2x,$
   (c) $X = \{$ Menge aller differenzierbaren Funktionen von $\mathbb{R}$ in $\mathbb{R}\ \}, \quad Y = X,$
   $f :$ jede Funktion aus X wird auf ihre Ableitung abgebildet.

36. Es seien $f : M \to N$ und $g : N \to L$ Abbildungen. Für die Verknüpfung $g \circ f$ dieser Abbildungen zeige man:
   (a) Sind f und g surjektiv, so ist auch $g \circ f$ surjektiv.
   (b) Sind f und g injektiv, so ist auch $g \circ f$ injektiv.
   (c) Sind f und g bijektiv, so ist auch $g \circ f$ bijektiv.

37. Gegeben seien die folgenden Abbildungen:
   - $f : \mathbb{R} \to [0,1] \quad$ mit $\quad f(x) = \sin^2 x,$
   - $g : [0,\infty) \to [0,\infty) \quad$ mit $\quad g(x) = \sqrt{x},$
   - $h : \mathbb{R}_+ \to \mathbb{R} \quad$ mit $\quad h(x) = \ln x,$
   - $p : \mathbb{R} \to [-1,1] \quad$ mit $\quad p(x) = \sin 2x.$

   Man bilde alle möglichen Verknüpfungen dieser Abbildungen bzw. geeigneter Einschränkungen dieser Abbildungen und ermittle deren Eigenschaften.

38. $\mathcal{A}$ sei eine σ-Algebra. Man beweise folgende Eigenschaft:
$$A \in \mathcal{A} \wedge B \in \mathcal{A} \Longrightarrow (A \cap B) \in \mathcal{A}.$$

39. Gegeben seien die algebraischen Strukturen $\mathcal{S}_1 = (\mathbb{R}_+; 1; <; \cdot)$ und $\mathcal{S}_2 = (\mathbb{R}; 0; <; +)$. Man zeige, daß die Abbildung $f : \mathbb{R}_+ \to \mathbb{R}$ mit $f(x) = \ln x$ ein Isomorphismus ist.

40. Gegeben seien die algebraischen Strukturen $\mathcal{S}_1 = (M, \circ)$ mit

$$M = \{\, f_1(x) = x,\ f_2(x) = 1/x,\ f_3(x) = -x,\ f_4(x) = -1/x \,\}$$

und

$$\circ : (f_i \circ f_j)(x) = f_i(f_j(x)) \quad \text{für} \quad i, j = 1, 2, 3, 4$$

sowie $\mathcal{S}_2 = (N, *)$ mit

$$N = \{\, (1,1), (1,-1), (-1,1), (-1,-1) \,\}$$

und

$$* : (i, j) * (k, l) = (ik, jl).$$

Man gebe einen Isomorphismus $\varrho : M \to N$ in der Form $\varrho(f) = (\varrho_1(f), \varrho_2(f))$ an.
Weiterhin überlege man sich zweistellige Relationen R_1 und R_2, so daß der gefundene Isomorphismus auch Isomorphismus von $\bar{S}_1 = (M, R_1, \circ)$ auf $\bar{S}_2 = (N, R_2, *)$ ist.

41. Gegeben seien die algebraischen Strukturen $\mathcal{S}_1 = (X, \circ)$ mit

$$X = \left\{ (1,0), \left(-\frac{1}{2}, \frac{1}{2}\sqrt{3}\right), \left(-\frac{1}{2}, -\frac{1}{2}\sqrt{3}\right) \right\}$$

und

$$\circ :\ (a, b) \circ (c, d) = (ac - bd, ad + bc)$$

sowie

$$\mathcal{S}_2 = (\mathbb{Z}, +).$$

Weiterhin sei die Abbildung $f : \mathbb{Z} \to X$ mit

$$f(n) = \left(\cos\left(\frac{2n\pi}{3}\right), \sin\left(\frac{2n\pi}{3}\right)\right)$$

gegeben.

(a) Man zeige, daß f ein Homomorphismus von $\mathbb{Z}$ auf X ist.

(b) Welche Relation induziert f?

(c) Man beschreibe die durch f erzeugte Faktorstruktur.

(d) Man zeige die Isomorphie zwischen der Faktorstruktur und $S_1 = (X, \circ)$.

42. Man zeige, daß die Menge $\mathbb{N}$ der natürlichen Zahlen mit der Operation

$$\circ : m \circ n = ggT(m, n)$$

eine Halbgruppe bildet. (Mit $ggT(m, n)$ wird der größte gemeinsame Teiler der Zahlen m und n bezeichnet.) Besitzt $\mathbb{N}(\circ)$ ein neutrales Element?

43. Eine Restklasse $[k]_R$ bezüglich der Division durch m sei die Menge der ganzen Zahlen, die bei der Division durch m denselben Rest lassen wie die Zahl k. Die Menge dieser Restklassen bezeichnet man mit $\mathbb{Z}_m$ und definiert die Operationen $\oplus$ und $\odot$ folgendermaßen:

$$[m]_R \oplus [n]_R = [m + n]_R,$$
$$[m]_R \odot [n]_R = [m \cdot n]_R.$$

(a) Ist $\mathbb{Z}_7(\oplus)$ eine Gruppe?

(b) Ist $\mathbb{Z}_7(\odot)$ eine Gruppe oder eine Halbgruppe?

(c) Für welche m hat $\mathbb{Z}_m(\odot)$ Nullteiler?

44. Man beweise:

(a) Das neutrale Element eines Monoids ist eindeutig bestimmt.

(b) Ist $G(\circ)$ eine Gruppe, e das neutrale Element und gilt für alle $a \in G$ $a \circ a = e$, so handelt es sich um eine kommutative Gruppe.

45. Gegeben seien die Permutationen

$$s_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 5 & 4 \end{pmatrix}, \quad s_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 5 & 4 & 2 & 3 \end{pmatrix},$$

$$s_3 = (\ 2 \quad 4 \quad 3 \quad 5\ ) \quad \text{und} \quad s_4 = (\ 2 \quad 4 \quad 5\ )(\ 3 \quad 1\ ).$$

   (a) Man wandle s_1 und s_2 in die Zyklendarstellung und s_3 und s_4 in die ausführliche Schreibweise um.

   (b) Man bilde alle Permutationen $s_i \circ s_j$ mit $i, j = 1, 2, 3, 4$ unter Zuhilfenahme
      i. der Zyklendarstellung,
      ii. der ausführliche Schreibweise.

   (c) Man bestimme für alle Permutationen aus a) und b) ihr Signum.

46. Man stelle alle Permutationen von 3 Elementen als Zyklen dar und bilde die Verknüpfungstabelle. Man kennzeichne die geraden Permutationen in der Tabelle.

47. Gibt es eine Gruppe, in der zu einem gewissen Teiler der Gruppenordnung keine Untergruppe mit dieser Ordnung existiert?

48. Man finde eine Gruppe, die eine echte, zu ihr isomorphe Untergruppe enthält.

49. Es sei $M = \left\{\, m + n\sqrt{5} \mid m, n \in \mathbb{Z} \,\right\}$. Man zeige:

   (a) $M(+)$ ist eine kommutative Gruppe.

   (b) $M(\cdot)$ ist Halbgruppe.

50. Man untersuche die Struktur $\mathcal{S} = M(\circ)$ mit

   - $M = \mathbb{R}\backslash\{1\}$ und
   - $\circ : M \times M \to M$ mit $a \circ b = a + b - ab$.

51. Es seien $p_1 = (1 \quad 4 \quad 3 \quad 2)$ und $p_2 = (1 \quad 3)$ Permutationen aus der Gruppe $\mathbb{S}_4$. Welche Gruppe wird von $\{\, p_1, p_2 \,\}$ erzeugt? Man gebe die Strukturtafel an und bestimme alle Untergruppen und Normalteiler.

52. (a) Man zeige, daß die Menge $M = \{\, (1), (1 \quad 3), (2 \quad 4), (1 \quad 3)(2 \quad 4) \,\}$ eine Untergruppe von $\mathbb{S}_4(\circ)$ bildet. Man gebe ein Erzeugendensystem für M an.

   (b) Man gebe für die Untergruppe der geraden Permutationen von $\mathbb{S}_3(\circ)$ ein Erzeugendensystem an.

   (c) Man bestimme 4 dreielementige Untergruppen von $\mathbb{S}_4(\circ)$ und zeige, daß für jede dieser Untergruppen ein Isomorphismus auf $\mathbb{Z}_3(\oplus)$ existiert.

   (d) Gibt es einen Isomorphismus von $\mathbb{Z}_3(\odot)$ nach $\mathbb{Z}_3(\oplus)$?

53. Es sei F die Menge aller rationalen Funktionen $f : \mathbb{R} \to \mathbb{R}$ mit

$$f(x) = \frac{ax + b}{cx + d}, \quad a, b, c, d \in \mathbb{Z}, \quad ggT(a, b, c, d) = 1 \quad \text{und} \quad ad - bc \neq 0.$$

   Man untersuche die Struktur $\mathcal{S} = F(\circ)$ mit $(f \circ g)(x) = f(g(x))$ hinsichtlich ihrer Eigenschaften und finde vier nichttriviale Unterstrukturen.

54. Gegeben seien die Gruppen $\mathbb{Z}(+)$ und $\mathbb{Z}_3(\oplus)$, sowie die Funktionen

$$f, g : \mathbb{Z} \to \mathbb{Z}_3$$

   mit

$$f(k) = [2k]_R \quad \text{und} \quad g(k) = [2 + k]_R.$$

   (a) Sind f und g Homomorphismen?

   (b) Man bilde die Urbilder $f^{-1}([0]_R)$ und $g^{-1}([0]_R)$ der Restklasse $[0]_R$.

   (c) Es sei $N = f^{-1}([0]_R)$. Man zeige, daß N eine Untergruppe von $\mathbb{Z}(+)$ ist.

   (d) Wie lautet die durch f induzierte Kongruenzrelation R_f? Man gebe die Elemente der Faktorgruppe $\mathbb{Z}/R_f$ an.

55. Man untersuche die folgenden Strukturen auf ihre algebraischen Eigenschaften (Ring, kommutativer Ring, Ring mit Einselement, Körper).

(a) $M(+,\cdot)$ mit $M = \left\{ m + n\sqrt{5} \mid m,n \in \mathbb{Z} \right\}$ und $+$ und $\cdot$ als gewöhnliche Addition und Multiplikation reeller Zahlen.

(b) $C(\oplus,\odot)$ mit $C = \{ (a,b) \mid a,b \in \mathbb{R} \}$ und
$(a,b) \oplus (c,d) = (a+c, b+d)$ und $(a,b) \odot (c,d) = (ac - bd, ad + bc)$.

56. Es seien die Untermengen $C_R = \{ (a,0) \mid a \in \mathbb{R} \}$ und $C_L = \{ (0,b) \mid b \in \mathbb{R} \}$ der Menge C aus Aufgabe 55(b) gegeben. Welche der beiden Mengen bildet einen Unterkörper von $C(\oplus,\odot)$?

57. Es seien $Re : C \to \mathbb{R}$ mit $Re\,((a,b)) = a$ und $Id : C_R \to \mathbb{R}$ mit $Id\,((a,0)) = a$ Abbildungen von C bzw. C_R in $\mathbb{R}$. Welche der Abbildungen ist ein Homomorphismus bzw. Isomorphismus?

58. Man zeige, daß
$$M_5 = \left\{ m + n\sqrt{5} \mid m,n \in \mathbb{Z}, 5|m, 5|n \right\}$$
ein Ideal in $M(+,\cdot)$ (siehe Aufgabe 55(a)) ist und gebe die zugehörigen Kongruenzklassen an.

59. Es sei
$$O_2 = \{ r = k \cdot 2^n \mid k \in \mathbb{N}, n \in \mathbb{Z} \}$$
eine Teilmenge der rationalen Zahlen. Weiterhin sei die Abbildung $\varphi : O_2 \to \mathbb{Z}$ durch
$$\varphi(r) = \max \{ n \mid r = k \cdot 2^n, k \in \mathbb{N}, n \in \mathbb{Z} \}$$
gegeben.

(a) Man untersuche $O_2(+)$ und $O_2(\cdot)$ hinsichtlich ihrer algebraischen Eigenschaften.

(b) Man zeige, daß φ als Abbildung von $O_2(\cdot)$ in $\mathbb{Z}(\cdot)$ ein Homomorphismus ist, als Abbildung von $O_2(+)$ in $\mathbb{Z}(+)$ jedoch nicht.

(c) Man bestimme den Kern des Homomorphismus $\varphi : O_2(\cdot) \to \mathbb{Z}(\cdot)$.

(d) Welches sind die Elemente der zugehörigen Faktorgruppe O_2/R_φ ?

Kapitel 2

Lineare Algebra

2.1. Vektorräume

Das wohl wichtigste Beispiel einer algebraischen Struktur ist der Vektorraum; er bildet die Grundlage für die gesamte lineare Algebra, für die Analysis und andere mathematische Gebiete. Eine algebraische Struktur $V(\oplus;\odot,K(+,\cdot))$ heißt **Vektorraum** (**linearer Vektorraum** oder **linearer Raum**) über dem Körper $K(+,\cdot)$, wenn $V(\oplus)$ ein Modul ist und zwischen den Elementen des Körpers und den Elementen von V eine binäre Operation

$$\odot: \quad K \times V \longmapsto V$$

mit Werten in V erklärt ist, die für alle $\lambda, \mu \in K$ und alle $\mathbf{x}, \mathbf{y} \in V$ folgende Bedingungen erfüllt:

1. $\lambda \odot (\mathbf{x} \oplus \mathbf{y}) = \lambda \odot \mathbf{x} \oplus \lambda \odot \mathbf{y}$,
2. $(\lambda + \mu) \odot \mathbf{x} = \lambda \odot \mathbf{x} \oplus \mu \odot \mathbf{x}$,
3. $\lambda \odot (\mu \odot \mathbf{x}) = (\lambda \cdot \mu) \odot \mathbf{x}$,
4. $1 \odot \mathbf{x} = \mathbf{x}$.

Daß es sich hier wirklich um eine algebraische Struktur im Sinne unserer Definition handelt, sieht man wie folgt ein: Wir haben es hier zunächst mit einer Trägermenge V zu tun, auf der eine Operation $\oplus$ erklärt ist. Die Operation $\odot$ zwischen den Elementen des Körpers und den Elementen aus V nennt man Multiplikation. Genau genommen definiert jedes Element λ des Körpers $K(+,\cdot)$ mittels der Multiplikation $\odot$ eine binäre Relation auf V, die wir mit $\lambda_\odot$ bezeichnen:

$$\lambda_\odot: \quad V \times V \longmapsto V \text{ mit } (\mathbf{x},\mathbf{y}) \in \lambda_\odot \iff \mathbf{y} = \lambda \odot \mathbf{x}.$$

In Abhängigkeit von der Mächtigkeit von K können dies auch überabzählbar viele Relationen sein.
Die Elemente eines Vektorraumes nennt man **Vektoren**. Als Modul enthält ein Vektorraum auch ein neutrales Element, den Nullvektor; es wird mit $\mathbf{o}$ bezeichnet. Aus ersichtlichen Gründen schreibt man die Operation '$\oplus$' als Addition '$+$' und die Operation '$\odot$' als Multiplikation '$\cdot$' zwischen den Körperelementen und den Vektoren, wobei das Multiplikationszeichen oft weggelassen wird. Bei einem Vektor $\lambda\mathbf{x}$ werden wir vom Faktor λ sprechen, mit dem der Vektor $\mathbf{x}$ multipliziert wurde. Falls aus dem Zusammenhang klar sein sollte, über welchem Körper $K(+,\cdot)$ der betreffende Vektorraum definiert ist, verwenden wir einfach die Bezeichnung V für einen Vektorraum. Meist werden wir hier als Körper den Körper der reellen Zahlen benutzen.
Aus den einen Vektorraum definierenden Bedingungen zeigen wir z. B., daß für alle Vektoren $\mathbf{x} \in V$ stets $0\mathbf{x} = \mathbf{o}$ gilt:

$$0\mathbf{x} = (0+0)\mathbf{x} = 0\mathbf{x} + 0\mathbf{x} \Longrightarrow \mathbf{o} = 0\mathbf{x}.$$

Aus der Tatsache, daß $0\mathbf{x} = \mathbf{o}$ gilt, folgert man leicht $(-1)\mathbf{x} = -\mathbf{x}$, wobei hier einmal -1 das zum Einselement $1 \in K$ hinsichtlich der Addition im Körper inverse Element darstellt und andererseits $-\mathbf{x}$ das hinsichtlich der Vektoraddition inverse Element bedeutet:

$$(-1)\mathbf{x} + \mathbf{x} = (-1)\mathbf{x} + 1\mathbf{x} = (1 + (-1)\mathbf{x} = 0\mathbf{x} = \mathbf{o} \Longrightarrow -\mathbf{x} = (-1)\mathbf{x}.$$

Anstelle von $\mathbf{x} + (-\mathbf{y})$ schreiben wir $\mathbf{x} - \mathbf{y}$. Hier drei Standardbeispiele für Vektorräume.

1. Es sei V die Menge aller n-Tupel von reellen Zahlen

$$V = \{\ (x_1, x_2, \ldots, x_n) \mid x_i \in \mathbb{R}, i = 1, \ldots, n\ \}$$

mit der Addition

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n).$$

Die Multiplikation mit einer reellen Zahl wird ebenfalls komponentenweise erklärt:

$$\lambda(x_1, x_2, \ldots, x_n) = (\lambda x_1, \lambda x_2, \ldots, \lambda x_n).$$

Mit diesen Operationen bildet $V(+; \cdot, \mathbb{R})$ einen Vektorraum über dem Körper der reellen Zahlen, den man mit $\mathbb{R}^n$ bezeichnet. Die Zahlen x_i $(i = 1, \ldots, n)$ heißen **Komponenten** des Vektors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

2. Es sei Π_n die Menge aller reellen Polynome vom Grade höchstens n. Jedes Polynom p mit

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

ist durch seine Koeffizienten wohlbestimmt, also durch ein $(n+1)$-Tupel $(a_n, \ldots, a_0)$ von reellen Zahlen. Umgekehrt entspricht jedem solchen Tupel von reellen Zahlen genau ein Polynom vom Grade höchstens n. Zwei Polynome aus Π_n kann man addieren und mit einer reellen Zahl multiplizieren. Damit bildet Π_n einen Vektorraum über dem Körper der reellen Zahlen.

3. Es sei $F(a, b)$ die Menge aller reellwertigen Funktionen auf dem Intervall $[a, b]$ mit der Addition

$$(f + g)(x) = f(x) + g(x) \quad \forall x \in [a, b]$$

und der Multiplikation mit einer reellen Zahl gemäß

$$(\lambda f)(x) = \lambda \cdot f(x) \quad \forall x \in [a, b].$$

Man überzeugt sich sofort, daß $F(a, b)$ Vektorraum über dem Körper der reellen Zahlen ist.

Eine Untermenge $U \subseteqq V$ eines Vektorraumes $V(+; \cdot, K)$ über einem Körper $K(+, \cdot)$ heißt **Unterraum**, wenn U mit der Vektoraddition in V und der Multiplikation zwischen den Elementen aus V und K einen Vektorraum bildet.

Satz 27. *Eine Menge $U \subseteqq V$ eines Vektorraumes $V(+; \cdot, K(+, \cdot))$ ist genau dann ein Unterraum, wenn $\lambda\mathbf{x} + \mu\mathbf{y} \in U$ für alle Vektoren $\mathbf{x}, \mathbf{y} \in U$ und alle Körperelemente $\lambda, \mu \in K$ gilt.*

Beweis. Zum Beweis erwähnen wir nur, daß die Abgeschlossenheit gegenüber der Vektoraddition und der Multiplikation mit einem Körperelement notwendig und hinreichend für einen Unterraum ist, da sich die anderen Bedingungen damit automatisch übertragen. □

So ist z. B. die Menge aller Polynome auf einem gegebenen Intervall $[a, b]$ ein Unterraum des Vektorraumes aller reellwertigen Funktionen auf dem Intervall $[a, b]$.

Es sei X eine beliebige Untermenge des Vektorraumes $V(+; \cdot, K)$. Ein Unterraum U von V heißt **lineare Überdeckung** von X, wenn die Menge X vollständig im Unterraum U liegt: $X \subseteqq U$. Der Durchschnitt von beliebig vielen linearen Überdeckungen ist wieder eine lineare Überdeckung. Den Durchschnitt aller linearen Überdeckungen einer Menge X nennt man **lineare Hülle** und bezeichnet ihn mit $\lin(X)$:

$$\lin(X) = \bigcap_{U : X \subseteqq U} U.$$

Damit ist die lineare Hülle der kleinste Unterraum aus V, in dem die Menge X liegt. Um die lineare Hülle (insbesondere von endlich vielen Vektoren) berechnen zu können, müssen wir die Struktur von Vektorräumen genauer untersuchen. Aus der Definition eines Vektorraumes folgt sofort, daß mit r Vektoren $\mathbf{x}_1, \ldots, \mathbf{x}_r$ aus V jeder Vektor $\mathbf{x}$ der Form

$$\mathbf{x} = \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \cdots + \lambda_r \mathbf{x}_r$$

mit beliebigen Körperelementen $\lambda_i \in K, i = 1, \ldots, r$ auch in V liegt. Einen solchen Vektor $\mathbf{x}$ nennt man **Linearkombination** der Vektoren $\mathbf{x}_1, \ldots, \mathbf{x}_r$. Liegen nun alle Vektoren einer Menge X in einem Unterraum U, so enthält dieser auch alle Linearkombinationen von Vektoren aus X. Daher enthält jede lineare Überdeckung von X alle Linearkombinationen von Vektoren aus X; das gilt auch für die lineare Hülle. Da die lineare Hülle $\lin(X)$ einer Menge X die kleinste lineare Überdeckung von X ist und die Menge aller Linearkombinationen von Vektoren aus X einen Unterraum bildet, folgt

Satz 28. *Die lineare Hülle* $\lin(X)$ *einer Menge* X *von Vektoren aus einem Vektorraum* $V(+;\cdot,K)$ *ist die Menge aller Linearkombinationen von Vektoren aus der Menge* X*, also die Menge*

$$\lin(X) = \left\{ \mathbf{x} \;\middle|\; \mathbf{x} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i, \lambda_i \in K, \mathbf{x}_i \in X, i = 1, \ldots, r, r \in \mathbb{N} \right\}.$$

Dieser Satz macht es möglich, Vektoren der linearen Hülle zu berechnen. Im Falle $\lin(X) = V$ ist die Menge X ein Erzeugendensystem des Vektorraumes V. Es wird sicher viele Erzeugendensysteme für einen Vektorraum geben. Die kleinsten unter ihnen sind die Basen. Eine Untermenge $B \subseteqq V$ heißt **Basis** (**Fundamentalsystem**) für den Vektorraum V, wenn B ein Erzeugendensystem für V ist und kein Erzeugendensystem als echte Untermenge hat, also ein minimales Erzeugendensystem für den Vektorraum V darstellt.
Es sei X ein Erzeugendensystem von V, aber keine Basis: $\lin(X) = V$. Dann gibt es in X einen Vektor $\mathbf{x}$ und $X \setminus \{\mathbf{x}\}$ ist auch ein Erzeugendensystem für V: $\lin(X \setminus \{\mathbf{x}\}) = V$. Folglich gibt es r Vektoren $\mathbf{x}_1, \ldots, \mathbf{x}_r$ aus $X \setminus \{\mathbf{x}\}$ und Elemente $\lambda_1, \ldots, \lambda_r$ aus K mit

$$\mathbf{x} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i.$$

Ist umgekehrt ein Vektor $\mathbf{x} \in X$ auf diese Weise darstellbar, so kann man ihn in jeder Linearkombination von Elementen aus der Menge X durch die rechte Seite ersetzen und erhält so nur Linearkombinationen, die den Vektor $\mathbf{x}$ nicht mehr enthalten, also nur noch Linearkombinationen aus $X \setminus \{\mathbf{x}\}$; folglich gilt $\lin(X) = \lin(X \setminus \{\mathbf{x}\})$. Aus dieser Überlegung folgt der fundamentale Begriff der linearen Algebra: Wir nennen r Vektoren $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ **linear abhängig**, wenn sich mindestens einer von ihnen als Linearkombination der anderen darstellen läßt; andernfalls heißen sie **linear unabhängig**. Häufig verwendet man das folgende Kriterium für die lineare Unabhängigkeit von Vektoren.

Satz 29. *Aus einem Vektorraum* V *über einem Körper* $K(+,\cdot)$ *sind genau dann* r *Vektoren* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ *linear unabhängig, wenn die Gleichung*

$$\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \cdots + \lambda_r \mathbf{x}_r = \mathbf{o}$$

nur die triviale Lösung (d. h. Nullösung) $\lambda_i = 0, i = 1, \ldots, r$ *hat.*

Beweis. Es seien $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ linear unabhängige Vektoren. Angenommen, die Gleichung

$$\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \cdots + \lambda_r \mathbf{x}_r = \mathbf{o}$$

hätte eine nichttriviale Lösung $\lambda_1, \lambda_2, \ldots, \lambda_r$, d. h. mindestens eines der Körperelemente - sagen wir λ_{i_*} - muß vom Nullelement des Körpers K verschieden sein: $\lambda_{i_*} \neq 0$. Dann können wir die Gleichung nach dem Vektor $\mathbf{x}_{i_*}$ auflösen:

$$\mathbf{x}_{i_*} = \frac{\lambda_1}{-\lambda_{i_*}} \mathbf{x}_1 + \cdots + \frac{\lambda_{i_*-1}}{-\lambda_{i_*}} \mathbf{x}_{i_*-1} + \frac{\lambda_{i_*+1}}{-\lambda_{i_*}} \mathbf{x}_{i_*+1} + \cdots + \frac{\lambda_r}{-\lambda_{i_*}} \mathbf{x}_r,$$

also ist der Vektor $\mathbf{x}_{i_*}$ Linearkombination der anderen $r-1$ Vektoren; folglich sind die gegebenen r Vektoren im Widerspruch zur Voraussetzung linear abhängig. Dieser Widerspruch kann nur dadurch aufgelöst werden, daß wir die Annahme fallenlassen, also die fragliche Gleichung nur die triviale Lösung besitzt.
Umgekehrt habe die Gleichung

$$\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \cdots + \lambda_r \mathbf{x}_r = \mathbf{o}$$

nur die triviale Lösung und wir nehmen an, daß die Vektoren $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ linear abhängig sind, also etwa o. B. d. A.

$$\mathbf{x}_1 = \mu_2 \mathbf{x}_2 + \cdots + \mu_r \mathbf{x}_r$$

mit gewissen Elementen $\mu_2, \ldots, \mu_r \in K$ gilt. Aus dieser Gleichung folgt sofort, daß die Gleichung

$$\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \cdots + \lambda_r \mathbf{x}_r = \mathbf{o}$$

die nichttriviale Lösung $\lambda_1 = -1, \lambda_i = \mu_i, i = 2, \ldots, r$ hat, was der Voraussetzung widerspricht. □
Im allgemeinen wird die Darstellung eines Vektors als Linearkombination gewisser anderer nicht eindeutig sein, sofern eine solche Darstellung überhaupt existiert. Bei linear unabhängigen Vektoren liegt aber Eindeutigkeit vor.

Satz 30. *Jeder Vektor eines Vektorraumes läßt sich auf höchstens eine Weise als Linearkombination von r gegebenen linear unabhängigen Vektoren darstellen.*

Beweis. Sind nämlich die Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_r$ linear unabhängig und gilt

$$\mathbf{x} = \sum_{i=1}^{r} \lambda_i \mathbf{a}_i = \sum_{i=1}^{r} \mu_i \mathbf{a}_i,$$

so folgt daraus

$$\mathbf{o} = \sum_{i=1}^{r} \lambda_i \mathbf{a}_i - \sum_{i=1}^{r} \mu_i \mathbf{a}_i = \sum_{i=1}^{r} (\lambda_i - \mu_i) \mathbf{a}_i,$$

woraus sich $\lambda_i - \mu_i = 0, i = 1, \ldots, r$ ergibt. □

Hat man eine beliebige Menge linear unabhängiger Vektoren gegeben, so sind die Vektoren jeder Untermenge davon auch linear unabhängig. Hat man dagegen eine Menge von linear abhängigen Vektoren gegeben, so sind die Vektoren jeder Obermenge linear abhängig. Unsere Überlegungen gestatten es nun, zwei Charakterisierungen für eine Basis auszusprechen.

Satz 31. *Eine Menge $B \subseteq V$ ist genau dann Basis des Vektorraumes V über einem Körper $K(+, \cdot)$, wenn eine der folgenden Bedingungen erfüllt ist:*

1. *Der Vektorraum V wird von B erzeugt, und je endlich viele Vektoren aus B sind linear unabhängig.*
2. *Jeder Vektor aus dem Vektorraum V läßt sich auf genau eine Weise als Linearkombination von endlich vielen Vektoren aus B darstellen.*

Beweis. Die Definition einer Basis können wir wie folgt schreiben:

$$\operatorname{lin}(B) = V, \quad \mathbf{x} \notin \operatorname{lin}(B \setminus \{\mathbf{x}\}) \quad \forall \mathbf{x} \in B.$$

Um die erste Charakterisierung zu beweisen, ist also nur zu zeigen: Je endlich viele Vektoren aus B sind genau dann linear unabhängig, wenn

$$\mathbf{x} \notin \operatorname{lin}(B \setminus \{\mathbf{x}\}) \quad \forall \mathbf{x} \in B$$

gilt. Es seien je endlich viele Vektoren aus B linear unabhängig; angenommen, die Bedingung gilt nicht, d. h. es gibt ein $\mathbf{x} \in B$ und $\mathbf{x} \in \operatorname{lin}(B \setminus \{\mathbf{x}\})$. Dann ist aber der Vektor $\mathbf{x} \in B$ eine Linearkombination von endlich vielen Vektoren aus $B \setminus \{\mathbf{x}\}$, was der Voraussetzung widerspricht. Setzen wir umgekehrt voraus, daß für B die Bedingung

$$\mathbf{x} \notin \operatorname{lin}(B \setminus \{\mathbf{x}\}) \quad \forall \mathbf{x} \in B$$

gilt. Diese Bedingung gilt dann auch für jede endliche Untermenge von B, was gerade die lineare Unabhängigkeit der Vektoren jeder endlichen Untermenge von B bedeutet.
Für die zweite Charakterisierung haben wir zunächst zu zeigen: Wenn B eine Basis von V ist, so ist jeder Vektor $\mathbf{x} \in V$ auf genau eine Weise als Linearkombination von endlich vielen Vektoren aus B darstellbar. Aus der Basiseigenschaft folgt, daß jeder Vektor aus V als endliche Linearkombination von Vektoren aus B darstellbar ist. Nach der 1. Charakterisierung sind aber je endlich viele Vektoren aus B linear unabhängig; mit dem vorangegangenen Satz schließen wir, daß die Darstellung eindeutig ist. Setzen wir nun umgekehrt voraus, daß sich jeder Vektor aus V eindeutig als endliche Linearkombination von Vektoren aus B darstellen läßt. Dann wird V sicherlich von B erzeugt. Angenommen, es gibt in B r linear abhängige Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_r$. Es sei

$$\mathbf{x} = \mu_1 \mathbf{a}_1 + \cdots + \mu_r \mathbf{a}_r$$

mit $\mu_1 \neq 0$ und

$$\mathbf{a}_1 = \lambda_2 \mathbf{a}_2 + \cdots + \lambda_r \mathbf{a}_r.$$

Dann gilt auch

$$\mathbf{x} = 0\mathbf{a}_1 + (\mu_1 \lambda_2 + \mu_2)\mathbf{a}_2 + \cdots + (\mu_1 \lambda_r + \mu_r)\mathbf{a}_r.$$

Folglich ist der Vektor $\mathbf{x}$ auf zwei verschiedene Arten als Linearkombination von Vektoren aus V darstellbar, was der Voraussetzung widerspricht. □

Die eindeutig bestimmten Faktoren in der Darstellung eines Vektors durch Basisvektoren nennt man die **Koordinaten** des Vektors bezüglich der betreffenden Basis. Im $\mathbb{R}^n$ gibt es eine besonders einfache Basis, gegeben durch die sog. n natürlichen Ainheizwäktorn

$$\mathbf{e}_1 = (1,0,0,\ldots,0),\ \mathbf{e}_2 = (0,1,0\ldots,0),\ldots,\mathbf{e}_n = (0,0,0,\ldots,1).$$

Diese Basis wollen wir die **natürliche Basis** des $\mathbb{R}^n$ nennen. Die Koordinaten jedes Vektors aus dem $\mathbb{R}^n$ stimmen mit seinen Komponenten überein, falls man die natürliche Basis wählt.
Aber auch je n Vektoren der Form

$$\mathbf{x}_1 = (x_{11},0,0,\ldots,0), \mathbf{x}_2 = (x_{12},x_{22},0,\ldots,0),\ldots,\mathbf{x}_n = (x_{1n},x_{2n},\ldots,x_{nn})$$

mit $x_{ii} \neq 0$ $(i = 1,\ldots,n)$ bilden eine Basis des $\mathbb{R}^n$. Wenn wir nämlich die Gleichung

$$\sum_{j=1}^{n} \lambda_j \mathbf{x}_j = \mathbf{o}$$

betrachten, so bedeutet sie in Komponentenschreibweise

$$\sum_{j=1}^{n} \lambda_j x_{ij} = 0 \quad i = 1,\ldots,n,$$

und wegen $x_{ij} = 0$ für $j < i$ lauten diese Gleichungen

$$\sum_{j=i}^{n} \lambda_j x_{ij} = 0 \quad i = 1,\ldots,n;$$

die letzte Gleichung $(i = n)$ liefert $\lambda_n x_{nn} = 0$, also $\lambda_n = 0$; setzen wir dies in alle anderen Gleichungen ein, so lautet die $(n-1)$-te Gleichung $\lambda_{n-1}x_{n-1,n-1} = 0$, woraus $\lambda_{n-1} = 0$ folgt usw. bis $\lambda_1 = 0$. Folglich sind die Vektoren $\mathbf{x}_1,\ldots,\mathbf{x}_n$ linear unabhängig.
Grundlegend für die gesamte lineare Algebra ist der **Austauschsatz von Steinitz**:

Satz 32. *Es seien*

$$\mathbf{v}_1,\ldots,\mathbf{v}_n$$

gegebene, linear unabhängige Vektoren eines Vektorraumes $V(+;\cdot,K(+,\cdot))$ *und*

$$\mathbf{w}_1,\ldots,\mathbf{w}_m$$

linear unabhängige Vektoren aus der linearen Hülle $\lin(\mathbf{v}_1,\ldots,\mathbf{v}_n)$.
Dann kann man m *Vektoren*

$$\mathbf{v}_{i_1},\ldots,\mathbf{v}_{i_m}$$

aus $\{\ \mathbf{v}_1,\ldots,\mathbf{v}_n\ \}$ *durch die Vektoren* $\mathbf{w}_1,\ldots,\mathbf{w}_m$ *so ersetzen, daß die Vektoren aus der Menge*

$$(\{\ \mathbf{v}_1,\ldots,\mathbf{v}_n\ \} \setminus \{\ \mathbf{v}_{i_1},\ldots,\mathbf{v}_{i_m}\ \}) \cup \{\ \mathbf{w}_1,\ldots,\mathbf{w}_m\ \}$$

linear unabhängig sind.

Beweis. Der Beweis ist konstruktiv, d. h. es werden auf algorithmischem Wege solche Vektoren aus $\{\ \mathbf{v}_1,\ldots,\mathbf{v}_n\ \}$ gefunden, die durch die Vektoren $\mathbf{w}_l,\ l = 1,\ldots,m$ ersetzbar sind; die Ersetzung wird ausgeführt. Da die Vektoren $\mathbf{w}_1,\ldots,\mathbf{w}_m$ in der linearen Hülle der Vektoren $\mathbf{v}_1,\ldots,\mathbf{v}_n$ liegen, gibt es eine eindeutige Darstellung der Form

$$\mathbf{w}_i = \sum_{j=1}^{n} a_{ij}\mathbf{v}_j,\ i = 1,\ldots,m.$$

In der Gleichung für $\mathbf{w}_1$ ist mindestens eine der Zahlen $a_{1j},\ j = 1,\ldots,n$ ungleich Null; durch Umnumerieren der Vektoren $\mathbf{v}_j$ kann man errreichen, daß $a_{11} \neq 0$ gilt. Wir lösen nun die Gleichung für $\mathbf{w}_1$ nach dem Vektor $\mathbf{v}_1$ auf

$$\mathbf{v}_1 = \frac{1}{a_{11}}\mathbf{w}_1 - \frac{a_{12}}{a_{11}}\mathbf{v}_2 - \cdots - \frac{a_{1n}}{a_{11}}\mathbf{v}_n$$

und setzen das Ergebnis in die übrigen Gleichungen ($i = 2, \ldots, m$) ein:

$$\mathbf{w}_i = \frac{a_{i1}}{a_{11}}\mathbf{w}_1 + \left(a_{i2} - \frac{a_{i1}a_{12}}{a_{11}}\right)\mathbf{v}_2 + \cdots + \left(a_{in} - \frac{a_{i1}a_{1n}}{a_{11}}\right)\mathbf{v}_n.$$

Zusammen erhalten wir damit Gleichungen der Form

$$\mathbf{v}_1 = a_{11}^{(1)}\mathbf{w}_1 + \sum_{j=2}^{n} a_{1j}^{(1)}\mathbf{v}_j,$$
$$\mathbf{w}_i = a_{i1}^{(1)}\mathbf{w}_1 + \sum_{j=2}^{n} a_{ij}^{(1)}\mathbf{v}_j,\ i = 2, \ldots, m$$

mit $a_{11}^{(1)} \neq 0$. Nehmen wir für einen Augenblick an, daß die Vektoren $\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ linear abhängig sind. Dann muß $\mathbf{w}_1$ Linearkombination der übrigen Vektoren sein, was aber der obigen eindeutigen Darstellung von $\mathbf{w}_1$ als Linearkombination der Vektoren $\mathbf{v}_1, \ldots, \mathbf{v}_n$ widerspricht, in der der Faktor am Vektor $\mathbf{v}_1$ ungleich Null ist. Dieser Widerspruch zeigt uns, daß die Vektoren $\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ linear unabhängig sind:

$$\operatorname{lin}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \operatorname{lin}(\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n).$$

Betrachten wir die obige neue Gleichung für $\mathbf{w}_2$; in ihr muß eine der Zahlen $a_{2j}^{(1)}$ ($j = 2, \ldots, n$) ungleich Null sein, da andernfalls die Vektoren $\mathbf{w}_1, \mathbf{w}_2$ linear abhängig wären, was unserer Voraussetzung widerspricht. Durch Umnumerieren der Vektoren $\mathbf{v}_2, \ldots, \mathbf{v}_n$ kann man erreichen, daß $a_{22}^{(1)} \neq 0$ gilt. Somit ist es möglich, die Gleichung für $\mathbf{w}_2$ nach $\mathbf{v}_2$ aufzulösen und das Ergebnis in alle übrigen Gleichungen einzusetzen, womit wir Gleichungen der Form

$$\mathbf{v}_1 = a_{11}^{(2)}\mathbf{w}_1 + a_{12}^{(2)}\mathbf{w}_2 + \sum_{j=3}^{n} a_{1j}^{(2)}\mathbf{v}_j,$$
$$\mathbf{v}_2 = a_{21}^{(2)}\mathbf{w}_1 + a_{22}^{(2)}\mathbf{w}_2 + \sum_{j=3}^{n} a_{2j}^{(2)}\mathbf{v}_j,$$
$$\mathbf{w}_i = a_{i1}^{(2)}\mathbf{w}_1 + a_{i2}^{(2)}\mathbf{w}_2 + \sum_{j=3}^{n} a_{ij}^{(2)}\mathbf{v}_j,\ i = 3 \ldots, m$$

erhalten. Nach r Schritten entstehen Gleichungen der Form

$$\mathbf{v}_i = \sum_{j=1}^{r} a_{ij}^{(r)}\mathbf{w}_j + \sum_{j=r+1}^{n} a_{ij}^{(r)}\mathbf{v}_j,\ i = 1, \ldots, r,$$
$$\mathbf{w}_i = \sum_{j=1}^{r} a_{ij}^{(r)}\mathbf{w}_j + \sum_{j=r+1}^{n} a_{ij}^{(r)}\mathbf{v}_j,\ i = r+1, \ldots, m,$$

und die Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n$ sind linear unabhängig. Wir betrachten die Gleichung für den Vektor $\mathbf{w}_{r+1}$; unter den Faktoren an den Vektoren $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n$ ist mindestens einer ungleich Null, da sonst die Gleichung besagen würde, daß der Vektor $\mathbf{w}_{r+1}$ linear abhängig von den Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_r$ ist, was wegen der Voraussetzung aber ausgeschlossen ist. Ohne Beschränkung der Allgemeinheit sei der Faktor vor dem Vektor $\mathbf{v}_{r+1}$ ungleich Null, was durch Umnumerieren stets erreichbar ist. Damit können wir die Gleichung für den Vektor $\mathbf{w}_{r+1}$ nach dem Vektor $\mathbf{v}_{r+1}$ auflösen und das Ergebnis in alle anderen Gleichungen einsetzen; danach tritt auf den rechten Seiten aller m Gleichungen anstelle des Vektors $\mathbf{v}_{r+1}$ der Vektor $\mathbf{w}_{r+1}$ auf. Angenommen, die Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_r, \mathbf{w}_{r+1}, \mathbf{v}_{r+2}, \ldots, \mathbf{v}_n$ sind linear abhängig, d. h. es gelte eine Gleichung der Form

$$\lambda\mathbf{w}_{r+1} + \sum_{j=1}^{r} \lambda_j\mathbf{w}_j + \sum_{j=r+2}^{n} \lambda_j\mathbf{v}_j = \mathbf{o},$$

wobei nicht alle Faktoren an den Vektoren gleich Null sind. Da die Vektoren

$$\mathbf{w}_1, \ldots, \mathbf{w}_r, \mathbf{v}_{r+2}, \ldots, \mathbf{v}_n$$

linear unabhängig sind, muß $\lambda \neq 0$ sein; wir können die Gleichung nach dem Vektor $\mathbf{w}_{r+1}$ auflösen und erhalten ihn als Linearkombination der $n-1$ Vektoren

$$\mathbf{w}_1, \ldots, \mathbf{w}_r, \mathbf{v}_{r+2}, \ldots, \mathbf{v}_n,$$

was aber unmöglich sein kann, da in der obigen Darstellung von $\mathbf{w}_{r+1}$ durch die Vektoren

$$\mathbf{w}_1, \ldots, \mathbf{w}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n$$

der Faktor vor dem Vektor $\mathbf{v}_{r+1}$ ungleich Null ist. Dieser Widerspruch beweist, daß die genannten Vektoren linear unabhängig sind.
Nach m Schritten erhalten wir so ein Gleichungssystem der Form

$$\mathbf{v}_i = \sum_{j=1}^{m} a_{ij}^{(m)} \mathbf{w}_j + \sum_{j=m+1}^{n} a_{ij}^{(m)} \mathbf{v}_j,\ i = 1, \ldots, m,$$

und die Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{v}_{m+1}, \ldots, \mathbf{v}_n$ sind linear unabhängig. □
Der Austauschsatz von Steinitz hilft uns z. B. bei folgender Frage: Wie kann man vorgegebene linear unabhängige Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_m$ zu einer Basis ergänzen? Nach dem Austauschsatz bietet sich das folgende Verfahren an: Man nehme eine Basis B des Vektorraumes; die Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_m$ sind dann Linearkombinationen aus endlich vielen Vektoren $\mathbf{v}_1, \ldots, \mathbf{v}_n$ aus B. Nach dem Austauschsatz kann man daraus m Vektoren gegen $\mathbf{w}_1, \ldots, \mathbf{w}_m$ austauschen; das so entstehende neue System von Vektoren bildet dann eine Basis des betreffenden Vektorraumes.
Eine weitere Folge aus dem Austauschsatz ist der Dimensionsbegriff. Nach dem Austauschsatz gilt für einen Vektorraum $V(+; \cdot, K(+, \cdot))$ mit endlicher Basis $B = \{ \mathbf{b}_1, \ldots, \mathbf{b}_n \}$, daß nur höchstens n beliebig ausgewählte Vektoren linear unabhängig sein können, also je $n+1$ Vektoren linear abhängig sind. Gäbe es nämlich $n+1$ linear unabhängige Vektoren $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{n+1}$, so könnte man nach dem Austauschsatz n Vektoren davon durch die Vektoren $\mathbf{b}_1, \ldots, \mathbf{b}_n$ austauschen; seien dies etwa die Vektoren $\mathbf{x}_1, \ldots, \mathbf{x}_n$; die Vektoren $\mathbf{b}_1, \ldots, \mathbf{b}_n, \mathbf{x}_{n+1}$ wären dann linear unabhängig, was aber unmöglich sein kann, da $\mathbf{x}_{n+1} \in \operatorname{lin}(\mathbf{b}_1, \ldots, \mathbf{b}_n) = V$ gilt. Also sind nur zwei sich ausschließende Fälle möglich:

1. Der Vektorraum $V(+; \cdot, K)$ hat eine endliche Basis mit n Vektoren. In diesem Falle hat jede Basis aus $V(+; \cdot, K)$ genau n Vektoren.

2. Es gibt keine endliche Basis in $V(+; \cdot, K)$; jede endliche Untermenge erzeugt einen echten Unterraum von V.

Im ersten Falle sagen wir, daß der Vektorraum die **Dimension** n hat, d. h. die Dimension ist in diesem Falle die maximale Anzahl linear unabhängiger Vektoren:

$$\dim_K V = n.$$

Im zweiten Falle ist der Vektorraum unendlichdimensional:

$$\dim_K V = \infty.$$

Im unendlichdimensionalen Fall unterscheidet man noch zwischen abzählbar und überabzählbar unendlich. Im abzählbaren Fall existiert eine abzählbare Basis, d. h. eine Basis mit der Mächtigkeit der natürlichen Zahlen. So ist z. B. der Vektorraum aller n-Tupel von reellen Zahlen über dem Körper der reellen Zahlen n-dimensional, während der Vektorraum aller auf einem Intervall reellwertigen Funktionen über dem Körper der reellen Zahlen sicher unendlichdimensional ist. Der Vektorraum aller Polynome über dem Körper der reellen Zahlen hat eine abzählbare Basis, nämlich die Vektoren

$$1, x, x^2, x^3, \ldots, x^n, \ldots.$$

Es sei nun U ein Unterraum eines Vektorraumes $V(+; \cdot, K)$, B eine Basis von U und $\overline{B}$ eine Ergänzung von B zu einer Basis von $V(+; \cdot, K)$:

$$B \cap \overline{B} = \emptyset,\ \operatorname{lin}\left(B \cup \overline{B}\right) = V.$$

Dann ist $\operatorname{lin}\left(\overline{B}\right)$ ein Unterraum von $V(+; \cdot, K)$ mit folgenden Eigenschaften:

$$U + \operatorname{lin}\left(\overline{B}\right) = \left\{ \mathbf{u} + \mathbf{v} \mid \mathbf{u} \in U, \mathbf{v} \in \operatorname{lin}\left(\overline{B}\right) \right\} = V,\ \operatorname{lin}\left(\overline{B}\right) \cap U = \{\mathbf{o}\}.$$

Der Unterraum $\operatorname{lin}\left(\overline{B}\right)$ ergänzt also den Unterraum U zum gesamten Vektorraum; daher heißt $\operatorname{lin}\left(\overline{B}\right)$ **Komplementraum** (**algebraisches Komplement**) zu U in V. Ein Komplementraum zu U ist einerseits abhängig von der gewählten Basis in U und andererseits auch abhängig von der gewählten Ergänzung zu einer Basis des Raumes $V(+; \cdot, K)$; daher gibt es zu einem Unterraum i. a. mehrere Komplementräume. Für den Faktorraum $V/U(+; \cdot, K)$ gilt dann für jeden Komplementraum $\overline{U}$ zu U:

$$V/U = \left\{ \mathbf{x} + U \mid \mathbf{x} \in \overline{U} \right\}.$$

Beispiel. Die Menge $U = \{ \lambda(1,1,1) \mid \lambda \in \mathbb{R} \}$ ist ein Unterraum des $\mathbb{R}^3$ und hat z. B. die Basis $B = \{ (1,1,1) \}$; durch die natürlichen Einheitsvektoren $\mathbf{e}_2, \mathbf{e}_3$ kann man sie zu einer Basis von $\mathbb{R}^3$ ergänzen:

$$\overline{B} = \{ \mathbf{e}_2, \mathbf{e}_3 \} \text{ und } \operatorname{lin}\left(\overline{B}\right) = \{ (0,\lambda,\mu) \mid \lambda,\mu \in \mathbb{R} \}.$$

Andererseits kann man auch die Vektoren $\mathbf{e}_1, \mathbf{e}_2$ als Ergänzung wählen und erhält als Komplementraum

$$\operatorname{lin}\left(\overline{B}\right) = \{ (\lambda,\mu,0) \mid \lambda,\mu \in \mathbb{R} \}.$$

Die Konstruktion eines Komplementraumes $\overline{U}$ eines Unterraumes U von $V(+;\cdot,K)$ liefert:

$$\dim_K U + \dim_K \overline{U} = \dim_K V.$$

Allgemein gilt

Satz 33 (Dimensionssatz für Unterräume). *Für endlichdimensionale Unterräume U, V eines Vektorraumes gilt*

$$\dim(U+V) + \dim(U \cap V) = \dim U + \dim V.$$

Beweis. Es sei $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ eine Basis von $U \cap V$; diese ergänzen wir zu einer Basis von U:

$$\operatorname{lin}(\mathbf{b}_1, \ldots, \mathbf{b}_n, \mathbf{c}_1, \ldots, \mathbf{c}_l) = U.$$

Nun ergänzen wir $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ zu einer Basis von V:

$$\operatorname{lin}(\mathbf{b}_1, \ldots, \mathbf{b}_n, \mathbf{d}_1, \ldots, \mathbf{d}_m) = V.$$

Wir zeigen, daß $\{\mathbf{b}_1, \ldots, \mathbf{b}_n, \mathbf{c}_1, \ldots, \mathbf{c}_l, \mathbf{d}_1, \ldots, \mathbf{d}_m\}$ eine Basis von $U+V$ ist, womit die Behauptung bewiesen wäre. Zunächst ist klar, daß diese Vektoren den Raum $U+V$ erzeugen. Es bleibt also nur die lineare Unabhängigkeit zu zeigen. Es sei

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{b}_i, \quad \mathbf{y} = \sum_{j=1}^{l} y_j \mathbf{c}_j, \quad \mathbf{z} = \sum_{k=1}^{m} z_k \mathbf{d}_k, \quad \mathbf{x}+\mathbf{y}+\mathbf{z} = \mathbf{o}$$

und wir haben $\mathbf{x} = \mathbf{y} = \mathbf{z} = \mathbf{o}$ zu zeigen. Es gilt $\mathbf{x} \in U \cap V, \mathbf{y} \in U, \mathbf{z} \in V$ und $-\mathbf{z} = \mathbf{x}+\mathbf{y}$, also $\mathbf{z} \in U$ und damit $\mathbf{z} \in U \cap V$. Nun ist $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ aber eine Basis von $U \cap V$, woraus $\mathbf{z} = \mathbf{o}$ folgt. Analog erhält man $\mathbf{y} = \mathbf{o}$, so daß $\mathbf{x} = \mathbf{o}$ folgt. □

Wir fragen nun danach, welche Eigenschaften Homomorphismen auf Vektorräumen haben. Es seien also zwei Vektorräume

$$V(+;\cdot,K),\ \overline{V}(+;\cdot,\overline{K})$$

und ein Homomorphismus φ von V in $\overline{V}$ gegeben. Die Operationen bezeichnen wir in beiden Strukturen mit dem gleichen Symbol. Nach der Homomorphiebedingung muß zunächst

$$\varphi(\mathbf{x}+\mathbf{y}) = \varphi(\mathbf{x}) + \varphi(\mathbf{y})$$

für alle $\mathbf{x}, \mathbf{y} \in V$ gelten. Außerdem muß eine Verträglichkeitsbedingung für die Multiplikation von Vektoren mit Elementen aus den Körpern erfüllt sein:

$$\varphi(\lambda \cdot \mathbf{x}) = \varphi(\lambda) \cdot \varphi(\mathbf{x}).$$

Mit dieser Verträglichkeitsbedingung schließen wir:

$$\begin{aligned} \varphi(\lambda+\mu) \cdot \varphi(\mathbf{x}) &= \varphi\left((\lambda+\mu)\cdot\mathbf{x}\right) \\ &= \varphi(\lambda\mathbf{x}+\mu\mathbf{x}) = \varphi(\lambda\mathbf{x}) + \varphi(\mu\mathbf{x}) \\ &= \varphi(\lambda)\cdot\varphi(\mathbf{x}) + \varphi(\mu)\cdot\varphi(\mathbf{x}) \\ &= (\varphi(\lambda)+\varphi(\mu))\cdot\varphi(\mathbf{x}) \end{aligned}$$

also

$$(\varphi(\lambda+\mu) - (\varphi(\lambda)+\varphi(\mu)))\,\varphi(\mathbf{x}) = \mathbf{o},$$

was die Bedingung

$$\varphi(\lambda+\mu) = \varphi(\lambda) + \varphi(\mu)$$

erzwingt. Analog folgt die Bedingung

$$\varphi(\lambda \cdot \mu) = \varphi(\lambda) \cdot \varphi(\mu).$$

Wenn wir berücksichtigen, daß die Strukturabbildung φ bijektiv zwischen beiden Körpern wirkt, so erhalten wir, daß ein Homomorphismus nur zwischen solchen Vektorräumen existieren kann, wo die betreffenden Körper isomorph sind. Wir betrachten daher nur Homomorphismen zwischen Vektorräumen über den gleichen Körpern K. Die Verträglichkeitsbedingung lautet damit:

$$\varphi(\lambda \cdot \mathbf{x}) = \lambda \cdot \varphi(\mathbf{x}) \; \forall \mathbf{x} \in V, \forall \lambda \in K.$$

Beide Bedingungen fassen wir zu einer zusammen und führen dafür den Begriff **lineare Abbildung** ein. Eine Strukturabbildung φ eines Vektorraumes $V(+; \cdot, K)$ in einen Vektorraum $\overline{V}(+; \cdot, K)$ heißt **linear**, wenn für alle $\mathbf{x}, \mathbf{y} \in V$ und alle $\lambda, \mu \in K$ gilt:

$$\varphi(\lambda \cdot \mathbf{x} + \mu \cdot \mathbf{y}) = \lambda \cdot \varphi(\mathbf{x}) + \mu \cdot \varphi(\mathbf{y}).$$

Es sei noch einmal bemerkt, daß die linearen Abbildungen gerade die Homomorphismen auf Vektorräumen sind und daher die für Homomorphismen gefundenen Eigenschaften direkt übertragen werden können. Unseren Studien zu Homomorphismen folgend gilt somit der folgende Satz für lineare Abbildungen.

Satz 34. *Lineare Abbildungen zwischen Vektorräumen haben folgende Eigenschaften.*

1. *Das Bild eines Vektorraumes bei einer linearen Abbildung ist wieder ein Vektorraum.*
2. *Eine lineare Abbildung überführt Unterräume in Unterräume.*
3. *Das Urbild eines Unterraumes ist ein Unterraum im Urbildraum.*

Weiterhin können wir aus den allgemeinen Betrachtungen über Unterstrukturen und Kongruenzrelationen sofort den nächsten Satz aussprechen.

Satz 35. *Die Unterräume und die Kongruenzen eines Vektorraumes entsprechen einander umkehrbar eindeutig: Ist R eine Kongruenz auf dem Raum $V(+; \cdot, K)$, so entspricht ihr der Unterraum aller zum Nullvektor kongruenten Vektoren: $R \longmapsto U = [\mathbf{o}]_R$.*
Jedem Unterraum U entspricht die Kongruenz R_U auf $V(+; \cdot, K)$ mit

$$\mathbf{x} R_U \mathbf{y} \text{ genau dann, wenn } \mathbf{x} + U = \mathbf{y} + U.$$

Nach dem Homomorphiesatz ist das Bild $\varphi(V)$ eines gegebenen Raumes $V(+; \cdot, K)$ bei einer linearen Abbildung φ isomorph zum Faktorraum V/U mit

$$U = \ker(\varphi) = \{ \mathbf{x} \mid \varphi(\mathbf{x}) = \mathbf{o} \};$$

dabei sind die Operationen auf

$$V/U = \{ \mathbf{x} + U \mid \mathbf{x} \in V \}$$

wie folgt definiert:

$$(\mathbf{x} + U) \bar{+} (\mathbf{y} + U) = (\mathbf{x} + \mathbf{y}) + U, \; \lambda \bar{\cdot} (\mathbf{x} + U) = \lambda \cdot \mathbf{x} + U.$$

Daraus folgern wir, daß eine lineare Abbildung genau dann injektiv ist, wenn ihr Kern nur aus dem Nullvektor besteht.

Satz 36. *Es sei $V(+; \cdot, K)$ ein Vektorraum mit einer Basis. Jede lineare Abbildung, die den Vektorraum V in einen Vektorraum $\overline{V}$ abbildet, ist eindeutig bestimmt durch die Vorgabe der Bildvektoren für alle Vektoren einer beliebig fixierten Basis.*

Beweis. Es sei B eine Basis aus V, der Vektor $\mathbf{x}$ beliebig aus V gewählt und φ eine lineare Abbildung auf V. Wir haben zu zeigen, daß das Bild $\varphi(\mathbf{x})$ eindeutig bestimmt ist, wenn man $\varphi(B)$ vorgibt.
Zum Vektor $\mathbf{x}$ gibt es r Vektoren $\mathbf{b}_1, \ldots, \mathbf{b}_r \in B$, so daß sich $\mathbf{x}$ eindeutig als Linearkombination dieser r Vektoren darstellen läßt

$$\mathbf{x} = \sum_{i=1}^{r} \lambda_i \mathbf{b}_i, \; \lambda_i \in K, \; i = 1, \ldots, r.$$

Mit der Linearität von φ schließen wir

$$\varphi(\mathbf{x}) = \varphi(\sum_{i=1}^{r} \lambda_i \mathbf{b}_i) = \sum_{i=1}^{r} \lambda_i \varphi(\mathbf{b}_i).$$

Sind also die Bilder einer Basis vorgegeben, so kann man über die Koordinaten eines Vektors $\mathbf{x}$ bezüglich einer Basis das Bild $\varphi(\mathbf{x})$ ermitteln, denn es ist die Linearkombination der Bildvektoren aus der Basis mit den Koordinaten aus dem Urbildraum als Faktoren. □
Da man bei Vorgabe der Bilder einer beliebig gewählten, dann aber fixierten Basis nach der obigen Formel das Bild jedes Vektors berechnen kann, gilt auch die Umkehrung.

Satz 37. *Durch die Vorgabe der Bilder einer beliebig gewählten Basis aus einem Vektorraum ist genau eine lineare Abbildung vollständig charakterisiert.*

Wie wir bereits wissen, besteht bei einer linearen Abbildung φ auf einem Vektorraum V der Kern von φ aus genau den Vektoren, die auf den Nullvektor im Bildraum abgebildet werden:

$$\ker(\varphi) = \{\, \mathbf{x} \in V \mid \varphi(\mathbf{x}) = \mathbf{o} \,\}.$$

Der Kern einer linearen Abbildung ist ein Unterraum U von V. Es sei nun $\overline{U}$ ein Komplementraum in V zum Kern U und B eine Basis von $\overline{U}$. Jeder Vektor $\mathbf{x}$ aus V läßt sich dann als Summe $\mathbf{u} + \overline{\mathbf{u}}$ mit $\mathbf{u} \in U, \overline{\mathbf{u}} \in \overline{U}$ darstellen und $\overline{\mathbf{u}}$ ist Linearkombination von Basisvektoren aus B. Also ist jeder Vektor $\mathbf{y}$ aus dem Bildraum $\varphi(V)$ von der Form

$$\mathbf{y} = \varphi(\mathbf{x}) = \varphi(u + \sum_{i=1}^{r} \lambda_i \overline{\mathbf{u}}_i) = \sum_{i=1}^{r} \lambda_i \varphi(\overline{\mathbf{u}}_i)$$

mit $\mathbf{u} \in U = \ker(\varphi), \overline{\mathbf{u}}_1, \ldots, \overline{\mathbf{u}}_r \in B$, d. h. $\varphi(B)$ ist ein Erzeugendensystem von $\varphi(V)$.

Satz 38. *Das Bild jeder Basis eines Komplementraumes zum Kern einer linearen Abbildung ist eine Basis im Bildraum.*

Beweis. Es sei $\overline{U}$ ein Komplementraum zum Kern $\ker(\varphi)$ und B eine Basis von $\overline{U}$. Wir haben zu zeigen, daß $\varphi(B)$ eine Basis im Bildraum $\varphi(V)$ ist. Es seien dazu $\varphi(\mathbf{b}_1)), \ldots, \varphi(\mathbf{b}_r)) \in \varphi(B)$ beliebig gewählt. Dann gilt

$$\sum_{i=1}^{r} \lambda_i \varphi(\mathbf{b}_i)) = \mathbf{o} \iff \sum_{i=1}^{r} \lambda_i \mathbf{b}_i \in \ker(\varphi) = U.$$

Da die Vektoren $\mathbf{b}_1, \ldots, \mathbf{b}_r$ aus dem Komplementraum $\overline{U}$ von U sind und beide nur den Nullvektor gemeinsam haben, folgt

$$\sum_{i=1}^{r} \lambda_i \mathbf{b}_i = \mathbf{o},$$

woraus wir mit der linearen Unabhängigkeit der Vektoren $\mathbf{b}_1, \ldots, \mathbf{b}_r$ schließen, daß

$$\lambda_i = 0, i = 1, \ldots, r$$

sein muß. Damit haben wir gezeigt, daß beliebig ausgewählte Vektoren aus $\varphi(B)$ linear unabhängig sind. □

Satz 39. *Für jeden endlichdimensionalen Vektorraum $V(+;\cdot,K)$ und jede lineare Abbildung φ auf ihm gilt:*

$$\dim_K V = \dim_K \ker(\varphi) + \dim_K \varphi(V) \geqq \dim_K \varphi(V).$$

Beweis. Die Aussage folgt aus den vorangegangenen Überlegungen. Wegen

$$V/U = \left\{\, \mathbf{x} + U \mid \mathbf{x} \in \overline{U} \,\right\}$$

und des Homomorphisatzes ergibt sich

$$\dim_K V = \dim_K U + \dim_K \overline{U} = \dim_K \ker(\varphi) + \dim_K \varphi(V) \geqq \dim_K \varphi(V),$$

was gerade im Satz behauptet wurde. □
Bei einer linearen Abbildung kann also ein Dimensionsverlust, niemals ein Dimensionsgewinn eintreten. Diese Aussage bezieht sich auf das Bild, nicht auf den Raum, in den abgebildet wird. Sehr wohl kann man von einem Vektorraum niederer Dimension in einen Vektorraum höherer Dimension abbilden. Eine solche Abbildung ist jedoch nicht mit einem Informationsgewinn verbunden.

2.2. Algorithmen zum Austauschsatz

Der Beweis des Austauschsatzes ist konstruktiv. Mathematische Sätze, zu denen man einen konstruktiven Beweis hat, haben oft wichtige Anwendungen, weil man direkt aus dem Beweis einen Algorithmus ziehen kann. Es gibt sogar Mathematiker, die nur mathematische Sätze mit einem konstruktiven Beweis akzeptieren. Die Bedeutung des Austauschsatzes von Steinitz liegt darin, daß der Beweis des Satzes konstruktive Methoden der linearen Algebra begründet. Im algorithmischen Teil des Beweises wird die eindeutige Darstellung

$$\mathbf{w}_i = \sum_{j=1}^{n} a_{ij}\mathbf{v}_j,\ i = 1, \ldots, m.$$

nach gewissen m Vektoren $\mathbf{v}_{j_i}, (i = 1, \ldots, m)$ aus $\mathbf{v}_1, \ldots, \mathbf{v}_n$ aufgelöst, so daß diese Vektoren als Linearkombination der Vektoren $\mathbf{w}_1, \ldots \mathbf{w}_m$ und der Vektoren aus $\{ \mathbf{v}_1, \ldots, \mathbf{v}_n \} \setminus \{ \mathbf{v}_{j_1}, \ldots, \mathbf{v}_{j_m} \}$ dargestellt sind. Im Beweis ist enthalten, welche Vektoren $\mathbf{v}_{j_i}$ dafür genommen werden dürfen. Folglich besteht der Beweis aus 2 Teilen, die gemischt auftreten: einem Begründungsteil und einem algorithmischen Teil. Im Begründungsteil wird nachgewiesen, warum gewisse Operationen bzw. Beweisschritte ausführbar sind. Der algorithmische Teil vollzieht sich hier ausschließlich auf den obigen Linearkombinationen. Diese stellen wir zweckmäßigerweise in Tabellenform dar:

	$\mathbf{v}_1$	$\mathbf{v}_2$	$\ldots$	$\mathbf{v}_n$
$\mathbf{w}_1$	a_{11}	a_{12}	$\ldots$	a_{1n}
$\mathbf{w}_2$	a_{21}	a_{22}	$\ldots$	a_{2n}
$\ldots$	$\ldots\ldots\ldots\ldots\ldots\ldots$			
$\mathbf{w}_m$	a_{m1}	a_{m2}	$\ldots$	a_{mn} .

Hierin ist die $\mathbf{w}_i$-Zeile nur eine andere Darstellungsform für die Linearkombination

$$\mathbf{w}_i = a_{i1}\mathbf{v}_1 + a_{i2}\mathbf{v}_2 + \cdots + a_{in}\mathbf{v}_n = \sum_{j=1}^{n} a_{ij}\mathbf{v}_j.$$

Dem Beweis des Austauschsatzes folgend ist die Gleichung für $\mathbf{w}_1$ nach einem Vektor $\mathbf{v}_{j_1}$ aufzulösen, was natürlich nur dann möglich ist, wenn der entsprechende Faktor a_{1j_1} ungleich Null ist. Es sei etwa $a_{11} \neq 0$, also $j_1 = 1$. Auflösen der 1. Gleichung nach $\mathbf{v}_1$ und einsetzen in die übrigen liefert:

$$\mathbf{v}_1 = \frac{1}{a_{11}}(\mathbf{w}_1 - a_{12}\mathbf{v}_2 - a_{13}\mathbf{v}_3 - \cdots - a_{1n}\mathbf{v}_n)$$

$$\mathbf{w}_i = \frac{1}{a_{11}}(a_{i1}\mathbf{w}_1 + (a_{11}a_{i2} - a_{i1}a_{12})\mathbf{v}_2 + \cdots + (a_{11}a_{in} - a_{i1}a_{1n})\mathbf{v}_n),$$

wobei die letzte Gleichung für $i = 2, \ldots, m$ gilt. In Tabellenform geschrieben lauten diese Linearkombinationen:

	$\mathbf{w}_1$	$\mathbf{v}_2$	$\ldots$	$\mathbf{v}_n$
$\mathbf{v}_1$	$a_{11}^{(1)}$	$a_{12}^{(1)}$	$\ldots$	$a_{1n}^{(1)}$
$\mathbf{w}_2$	$a_{21}^{(1)}$	$a_{22}^{(1)}$	$\ldots$	$a_{2n}^{(1)}$
$\ldots$	$\ldots\ldots\ldots\ldots\ldots\ldots$			
$\mathbf{w}_m$	$a_{m1}^{(1)}$	$a_{m2}^{(1)}$	$\ldots$	$a_{mn}^{(1)}$

mit

$$a_{11}^{(1)} = \frac{1}{a_{11}}, \quad a_{1j}^{(1)} = -a_{1j}a_{11}^{(1)}, \qquad j = 2, \ldots, n$$

$$a_{ij}^{(1)} = a_{ij} + a_{i1}a_{1j}^{(1)}, \qquad i = 2, \ldots, m; j = 2, \ldots, n.$$

$$a_{i1}^{(1)} = a_{i1}a_{11}^{(1)}, \qquad i = 2, \ldots, m.$$

Nach r Schritten haben wir ein System von Linearkombinationen erreicht, das durch folgende Tabelle repräsentiert wird:

	$\mathbf{w}_1$	$\mathbf{w}_2$	$\cdots$	$\mathbf{w}_r$	$\mathbf{v}_{r+1}$	$\cdots$	$\mathbf{v}_n$
$\mathbf{v}_1$	$a_{11}^{(r)}$	$a_{12}^{(r)}$	$\cdots$	$a_{1r}^{(r)}$	$a_{1,r+1}^{(r)}$	$\cdots$	$a_{1n}^{(r)}$
$\mathbf{v}_2$	$a_{21}^{(r)}$	$a_{22}^{(r)}$	$\cdots$	$a_{2r}^{(r)}$	$a_{2,r+1}^{(r)}$	$\cdots$	$a_{2n}^{(r)}$
$\cdots$	$\cdots$						
$\mathbf{v}_r$	$a_{r1}^{(r)}$	$a_{r2}^{(r)}$	$\cdots$	$a_{rr}^{(r)}$	$a_{r,r+1}^{(r)}$	$\cdots$	$a_{rn}^{(r)}$
$\mathbf{w}_{r+1}$	$a_{r+1,1}^{(r)}$	$a_{r+1,2}^{(r)}$	$\cdots$	$a_{r+1,r}^{(r)}$	$a_{r+1,r+1}^{(r)}$	$\cdots$	$a_{r+1,n}^{(r)}$
$\cdots$	$\cdots$						
$\mathbf{w}_m$	$a_{m1}^{(r)}$	$a_{m2}^{(r)}$	$\cdots$	$a_{mr}^{(r)}$	$a_{m,r+1}^{(r)}$	$\cdots$	$a_{mn}^{(r)}$.

Dem Beweis des Austauschsatzes folgend haben wir nun die Linearkombination für einen der Vektoren $\mathbf{w}_i, i = r+1, \ldots, m$ nach einem der Vektoren $\mathbf{v}_j, j = r+1, \ldots, n$ aufzulösen. Für diesen Akt dürfen alle jene Vektorpaare $(\mathbf{w}_i, \mathbf{v}_j)$ benutzt werden, bei denen der Faktor $a_{ij}^{(r)} \neq 0$ ist. Es sei etwa $a_{r+1,r+1}^{(r)} \neq 0$. Wir lösen die Linearkombination für den Vektor $\mathbf{w}_{r+1}$ nach dem Vektor $\mathbf{v}_{r+1}$ auf und setzen das Ergebnis in die übrigen Gleichungen ein. So erhalten wir die neue Tabelle

	$\mathbf{w}_1$	$\mathbf{w}_2$	$\cdots$	$\mathbf{w}_{r+1}$	$\mathbf{v}_{r+2}$	$\cdots$	$\mathbf{v}_n$
$\mathbf{v}_1$	$a_{11}^{(r+1)}$	$a_{12}^{(r+1)}$	$\cdots$	$a_{1,r+1}^{(r+1)}$	$a_{1,r+2}^{(r+1)}$	$\cdots$	$a_{1n}^{(r+1)}$
$\mathbf{v}_2$	$a_{21}^{(r+1)}$	$a_{22}^{(r+1)}$	$\cdots$	$a_{2,r+1}^{(r+1)}$	$a_{2,r+2}^{(r+1)}$	$\cdots$	$a_{2n}^{(r+1)}$
$\cdots$	$\cdots$						
$\mathbf{v}_{r+1}$	$a_{r+1,1}^{(r+1)}$	$a_{r+1,2}^{(r+1)}$	$\cdots$	$a_{r+1,r+1}^{(r+1)}$	$a_{r+1,r+2}^{(r+1)}$	$\cdots$	$a_{r+1,n}^{(r+1)}$
$\mathbf{w}_{r+2}$	$a_{r+2,1}^{(r+1)}$	$a_{r+2,2}^{(r+1)}$	$\cdots$	$a_{r+2,r+1}^{(r+1)}$	$a_{r+2,r+2}^{(r+1)}$	$\cdots$	$a_{r+2,n}^{(r+1)}$
$\cdots$	$\cdots$						
$\mathbf{w}_m$	$a_{m1}^{(r+1)}$	$a_{m2}^{(r+1)}$	$\cdots$	$a_{m,r+1}^{(r+1)}$	$a_{m,r+2}^{(r+1)}$	$\cdots$	$a_{mn}^{(r+1)}$,

wobei sich die Faktoren in den neuen Linearkombinationen nach folgenden Formeln berechnen:

$$a_{r+1,r+1}^{(r+1)} = \frac{1}{a_{r+1,r+1}^{(r)}},\ a_{i,r+1}^{(r+1)} = a_{i,r+1}^{(r)} a_{r+1,r+1}^{(r+1)},\ i = 1, \ldots, m; i \neq r+1,$$

$$a_{ij}^{(r+1)} = a_{ij}^{(r)} - a_{i,r+1}^{(r+1)} a_{r+1,j}^{(r)}, i = 1, \ldots, m; i \neq r+1; j = 1, \ldots, n; j \neq r+1,$$

$$a_{r+1,j}^{(r+1)} = -a_{r+1,j}^{(r)} a_{r+1,r+1}^{(r+1)},\ j = 1, \ldots, n; j \neq r+1.$$

Für $r = m$ endet der Algorithmus. Dieser algorithmische Extrakt aus dem Beweis des Austauschsatzes von Steinitz zeigt uns noch etwas: Bei den Operationen auf der Tabellenform der Linearkombinationen kann man von der expliziten Existenz aller Vektoren abstrahieren, da sie für den eigentlichen Algorithmus nicht interessieren. Im Ergebnis des Algorithmus ist nur wichtig, welche Vektoren jeweils für den Austausch ausgewählt wurden; diese können wir durch ihre Indices repräsentieren, indem wir z. B. am Anfang an die **v**-Vektoren die Indices $1, 2, \ldots, n$ vergeben und an die **w**-Vektoren die Indices $n+1, \ldots, n+m$. Diese kleine Tatsache ist wesentlich, weil man Vektoren nicht nach einer einheitlichen Methode in den Rechner eingeben kann. Im Austauschsatz haben wir vorausgesetzt, daß die Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_m$ linear unabhängig sind. Diese Bedingung brauchen wir zunächst nicht, um den Algorithmus auszuführen. Sollte sie verletzt sein, wird man in einem gewissen Schritt r des Algorithmus kein geeignetes Vektorpaar $(\mathbf{w}_i, \mathbf{v}_j)$ mit $a_{ij}^{(r)} \neq 0$ finden. In einem solchen Falle würde der Algorithmus mit dem r-ten Schritt enden. Daher sollte ein formaler Algorithmus auch ausgeben, welche Vektoren und wieviele ausgetauscht wurden. Die in einem Austauschschritt ausgewählte Zeile in der Tabelle nennt man auch **Pivotzeile**, und die ausgewählte Spalte heißt **Pivotspalte**; das auf der Kreuzung von Pivotzeile und Pivotspalte stehende Element heißt **Pivotelement**. *Beispiel.* Im $\mathbb{R}^4$ seien die Vektoren

$$\mathbf{b}_1 = (4; 0; -1; 2), \mathbf{b}_2 = (3; 2; -2; 1), \mathbf{b}_3 = (-1; 2; 0; 0)$$

zu einer Basis zu ergänzen. Als Ausgangsbasis des $\mathbb{R}^4$ wählen wir die natürlichen Einheitsvektoren $B = \{ \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4 \}$. Damit lautet das Anfangstableau für den Austauschalgorithmus:

	1	2	3	4
5	$*4$	0	-1	2
6	3	2	-2	1
7	-1	2	0	0

.

Der Austausch von $\mathbf{b}_1$ gegen $\mathbf{e}_1$ liefert die neue Tabelle

	5	2	3	4
1	$\frac{1}{4}$	0	$\frac{1}{4}$	$-\frac{1}{2}$
6	$\frac{3}{4}$	$*2$	$-\frac{5}{4}$	$-\frac{1}{2}$
7	$-\frac{1}{4}$	2	$-\frac{1}{4}$	$\frac{1}{2}$

.

Nun können wir $\mathbf{b}_2$ gegen $\mathbf{e}_2$ austauschen und erhalten die neue Tabelle

	5	6	3	4
1	$\frac{1}{4}$	0	$\frac{1}{4}$	$-\frac{1}{2}$
2	$-\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{1}{4}$
7	-1	1	$*1$	1

.

Im letzten Schritt tauschen wir $\mathbf{b}_3$ gegen $\mathbf{e}_3$ und erhalten die Endtabelle

	5	6	7	4
1	$\frac{1}{2}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{3}{4}$
2	$\frac{1}{4}$	$-\frac{1}{8}$	$\frac{5}{8}$	$-\frac{3}{8}$
3	1	-1	1	-1

.

Insbesondere entnehmen wir dieser Tabelle, daß die Vektoren $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{e}_4$ eine Basis des $\mathbb{R}^4$ bilden. Gleichzeitig wurden uns die Koordinaten der natürlichen Einheitsvektoren $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ bezüglich dieser Basis geliefert. Im folgenden Programm ist dieser Algorithmus implementiert.

```
//=============================================================================
//     Austausch von m Vektoren aus n Vektoren
// Rückkehrwert: Anzahl der ausgetauschten Vektoren.
//=============================================================================
#include "ls.h"
int ls_austausch(int m,    //  Anzahl der auszutauschenden Vektoren
                 int n,    //  Anzahl der Quell-Vektoren
                 REAL *A,  //  zweidimensionales Feld,
                           //  I: zeilenweise die Linearfaktoren
                           //  O: zeilenweise die neuen Linearfaktoren
                 int **iw, //  Hilfsfeld (oder NULL)
                           //  O: im Falle j=iw[i]<=0 ist der i-te Vektor
                           //     gegen den -j-ten ausgetauscht worden;
                           //     andernfalls erfolgte wegen linearer
                           //     Abhängigkeit kein Austausch
                 int **jv) //  Hilfsfeld (oder NULL)
                           //  O: im Falle i=jv[j]<=0 ist der -i-te Vektor
                           //     gegen den j-ten ausgetauscht worden
{ REAL epsaustausch=1.e-10,t,piv,klein=epsaustausch,*a,*aa,*ae=A+m*n;
  int i, js, rc=0, j, *iww=*iw, *jvv=*jv;
  if(!iww) *iw=iww=new int[n]; if(!jvv) *jv=jvv=new int[n];
  for(j=0; j<n; jvv[j++]=j);
  for(i=0, a=A; a<ae; a+=n, i++)
  { piv=klein;            // Pivotisierung
    for(j=0; j<n; j++)
    { t=fabs(a[j]); if((jvv[j]>0)&&(piv<t-epsaustausch)) piv=t, js=j;}
    if(piv==klein){ iww[i]=i+1; continue;} else klein=piv*epsaustausch;
    iww[i]=-js, jvv[js]=-i, piv=1/a[js], a[js]=1;     // Transformation
    for(aa=A+js; aa<ae; *aa*=piv, aa+=n);
    for(j=0; j<n; j++)
    { t=a[j]; if((j==js)||(fabs(t)<klein)) continue;
      a[j]=0; for(aa=A; aa<ae; aa[j]-=aa[js]*t, aa+=n);
    }
    rc++;
```

```
  }
  return rc;
}
```

Der Algorithmus AUSTAUSCH operiert nur auf den Faktoren a_{ij} der Linearkombinationen. Ein Rechteckschema von $m \cdot n$ Zahlen a_{ij} werden wir **Matrix A** nennen:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = (a_{ij})_{m,n}.$$

Eine Matrix hat Zeilen und Spalten (i-te Zeile, j-te Spalte). Jede Zeile kann als Vektor des $\mathbb{R}^n$ (**Zeilenvektor**) und jede Spalte als Vektor des $\mathbb{R}^m$ (**Spaltenvektor**) angesehen werden; a_{ij} heißt Matrixelement, die Zahlen a_{ii} nennt man **Hauptdiagonalelemente**. Sind eine Basis $B = \{ \mathbf{b}_1, \ldots, \mathbf{b}_n \}$ und eine Matrix $\mathbf{A} = (a_{ij})_{m,n}$ gegeben, so werden durch

$$\mathbf{w}_i = \sum_{j=1}^{n} a_{ij}\mathbf{b}_j, \quad i = 1, \ldots, m$$

m Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_m$ definiert, und wir können mit dem Austauschalgorithmus entscheiden, ob sie linear unabhängig sind oder nicht. Im ersten Falle endet der Algorithmus bei $r = m$, sonst früher. Aus der Kette

$$\begin{aligned} \sum_{i=1}^{m} \lambda_i \mathbf{w}_i = \mathbf{o} &\iff \sum_{i=1}^{m} \lambda_i (\sum_{j=1}^{n} a_{ij}\mathbf{b}_j) = \mathbf{o} \\ &\iff \sum_{j=1}^{n} (\sum_{i=1}^{m} a_{ij}\lambda_i)\mathbf{b}_j = \mathbf{o} \\ &\iff \sum_{i=1}^{m} a_{ij}\lambda_i = 0, \quad j = 1, \ldots, n \end{aligned}$$

schließen wir, daß die Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_m$ genau dann linear unabhängig sind, wenn die Zeilenvektoren der Matrix $\mathbf{A}$ diese Eigenschaft haben. Genauer gesagt: Unter den Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_m$ gibt es genau dann r linear unabhängige, wenn es unter den Zeilenvektoren der Matrix $\mathbf{A}$ r linear unabhängige gibt. Die maximale Anzahl linear unabhängiger Zeilenvektoren einer Matrix $\mathbf{A}$ nennt man den **Zeilenrang** der Matrix $\mathbf{A}$. Es sei eine Matrix $\mathbf{A} = (a_{ij})_{m,n}$ mit den Zeilenvektoren

$$\mathbf{w}_i = \sum_{j=1}^{n} a_{ij}\mathbf{e}_j, \quad i = 1, \ldots, m$$

gegeben. Sie möge den Zeilenrang r haben; ohne Beschränkung der Allgemeinheit nehmen wir an, daß die ersten r Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_r$ linear unabhängig sind und mit dem Austauschalgorithmus die Vektoren $\mathbf{e}_1, \ldots, \mathbf{e}_r$ gegen $\mathbf{w}_1, \ldots, \mathbf{w}_r$ ausgetauscht werden. Dann sind

$$\overline{U} = \text{lin}\,(\mathbf{w}_1, \ldots, \mathbf{w}_r)$$

und

$$U = \text{lin}\,(\mathbf{e}_{r+1}, \ldots, \mathbf{e}_n)$$

Komplementräume und für den Restklassenraum gilt

$$\mathbb{R}^n/U = \left\{ \mathbf{x} + U \mid \mathbf{x} \in \overline{U} \right\},$$

woraus

$$\dim(\mathbb{R}^n/U) = r$$

folgt.
Zu einer gegebenen Matrix $\mathbf{A} = (a_{ij})_{m,n}$ kann man mit dem Austauschalgorithmus den Zeilenrang berechnen. Dazu wähle man

$$\{ \mathbf{v}_1, \ldots, \mathbf{v}_n \} = \{ \mathbf{e}_1, \ldots, \mathbf{e}_n \}$$

und

$$\mathbf{w}_i = \sum_{j=1}^{n} a_{ij}\mathbf{e}_j, \quad i = 1, \ldots, m.$$

Dann ist $\mathbf{w}_i$ gerade der i-te Zeilenvektor der Matrix $\mathbf{A}$, und der Austauschalgorithmus liefert die maximale Anzahl ausgetauschter Vektoren, also die maximale Anzahl linear unabhängiger Zeilenvektoren, d. h. den Zeilenrang. Natürlich nutzen wir damit den Algorithmus zweckentfremdet, da ja der Zeilenrang nur ein Nebenprodukt ist. Wir wollen aus diesem Grunde im Austauschalgorithmus jene Operationen einsparen, die für die Berechnung des Zeilenranges unnötig sind. Betrachten wir den ersten Schritt und es sei etwa $\mathbf{w}_1$ gegen $\mathbf{v}_1$ zu tauschen, also $a_{11} \neq 0$. Diese Situation ist durch eventuelles Vertauschen der Pivotzeile mit der ersten Zeile und der Pivotspalte mit der ersten Spalte zu erreichen. Offenbar müssen wir nur die Transformation für die Elemente der Untermatrix

$$\begin{bmatrix} a_{22} & a_{23} & \cdots & a_{2n} \\ a_{32} & a_{33} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

ausführen. Die unvollständigen Transformationsformeln lauten hier:

$$l_{i,1}^{(1)} = \frac{a_{i1}}{a_{11}},\ i = 2, \ldots, m;$$

$$a_{ij}^{(1)} = a_{ij} - l_{i1}^{(1)} a_{1j},\ i = 2, \ldots, m; j = 2, \ldots, n.$$

Wir entnehmen sie unmittelbar den Transformationsformeln aus dem Austauschalgorithmus, wobei wir die Transformation der Pivotzeile (hier die erste Zeile) weggelassen und die transformierten Elemente der Pivotspalte mit $l_{i1}, i = 2, \ldots, m$ bezeichnet haben. Es sei noch bemerkt, daß wir auch keinen Austausch der Indices aus der 0-ten Zeile und 0-ten Spalte vorzunehmen brauchen, da wir ja nur am Zeilenrang der Matrix interessiert sind. Nach diesen Bemerkungen können wir den unvollständigen Austauschalgorithmus schon verbal beschreiben, wobei wir die folgende Tabellenform verwenden wollen:

	1	2	...	n
1	a_{11}	a_{12}	...	a_{1n}
2	a_{21}	a_{22}	...	a_{2n}
...	.....			
m	a_{m1}	a_{m2}	...	a_{mn} .

Die zu lösende Aufgabe besteht darin, die maximale Anzahl linear unabhängiger Zeilen der Matrix $\mathbf{A}$ zu bestimmen. Das folgende unvollständige Austauschverfahren wurde wohl zuerst von K. F. Gauß (1777-1855) angegeben, jedoch mit einem anderen Ziel.

Schritt 0: r:=0.

Schritt 1: Man suche eine Zeile i $(i > r)$ in der aktuellen Tabelle, die ein $a_{ij} \neq 0$ $(j > r)$ enthält; falls keine solche Zeile existiert, ist das Verfahren beendet; andernfalls vertausche man Zeile i mit Zeile $r+1$ und Spalte j mit Spalte $r+1$, so daß danach $a_{r+1,r+1} \neq 0$ gilt.

Schritt 2: Die Tabelle wird transformiert gemäß der folgenden Formeln:

$$l_{i,r+1} = \frac{a_{i,r+1}}{a_{r+1,r+1}},\ i = r+2, \ldots, m;$$

$$a_{ij} := a_{ij} - l_{i,r+1} a_{r+1,j},\ i = r+2, \ldots, m; j = r+2, \ldots, n.$$

Schritt 3: $r := r+1$; man wiederhole Schritt 1.

Das Verfahren endet offenbar, falls $r = m$ ist oder die aktuelle Matrix ab Zeile $r+1$ und Spalte $r+1$ nur noch Nullelemente enthält. Die letzte aktuelle Zahl r ist der Zeilenrang der Ausgangsmatrix $\mathbf{A}$. Im Interesse der vollständigen Reproduzierbarkeit der Ausgangsmatrix aus der Endtabelle speichern wir die Faktoren l_{ij} auf den entsprechenden Elementen a_{ij} ab:

$$a_{i,r+1} := l_{i,r+1},\ i = r+2, \ldots, m.$$

Zur Illustration betrachten wir das folgende Beispiel:

	1	2	3	4	5
1	0	0	1	2	0
2	1	−1	3	0	−2
3	0	−2	0	0	−1

.

Zunächst vertauschen wir die Spalten 1 und 3:

	3	2	1	4	5
1	∗1	0	0	2	0
2	3	−1	1	0	−2
3	0	−2	0	0	−1

.

Nach dem ersten Durchlauf haben wir die folgende Tabelle:

	3	2	1	4	5
1	1	0	0	2	0
2	3	∗ − 1	1	−6	−2
3	0	−2	0	0	−1

.

Der zweite Durchlauf liefert die Endtabelle:

	3	2	1	4	5
1	1	0	0	2	0
2	3	−1	1	−6	−2
3	0	−2	−2	12	3

.

Ein entsprechendes Programm sei ebenfalls angegeben.

```
//=============================================================================
//    Transformation einer Matrix auf Halbdiagonalform mit Spalten-Auswahl
// Rückkehrwert: Zeilenrang der Matrix.
//=============================================================================
#include "ls.h"
ushort ls_gauss(ushort m,  // Zeilenanzahl
             ushort n,     // Spaltenanzahl
             ushort nn,    // Anzahl der ersten Spalten, die in die Auswahl
                           // einbezogen werden sollen
             REAL *A,      // zu transformierende (m,n)-Matrix
                           // O: transformierte Matrix
             ushort **inds)// n-dimen. Hilfsfeld (oder NULL)
                           // O: Spaltenreihenfolge
{ ushort i, j, js, rc=0, *ind=*inds;
  REAL epsgauss=1.e-10, piv ,t, klein=epsgauss, *scal, *a, *aa, *ae=A+m*n;
  scal=new REAL[n]; if(!ind) ind=*inds=new ushort[n];
  for(i=0, a=A; i<nn; scal[i++]=t, a+=n)
    for(t=0, j=0; j<n; t+=fabs(a[j++]));
  for(j=0; j<n; ind[j++]=j);
  for(i=0, a=A; a<ae; i++, a+=n)
  { piv=klein=epsgauss;                        // Spalten-Auswahl
    for(j=i; j<nn; j++){ t=fabs(a[ind[j]])*scal[j];
    if(piv<t-klein) piv=t, js=j;}
    if(piv==klein){ rc++; continue;} else klein=piv*epsgauss;
    j=ind[js], ind[js]=ind[i], ind[i]=js=j, piv=1/a[js];// Transformation
    for(aa=a+n; aa<ae; aa+=n)
    { t=aa[js]*=piv; if(fabs(t)<klein){ aa[js]=0; continue;}
      for(j=i; ++j<n; aa[ind[j]]-=a[ind[j]]*t);}
  }
  delete []scal;
  return m-rc;
}
```

Wir wollen noch eine Interpretation der Operationen, die der Algorithmus auf der Matrix ausführt, betrachten. Dazu nehmen wir zur Vereinfachung der Darstellung an, daß keine Zeilen- und Spaltenvertauschungen vorgenommen werden. Die erste Transformation der Matrixelemente lautet

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j},\ i = 2, \ldots, m, j = 2, \ldots, n.$$

Für $j = 1$ erzeugen diese Formeln in der ersten Spalte unterhalb des Pivotelementes a_{11} Nullelemente. Dabei wird das $\frac{a_{i1}}{a_{11}}$-fache der ersten Zeile von der i-ten subtrahiert ($i = 2, \ldots, m$). Folglich überführt die erste Transformation die Matrix $\mathbf{A}$ in die Matrix

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & a_{m2}^{(1)} & \cdots & a_{mn}^{(1)} \end{bmatrix}.$$

Im zweiten Transformationsschritt werden die Elemente in der 2. Spalte unterhalb von $a_{22}^{(1)}$ zu Null gemacht, indem das $\frac{a_{i2}^{(1)}}{a_{22}^{(1)}}$-fache der zweiten Zeile von der i-ten subtrahiert wird $(i = 3, \ldots, m)$ usw. Wenn die Matrix $\mathbf{A}$ den Zeilenrang r hat, so wird sie mit dem Algorithmus in die Matrix

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} & a_{1,r+1} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2r}^{(1)} & a_{2,r+1}^{(1)} & \cdots & a_{2n}^{(1)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{rr}^{(r-1)} & a_{r,r+1}^{(r-1)} & \cdots & a_{rn}^{(r-1)} \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}$$

überführt, wobei in jedem Transformationsschritt in geeigneter Weise ein gewisses Vielfaches einer Zeile zu anderen addiert wird. In der Endmatrix kann der untere Nullteil auch fehlen; die Hauptdiagonalelemente bis zur r-ten Zeile sind ungleich Null. Man sagt, daß die Matrix $\mathbf{A}$ auf **Halbdiagonalform** transformiert wurde. Wir konstatieren zwei Beobachtungen: Der Algorithmus transformiert eine Matrix auf Halbdiagonalform. Die Addition einer Linearkombination von Zeilen zu einer anderen ändert den Zeilenrang der Matrix nicht.

2.3. Lineare Abbildungen und Matrizen

Es sei $V(+;\cdot,\mathbb{R})$ ein endlichdimensionaler Vektorraum der Dimension n. In V sei eine Basis $B = \{\, \mathbf{b}_1, \ldots, \mathbf{b}_n \,\}$ gegeben. Da jede lineare Abbildung φ auf V durch Vorgabe der Bilder einer fixierten Basis vollständig charakterisiert ist, definieren wir φ so, daß die Bilder der Basisvektoren gerade die natürlichen Einheitsvektoren des $\mathbb{R}^n$ sind:

$$\varphi(\mathbf{b}_i) = \mathbf{e}_i,\ i = 1, \ldots, n.$$

Es seien $\mathbf{x}$ ein Vektor aus V und $x_1, \ldots, x_n$ seine Koordinaten bezüglich der gewählten Basis B:

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{b}_i$$

Im Falle $\mathbf{x} \in \ker(\varphi)$ folgt

$$\mathbf{o} = \varphi(x) = \varphi(\sum_{i=1}^{n} x_i \mathbf{b}_i) = \sum_{i=1}^{n} x_i \varphi(\mathbf{b}_i) = \sum_{i=1}^{n} x_i \mathbf{e}_i,$$

woraus sich $x_i = 0, i = 1, \ldots, n$ ergibt, da die natürlichen Einheitsvektoren linear unabhängig sind. Damit ist die so definierte lineare Abbildung ein Isomorphismus und wir haben

Satz 40. *Jeder n-dimensionale Vektorraum über dem Körper der reellen Zahlen ist isomorph zum $\mathbb{R}^n$.*

Wir können daher unsere Untersuchungen auf den $\mathbb{R}^n$ einschränken. Im Mittelpunkt der Untersuchungen steht dabei das Studium der linearen Abbildungen zwischen Vektorräumen. Bisher haben wir lineare Abbildungen abstrakt behandelt. Nun wollen wir untersuchen, wie sich lineare Abbildungen berechnen lassen, d. h. wie man das Bild eines beliebigen Vektors bei einer linearen Abbildung berechnet. Im Vektorraum $\mathbb{R}^n$ sei eine Basis $B = \{\, \mathbf{b}_1, \ldots, \mathbf{b}_n \,\}$ und im Vektorraum $\mathbb{R}^m$ eine Basis $C = \{\, \mathbf{c}_1, \ldots, \mathbf{c}_m \,\}$ gegeben; ferner sei φ eine lineare Abbildung des $\mathbb{R}^n$ in den $\mathbb{R}^m$. Die Bilder $\varphi(\mathbf{b}_1), \ldots, \varphi(\mathbf{b}_n)$ der Basisvektoren aus B lassen sich dann auf genau eine Weise als Linearkombinationen der Basisvektoren aus C darstellen:

$$\varphi(\mathbf{b}_j) = \sum_{i=1}^{m} a_{ij} \mathbf{c}_i = a_{1j}\mathbf{c}_1 + a_{2j}\mathbf{c}_2 + \cdots + a_{mj}\mathbf{c}_m,\ j = 1, \ldots, n.$$

Die Faktoren in der j-ten Linearkombination sind die Koordinaten des Vektors $\varphi(\mathbf{b}_j)$ bezüglich der Basis C.

Für das Bild eines beliebigen Vektors $\mathbf{x} = x_1\mathbf{b}_1 + \cdots + x_n\mathbf{b}_n \in \mathbb{R}^n$ folgt daraus

$$\begin{aligned}
\varphi(\mathbf{x}) &= \varphi(x_1\mathbf{b}_1 + \cdots + x_n\mathbf{b}_n) \\
&= x_1\varphi(\mathbf{b}_1) + \cdots + x_n\varphi(\mathbf{b}_n) \\
&= x_1\sum_{i=1}^{m} a_{i1}\mathbf{c}_i + \cdots + x_n\sum_{i=1}^{m} a_{in}\mathbf{c}_i \\
&= a_{11}x_1\mathbf{c}_1 + a_{21}x_1\mathbf{c}_2 + \cdots + a_{m1}x_1\mathbf{c}_m + \\
&\quad a_{12}x_2\mathbf{c}_1 + a_{22}x_2\mathbf{c}_2 + \cdots + a_{m2}x_2\mathbf{c}_m + \\
&\quad + \cdots + \\
&\quad a_{1n}x_n\mathbf{c}_1 + a_{2n}x_n\mathbf{c}_2 + \cdots + a_{mn}x_n\mathbf{c}_m \\
&= (a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n)\mathbf{c}_1 + \\
&\quad (a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n)\mathbf{c}_2 + \\
&\quad + \cdots + \\
&\quad (a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n)\mathbf{c}_m.
\end{aligned}$$

Das Bild $\varphi(\mathbf{x})$ hat also bezüglich der fixierten Basis C die Koordinaten $y_1, \ldots, y_m$ mit

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n,\ i = 1, \ldots, m.$$

Folglich ist bei fixierten Basen in Bild- und Urbildraum jeder linearen Abbildung eine wohlbestimmte Matrix zugeordnet. Die Matrix enthält in der j-ten Spalte die Koordinaten des Bildvektors vom j-ten Basisvektor bezüglich der im Bildraum gewählten Basis. Besteht die gewählte Basis im Bildraum aus den natürlichen Einheitsvektoren, so stimmen die Koordinaten der Bildvektoren mit den Komponenten überein. Folglich enthält in diesem Falle die j-te Spalte der Matrix den Bildvektor des j-ten Basisvektors aus dem Urbildraum.
Umgekehrt definiert jede (m, n)-Matrix

$$\mathbf{A} = (a_{ij})_{m,n}$$

bei fixierten Basen B und C in Bild- und Urbildraum durch

$$\varphi(\mathbf{b}_j) = \sum_{i=1}^{m} a_{ij}\mathbf{c}_i,\ j = 1, \ldots, n$$

genau eine lineare Abbildung φ des $\mathbb{R}^n$ in den $\mathbb{R}^m$. Es gibt daher eine umkehrbar eindeutige Abbildung zwischen der Menge $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ aller linearen Abbildungen des $\mathbb{R}^n$ in den $\mathbb{R}^m$ und der Menge $\mathcal{M}_{mn}(\mathbb{R})$ aller (m, n)-Matrizen. Die (m, n)-Matrix mit nur Nullelementen heißt **Nullmatrix**. Im Falle $m = n$ spricht man von einer quadratischen Matrix. Eine quadratische Matrix nennt man **Einheitsmatrix**, falls die i-te Zeile den i-ten Einheitsvektor enthält. An einem Beispiel soll die Zuordnung einer Matrix zu einer linearen Abbildung demonstriert werden. Im $\mathbb{R}^2$ $(n = 2)$ sei die Basis

$$B = \{\ (1; 2), (0; 1)\ \}$$

und im $\mathbb{R}^3$ $(m = 3)$ sei die Basis

$$C = \{\ (0; 2; -1), (1; 1; 1), (2; 0; -3)\ \}$$

gegeben. Wir definieren eine lineare Abbildung φ durch $\varphi(B)$:

$$\varphi(1; 2) = (1; 1; 0), \varphi(0; 1) = (-2; 2; 3).$$

Man bestimme die der Abbildung zugeordnete Matrix. Die Koeffizienten der Matrix sind gerade die Koordinaten der Bildvektoren bezüglich der Basis C:

$$\begin{aligned}
(1; 1; 0) &= a_{11}(0; 2; -1) + a_{21}(1; 1; 1) + a_{31}(2; 0; -3), \\
(-2; 2; 3) &= a_{12}(0; 2; -1) + a_{22}(1; 1; 1) + a_{32}(2; 0; -3).
\end{aligned}$$

In Komponentenschreibweise lauten diese Gleichungen:

$$\begin{aligned}
1 &= a_{21} + 2a_{31} \\
1 &= 2a_{11} + a_{21} \\
0 &= -a_{11} + a_{21} - 3a_{31} \\
-2 &= a_{22} + 2a_{32} \\
2 &= 2a_{12} + a_{22} \\
3 &= -a_{12} + a_{22} - 3a_{32}.
\end{aligned}$$

Wir haben also ein lineares Gleichungssystem zu lösen, um die zugeordnete Matrix zu erhalten. Dies verschieben wir auf den nächsten Abschnitt. Hätten wir im Bildraum $\mathbb{R}^3$ die natürlichen Einheitsvektoren als Basis gewählt, bestünde die zugeordnete Matrix einfach aus den Bildvektoren der gewählten Basis des Urbildraumes:

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 1 & 2 \\ 0 & 3 \end{bmatrix}.$$

Die Menge $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ aller linearen Abbildungen des $\mathbb{R}^n$ in den $\mathbb{R}^m$ bildet mit den Operationen:

$$\varphi + \psi : \mathbf{x} \longmapsto \varphi(\mathbf{x}) + \psi(\mathbf{x}) \ \forall \mathbf{x} \in \mathbb{R}^n,$$

$$\lambda\varphi : \mathbf{x} \longmapsto \lambda\varphi(\mathbf{x}) \ \forall \mathbf{x} \in \mathbb{R}^n \ \forall \lambda \in \mathbb{R}$$

einen Vektorraum über $\mathbb{R}$. Wegen der umkehrbar eindeutigen Abbildung zwischen der Menge $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ und der Menge $\mathcal{M}_{mn}(\mathbb{R})$ aller (m,n)-Matrizen mit Koeffizienten aus $\mathbb{R}$ können wir untersuchen, welche Operationen zwischen Matrizen mit der Addition von linearen Abbildungen in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ und mit der Multiplikation einer linearen Abbildung mit einer reellen Zahl verträglich sind, so daß $\mathcal{M}_{mn}(\mathbb{R})$ ein Vektorraum wird und außerdem Isomorphie zwischen $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ und $\mathcal{M}_{mn}(\mathbb{R})$ besteht.
Dazu seien $\varphi, \psi \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ lineare Abbildungen und

$$\mathbf{A}^\varphi = (a_{ij}^\varphi)_{m,n}, \ \mathbf{A}^\psi = (a_{ij}^\psi)_{m,n}$$

die den Abbildungen zugeordneten Matrizen bezüglich vorgegebener Basen

$$B = \{ \mathbf{b}_1, \ldots, \mathbf{b}_n \}$$

des $\mathbb{R}^n$ bzw.

$$C = \{ \mathbf{c}_1, \ldots, \mathbf{c}_m \}$$

des $\mathbb{R}^m$; sei $\mathbf{x}$ ein Vektor aus dem $\mathbb{R}^n$ mit den Koordinaten $x_1, \ldots, x_n$ bezüglich der Basis B. Dann gilt nach den obigen Überlegungen

$$\varphi(\mathbf{x}) = \sum_{i=1}^{m}(\sum_{j=1}^{n} a_{ij}^\varphi x_j)\mathbf{c}_i, \ \psi(\mathbf{x}) = \sum_{i=1}^{m}(\sum_{j=1}^{n} a_{ij}^\psi x_j)\mathbf{c}_i,$$

und daher

$$\begin{aligned} \varphi(\mathbf{x}) + \psi(\mathbf{x}) &= \sum_{i=1}^{m}(\sum_{j=1}^{n} a_{ij}^\varphi x_j + \sum_{j=1}^{n} a_{ij}^\psi x_j)\mathbf{c}_i \\ &= \sum_{i=1}^{m}(\sum_{j=1}^{n}(a_{ij}^\varphi + a_{ij}^\psi)x_j)\mathbf{c}_i. \end{aligned}$$

Also ist der linearen Abbildung $\varphi + \psi$ die Matrix

$$\mathbf{A}^\varphi + \mathbf{A}^\psi = (a_{ij}^\varphi + a_{ij}^\psi)_{m,n}$$

zugeordnet, was wir daher als eine sinnvolle Definition der Matrizenaddition ansehen können. Ganz ähnlich rechnet man aus, daß der linearen Abbildung $\lambda\varphi, \lambda \in \mathbb{R}$ die Matrix $(\lambda \cdot a_{ij}^\varphi)_{m,n}$ zugeordnet ist, woraus wir schließen, daß die Multiplikation einer Matrix $\mathbf{A}$ mit einer reellen Zahl λ durch

$$\lambda \cdot \mathbf{A} = (\lambda \cdot a_{ij})_{m,n}$$

zu definieren ist. Die so erklärten Operationen auf der Menge $\mathcal{M}_{mn}(\mathbb{R})$ aller (m,n)-Matrizen machen diesen zu einem Vektorraum über dem Körper der reellen Zahlen, und die beiden Vektorräume $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ und $\mathcal{M}_{mn}(\mathbb{R})$ sind isomorph. Insbesondere bildet die Menge $\mathcal{M}_{1n}(\mathbb{R})$ aller $(1,n)$-Matrizen einen Vektorraum über $\mathbb{R}$, der isomorph zum $\mathbb{R}^n$ ist; entsprechend auch die Menge $\mathcal{M}_{m1}(\mathbb{R})$ aller $(m,1)$-Matrizen. Der Vektorraum $\mathcal{M}_{nm}(\mathbb{R})$ heißt der zu $\mathcal{M}_{mn}(\mathbb{R})$ **transponierte Vektorraum**; entsprechend für Matrizen: zu einer Matrix $\mathbf{A} = (a_{ij})_{m,n}$ heißt die Matrix

$$\mathbf{A}^\mathsf{T} = (a_{ji})_{n,m}$$

die zu **A transponierte Matrix**. Sie entsteht aus der Matrix dadurch, daß die Zeilen der einen zu den Spalten der anderen werden.

Als nächstes wollen wir ausrechnen, welche zweistellige Matrizenoperation die Verknüpfung von zwei linearen Abbildungen liefert. Um zwei Abbildungen φ, ψ zu verknüpfen, muß der Bildraum der einen gerade der Urbildraum der anderen sein. Es seien im $\mathbb{R}^n$ eine Basis $B = \{ \mathbf{b}_1, \ldots, \mathbf{b}_n \}$, im $\mathbb{R}^m$ eine Basis $C = \{ \mathbf{c}_1, \ldots, \mathbf{c}_m \}$, im $\mathbb{R}^l$ eine Basis $D = \{ \mathbf{d}_1, \ldots, \mathbf{d}_l \}$ gegeben und

$$\varphi : \mathbb{R}^n \longmapsto \mathbb{R}^m,\ \varphi \longmapsto \mathbf{A}^\varphi = (a_{ij}^\varphi)_{m,n},$$

$$\psi : \mathbb{R}^m \longmapsto \mathbb{R}^l,\ \psi \longmapsto \mathbf{A}^\psi = (a_{ij}^\psi)_{l,m}.$$

Für die Basisvektoren $\mathbf{b}_j$ aus B berechnen wir

$$\begin{aligned}\psi(\varphi(\mathbf{b}_j)) &= \psi(\sum_{i=1}^{m} a_{ij}^\varphi \mathbf{c}_i) = \sum_{i=1}^{m} a_{ij}^\varphi \psi(\mathbf{c}_i) \\ &= \sum_{i=1}^{m} a_{ij}^\varphi (\sum_{k=1}^{l} a_{ki}^\psi \mathbf{d}_k) \\ &= \sum_{k=1}^{l} (\sum_{i=1}^{m} a_{ki}^\psi a_{ij}^\varphi) \mathbf{d}_k.\end{aligned}$$

Folglich ist der Verknüpfung $\psi \circ \varphi$ die Matrix

$$(a_{kj}^{\psi\circ\varphi})_{l,n} = (\sum_{i=1}^{m} a_{ki}^\psi a_{ij}^\varphi)_{l,n}$$

zugeordnet und wir haben die Multiplikation zweier Matrizen in entsprechender Weise zu definieren:

$$\mathbf{A} \cdot \mathbf{B} = (a_{ij})_{m,n} \cdot (b_{jk})_{n,l} = (c_{ik})_{m,l}$$

mit

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk},\ i = 1, \ldots, m,\ k = 1, \ldots, l.$$

Beispiel:

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}, \mathbf{B} = \begin{bmatrix} g & h \\ i & j \\ k & l \end{bmatrix}, \mathbf{AB} = \begin{bmatrix} ag+bi+ck & ah+bj+cl \\ dg+ei+fk & dh+ej+fl \end{bmatrix}.$$

Durch einfaches Ausrechnen zeigt man, daß die Matrizenmultiplikation assoziativ und mit der Matrizenaddition distributiv ist:

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}),\ (\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC},\ \mathbf{C}(\mathbf{A}+\mathbf{B}) = \mathbf{CA}+\mathbf{CB},$$

wobei natürlich die Zeilen- und Spaltenanzahlen so gewählt sein müssen, daß die Operationen auch ausführbar sind.
Im Falle $m = n$ ist $\mathbf{AB} \in \mathcal{M}_{nn}(\mathbb{R})$. Hier entspricht der identischen Abbildung die Einheitsmatrix

$$\mathbf{E} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}.$$

Die Menge $\mathcal{M}_{nn}(\mathbb{R})$ aller (n,n)-Matrizen bildet somit einen Ring mit Einselement bezüglich der Matrizenaddition und der Matrizenmultiplikation und dieser ist isomorph zum Ring der linearen Abbildungen $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ mit der Addition und der Nacheinanderausführung von Abbildungen. Der Ring ist nicht kommutativ und enthält Nullteiler, z. B.

$$\begin{bmatrix} 1 & -2 & 4 \\ 3 & 1 & 5 \\ 2 & 4 & 0 \end{bmatrix} \begin{bmatrix} 2 & 4 & -2 \\ -1 & -2 & 1 \\ -1 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Im Vektorraum $\mathbb{R}^n$ nehmen wir nun als Basis die natürlichen Einheitsvektoren, entsprechend auch im $\mathbb{R}^m$. In diesem Falle stimmen die Komponenten mit den Koordinaten sowohl im Bild- als auch im Urbildraum überein.

Es sei φ eine lineare Abbildung und $\mathbf{A}$ die ihr zugeordnete Matrix bezüglich der beiden natürlichen Basen. Mit $\mathbf{A}_1, \ldots, \mathbf{A}_n$ bezeichnen wir die Spalten der Matrix $\mathbf{A}$, also die Bilder der Basisvektoren des Urbildraumes:

$$\varphi(\mathbf{e}_j) = \mathbf{A}_j,\ j = 1, \ldots, n.$$

Für die Dimension des Bildes gilt dann

$$\dim \varphi(\mathbb{R}^n) = \dim \operatorname{lin}(\mathbf{A}_1, \ldots, \mathbf{A}_n).$$

Die Dimension des Bildraumes ist natürlich unabhängig von der Matrix $\mathbf{A}$; folglich hat der Unterraum

$$\operatorname{lin}(\mathbf{A}_1, \ldots, \mathbf{A}_n)$$

des $\mathbb{R}^m$ für jede Matrix $\mathbf{A}$, die man der Abbildung φ zuordnen kann, die gleiche Dimension; anders gesagt: Jede dieser Matrizen hat die gleiche maximale Anzahl linear unabhängiger Spaltenvektoren. Die maximale Anzahl linear unabhängiger Spaltenvektoren einer Matrix nennt man **Spaltenrang**. Im Abschnitt 2.2. haben wir gelernt, daß die Dimension des durch die lineare Abbildung induzierten Restklassenklassenraumes gerade der Zeilenrang einer beliebigen, ihr zugeordneten Matrix ist. Nach Homomorphiesatz sind aber der Bildraum $\varphi(\mathbb{R}^n)$ und der Restklassenraum isomorph. Somit stimmen Zeilenrang und Spaltenrang einer Matrix überein und wir können vom **Rang** $\operatorname{rg}(\mathbf{A})$ einer Matrix $\mathbf{A}$ sprechen. Der Algorithmus GAUSS bestimmt also den Rang einer Matrix (indem er sie auf sog. Halbdiagonalform transformiert) und damit die Dimension des Bildes eines Vektorraumes bei einer linearen Abbildung.

2.4. Lineare Gleichungssysteme

Nachdem wir im vorangegangenen Abschnitt gelernt haben, wie man lineare Abbildungen berechnet, wollen wir hier die Umkehrung der Aufgabe behandeln. Die Aufgabe lautet: Bei gegebener Abbildung $\varphi \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ und gegebenem Vektor $\mathbf{y} \in \mathbb{R}^m$ finde man alle Vektoren $\mathbf{x} \in \mathbb{R}^n$, die die Gleichung $\varphi(\mathbf{x}) = \mathbf{y}$ erfüllen. Zur Lösung dieser Aufgabe benutzen wir in diesem Abschnitt in den Vektorräumen $\mathbb{R}^n$ und $\mathbb{R}^m$ jeweils die natürlichen Einheitsbasen. Bezüglich dieser Basen sei der linearen Abbildung φ die Matrix $\mathbf{A} = (a_{ij})_{m,n}$ zugeordnet. Dann bedeutet die Gleichung $\varphi(\mathbf{x}) = \mathbf{y}$ in Komponentenschreibweise

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= y_m.
\end{aligned}$$

Dies ist ein lineares Gleichungssystem mit m Gleichungen und n Unbekannten $x_1, \ldots, x_n$. Für ein Gleichungssystem führen wir die sog. Matrizenschreibweise ein. Dazu fassen wir die Vektoren $\mathbf{x} = (x_1, \ldots, x_n)$ und $\mathbf{y} = (y_1, \ldots, y_m)$ als $(n,1)$- bzw. $(m,1)$-Matrizen auf:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},\ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Diese Schreibweise ist dadurch gerechtfertigt, daß der Vektorraum $\mathcal{M}_{n1}(\mathbb{R})$ aller $(n,1)$-Matrizen isomorph zum $\mathbb{R}^n$ ist; entsprechend ist der Vektorraum $\mathcal{M}_{1n}(\mathbb{R})$ aller $(1,n)$-Matrizen isomorph zum $\mathcal{M}_{n1}(\mathbb{R})$. Damit können wir das Gleichungssytem in der Kurzform $\mathbf{Ax} = \mathbf{y}$ schreiben. Die Matrix $\mathbf{A}$ nennt man in diesem Zusammenhang auch **Koeffizientenmatrix**. Unsere Ausgangsfrage nach den Lösungen der Gleichung $\varphi(\mathbf{x}) = \mathbf{y}$ ist so gleichbedeutend mit der Frage nach allen Lösungen des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{y}$ und wir können unsere Erkenntnisse über lineare Abbildungen auf das lineare Gleichungssystem anwenden.
Zunächst wollen wir uns mit der sog. homogenen Gleichung $\varphi(\mathbf{x}) = \mathbf{o}$ beschäftigen, d. h. mit dem homogenen linearen Gleichungssystem $\mathbf{Ax} = \mathbf{o}$. Wegen

$$\ker(\varphi) = \{\, \mathbf{x} \mid \mathbf{Ax} = \mathbf{o} \,\}$$

und

$$n = \dim \mathbb{R}^n = \dim \ker(\varphi) + \dim \varphi(\mathbb{R}^n = \dim \ker(\varphi) + \operatorname{rg}(\mathbf{A})$$

folgt:

Satz 41. *Die Anzahl der linear unabhängigen Lösungen eines homogenen linearen Gleichungssystems* $\mathbf{Ax} = \mathbf{o}$ *ist gleich der Anzahl der Unbekannten minus Rang der Koeffizientenmatrix.*

Das sog. inhomogene System $\mathbf{Ax} = \mathbf{y}$ ist offenbar genau dann lösbar, wenn $\mathbf{y} \in \varphi(\mathbb{R}^n)$ gilt. Da die Spaltenvektoren $\mathbf{A}_1, \ldots, \mathbf{A}_n$ der Matrix $\mathbf{A}$ die Bildvektoren der fixierten Basis im $\mathbb{R}^n$ sind, gehört der Vektor $\mathbf{y}$ genau dann zum Bild $\varphi(\mathbb{R}^n)$, wenn $\mathbf{y}$ in der linearen Hülle der Spaltenvektoren der Matrix $\mathbf{A}$ liegt. Zusammen: Das System $\mathbf{Ax} = \mathbf{y}$ ist genau dann lösbar, wenn

$$\operatorname{lin}(\mathbf{A}_1, \ldots, \mathbf{A}_n) = \operatorname{lin}(\mathbf{A}_1, \ldots, \mathbf{A}_n, \mathbf{y})$$

gilt. Wenn wir noch die sog. **erweiterte Koeffizientenmatrix**

$$(\mathbf{A}, \mathbf{y}) = \begin{bmatrix} a_{11} & \cdots & a_{1n} & y_1 \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} & y_m \end{bmatrix}$$

einführen und den Dimensionsbegriff berücksichtigen, haben wir den folgenden Satz bewiesen.

Satz 42. *Das lineare Gleichungssystem $\mathbf{Ax} = \mathbf{y}$ ist genau dann lösbar, wenn die Koeffizientenmatrix und die erweiterte Koeffizientenmatrix den gleichen Rang haben:*

$$\operatorname{rg}(\mathbf{A}) = \operatorname{rg}(\mathbf{A}, \mathbf{y}).$$

Die Menge aller Lösungen eines linearen Gleichungssystems wird oft auch als **allgemeine Lösung** bezeichnet. Da die lineare Abbildung φ ein Homomorphismus ist und somit eine Restklassenzerlegung des $\mathbb{R}^n$ nach dem Normalteiler $\ker(\varphi)$ induziert, bedeutet das Aufsuchen aller Lösungen von $\mathbf{Ax} = \mathbf{y}$ die Ermittlung jener Restklasse, deren Vektoren sämtlich auf $\mathbf{y}$ abgebildet werden. Die Restklassen sind von der Form

$$\mathbf{x}^* + \ker(\varphi) = \{\, \mathbf{x}^* + \mathbf{x} \mid \mathbf{Ax} = \mathbf{o} \,\}.$$

Also hat im Falle der Lösbarkeit von $\mathbf{Ax} = \mathbf{y}$ die Menge X^* aller Lösungen die Form

$$X^* = \{\, \mathbf{x}^* + \mathbf{x} \mid \mathbf{Ax} = \mathbf{o} \,\},$$

wobei $\mathbf{x}^*$ eine spezielle Lösung von $\mathbf{Ax} = \mathbf{y}$ darstellt. Nach dem vorletzten Satz wird der Unterraum

$$\ker(\varphi) = \{\, \mathbf{x} \mid \mathbf{Ax} = \mathbf{o} \,\}$$

durch $n - r$ linear unabhängige Lösungen des homogenen Systems erzeugt, wobei r der Rang der Matrix $\mathbf{A}$ ist. Zusammenfassend können wir also den folgenden Satz aussprechen.

Satz 43. *Die (m, n)-Matrix $\mathbf{A}$ habe den Rang r; es seien $\mathbf{a}_1, \ldots, \mathbf{a}_{n-r}$ linear unabhängige Lösungen des homogenen Systems $\mathbf{Ax} = \mathbf{o}$ und $\mathbf{x}^*$ eine spezielle Lösung des inhomogenen Systems $\mathbf{Ax} = \mathbf{y}$. Dann läßt sich die Menge X^* aller Lösungen des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{y}$ in der Form*

$$X^* = \{\, \mathbf{x}^* + \lambda_1 \mathbf{a}_1 + \lambda_2 \mathbf{a}_2 + \cdots + \lambda_{n-r} \mathbf{a}_{n-r} \mid \lambda_i \in \mathbb{R}, i = 1, \ldots, n - r \,\}$$

darstellen, d. h. zu jeder Lösung $\overline{\mathbf{x}}$ des Systems $\mathbf{Ax} = \mathbf{y}$ gibt es reelle Zahlen $\lambda_1, \ldots, \lambda_{n-r}$, so daß

$$\overline{\mathbf{x}} = \mathbf{x}^* + \lambda_1 \mathbf{a}_1 + \lambda_2 \mathbf{a}_2 + \cdots + \lambda_{n-r} \mathbf{a}_{n-r}$$

gilt.

Wir überlegen uns nun, daß man mit dem Algorithmus GAUSS sowohl die Lösbarkeit eines linearen Gleichungssystems entscheiden als auch die allgemeine Lösung bestimmen kann. Wir bringen die erweiterte Koeffizientenmatrix $(\mathbf{A}, \mathbf{y})$ mit dem Algorithmus GAUSS auf Halbdiagonalform $(\mathbf{U}, \mathbf{z})$, wobei die letzte Spalte nicht in eventuelle Spaltenvertauschungen einbezogen wird. Bei dieser Transformation werden schrittweise Linearkombinationen von Zeilen zu anderen addiert. Das neue Gleichungssystem mit der erweiterten Koeffizientenmatrix $(\mathbf{U}, \mathbf{z})$ lautet, falls keine Spaltenvertauschungen vorgenommen wurden:

$$\begin{array}{rcrcrcrcl}
u_{11}x_1 & + & u_{12}x_2 & + \cdots + & u_{1r}x_r & + \cdots + & u_{1n}x_n & = & z_1 \\
 & & u_{22}x_2 & + \cdots + & u_{1r}x_r & + \cdots + & u_{2n}x_n & = & z_2 \\
 & & & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & & \\
 & & & & u_{rr}x_r & + \cdots + & u_{rn}x_n & = & z_r \\
 & & & & 0x_r & + \cdots + & 0x_n & = & z_{r+1} \\
 & & & & \cdots\cdots & \cdots\cdots & \cdots\cdots & & \\
 & & & & 0x_r & + \cdots + & 0x_n & = & z_m.
\end{array}$$

Dieses Gleichungssystem ist wegen der Transformationsformeln im Algorithmus GAUSS aus $\mathbf{Ax} = \mathbf{y}$ durch mehrfache Anwendung folgender Operationen hervorgegangen:

- Multiplikation einer Gleichung mit einer Zahl (ungleich Null),
- Addition zweier Gleichungen.

Daher hat das neue Gleichungssystem die gleichen Lösungen wie das alte. Im transformierten System $\mathbf{Ux} = \mathbf{z}$ sind zwei Fälle möglich:
Fall 1: Es ist $r < m$ und für ein $l \in \{ r+1, \ldots, m \}$ gilt $z_l \neq 0$. Dann ist

$$\mathrm{rg}(\mathbf{U}, \mathbf{z}) = \mathrm{rg}(\mathbf{A}, \mathbf{y}) = r + 1 > r = \mathrm{rg}(\mathbf{A}),$$

was uns zeigt, daß das System unlösbar ist.
Fall 2: Es ist $r \leqq m, z_i = 0, i = r+1, \ldots, m$. In diesem Falle gilt

$$\mathrm{rg}(\mathbf{A}, \mathbf{y}) = \mathrm{rg}(\mathbf{U}, \mathbf{z}) = \mathrm{rg}(\mathbf{U}) = r = \mathrm{rg}(\mathbf{A}),$$

das System ist lösbar und kann in folgender Form geschrieben werden:

$$\begin{array}{llllll} u_{11}x_1 + & u_{12}x_2 & + \cdots + & u_{1r}x_r = z_1 & - \; u_{1,r+1}x_{r+1} & - \cdots - \; u_{1n}x_n \\ & u_{22}x_2 & + \cdots + & u_{2r}x_r = z_2 & - \; u_{2,r+1}x_{r+1} & - \cdots - \; u_{2n}x_n \\ \multicolumn{6}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots} \\ & & & u_{rr}x_r = z_r & - \; u_{r,r+1}x_{r+1} & - \cdots - \; u_{rn}x_n. \end{array}$$

Zu jeder Wahl von $(x^*_{r+1}, \ldots, x^*_n) \in \mathbb{R}^{n-r}$ erhält man genau eine spezielle Lösung $\mathbf{x}^* = (x^*_1, \ldots, x^*_n)$, indem man die restlichen Werte $x^*_r, x^*_{r-1}, \ldots, x^*_1$ sukzessiv von der letzten zur ersten Gleichung berechnet, was man üblicherweise **Rückwärtselimination** nennt. Speziell kann man natürlich

$$x^*_{r+1} = \cdots = x^*_n = 0$$

wählen. Im homogenen Fall ($\mathbf{z} = \mathbf{y} = \mathbf{o}$) folgt daraus: Für jede Wahl

$$(x^0_{r+1}, \ldots, x^0_n) \in \mathbb{R}^{n-r}$$

erhält man genau eine Lösung $(x^0_1, \ldots, x^0_n) \in \ker(\varphi)$. Wählt man insbesondere $n-r$ linear unabhängige Lösungen aus dem $\mathbb{R}^{n-r}$ (etwa die natürlichen Einheitsvektoren des $\mathbb{R}^{n-r}$), so kann man durch Rückwärtselimination $n-r$ linear unabhängige Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_{n-r}$ aus $\ker(\varphi)$ berechnen. Die Lösungen des Systems $\mathbf{Ax} = \mathbf{y}$ sind also genau alle Vektoren der Form

$$\mathbf{x} = \mathbf{x}^* + \lambda_1 \mathbf{a}_1 + \cdots + \lambda_{n-r}\mathbf{a}_{n-r}, \; \lambda_i \in \mathbb{R}.$$

Als Beispiel sei das System

$$\begin{array}{rcrcrcr} 2x_1 & + & 3x_2 & - & x_3 & + & x_4 = 5 \\ 4x_1 & + & 6x_2 & - & 3x_3 & + & 2x_4 = 0 \\ 4x_1 & + & 3x_2 & - & x_3 & - & x_4 = 4 \end{array}$$

gegeben. Durch den Algorithmus GAUSS transformieren wir das System auf die Form

$$\begin{array}{rcrcrcrl} 2x_1 & + & 3x_2 & - & x_3 & + & x_4 = & 5 \\ & - & 3x_2 & + & x_3 & - & 3x_4 = & -6 \; . \\ & & & - & x_3 & & = & -10 \end{array}$$

Hieraus entnehmen wir, daß die Koeffizientenmatrix den Rang 3 hat; der Kern ist eindimensional:

$$\begin{array}{rcrcrlcr} 2x_1 & + & 3x_2 & - & x_3 = & 5 & - & x_4 \\ & - & 3x_2 & + & x_3 = & -6 & + & 3x_4 \; . \\ & & & - & x_3 = & -10 & & \end{array}$$

Um eine spezielle Lösung $\mathbf{x}^* = (x^*_1, x^*_2, x^*_3, x^*_4)$ zu berechnen, wählen wir $x^*_4 = 0$ und erhalten

$$\mathbf{x}^* = (-\frac{1}{2}; \frac{16}{3}; 10; 0).$$

Um eine Basis vom Kern der entsprechenden linearen Abbildung zu gewinnen, haben wir das zugeordnete homogene System zu betrachten:

$$\begin{array}{rcrcrl} 2x_1 & + & 3x_2 & - & x_3 = & -x_4 \\ & - & 3x_2 & + & x_3 = & 3x_4 \; . \\ & & & - & x_3 = & 0 \end{array}$$

Mit $x_4 = 1$ erhalten wir den Vektor $\mathbf{a}_1 = (1; -1; 0; 1)$. Damit lautet die allgemeine Lösung

$$\mathbf{x} = (-\frac{1}{2}; \frac{16}{3}; 10; 0) + \lambda(1; -1; 0; 1),\ \lambda \in \mathbb{R}.$$

Abschließend wollen wir noch den wichtigen Spezialfall $m = n$ studieren. Es möge eine lineare Abbildung $\varphi \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ mit $\ker(\varphi) = \{\mathbf{o}\}$ gegeben sein. Wir untersuchen die Gleichung $\varphi(\mathbf{x}) = \mathbf{y}$ mit gegebenem Vektor $\mathbf{y} \in \mathbb{R}^n$. Der Gleichung entspricht bei gegebener Basis ein lineares Gleichungssystem $\mathbf{Ax} = \mathbf{y}$ mit einer (n, n)-Matrix $\mathbf{A}$. Wegen

$$\dim \mathbb{R}^n = \dim \ker(\varphi) + \dim \varphi(\mathbb{R}^n = \mathrm{rg}(\mathbf{A})$$

ist $\mathrm{rg}(\mathbf{A}) = n$. Eine (n, n)-Matrix $\mathbf{A}$, die den maximalen Rang n hat, heißt **regulär**; falls $\mathrm{rg}(\mathbf{A}) < n$ ausfällt, heißt die Matrix **singulär**. Es sei erwähnt, daß die Menge aller regulären (n, n)-Matrizen mit der Matrizenmultiplikation eine Gruppe bildet. Im regulären Fall bilden die Spaltenvektoren $\mathbf{A}_1, \ldots, \mathbf{A}_n$ der Matrix $\mathbf{A}$ eine Basis des $\mathbb{R}^n$; daher hat das System $\mathbf{Ax} = \mathbf{y}$ für jeden Vektor $\mathbf{y} \in \mathbb{R}^n$ genau eine Lösung $\mathbf{x}^*$. Diese Lösung können wir sowohl mit Hilfe des Algorithmus AUSTAUSCH als auch mit dem Algorithmus GAUSS berechnen. Zunächst benutzen wir den Algorithmus AUSTAUSCH und setzen

$$\mathbf{v}_j = \mathbf{e}_j,\ j = 1, \ldots, n,\ \mathbf{w}_i = \sum_{j=1}^{n} a_{ij}\mathbf{e}_j,\ i = 1, \ldots, n.$$

Dann ist $\mathbf{w}_i$ gerade der i-te Zeilenvektor der Matrix $\mathbf{A}$. Mit dem Austauschalgorithmus werden nun die Vektoren $\mathbf{e}_1, \ldots, \mathbf{e}_n$ gegen die Vektoren $\mathbf{w}_1, \ldots, \mathbf{w}_n$ ausgetauscht, und die Endtabelle liefert eine Darstellung der Form

$$\mathbf{e}_i = \sum_{j=1}^{n} \overline{a}_{ij}\mathbf{w}_j,\ i = 1, \ldots, n.$$

Dabei ist $\overline{\mathbf{A}} = (\overline{a}_{ij})_{n,n}$ die Matrix aus der Endtabelle, wobei wir ohne Beschränkung der Allgemeinheit annehmen, daß $\mathbf{e}_i$ gegen $\mathbf{w}_i$ $(i = 1, \ldots, n)$ ausgetauscht wurde. Ist nun ψ die dieser Matrix entsprechende lineare Abbildung, so folgt, daß $\varphi \circ \psi$ die identische Abbildung ist, da bei Nacheinanderausführung aus den natürlichen Einheitsvektoren wieder die natürlichen Einheitsvektoren werden. Entsprechend überführt die Abbildung $\psi \circ \varphi$ jeden Zeilenvektor der Matrix in sich. Beiden Verknüpfungen ist folglich die Einheitsmatrix zugeordnet. Nun haben wir die Matrizenmultiplikation gerade so definiert, wie es der Verknüpfung von linearen Abblidungen entspricht. Also erhalten wir

$$\mathbf{A} \cdot \overline{\mathbf{A}} = \overline{\mathbf{A}} \cdot \mathbf{A} = \mathbf{E},$$

wobei $\mathbf{E}$ die (n, n)-Einheitsmatrix darstellt.
Die Matrix $\overline{\mathbf{A}}$ nennt man **invers** zur Matrix $\mathbf{A}$; sie wird mit $\mathbf{A}^{-1}$ bezeichnet. Wir können daher sagen, daß der Algorithmus AUSTAUSCH im Falle einer regulären Matrix $\mathbf{A}$ die zu $\mathbf{A}$ inverse Matrix berechnet. Multiplizieren wir nun die Gleichung $\mathbf{Ax} = \mathbf{y}$ von links mit der inversen Matrix $\mathbf{A}^{-1}$, so folgt

$$\mathbf{x} = \mathbf{E} \cdot \mathbf{x} = \mathbf{A}^{-1} \cdot \mathsf{A}\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{y}$$

und die gesuchte Lösung des Systems $\mathbf{Ax} = \mathbf{y}$ ist berechnet. Man beachte dabei, daß die Berechnung der inversen Matrix ca. n^3 Operationen benötigt, wobei als Operation eine Addition plus einer Multiplikation gerechnet wird. Als Beispiel nehmen wir die Matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ -2 & 0 & -1 \end{bmatrix}.$$

Der Austauschalgorithmus liefert die inverse Matrix

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ -1 & \frac{1}{2} & -1 \end{bmatrix}.$$

Hat man die inverse Matrix einmal bestimmt, kann man das System $\mathbf{Ax} = \mathbf{y}$ für jede rechte Seite $\mathbf{y}$ sofort durch $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$ lösen. Dabei sind n^2 Operationen auszuführen.
Wenden wir uns nun dem Algorithmus GAUSS zu, um das System $\mathbf{Ax} = \mathbf{y}$ zu lösen. Wir wenden also auf die Matrix $\mathbf{A}$ den Algorithmus GAUSS an. Zur Vereinfachnung der Darlegungen nehmen wir zunächst an, daß keine

Zeilen- oder Spaltenvertauschungen ausgeführt werden müssen. Aus der Endtabelle des Algorithmus bilden wir zwei Matrizen:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{bmatrix}, \mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}.$$

Dabei sind in der unteren Dreiecksmatrix $\mathbf{L}$ die Faktoren zusammengefaßt, mit denen die entsprechenden Zeilen multipliziert wurden. Die obere Dreiecksmatrix $\mathbf{U}$ enthält die transformierte Matrix. Die Matrix $\mathbf{L}$ kann man als Produkt

$$\mathbf{L} = \mathbf{L}_1 \mathbf{L}_2 \cdots \mathbf{L}_{n-1}$$

darstellen mit

$$\mathbf{L}_i = \begin{bmatrix} 1 & 0 & 0 \cdots & 0 & 0 \cdots 0 \\ 0 & 1 & 0 \cdots & 0 & 0 \cdots 0 \\ 0 & 0 & 1 \cdots & 0 & 0 \cdots 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 \cdots & 1 & 0 \cdots 0 \\ 0 & 0 & 0 \cdots & l_{i+1,i} & 1 \cdots 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 \cdots & l_{ni} & 0 \cdots 1 \end{bmatrix}, \ (i = 1, \ldots, n-1).$$

Die inverse Matrix dazu lautet

$$\mathbf{L}_i^{-1} = \begin{bmatrix} 1 & 0 & 0 \cdots & 0 & 0 \cdots 0 \\ 0 & 1 & 0 \cdots & 0 & 0 \cdots 0 \\ 0 & 0 & 1 \cdots & 0 & 0 \cdots 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 \cdots & 1 & 0 \cdots 0 \\ 0 & 0 & 0 \cdots & -l_{i+1,i} & 1 \cdots 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 \cdots & -l_{ni} & 0 \cdots 1 \end{bmatrix}, \ (i = 1, \ldots, n-1),$$

was sich durch Ausrechnen sofort feststellen läßt. Den Algorithmus GAUSS können wir mit diesen Matrizen in Matrixform schreiben. Der 1. Schritt im Algorithmus besteht in der Multiplikation der Matrix $\mathbf{A}$ mit $\mathbf{L}_1^{-1}$:

$$\begin{bmatrix} 1 & 0 & 0 \cdots 0 \\ -l_{21} & 1 & 0 \cdots 0 \\ -l_{31} & 0 & 1 \cdots 0 \\ \cdots & \cdots & \cdots \\ -l_{n1} & 0 & 0 \cdots 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & \cdots & a_{3n}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{bmatrix}.$$

Entsprechend haben wir nach dem letzten Schritt

$$\mathbf{L}_{n-1}^{-1} \mathbf{L}_{n-2}^{-1} \cdots \mathbf{L}_2^{-1} \mathbf{L}_1^{-1} \mathbf{A} = \mathbf{U}.$$

Aus dieser Gleichung folgt:

$$\mathbf{L}_1 \mathbf{L}_2 \cdots \mathbf{L}_{n-1} (\mathbf{L}_{n-1}^{-1} \cdots \mathbf{L}_2^{-1} \mathbf{L}_1^{-1} \mathbf{A}) = \mathbf{A} = \mathbf{L}_1 \mathbf{L}_2 \cdots \mathbf{L}_{n-1} \mathbf{U} = \mathbf{L} \cdot \mathbf{U}.$$

Damit haben wir gezeigt, daß der Algorithmus GAUSS im Falle einer regulären (n, n)-Matrix $\mathbf{A}$ eine Zerlegung dieser Matrix in das Produkt einer unteren Dreiecksmatrix $\mathbf{L}$ und einer oberen Dreiecksmatrix $\mathbf{U}$ liefert. Eine solche Zerlegung nennt man kurz **LU-Zerlegung** der Matrix $\mathbf{A}$. Ist nun eine **LU**-Zerlegung der Matrix $\mathbf{A}$ gegeben, so hat man anstelle des Gleichungssystems $\mathbf{Ax} = \mathbf{y}$ das System $\mathbf{LUx} = \mathbf{y}$ zu lösen. Dieses System wird in zwei Schritten gelöst: Zunächst löst man das System $\mathbf{Lz} = \mathbf{y}$ und danach das System $\mathbf{Ux} = \mathbf{z}$. Das System $\mathbf{Lz} = \mathbf{y}$ hat die Form

$$\begin{array}{lllllllll} z_1 & & & & & & & & = y_1 \\ l_{21} z_1 & + & z_2 & & & & & & = y_2 \\ l_{31} z_1 & + & l_{32} z_2 & + & z_3 & & & & = y_3 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n1} z_1 & + & l_{n2} z_2 & + & l_{n3} z_3 & + \cdots + & l_{n-1} z_{n-1} & + & z_n = y_n \end{array},$$

was man durch Vorwärtselimination lösen kann. Die so erhaltene Lösung wird zur rechten Seite für das System $\mathbf{Ux} = \mathbf{z}$:

$$\begin{array}{ccccccccc} u_{11}x_1 & + & u_{12}x_2 & + & u_{13}x_3 & + \cdots + & u_{1n}x_n = z_1 \\ & & u_{22}x_2 & + & u_{23}x_3 & + \cdots + & u_{2n}x_n = z_2 \\ & & & & u_{33}x_3 & + \cdots + & u_{3n}x_n = z_3 \\ & & & & \cdots\cdots\cdots\cdots\cdots\cdots & & \\ & & & & & & u_{nn}x_n = z_n \end{array},$$

woraus wir durch Rückwärtselimination die Lösung des Systems $\mathbf{Ax} = \mathbf{y}$ erhalten. Wir erwähnen noch, daß sich bei Zeilen- oder Spaltenvertauschungen während der Arbeit des Algorithmus GAUSS die hier dargelegten Überlegungen prinzipiell nicht ändern, da solche Vertauschungen durch Matrizen beschrieben werden können, die bis auf Zeilen- bzw. Spaltenvertauschungen mit der Einheitsmatrix übereinstimmen. Solche Matrizen nennt man **Anordnungsmatrizen** oder auch **Permutationsmatrizen**. Das Produkt zweier Anordnungsmatrizen ist wieder eine. Die Anordnungsmatrizen bilden eine Untergruppe der regulären (n, n)-Matrizen. Multiplizieren wir etwa die Matrix $\mathbf{A}$ von rechts mit der Anordnungsmatrix

$$\mathbf{P}_{21} = \begin{bmatrix} 0 & 1 & 0 \ldots 0 \\ 1 & 0 & 0 \ldots 0 \\ 0 & 0 & 1 \ldots 0 \\ \cdots\cdots\cdots\cdots \\ 0 & 0 & 0 \ldots 1 \end{bmatrix},$$

so erhält man eine Matrix, die durch Vertauschen der ersten und zweiten Spalte aus der Matrix $\mathbf{A}$ entsteht. Bei Multiplikation von links werden die erste und zweite Zeile vertauscht. Allgemein liefert daher der Algorithmus GAUSS bei Eingabe einer regulären (n, n)-Matrix $\mathbf{A}$ eine untere Dreiecksmatrix $\mathbf{L}$, eine obere Dreiecksmatrix $\mathbf{U}$, eine linke Anordnungsmatrix $\mathbf{P}$ und eine rechte Anordnungsmatrix $\mathbf{Q}$, so daß $\mathbf{LU} = \mathbf{PAQ}$ gilt. Zur Vervollständigung haben wir den Algorithmus LU angegeben, der mit GAUSS im Falle einer regulären (n, n)-Matrix $\mathbf{A}$ übereinstimmt. Nachdem man eine LU-Zerlegung erfolgreich ermittelt hat, kann man mit dem LU_SOLVE eine Loesung zu jeder rechten Seite berechnen. Abschließend schätzen wir noch den Rechenaufwand für die $\mathbf{LU}$-Zerlegung. Im r-ten Schritt werden $(n - r)^2$ Operationen ausgeführt, also insgesamt

$$\sum_{r=1}^{n-1}(n-r)^2 = \frac{(2n-1)(n-1)n}{6} = \frac{n^3}{3} - \frac{n^2}{2} + \frac{n}{6}$$

Operationen. Der Aufwand für die Vorwärts- und Rückwärtselimination ist nur von quadratischer Ordnung. Das zeigt uns, daß das Lösen eines linearen Gleichungssystems über die $\mathbf{LU}$-Zerlegung wesentlich schneller abläuft als über die Berechnung der inversen Matrix. Die Berechnung der inversen Matrix ist auch mittels der $\mathbf{LU}$-Zerlegung möglich. Man überlege sich, daß dieser algorithmische Weg zur Berechnung der inversen Matrix ebenfalls n^3 Operationen benötigt. Es sei daher betont, daß sich die Berechnung der inversen Matrix selbst dann nicht lohnt, wenn man das System für viele rechte Seiten lösen muß. Die Berechnung von $\mathbf{A}^{-1}\mathbf{y}$ erfordert den gleichen Aufwand wie das Lösen der beiden Gleichungssysteme $\mathbf{Lz} = \mathbf{y}$ und $\mathbf{Ux} = \mathbf{z}$. Die Berechnung von $\mathbf{A}^{-1}$ kostet aber dreimal mehr als die $\mathbf{LU}$-Zerlegung.

2.5. Determinanten

Über der Menge $\mathcal{M}_{nn}(\mathbb{R})$ aller (n, n)-Matrizen führen wir eine reellwertige Funktion $\mathrm{Det}(\mathbf{A})$ ein, die früher im Zusammenhang mit dem Lösen linearer Gleichungssysteme wichtig war. Im Gegensatz zur meist üblichen Definition wählen wir hier eine solche, die uns sogleich effektive Möglichkeiten ihrer Berechnung aufzeigt. Es sei $\mathbf{A}$ eine (n, n)-Matrix mit den Spalten $\mathbf{A}_1, \ldots, \mathbf{A}_n$. Unter der **Determinante** $\mathrm{Det}(\mathbf{A})$ verstehen wir eine Funktion

$$\mathrm{Det} : \mathcal{M}_{nn}(\mathbb{R}) \longmapsto \mathbb{R}$$

mit folgenden Eigenschaften:

1. $\mathrm{Det}(\mathbf{E})$=1,
2. $\mathrm{Det}(\mathbf{A}_1, \ldots, \lambda\mathbf{A}_i, \ldots, \mathbf{A}_n) = \lambda\mathrm{Det}(\mathbf{A}_1, \ldots, \mathbf{A}_i, \ldots, \mathbf{A}_n)$,
3. $\mathrm{Det}(\mathbf{A}_1, \ldots, \mathbf{A}_{i-1}, \mathbf{a} + \mathbf{a}', \mathbf{A}_{i+1}, \ldots, \mathbf{A}_n) =$
   $\mathrm{Det}(\mathbf{A}_1, \ldots, \mathbf{A}_{i-1}, \mathbf{a}, \mathbf{A}_{i+1}, \ldots, \mathbf{A}_n) + \mathrm{Det}(\mathbf{A}_1, \ldots, \mathbf{A}_{i-1}, \mathbf{a}', \mathbf{A}_{i+1}, \ldots, \mathbf{A}_n)$,
4. $\mathrm{Det}(\mathbf{A}_1, \ldots, \mathbf{A}_i, \ldots, \mathbf{A}_j, \ldots, \mathbf{A}_n) = -\mathrm{Det}(\mathbf{A}_1, \ldots, \mathbf{A}_j, \ldots, \mathbf{A}_i, \ldots, \mathbf{A}_n)$.

Die Eigenschaften 2.-4. drückt man in Worten wie folgt aus: Die Determinantenfunktion ist homogen, additiv und alternierend in den Spalten. Aus dieser Definition ziehen wir einige Schlußfolgerungen.

Satz 44. *Die Determinate einer Matrix mit zwei gleichen Spalten ist gleich Null.*

Beweis. Da die Determinante alternierend in den Spalten ist, kann man die zwei gleichen Spalten vertauschen, ohne die Matrix selbst zu ändern; dabei ändert sich aber das Vorzeichen der Determinante, woraus $\mathrm{Det}(\mathbf{A}) = 0$ folgt. □

Satz 45. *Der Wert der Determinante ändert sich nicht, wenn man eine Linearkombination von Spaltenvektoren zu einer Spalte addiert, die nicht in der Linearkombination auftritt.*

Beweis. Es ist wegen des letzten Satzes, der Additivität, der Homogenität

$$\begin{aligned}\mathrm{Det}(\mathbf{A}_1 + \lambda\mathbf{A}_2, \mathbf{A}_2, \ldots, \mathbf{A}_n) &= \mathrm{Det}(\mathbf{A}_1, \ldots, \mathbf{A}_n) + \lambda\mathrm{Det}(\mathbf{A}_2, \mathbf{A}_2, \ldots, \mathbf{A}_n)\\ &= \mathrm{Det}(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n),\end{aligned}$$

womit der Satz bereits bewiesen ist, da die Schlußweise wiederholt anwendbar ist und für jede Spalte verwendet werden kann. □

Im Falle einer singulären Matrix muß mindestens eine Spalte Linearkombination gewisser anderer Spalten sein. Es sei dies etwa die erste. Indem wir eine gewisse Linearkombination anderer Spalten zur ersten addieren, erhalten wir eine Nullspalte. Also folgt mit dem letzten Satz, daß für eine solche Matrix

$$\mathrm{Det}(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n) = \mathrm{Det}(\mathbf{o}, \mathbf{A}_2, \ldots, \mathbf{A}_n)$$

gilt. Die Homogenität liefert für beliebiges λ

$$\mathrm{Det}(\mathbf{o}, \mathbf{A}_2, \ldots, \mathbf{A}_n) = \lambda\mathrm{Det}(\mathbf{o}, \mathbf{A}_2, \ldots, \mathbf{A}_n),$$

was nur gelten kann, wenn

$$\mathrm{Det}(\mathbf{o}, \mathbf{A}_2, \ldots, \mathbf{A}_n) = 0$$

gilt. Damit haben wir den folgenden Satz bewiesen.

Satz 46. *Die Determinante einer singulären Matrix ist gleich Null.*

Für den Fall, daß die Matrix eine spezielle Form hat, kann man den Wert der Determinante leicht berechnen.

Satz 47. *Bei einer oberen Dreiecksmatrix*

$$\mathbf{U} = (u_{ij})_{n,n},\ u_{ij} = 0, i > j$$

ist die Determinante gleich dem Produkt der Hauptdiagonalelemente:

$$\mathrm{Det}(\mathbf{U}) = \mathrm{Det}\begin{bmatrix} u_{11} & u_{12} & u_{13}\cdots u_{1n}\\ 0 & u_{22} & u_{23}\cdots u_{2n}\\ 0 & 0 & u_{33}\cdots u_{3n}\\ \cdots & \cdots & \cdots\\ 0 & 0 & 0 \quad \cdots u_{nn}\end{bmatrix} = u_{11}u_{22}u_{33}\cdots u_{nn}.$$

Beweis. Es sei eine obere Dreicksmatrix $\mathbf{U}$ gegeben. Wir bemerken zunächst, daß folgendes gilt:

$$\mathrm{Det}\begin{bmatrix} u_{11} & u_{12} & u_{13}\ldots u_{1n}\\ 0 & u_{22} & u_{23}\ldots u_{2n}\\ 0 & 0 & u_{33}\ldots u_{3n}\\ \cdots & \cdots & \cdots\\ 0 & 0 & 0 \quad \ldots u_{nn}\end{bmatrix} =$$

$$\mathrm{Det}\begin{bmatrix} u_{11} & 0 & u_{13}\cdots u_{1n}\\ 0 & u_{22} & u_{23}\cdots u_{2n}\\ 0 & 0 & u_{33}\cdots u_{3n}\\ \cdots & \cdots & \cdots\\ 0 & 0 & 0 \quad \cdots u_{nn}\end{bmatrix} + \mathrm{Det}\begin{bmatrix} u_{11} & u_{12} & u_{13}\cdots u_{1n}\\ 0 & 0 & u_{23}\cdots u_{2n}\\ 0 & 0 & u_{33}\cdots u_{3n}\\ \cdots & \cdots & \cdots\\ 0 & 0 & 0 \quad \cdots u_{nn}\end{bmatrix}.$$

Da im zweiten Summanden die erste und zweite Spalte der Matrix linear abhängig sind, ist dieser Summand gleich Null. Nach diesem Schema können wir suksessiv alle Elemente oberhalb der Hauptdiagonalen von $\mathbf{U}$ durch Nullen ersetzen, ohne den Wert der Determinante zu ändern. Folglich schließen wir mit der Homogenität:

$$\mathrm{Det}(\mathbf{U}) = \mathrm{Det}\begin{bmatrix} u_{11} & 0 & 0 & \dots 0 \\ 0 & u_{22} & 0 & \dots 0 \\ 0 & 0 & u_{33} & \dots 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots u_{nn} \end{bmatrix} = u_{11}u_{22}\cdots u_{nn}\mathrm{Det}(\mathbf{E}),$$

womit der Satz für eine obere Dreiecksmatrix bewiesen ist. □
Für eine untere Dreicksmatrix geht man ganz analog vor.

Satz 48. *Bei einer unteren Dreiecksmatrix ist die Determinante gleich dem Produkt der Hauptdiagonalelemente.*

Nun können wir leicht die Determinante einer Matrix berechnen.

Satz 49. *Wenn für eine quadratische (n,n)-Matrix $\mathbf{A}$ eine $\mathbf{LU}$-Zerlegung gegeben ist:*

$$\mathbf{A} = \mathbf{LU},$$

so ist der Wert der Determinante von $\mathbf{A}$ gleich dem Produkt der Hauptdiagonalelemente der oberen Dreiecksmatrix $\mathbf{U}$:

$$\mathrm{Det}(\mathbf{A}) = \mathrm{Det}(\mathbf{U}) = u_{11}\cdot u_{22}\cdots u_{nn}.$$

Beweis. Es sei die Matrix $\mathbf{A}$ regulär und $\mathbf{A} = \mathbf{LU}$. Bei der Matrizenmultiplikation von $\mathbf{L}$ mit $\mathbf{U}$ wird die i-te Spalte der Matrix $\mathbf{L}$ mit u_{ii} multipliziert und außerdem eine Linearkombination der nachfolgenden Spalten von $\mathbf{L}$ mit den Indices $i+1,\dots,n$ zur i-ten Spalte addiert. Also gilt

$$\mathrm{Det}(\mathbf{A}) = \mathrm{Det}(\mathbf{LU}) = u_{11}\cdot u_{22}\cdots u_{nn}\mathrm{Det}(\mathbf{L}).$$

Die untere Dreiecksmatrix $\mathbf{L}$ hat alle Hauptdiagonalelemente gleich 1. Wegen

$$\mathrm{Det}(\mathbf{L}) = \mathrm{Det}(\mathbf{E}\cdot\mathbf{L})$$

wird bei der Multiplikation der Einheitsmatrix mit der Matrix $\mathbf{L}$ zur i-ten Spalte der Einheitsmatrix eine Linearkombination der ersten $i-1$ Spalten addiert ($i = 1,\dots,n$). Diese Operation ändert den Wert der Determinante der Einheitsmatrix nicht, also gilt $\mathrm{Det}(\mathbf{L}) = 1$, womit der Satz für eine reguläre Matrix bewiesen ist. Im singulären Fall gilt die Aussage offenbar auch, da beide Seiten gleich Null sind. □

Satz 50. *Die Determinante des Produktes zweier Matrizen ist gleich dem Produkt der Determinanten beider Matrizen:*

$$\mathrm{Det}(\mathbf{A}\cdot\mathbf{B}) = \mathrm{Det}(\mathbf{A})\cdot\mathrm{Det}(\mathbf{B}).$$

Beweis. Ist $\mathbf{B}$ eine singuläre Matrix, so gilt die Aussage offenbar, da beide Seiten gleich Null sind. Es sei $\mathbf{B}$ eine reguläre Matrix und

$$\mathbf{B} = \mathbf{L}\cdot\mathbf{U}.$$

Dann können wir mit dem vorangegangenen Satz und seinem Beweis die folgenden Gleichungskette schließen:

$$\mathrm{Det}(\mathbf{A}\cdot\mathbf{B}) = \mathrm{Det}(\mathbf{A}\cdot\mathbf{L}\cdot\mathbf{U}) = u_{11}\cdots u_{nn}\mathrm{Det}(\mathbf{A}\cdot\mathbf{L}) = \mathrm{Det}(\mathbf{A})\mathrm{Det}(\mathbf{B}),$$

was zu beweisen war. □
Schließlich folgt sofort aus unseren Sätzen der

Satz 51. *Für jede quadratische Matrix $\mathbf{A}$ gilt:* $\mathrm{Det}(\mathbf{A}) = \mathrm{Det}(\mathbf{A}^\mathsf{T})$.

Beweis. Wir brauchen nur den Fall einer regulären Matrix $\mathbf{A}$ zu betrachten. Außerdem sei eine $\mathbf{LU}$-Zerlegung der Matrix $\mathbf{A}$ gegeben. Unsere Sätze erlauben es, die folgende Gleichungskette aufzuschreiben:

$$\mathrm{Det}(\mathbf{A}) = \mathrm{Det}(\mathbf{LU}) = \mathrm{Det}(\mathbf{L})\mathrm{Det}(\mathbf{U}) = \mathrm{Det}(\mathbf{U}^\mathsf{T})\mathrm{Det}(\mathbf{L}^\mathsf{T}) = \mathrm{Det}(\mathbf{U}^\mathsf{T}\mathbf{L}^\mathsf{T}) = \mathrm{Det}(\mathbf{A}^\mathsf{T}),$$

welche den Satz beweist. □
Aus unseren Überlegungen schließen wir insbesondere, daß

$$\mathrm{Det}(\mathbf{A}) = (-1)^k\mathrm{Det}(\mathbf{U})$$

gilt, wobei die Matrix $\mathbf{U}$ die sich aus dem Algorithmus GAUSS ergebende obere Dreiecksmatrix darstellt und k die Anzahl der Zeilen- und Spaltenvertauschungen ist. Unabhängig von den Vertauschungen liefert das Produkt der Hauptdiagonalelemente in der Endtabelle des Algorithmus den Betrag der Determinante.
Nach unseren Untersuchungen ist eine Matrix genau dann regulär, wenn ihre Determinante ungleich Null ist. Damit können wir den folgenden Satz aussprechen.

Satz 52. *Das lineare Gleichungssystem $\mathbf{Ax} = \mathbf{y}$ mit einer quadratischen Matrix $\mathbf{A}$ ist genau dann lösbar, wenn* $\mathrm{Det}(\mathbf{A}) \neq 0$ *gilt.*

2.6. Skalarprodukt und Orthogonalität

Es sei ein Vektorraum $V(+;\cdot,\mathbb{R})$ gegeben. Das **Skalarprodukt** ist eine auf $V \times V$ definierte reellwertige Abbildung

$$(\cdot,\cdot) : V \times V \longmapsto \mathbb{R}$$

mit den folgenden Eigenschaften, die für alle $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ und alle $\lambda \in \mathbb{R}$ gelten sollen:

Symmetrie: $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})$,

Additivität: $(\mathbf{x} + \mathbf{y}, \mathbf{z}) = (\mathbf{x}, \mathbf{z}) + (\mathbf{y}, \mathbf{z})$,

Homogenität: $(\lambda\mathbf{x}, \mathbf{y}) = \lambda(\mathbf{x}, \mathbf{y})$,

Nichtnegativität: $(\mathbf{x}, \mathbf{x}) \geqq 0,\ (\mathbf{x}, \mathbf{x}) = 0 \Longleftrightarrow \mathbf{x} = \mathbf{o}$.

Als Beipspiele erwähnen wir den Vektorraum aller über einem Intervall $[a, b]$ integrierbaren reellwertigen Funktionen; das Skalarprodukt ist hier durch

$$\int_a^b f(x)g(x)dx$$

gegeben. Ein weiteres, für uns wichtiges Beispiel ist der Vektorraum $\mathbb{R}^n$, über dem für $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ durch

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i y_i$$

ein Skalarprodukt definiert ist, was man leicht nachprüfen kann. Gelegentlich verwenden wir für dieses Skalarprodukt auch die Matrixschreibweise, indem wir die Vektoren als $(n, 1)$-Matrizen auffassen:

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathsf{T}}\mathbf{y}.$$

Zwei Vektoren $\mathbf{x}, \mathbf{y} \in V$ heißen **orthogonal** , wenn $(\mathbf{x}, \mathbf{y}) = 0$ gilt. Für die Orthogonalität verifiziert man schnell für alle $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V, \lambda, \mu \in \mathbb{R}$:

$$(\mathbf{x}, \mathbf{y}) = 0 \iff (\mathbf{y}, \mathbf{x}) = 0,$$
$$(\mathbf{x}, \mathbf{y}) = 0 \text{ und } (\mathbf{x}, \mathbf{z}) = 0, \text{ so } (\mathbf{x}, \lambda\mathbf{y} + \mu\mathbf{z}) = 0,$$
$$(\mathbf{x}, \mathbf{y}) = 0\ \forall \mathbf{y} \in V \Longleftrightarrow \mathbf{x} = \mathbf{o},$$
$$(\mathbf{x}, \mathbf{x}) = 0 \Longleftrightarrow \mathbf{x} = \mathbf{o}.$$

Aus diesen Eigenschaften erhält man z. B. , daß alle zu den Vektoren eines Unterraumes U orthogonalen Vektoren wieder einen Unterraum bilden, den **Orthogonalraum**

$$U^* = \{\, \mathbf{x} \mid (\mathbf{x}, \mathbf{y}) = 0\ \forall \mathbf{y} \in U \,\}.$$

Es sei nun $V = \mathbb{R}^n$ und $\{\, \mathbf{b}_1, \ldots, \mathbf{b}_r \,\}$ eine Basis von U. Dann folgt

$$\mathbf{x} \in U^* \Longleftrightarrow (\mathbf{x}, \mathbf{b}_i) = 0,\ i = 1, \ldots, r.$$

Die Basisvektoren $\mathbf{b}_1, \ldots, \mathbf{b}_r$ fassen wir als Zeilenvektoren einer Matrix $\mathbf{B}$ auf. Dann ist

$$\operatorname{rg}(\mathbf{B}) = r,$$

und die Bedingung für die Vektoren des Orthogonalraumes lautet

$$\mathbf{x} \in U^* \Longleftrightarrow \mathbf{B}\mathbf{x} = \mathbf{o}.$$

Der Orthogonalraum U^* ist also Kern einer gewissen linearen Abbildung

$$\varphi : \mathbb{R}^n \longmapsto \mathbb{R}^n,$$

die zur Matrix $\mathbf{B}$ gehört. Aus $\dim\ker(\varphi) + \operatorname{rg}(\mathbf{B}) = \dim\mathbb{R}^n = n$ folgt, daß

$$\dim\ker(\varphi) = \dim U^* = n - r$$

sein muß und daher

$$\dim U + \dim U^* = \dim\mathbb{R}^n.$$

Andererseits haben U und U^* nur den Nullvektor gemeinsam. Folglich ist der Orthogonalraum ein Komplementraum von U.
Ein Beispiel aus dem $\mathbb{R}^3$: Es sei $U = \lin(\mathbf{b}_1, \mathbf{b}_2)$ mit

$$\mathbf{b}_1 = (3, 2, -1),\ \mathbf{b}_2 = (0, 1, 2),$$

d. h.

$$U = \{\, \mathbf{x} \mid \mathbf{x} = \lambda(3, 2, -1) + \mu(0, 1, 2)\,, \lambda, \mu \in \mathbb{R} \,\}.$$

Geometrisch ist U eine Ebene durch den Ursprung und U^* die Gerade durch den Ursprung, die auf U senkrecht steht.
Es sei bemerkt, daß man nicht in allen Vektorräumen ein Skalarprodukt definieren kann. Einen unendlichdimensionalen Vektorraum, in dem ein Skalarprodukt existiert, nennt man Hilbertraum.
Fundamental ist die **Cauchy-Schwarzsche Ungleichung**:

Satz 53. *Für alle Vektoren $\mathbf{x}, \mathbf{y}$ eines Vektorraumes mit Skalarprodukt gilt*

$$|(\mathbf{x}, \mathbf{y})| \leqq \sqrt{(\mathbf{x}, \mathbf{x})} \cdot \sqrt{(\mathbf{y}, \mathbf{y})}.$$

Beweis. Für $\mathbf{x} = \mathbf{o}$ oder $\mathbf{y} = \mathbf{o}$ ist die Ungleichung offenbar richtig; seien also $\mathbf{x} \neq \mathbf{o}$ und $\mathbf{y} \neq \mathbf{o}$. Wir verwenden ausschließlich die das Skalarprodukt definierenden Eigenschaften. Für alle $\lambda \in \mathbb{R}$ gilt offenbar

$$0 \leqq (\mathbf{x} + \lambda\mathbf{y}, \mathbf{x} + \lambda\mathbf{y}) = (\mathbf{x}, \mathbf{x}) + 2\lambda(\mathbf{x}, \mathbf{y}) + \lambda^2(\mathbf{y}, \mathbf{y}).$$

Rechts steht in der Ungleichung eine quadratische Funktion in λ, die nach dieser Ungleichung keine negativen Werte annimmt. Dies ist aber genau dann erfüllt, wenn die Diskriminante der Funktion nicht positiv ist, d. h. es muß gelten:

$$(\mathbf{x}, \mathbf{y})^2 \leqq (\mathbf{x}, \mathbf{x}) \cdot (\mathbf{y}, \mathbf{y}).$$

Auf beiden Seiten der Ungleichung stehen nichtnegative Zahlen; folglich darf man die Quadratwurzel ziehen, ohne daß sich die Ungleichungsrichtung ändert, womit wir die Behauptung erhalten. □
Aus dem Beweis dieses Satzes können wir noch erkennen, wann in der Cauchy-Schwarzschen Ungleichung die Gleichheit gilt. Es gilt offenbar genau dann, wenn $0 = (\mathbf{x} + \lambda\mathbf{y}, \mathbf{x} + \lambda\mathbf{y})$ ausfällt, was wiederum genau dann gilt, wenn $\mathbf{x} + \lambda\mathbf{y} = \mathbf{o}$ gilt, also die beiden Vektoren $\mathbf{x}, \mathbf{y}$ linear abhängig sind. Eine weitere, wichtige reellwertige Funktion auf einem Vektorraum V ist die Länge oder **Norm** $\|\cdot\|$ eines Vektors:

$$\|\cdot\| :\ V \longmapsto \mathbb{R},$$

die durch die folgenden Eigenschaften charakterisiert wird:

1. $\|\mathbf{x}\| \geqq 0,\ \|\mathbf{x}\| = 0 \Longleftrightarrow \mathbf{x} = \mathbf{o}$,
2. $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$,
3. $\|\mathbf{x} + \mathbf{y}\| \leqq \|\mathbf{x}\| + \|\mathbf{y}\|$.

Die letzte Bedingung nennt man **Dreiecksungleichung**. Sie besagt, daß die Länge der Summe zweier Vektoren niemals größer sein kann als die Längensumme der einzelnen Vektoren.
In Vektorräumen, auf denen ein Skalarprodukt $(\cdot, \cdot)$ definiert ist, wird durch

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})}$$

auch eine Norm definiert. Um das einzusehen, brauchen wir nur die Dreiecksungleichung zu beweisen, da die anderen Eigenschaften offensichtlich sind. Diese folgt aus der folgenden Kette, in der die Cauchy-Schwarzsche Ungleichung angewendet wird:

$$\begin{aligned}\|\mathbf{x} + \mathbf{y}\|^2 &= (\mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y}) = \|\mathbf{x}\|^2 + 2(\mathbf{x}, \mathbf{y}) + \|\mathbf{y}\|^2 \\ &\leqq \|\mathbf{x}\|^2 + 2|(\mathbf{x}, \mathbf{y})| + \|\mathbf{y}\|^2 \leqq \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 \\ &= (\|\mathbf{x}\| + \|\mathbf{y}\|)^2.\end{aligned}$$

Der Vektorraum $\mathbb{R}^n$ mit der Norm

$$\|\mathbf{x}\|_2 = \sqrt{(\mathbf{x}, \mathbf{x})} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

heißt **euklidischer Vektorraum**; die Norm heißt **euklidische Norm** oder **euklidische Länge**.
Die Cauchy-Schwarzsche Ungleichung können wir mit der euklischen Norm auch in der Form

$$-1 \leqq (\frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|}) \leqq 1$$

schreiben. Dies gestattet es uns, einen Winkel $\alpha(\mathbf{x}, \mathbf{y})$ zwischen zwei Vektoren $\mathbf{x}, \mathbf{y}$ eines euklidischen Vektorraumes zu definieren, indem wir festsetzen:

$$\cos\alpha(\mathbf{x}, \mathbf{y}) = (\frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|}).$$

Daß diese Definition des Winkels zwischen zwei Vektoren unserer Anschauung entspricht, zeigt die folgende Überlegung: Offenbar ist der Winkel zwischen orthogonalen Vektoren gleich $\frac{\pi}{2}$; im Falle $\mathbf{y} = \mathbf{x}$ erhalten wir

$$\cos\alpha(\mathbf{x}, \mathbf{x}) = 1,$$

also $\alpha(\mathbf{x}, \mathbf{x}) = 0$; im Falle $\mathbf{y} = -\mathbf{x}$ folgt

$$\cos\alpha(\mathbf{x}, -\mathbf{x}) = -1$$

und damit $\alpha(\mathbf{x}, \mathbf{x}) = \pi$.
Neben der euklidischen Norm gibt es auf dem $\mathbb{R}^n$ noch andere Normen, so z. B. die **Maximumnorm**

$$\|\mathbf{x}\|_\infty = \max\{\, |x_1|, |x_2|, \ldots, |x_n| \,\} = \max_j |x_j|$$

und die p-Norm

$$\|\mathbf{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x|^p}.$$

Auch für (m, n)-Matrizen $\mathbf{A} \in \mathcal{M}_{mn}(\mathbb{R})$ kann man eine Norm $\|\mathbf{A}\|$ einführen, indem man analoge Eigenschaften fordert:

$$\begin{aligned} \|\mathbf{A}\| &> 0 \;\forall \mathbf{A} \neq \mathbf{o}, \\ \|\lambda\mathbf{A}\| &= |\lambda| \cdot \|\mathbf{A}\|, \\ \|\mathbf{A} + \mathbf{B}\| &\leqq \|\mathbf{A}\| + \|\mathbf{B}\|. \end{aligned}$$

Eine Matrixnorm wird meist im Zusammenhang mit Vektornormen verwendet. Die Matrixnorm $\|\cdot\|$ heißt mit der Vektornorm $\|\cdot\|_a$ auf dem $\mathbb{R}^n$ und der Vektornorm $\|\cdot\|_b$ auf dem $\mathbb{R}^m$ **verträglich**, falls gilt:

$$\|\mathbf{A}\mathbf{x}\|_b \leqq \|\mathbf{A}\| \cdot \|\mathbf{x}\|_a.$$

So ist die **Zeilensummennorm**

$$\|\mathbf{A}\|_\infty = \max\left\{ \sum_{j=1}^{n} |a_{1j}|, \ldots, \sum_{j=1}^{n} |a_{mj}| \right\}$$

mit der Maximumnorm verträglich, was aus

$$\begin{aligned} \|\mathbf{A}\mathbf{x}\|_\infty &= \max_i \left\{ \sum_{j=1}^{n} |a_{ij}x_j| \right\} \\ &\leqq \max_i \left\{ \sum_{j=1}^{n} |a_{ij}| \max_j |x_j| \right\} \\ &= \|\mathbf{A}\|_\infty \cdot \|\mathbf{x}\|_\infty \end{aligned}$$

folgt.
Mit der euklidischen Vektornorm ist die **Schur-Norm** (für (n, n)-Matrizen)

$$\|\mathbf{A}\|_2 = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2}$$

verträglich, was man unter Nutzung der Cauchy-Schwarzschen Ungleichung so einsieht:

$$\|\mathbf{Ax}\|_2 = \sqrt{\sum_{i=1}^{n}(\sum_{j=1}^{n} a_{ij}x_j)^2} \leqq \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 \sum_{j=1}^{n} x_j^2}$$
$$= \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2}\ \|\mathbf{x}\|_2 = \|\mathbf{A}\|_2 \cdot \|\mathbf{x}\|_2.$$

Eine Teilmenge $\{ \mathbf{b}_1, \ldots, \mathbf{b}_r \}$ von r Vektoren eines euklidischen Vektorraumes V heißt **Orthonormalsystem**, falls je zwei von ihnen orthogonal sind und jeder die Norm 1 hat:

$$(\mathbf{b}_i, \mathbf{b}_j) = \begin{cases} 1 & : \quad i = j \\ 0 & : \quad i \neq j \end{cases} \qquad i = 1, \ldots, r;\ j = 1, \ldots, r.$$

Satz 54. *Die Vektoren eines Orthonormalsystems sind linear unabhängig.*

Beweis. Es sei $\{ \mathbf{b}_1, \ldots, \mathbf{b}_r \}$ ein Orthonormalsystem. Wenn wir die Gleichung

$$\sum_{i=1}^{r} \lambda_i \mathbf{b}_i = \mathbf{o}$$

annehmen, so folgt für $j = 1, \ldots, r$:

$$0 = (\mathbf{o}, \mathbf{b}_j) = (\sum_{i=1}^{r} \lambda_i \mathbf{b}_i, \mathbf{b}_j) = \sum_{i=1}^{r} \lambda_i (\mathbf{b}_i, \mathbf{b}_j) = \lambda_j,$$

was gerade die lineare Unabhängigkeit bedeutet. □
Im Falle $\dim V = r$ heißt ein Orthonormalsystem auch **Orthonormalbasis**.

Satz 55. *Jede Basis eines Unterraumes U des euklidischen Vektorraumes V kann in eine Orthonormalbasis von U überführt werden.*

Beweis. Der Beweis dieses Satzes ist konstruktiv, d. h. wir geben ein Verfahren an, das eine gegebene Basis orthonormiert. Es sei dazu $\{ \mathbf{b}_1, \ldots, \mathbf{b}_r \}$ eine beliebige Basis von $U \subseteqq V$. Daraus werden wir ein System $\{ \mathbf{a}_1, \ldots, \mathbf{a}_r \}$ von paarweise orthogonalen Vektoren konstruieren. Indem wir abschließend jeden dieser Vektoren durch seine Länge dividieren, erhalten wir eine Orthonormalbasis.
Im 1. Schritt wählen wir als Vektor $\mathbf{a}_1$ einen beliebigen aus $\{ \mathbf{b}_1, \ldots, \mathbf{b}_r \}$, etwa $\mathbf{b}_1$:

$$\mathbf{a}_1 = \mathbf{b}_1.$$

Für den Vektor $\mathbf{a}_2$ machen wir den Ansatz

$$\mathbf{a}_2 = \lambda_{21}\mathbf{a}_1 + \mathbf{b}_2.$$

Zunächst ist klar, daß die Vektoren $\mathbf{a}_1, \mathbf{a}_2$ für jede Wahl von λ_{21} linear unabhängig sind; dies folgt aus der linearen Unabhängigkeit der Vektoren $\mathbf{b}_1, \mathbf{b}_2$. Da wir orthogonale Vektoren anstreben, muß der Faktor λ_{21} so gewählt werden, daß die Vektoren $\mathbf{a}_1$ und $\mathbf{a}_2$ orthogonal sind:

$$(\mathbf{a}_1, \mathbf{a}_2) = 0,$$

woraus sich der Faktor λ_{21} bestimmen läßt:

$$0 = (\mathbf{a}_2, \mathbf{a}_1) = \lambda_{21}(\mathbf{a}_1, \mathbf{a}_1) + (\mathbf{b}_2, \mathbf{a}_1),$$

also

$$\lambda_{21} = -\frac{(\mathbf{b}_2, \mathbf{a}_1)}{(\mathbf{a}_1, \mathbf{a}_1)}.$$

Nehmen wir nun an, wir hätten schon paarweise orthogonale Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_{l-1}$ konstruiert. Im l-ten Schritt machen wir den Ansatz

$$\mathbf{a}_l = \lambda_{l1}\mathbf{a}_1 + \lambda_{l2}\mathbf{a}_2 + \ldots + \lambda_{l,l-1}\mathbf{a}_{l-1} + \mathbf{b}_l.$$

Aus der Orthogonalitätsbedingung zu den bereits konstruierten Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_{l-1}$ bestimmen wir die unbekannten Parameter $\lambda_{li}, i = 1, \ldots, l-1$: Aus

$$\begin{aligned} 0 &= (\mathbf{a}_i, \mathbf{a}_l) \\ &= \lambda_{l1}(\mathbf{a}_i, \mathbf{a}_1) + \ldots + \lambda_{li}(\mathbf{a}_i, \mathbf{a}_i) + \ldots + \lambda_{l,l-1}(\mathbf{a}_i, \mathbf{a}_{l-1}) + (\mathbf{a}_i, \mathbf{b}_l) \\ &= \lambda_{li}(\mathbf{a}_i, \mathbf{a}_i) + (\mathbf{a}_i, \mathbf{b}_l) \end{aligned}$$

folgt

$$\lambda_{li} = -\frac{(\mathbf{a}_i, \mathbf{b}_l)}{(\mathbf{a}_i, \mathbf{a}_i)},\ i = 1, \ldots, l-1.$$

Damit sind die nach r Schritten entstandenen Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_r$ orthogonal und keine Nullvektoren, also linear unabhängig. Sie entstehen als Linearkombinationen aus den gegebenen Basisvektoren $\mathbf{b}_1, \ldots, \mathbf{b}_r$ des Unterraumes U und bilden daher selbst eine Basis dieses Unterraumes. Die Vektoren

$$\frac{\mathbf{a}_1}{\|\mathbf{a}_1\|}, \frac{\mathbf{a}_2}{\|\mathbf{a}_2\|}, \ldots, \frac{\mathbf{a}_r}{\|\mathbf{a}_r\|}$$

bilden somit eine Orthonormalbasis von U. □

Das im Beweis verwendete Verfahren heißt **Erhard-Schmidtsches Orthogonalisierungsverfahren** .
Ein Beispiel im $\mathbb{R}^3$: Als Basis nehmen wir die Vektoren

$$\mathbf{b}_1 = (1;2;2),\ \mathbf{b}_2 = (3;4;5),\ \mathbf{b}_3 = (7;1;1).$$

Nach Schritt 1 ist $\mathbf{a}_1 = (1;2;2)$ der erste neue (noch unnormierte) Basisvektor. Der Ansatz in Schritt 2

$$\mathbf{a}_2 = \lambda_{21}(1;2;2) + (3;4;5)$$

liefert aus der Orthogonalisierungsforderung

$$0 = (\mathbf{a}_1, \mathbf{a}_2) = \lambda_{21}(1+4+4) + (3+8+10),$$

daß $\lambda_{21} = -\frac{7}{3}$ sein muß, was

$$\mathbf{a}_2 = (\frac{2}{3}; -\frac{2}{3}; \frac{1}{3})$$

ergibt. Im letzten Schritt haben wir den Ansatz

$$\mathbf{a}_3 = \lambda_{31}(1;2;2) + \lambda_{32}(\frac{2}{3}; -\frac{2}{3}; \frac{1}{3}) + (7;1;1)$$

und mit den Orthogonalitätsforderungen folgt:

$$0 = (\mathbf{a}_3, \mathbf{a}_1) = \lambda_{31} 9 + 11 \Longrightarrow \lambda_{31} = -\frac{11}{9},$$

$$0 = (\mathbf{a}_3, \mathbf{a}_2) = \lambda_{32} 1 + \frac{13}{3} \Longrightarrow \lambda_{32} = -\frac{13}{3}.$$

Damit erhalten wir den zu $\mathbf{a}_1$ und $\mathbf{a}_2$ orthogonalen Vektor

$$\mathbf{a}_3 = -\frac{11}{9}(1;2;2) - \frac{13}{3}(\frac{2}{3}; -\frac{2}{3}; \frac{1}{3}) + (7;1;1) = (\frac{26}{9}; \frac{13}{9}; -\frac{26}{9})$$

und insgesamt die Orthonormalbasis des $\mathbb{R}^3$ (nach Division durch ihre Länge):

$$(\frac{1}{3}; \frac{2}{3}; \frac{2}{3}),\ (\frac{2}{3}; -\frac{2}{3}; \frac{1}{3}),\ (\frac{2}{3}; \frac{1}{3}; -\frac{2}{3}).$$

Eine formale Umsetzung des Verfahrens ist im folgenden Programm ORTHO angegeben.

```
//=========================================================================
//      Erhard-Schmidt-sches Orthogonalisierungsverfahren
// Rückkehrwert: Anzahl der orthonormierten Spalten.
//=========================================================================
```

```
#include "ls.h"
ushort ls_ortho(ushort m,// Zeilenanzahl
            ushort n,   // Spaltenanzahl
            REAL *A)    // (m,n)-Matrix; 0: orthonormierte Spalten
{ ushort i, l ,rc=0;
  REAL epsortho=1.e-10, s, *x=new REAL[n], *a, *ae=A+m*n;
  for(l=0; l<n; l++)
  { for(i=0; i<l;i++)
    { if(!x[i]) continue; for(a=A, s=0; a<ae; s+=a[i]*a[l], a+=n);
      for(a=A, s*=-x[i]; a<ae; a[l]+=a[i]*s, a+=n);
    }
    for(a=A, s=0; a<ae; s+=a[l]*a[l],a+=n); x[l]=(s>epsortho)? 1/s:0;
  }
  for(i=0; i<n; i++)
    if(x[i]) for(a=A, s=sqrt(x[i]) ;a<ae; a[i]*=s, a+=n);
    else for(a=A; a<ae; a[i]=0, a+=n);
  delete []x;
  return n-rc;
}
```

Der Algorithmus benötigt etwa n^3 Operationen für die Orthogonalisierung einer (n,n)-Matrix und entspricht daher im Aufwand dem Algorithmus AUSTAUSCH.
Eine lineare Abbildung φ des $\mathbb{R}^n$ in sich heißt **orthogonal**, wenn sie eine Orthonormalbasis auf eine Orthonormalbasis abbildet. Die einer orthogonalen Abbildung bezüglich einer Orthonormalbasis zugeordnete Matrix heißt **orthogonale Matrix**.

Satz 56. *Es sei φ eine lineare Abbildung des $\mathbb{R}^n$ in sich, $\{\mathbf{b}_1,\ldots,\mathbf{b}_n\}$ eine Orthonormalbasis des $\mathbb{R}^n$ und $\mathbf{A}$ die ihr zugeordnete (n,n)-Matrix. Dann sind die folgenden Aussagen äquivalent:*

1. *Die Matrix $\mathbf{A}$ ist orthogonal.*
2. *Die Spaltenvektoren der Matrix $\mathbf{A}$ bilden eine Orthonormalbasis des $\mathbb{R}^n$.*
3. *Die inverse Matrix von $\mathbf{A}$ ist gleich ihrer transponierten:*
$$\mathbf{A}^{-1} = \mathbf{A}^{\mathsf{T}}.$$
4. *Die Zeilenvektoren der Matrix $\mathbf{A}$ bilden eine Orthonormalbasis des $\mathbb{R}^n$.*
5. *Das Skalarprodukt bleibt invariant unter der Matrix $\mathbf{A}$:*
$$(\mathbf{Ax},\mathbf{Ay}) = (\mathbf{x},\mathbf{y}) \quad \forall \mathbf{x},\mathbf{y}\in\mathbb{R}^n.$$

Beweis. Wir zeigen zuerst, daß die 1. Aussage zur 2. Aussage äquivalent ist. Es seien $\mathbf{A}_1,\ldots,\mathbf{A}_n$ die Spaltenvektoren der Matrix $\mathbf{A}$. Dann erhält man die behauptete Äquivalenz aus der folgenden Gleichungskette:

$$\begin{aligned}(\varphi(\mathbf{b}_r),\varphi(\mathbf{b}_s)) &= (\sum_{i=1}^{n} a_{ir}\mathbf{b}_i, \sum_{j=1}^{n} a_{js}\mathbf{b}_j)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ir}a_{js}(\mathbf{b}_i,\mathbf{b}_j)\\ &= \sum_{i=1}^{n} a_{ir}a_{is} = (\mathbf{A}_r,\mathbf{A}_s).\end{aligned}$$

Für die 3. Aussage bemerken wir, daß der i-te Spaltenvektor der Matrix $\mathbf{A}$ gerade der i-te Zeilenvektor der transponierten Matrix ist und damit die 2. Aussage zu $\mathbf{AA}^{\mathsf{T}} = \mathbf{E}$ äquivalent ist.
Die Zeilen der Matrix $\mathbf{A}$ sind die Spalten der Matrix $\mathbf{A}^{\mathsf{T}}$, und $\mathbf{A}^{\mathsf{T}}$ ist auch eine orthogonale Matrix:

$$(\mathbf{A}^{\mathsf{T}})^{\mathsf{T}} = \mathbf{A}, \quad (\mathbf{A}^{\mathsf{T}})^{\mathsf{T}}\mathbf{A}^{\mathsf{T}} = \mathbf{AA}^{\mathsf{T}} = \mathbf{E},$$

womit gezeigt ist, daß die 4. Aussage zur dritten äquivalent ist.
Abschließend zeigen wir, daß die 5. Aussage zur dritten äquivalent ist. Es gilt

$$(\mathbf{Ax},\mathbf{Ay}) = (\mathbf{Ax})^{\mathsf{T}}\mathbf{Ay} = \mathbf{x}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{Ay}.$$

Folglich gilt

$$(\mathbf{Ax},\mathbf{Ay}) = (\mathbf{x},\mathbf{y}) \quad\Longleftrightarrow\quad \mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{E},$$

womit die Äquivalenz aller Aussagen nachgewiesen ist. □
Unter Berücksichtigung der Winkeldefinition zwischen Vektoren eines euklidischen Vektorraumes folgt aus der 5. Aussage

Satz 57. *Eine orthogonale Abbildung des $\mathbb{R}^n$ auf sich ist längen- und winkeltreu.*

Orthogonale Abbildungen beschreiben Drehungen und Spiegelungen des Raumes. So ist z. B. im $\mathbb{R}^2$ einer Drehung φ der Vektoren um den Winkel α die Matrix

$$\mathbf{A} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

zugeordnet.

Satz 58. *Die Determinante einer orthogonalen Matrix* $\mathbf{A}$ *hat entweder den Wert 1 oder -1.*

Beweis. Mit den Orthogonalitätsaussagen und den Determinanteneigenschaften folgt:

$$1 = \mathrm{Det}(\mathbf{E}) = \mathrm{Det}(\mathbf{A}^{-1}\mathbf{A}) = \mathrm{Det}(\mathbf{A}^{\mathsf{T}}\mathbf{A}) = \mathrm{Det}(\mathbf{A})\mathrm{Det}(\mathbf{A}) = (\mathrm{Det}(\mathbf{A}))^2,$$

was die behauptete Aussage beinhaltet. □

2.7. Eigenwerte und Eigenvektoren

Es sei eine reelle, symmetrische (n,n)-Matrix $\mathbf{A} = (a_{ij})_{n,n}$ gegeben. Wir betrachten die folgende Aufgabenstellung:
Man finde eine orthogonale (n,n)-Matrix $\mathbf{Q}$ derart, daß die Matrix $\mathbf{Q}^{\mathsf{T}}\mathbf{A}\mathbf{Q}$ Diagonalgestalt hat.
Eine (n,n)-Matrix von Diagonalgestalt mit $\lambda_1, \ldots, \lambda_n$ als Hauptdiagonalelemente bezeichnen wir mit

$$\mathbf{diag}(\lambda_1, \ldots, \lambda_n).$$

Ist nun $\mathbf{Q}$ eine solche gesuchte Matrix, so folgt mit der orthogonalen Koordinatentransformation $\mathbf{x} = \mathbf{Q}\mathbf{y}$:

$$(\mathbf{x}, \mathbf{A}\mathbf{x}) = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} = (\mathbf{Q}\mathbf{y})^{\mathsf{T}}\mathbf{A}\mathbf{Q}\mathbf{y} = \mathbf{y}^{\mathsf{T}}\mathbf{Q}^{\mathsf{T}}\mathbf{A}\mathbf{Q}\mathbf{y} = \mathbf{y}^{\mathsf{T}}\mathbf{diag}(\lambda_1, \ldots, \lambda_n)\mathbf{y} = \sum_{j=1}^{n} \lambda_j y_j^2.$$

Aus dieser Gleichung schließen wir, daß

$$\{\, \mathbf{x} \mid \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} = \alpha \,\} = \left\{ \mathbf{y} \;\middle|\; \sum_{j=1}^{n} \lambda_j y_j^2 = \alpha \right\}$$

gilt (mit $\alpha > 0$). Im Falle $n = 2, \lambda_1 > 0, \lambda_2 > 0$ zeigt die Gleichung, daß durch $\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} = \alpha$ eine Ellipse beschrieben wird, deren Halbachsen die Längen λ_1, λ_2 haben; bei $\lambda_2 = 0$ wird eine Parabel beschrieben und bei $\lambda_1 = \lambda_2 = 1$ ein Kreis mit dem Durchmesser α. Bei dieser orthogonalen Koordinatentransformation wird also die quadratische Form $\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}$ in eine solche überführt, in der die gemischten Glieder nicht mehr auftreten. Daher spricht man hier von einer **Hauptachsentransformation**.
Wenn wir die Gleichung $\mathbf{Q}^{\mathsf{T}}\mathbf{A}\mathbf{Q} = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$ von links mit $\mathbf{Q}$ multiplizieren und die Spaltenvektoren der Matrix $\mathbf{Q}$ mit $\mathbf{Q}_1, \ldots, \mathbf{Q}_n$ bezeichnen, erhalten wir

$$\mathbf{A}\mathbf{Q} = \mathbf{Q}\mathbf{diag}(\lambda_1, \ldots, \lambda_n) = \mathbf{Q}(\lambda_1\mathbf{e}_1, \ldots, \lambda_n\mathbf{e}_n),$$

$$(\mathbf{A}\mathbf{Q}_1, \ldots, \mathbf{A}\mathbf{Q}_n) = (\lambda_1\mathbf{Q}_1, \ldots, \lambda_n\mathbf{Q}_n),$$

d. h.

$$\mathbf{A}\mathbf{Q}_j = \lambda_j\mathbf{Q}_j, \quad j = 1, \ldots, n$$

oder

$$(\mathbf{A} - \lambda_j\mathbf{E})\mathbf{Q}_j = \mathbf{o}, \quad j = 1, \ldots, n.$$

Diese Gleichung sagt uns, daß das homogene lineare Gleichungssystem $(\mathbf{A} - \lambda_j\mathbf{E})\mathbf{x} = \mathbf{o}$ eine nichttriviale Lösung $\mathbf{x} = \mathbf{Q}_j$ besitzt.
Eine reelle Zahl λ, zu der ein Vektor $\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{o}$ existiert mit $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, heißt **Eigenwert** der Matrix $\mathbf{A}$; jede nichttriviale Lösung des Gleichungssystems $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ nennt man **Eigenvektor** zum Eigenwert λ. Wegen $\lambda\mathbf{x} = \lambda\mathbf{E}\mathbf{x}$ kann man das System auch in der Form $(\mathbf{A} - \lambda\mathbf{E})\mathbf{x} = \mathbf{o}$, d. h.

$$\begin{array}{rcrcccrcl} (a_{11} - \lambda)x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & 0 \\ a_{21}x_1 & + & (a_{22} - \lambda)x_2 & + & \ldots & + & a_{2n}x_n & = & 0 \\ \multicolumn{9}{c}{\dotfill} \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \ldots & + & (a_{nn} - \lambda)x_n & = & 0 \end{array}$$

schreiben. Man kann verschiedene Aufgaben hinsichtlich Eigenwerten und Eigenvektoren formulieren; so z. B. Man finde einen absolut maximalen Eigenwert, man finde zu einem gegebenen Eigenwert alle Eigenvektoren, man finde alle Eigenwerte und alle Eigenvektoren usw.
Das Gleichungssystem hat genau dann eine nichttriviale Lösung, wenn

$$\mathrm{Det}(\mathbf{A} - \lambda\mathbf{E}) = 0$$

gilt. Nach unserer Theorie der linearen Gleichungssysteme bilden die Eigenvektoren zu einem Eigenwert λ, einen Unterraum U_λ, den man **Eigenraum** zum Eigenwert λ nennt.

Satz 59. *Eigenvektoren zu verschiedenen Eigenwerten sind linear unabhängig.*

Beweis. Ist nämlich $\mathbf{x} = \alpha\mathbf{y}$, und sind $\mathbf{x}$ Eigenvektor zum Eigenwert λ und $\mathbf{y}$ Eigenvektor zum Eigenwert μ, so folgt:

$$\lambda\mathbf{x} = \mathbf{A}\mathbf{x} = \mathbf{A}\alpha\mathbf{y} = \alpha\mathbf{A}\mathbf{y} = \alpha\mu\mathbf{y} = \mu\mathbf{x},$$

also $(\lambda - \mu)\mathbf{x} = o$, woraus sich $\lambda = \mu$ ergibt. □

Satz 60. *Zu jeder (n, n)-Matrix $\mathbf{A}$ gibt es höchstens n verschiedene Eigenwerte.*

Beweis. Nach dem vorangegangenen Satz sind Eigenvektoren zu verschiedenen Eigenwerten linear unabhängig. Im $\mathbb{R}^n$ gibt es aber höchstens n linear unabhängige Vektoren; folglich gibt es höchstens n verschiedene Eigenwerte. □
Die Determinantengleichung

$$\mathrm{Det}(\mathbf{A} - \lambda\mathbf{E}) = 0$$

für die Eigenwerte ist wegen des Zusammenhangs mit der **LU**-Zerlegung eine Polynomgleichung, die höchstens n verschiedene Lösungen hat. Die Determinante ist ein Polynom n-ten Grades in λ, wobei λ^n den Koeffizienten $(-1)^n$ hat. Man nennt sie **charakteristisches Polynom** der Matrix $\mathbf{A}$.

Satz 61. *Eigenvektoren zu verschiedenen Eigenwerten einer symmetrischen Matrix sind orthogonal.*

Beweis. Es seien λ, μ verschiedene Eigenwerte der Matrix $\mathbf{A}$ und $\mathbf{x}, \mathbf{y}$ entsprechende Eigenvektoren:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \quad \mathbf{A}\mathbf{y} = \mu\mathbf{y}.$$

Wir multiplizieren die erste Gleichung skalar mit $\mathbf{y}$ und die zweite skalar mit $\mathbf{x}$; dann folgt

$$\mathbf{y}^\mathsf{T}\mathbf{A}\mathbf{x} = \lambda\mathbf{y}^\mathsf{T}\mathbf{x} = \lambda(\mathbf{x}, \mathbf{y}),$$

$$\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{y} = \mu\mathbf{x}^\mathsf{T}\mathbf{y} = \mu(\mathbf{x}, \mathbf{y}).$$

Die linken Seiten stimmen wegen der Symmetrie von $\mathbf{A}$ überein:

$$\mathbf{y}^\mathsf{T}\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{y})^\mathsf{T}\mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{y},$$

also ergibt sich

$$\lambda(\mathbf{x}, \mathbf{y}) = \mu(\mathbf{x}, \mathbf{y}),$$

woraus wir wegen $\lambda \neq \mu$ schließen, daß $(\mathbf{x}, \mathbf{y}) = 0$ sein muß. □
Ohne Beweis erwähnen wir, daß alle Eigenwerte einer symmetrischen Matrix reell sind.
Sind $\lambda_1, \ldots, \lambda_r$ alle verschiedenen Eigenwerte ($r \leqq n$) der (n, n)-Matrix $\mathbf{A}$ und U_{λ_i} die entsprechenden Eigenräume, so folgt

$$\sum_{i=1}^{r} \dim U_{\lambda_i} \leqq n,$$

und im Falle $r = n$ gilt die Gleichheit.
Beispiel:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \quad \mathbf{A} - \lambda\mathbf{E} = \begin{bmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{bmatrix}.$$

Wir transformieren die Matrix $\mathbf{A} - \lambda\mathbf{E}$ auf Halbdiagonalform und erhalten die Matrix

$$\begin{bmatrix} 1 & 1 & -\lambda \\ 0 & \lambda + 1 & -\lambda^2 + 1 \\ 0 & 0 & -\lambda^2 + \lambda + 2 \end{bmatrix},$$

woraus sich ergibt:

$$\mathrm{Det}(\mathbf{A} - \lambda\mathbf{E}) = 1 \cdot (\lambda + 1) \cdot (-\lambda^2 + \lambda + 2) = 0.$$

Aus dieser Gleichung erhalten wir, daß die Matrix die beiden Eigenwerte -1 und 2 besitzt, wobei -1 zweifacher Eigenwert ist.

2.8. Übungen

1. Man finde unkonventionelle Beispiele für lineare Vektorräume.

2. Es sei $\mathbb{Z}^n$ die Menge aller n-Tupel $(x_1, \ldots, x_n)$ mit $x_i \in \{ 0, 1, \ldots, p-1 \}$, wobei p eine Primzahl ist. Man mache daraus einen Vektorraum über einem geeigneten Körper.

3. Gibt es einen Vektorraum ohne echte Basis? Man begründe die Antwort.

4. Man beweise den folgenden Satz:
   Die Komponenten und Koordinaten eines beliebigen Vektors bezüglich einer Basis aus dem $\mathbb{R}^n$ stimmen genau dann überein, wenn die Basis aus den natürlichen Einheitsvektoren gebildet wird.

5. Auf dem $\mathbb{R}^n$ ist für jedes m eine m-stellige Relation S^m erklärt:

   $$(\mathbf{x}_1, \ldots, \mathbf{x}_m) \in S^m \iff \mathbf{x}_1, \ldots, \mathbf{x}_m \text{ sind linear abhängig.}$$

   Welche Eigenschaften haben diese Relationen?

6. Man schreibe ein Programm für den Algorithmus AUSTAUSCH.

7. Man schreibe ein Programm für den Algorithmus GAUSS.

8. Warum multipliziert man zwei Matrizen in der angegebenen Weise und nicht nach der Regel

   $$\mathbf{AB} = \mathbf{C} \text{ mit } c_{ij} = a_{ij} b_{ij}$$

   oder einer anderen?

9. Es sei $A(\varphi)$ die Menge aller (n, n)-Matrizen, die einer gegebenen linearen Abbildung φ des $\mathbb{R}^n$ in sich zugeordnet sind, wenn man alle Basen des $\mathbb{R}^n$ durchläuft. Welche charakteristischen Eigenschaften haben die Matrizen dieser Menge?

10. Für welche Matrizen $\mathbf{B}$ gilt $\mathbf{AB} = \mathbf{BA}$, wobei

    $$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}?$$

11. Man untersuche, ob die folgenden Mengen Teilräume des $\mathbb{R}^3$ sind.

    (a)

    $$M = \left\{ \begin{pmatrix} x \\ y \\ x^2 + y^2 \end{pmatrix} \;\middle|\; x, y \in \mathbb{R} \right\},$$

    (b)

    $$M = \left\{ \begin{pmatrix} x \\ ax \\ b^2 x + c^2 y \end{pmatrix} \;\middle|\; x, y \in \mathbb{R} \right\} \quad a, b, c \in \mathbb{Z}.$$

12. Man untersuche, ob die folgenden Mengen Teilräume des $\mathbb{R}^3$ sind.

    (a)

    $$M = \left\{ \begin{pmatrix} \alpha \\ \beta \\ \alpha \cdot \beta \end{pmatrix} \;\middle|\; \alpha, \beta \in \mathbb{R} \right\},$$

    (b)

    $$M = \left\{ \mathbf{x} \in \mathbb{R}^3 \;\middle|\; (\mathbf{x}, \mathbf{y}) = 0 \right\},$$

    wobei $\mathbf{y}$ einen beliebigen aber festen Vektor des $\mathbb{R}^3$ bezeichnet.

13. Man prüfe auf lineare Unabhängigkeit.

(a) $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix};$

(b) $\mathbf{x}_1 = \begin{pmatrix} 4 \\ 3 \\ -5 \\ 5 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 2 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 2 \\ 1 \\ -1 \\ 3 \end{pmatrix};$

(c) $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 5 \\ 4 \end{pmatrix};$

(d) $\mathbf{x}_1 = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}.$

14. Für welche Werte von a und b sind die folgenden drei Vektoren linear unabhängig?

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ a \\ b \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ a \\ 1+a \\ 3 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 0 \\ -a \\ 2 \\ 1 \end{pmatrix}.$$

15. Man stelle das Element $\mathbf{x} \in \mathbb{R}^4$,

$$\mathbf{x} = \begin{pmatrix} 3 \\ -1 \\ -2 \\ -1 \end{pmatrix}$$

als Linearkombination der Basis $\{ \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4 \}$ mit

$$\mathbf{b}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ 0 \end{pmatrix}, \ \mathbf{b}_2 = \begin{pmatrix} -1 \\ 2 \\ 0 \\ 1 \end{pmatrix}, \ \mathbf{b}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \ \mathbf{b}_4 = \begin{pmatrix} 1 \\ 1 \\ -2 \\ -1 \end{pmatrix}.$$

dar.

16. Man ordne der linearen Abbildung

$$\varphi : \mathbb{R}^3 \longmapsto \mathbb{R}^4$$

mit

$$\varphi(\begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}) = \begin{pmatrix} -1 \\ 1 \\ -2 \\ 0 \end{pmatrix}, \quad \varphi(\begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix}) = \begin{pmatrix} 0 \\ 1 \\ -3 \\ -1 \end{pmatrix},$$

$$\varphi(\begin{pmatrix} -1 \\ 0 \\ -2 \end{pmatrix}) = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$$

eine Matrix zu.

17. Gegeben sei die lineare Abbildung

$$\varphi : \mathbb{R}^2 \longmapsto \mathbb{R}^3$$

mit

$$\varphi(\begin{pmatrix} 1 \\ 1 \end{pmatrix}) = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \quad \varphi(\begin{pmatrix} -1 \\ 2 \end{pmatrix}) = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}.$$

Welche Bilder haben die Vektoren

$$\begin{pmatrix} 5 \\ 7 \end{pmatrix}, \quad \begin{pmatrix} -4 \\ 3 \end{pmatrix} \quad ?$$

18. Eine lineare Abbildung $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$ ist gegeben durch:

$$\varphi(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}) = \begin{pmatrix} 6 \\ 9 \\ 8 \end{pmatrix}, \quad \varphi(\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}) = \begin{pmatrix} 3 \\ 7 \\ 7 \end{pmatrix}, \quad \varphi(\begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}) = \begin{pmatrix} 1 \\ 4 \\ 7 \end{pmatrix}.$$

(a) Man ermittle die Bilder von

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

(b) Wie lautet die zu φ gehörende Matrix, wenn als Basis in Urbild- und Bildraum jeweils

$$B = \{ \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \}$$

gewählt wird?

19. Es sei durch $\varphi : \mathbb{R}^4 \to \mathbb{R}^3$ eine lineare Abbildung gegeben. Es gelte

$$\varphi(\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}) = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}, \qquad \varphi(\begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}) = \begin{pmatrix} 4 \\ 1 \\ 5 \end{pmatrix},$$

$$\varphi(\begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}) = \begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix}, \qquad \varphi(\begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}) = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}.$$

Man bestimme, die zu φ gehörige Matrix $\mathbf{A}$, wenn im Urbild- bzw. Bildraum jeweils die Basisvektoren

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad \text{bzw.} \quad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

gewählt werden.

20. Durch $\mathbf{y} = \varphi(\mathbf{x}) = \mathbf{A}\mathbf{x}$ mit

$$\mathbf{A} = \begin{bmatrix} -1 & 3 & 2 \\ 2 & 0 & 1 \\ 4 & -2 & 0 \end{bmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

wird eine lineare Abbildung $\varphi : \mathbb{R}^3 \longmapsto \mathbb{R}^3$ beschrieben.

(a) Man bestimme den Kern von φ.

(b) Man bestimme das Bild der Menge $X = \{ \mathbf{x} \in \mathbb{R}^3 \,|\, (1,1,1)^\mathsf{T}\mathbf{x} = 1 \}$.

(c) Man bestimme das Urbild der Menge $Y = \{ \mathbf{y} \in \mathbb{R}^3 \,|\, (1,-2,1)^\mathsf{T}\mathbf{y} = 0 \}$.

21. Es sei $B = \{ \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3 \}$ eine beliebige Basis des $\mathbb{R}^3$ und $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$ eine lineare Abbildung mit

$$\varphi(\mathbf{b}_1) = \mathbf{b}_2, \quad \varphi(\mathbf{b}_2) = \mathbf{b}_3, \quad \varphi(\mathbf{b}_3) = \mathbf{b}_1.$$

Man bestimme die zu φ, $\varphi \circ \varphi$ und zu $\varphi \circ \varphi \circ \varphi$ gehörenden Matrizen, wenn als Basis immer B gewählt wird.

22. Es sei $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ eine lineare Abbildung. Die Bilder der Einheitsvektoren $\varphi(\mathbf{e}_i)$, $i = 1, \ldots, n$, seien linear unabhängig. Man beweise, daß dann die Bilder $\varphi(\mathbf{x}_k)$ linear unabhängiger Vektoren $\mathbf{x}_k \in \mathbb{R}^n$, $k = 1, \ldots, n$, linear unabhängige Vektoren im $\mathbb{R}^m$ sind.

23. Gegeben seien die Vektoren

$$\mathbf{b}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \\ 3 \\ -2 \end{pmatrix}, \quad \mathbf{b}_2 = \begin{pmatrix} 0 \\ 3 \\ -2 \\ 0 \\ 4 \end{pmatrix} \quad \text{und} \quad \mathbf{b}_3 = \begin{pmatrix} 0 \\ 2 \\ -2 \\ 1 \\ 0 \end{pmatrix}.$$

Man zeige die lineare Unabhängigkeit der drei Vektoren und ergänze sie zu einer Basis des $\mathbb{R}^5$. Wie lautet die Darstellung des Vektors

$$\mathbf{c} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

in der neuen Basis.

24. Man finde Algorithmen, die aus gegebenen m ($m \leqq n$) Vektoren des $\mathbb{R}^n$ linear unabhängige machen und schätze den Operationsaufwand für jeden Algorithmus ab. Welcher ist der beste?

25. Gegeben seien die Matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 0 & 2 \\ 1 & 1 & 1 \end{bmatrix}$$

und die durch sie gemäß $\varphi(\mathbf{x}) = \mathbf{Bx}$ vermittelte lineare Abbildung $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$.

(a) Man bestimme $\operatorname{rg}(\mathbf{B})$.

(b) Man beschreibe die Menge $\ker \varphi = \left\{ \mathbf{x} \in \mathbb{R}^3 : \mathbf{Bx} = \mathbf{o} \right\}$. Welche Dimension hat diese Menge?

(c) Ist φ bijektiv?

(d) Man löse

$$\varphi(\mathbf{x}) = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \quad \text{und} \quad \varphi(\mathbf{x}) = \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix}.$$

26. Man zeige:
Die Inverse der Transponierten einer regulären Matrix ist gleich der Transponierten der Inversen: $(\mathbf{A}^{\mathsf{T}})^{-1} = (\mathbf{A}^{-1})^{\mathsf{T}}$.

27. Man zeige:
Die Inverse des Produktes zweier regulärer Matrizen ist gleich dem Produkt der Inversen dieser Matrizen in umgekehrter Reihenfolge: $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

28. Man untersuche die Matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ \lambda & 1 & 2 \end{bmatrix}$$

in Abhängigkeit von λ. Insbesondere bestimme man alle jene Werte von λ, für die die Matrix regulär ist; gegebenenfalls berechne man die Inverse.

29. Die drei Gleichungen

$$2y - 5z = 4, \quad y + 2z = 1, \quad 3y - 3z = 5$$

beschreiben jeweils eine Ebene im $\mathbb{R}^3$. Man berechne den Durchschnitt dieser Ebenen.

30. Man untersuche, für welche Werte λ das Gleichungssystem

$$\begin{array}{rcrcrcl} x & + & y & + & z & = & 3 \\ 3x & + & 5y & + & z & = & 9 \\ 2x & + & 3y & + & z & = & \lambda^2 - 4\lambda + 6 \\ 5x & + & 6y & + & \lambda z & = & 15 \end{array}$$

lösbar ist und bestimme gegebenenfalls die allgemeine Lösung.

31. Für welche Werte von λ ist das System

$$\begin{array}{rcrcrcl} 7x & - & 2y & + & \lambda z & = & 3 \\ -4x & - & 6y & + & 3z & = & 2 \\ 10x & - & 10y & + & 13z & = & 0 \end{array}$$

unlösbar?

32. Man löse:

(a)
$$\begin{array}{rcrcrcrcl} 3x & + & y & + & 2z & + & 2u & = & 1 \\ 4x & + & 2y & + & 2z & + & u & = & 3 \\ x & + & 2y & + & z & + & u & = & 2, \end{array}$$

(b)
$$\begin{array}{rcrcrcl} 2x & + & 3y & + & 3z & = & 7 \\ -x & + & 2y & - & z & = & 1 \\ 2x & + & y & + & 3z & = & 5 \\ 3x & + & y & + & 4z & = & 6. \end{array}$$

33. Man untersuche, ob die folgenden Gleichungssysteme lösbar sind und bestimme gegebenenfalls die allgemeine Lösung:

(a)
$$\begin{array}{rcrcrcrcl} 2x & + & y & - & z & + & u & = & 5 \\ x & + & y & + & 2z & - & u & = & 1 \\ 3x & - & y & + & z & + & u & = & 0 \end{array}$$

(b)
$$\begin{array}{rcrcrcl} 2x & + & y & + & z & = & 5 \\ 2y & + & z & + & u & = & 5 \\ 2z & + & u & + & v & = & 7 \\ 2u & + & v & + & x & = & 12 \\ 2v & + & x & + & y & = & 11 \end{array}$$

(c)
$$\begin{array}{rcrcrcl} u & + & v & - & x & = & -2 \\ 2u & - & z & + & y & = & 5 \\ u & - & 2z & - & 2y & = & 0 \end{array}$$

34. Für $n \in \mathbb{N}$ sei

$$\mathbf{A}_n = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 1 & 1 \\ 1 & 2 & 2 & \dots & 2 & 2 & 2 \\ 1 & 2 & 3 & \dots & 3 & 3 & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 2 & 3 & \dots & n-2 & n-2 & n-2 \\ 1 & 2 & 3 & \dots & n-2 & n-1 & n-1 \\ 1 & 2 & 3 & \dots & n-2 & n-1 & n \end{bmatrix}$$

und

$$\mathbf{b}_n = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 2 \\ \vdots \\ \frac{n}{2} \end{pmatrix} \quad \text{für gerades } n \text{ bzw.} \quad \mathbf{b}_n = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 3 \\ \vdots \\ \frac{n+1}{2} \end{pmatrix} \quad \text{für ungerades } n.$$

Man löse $\mathbf{A}_n\mathbf{x} = \mathbf{b}_n$.

35. Ein Swimmingpool soll durch 10 gleichzeitig arbeitende Pumpen in 20 Stunden entleert werden. Dafür stehen 4 Pumpenarten bereit; die erste Sorte benötigt 800 Stunden pro Pumpe, die zweite 400 Stunden pro Pumpe, die dritte 200 Stunden pro Pumpe und die vierte 100 Stunden pro Pumpe. Man führe in einer Tabelle die verschiedenen Zusammenstellungen der Pumpen auf.

36. Die Summe der Hauptdiagonalelemente einer quadratischen Matrix $\mathbf{A}$ heißt **Spur** $\mathbf{sp}(\mathbf{A})$ der Matrix. Es sei

$$\mathbf{A} = (a_{ij})_{m,n}, \quad \mathbf{B} = (b_{ji})_{n,m}.$$

Man zeige, daß $\mathbf{sp}(\mathbf{AB}) = \mathbf{sp}(\mathbf{BA})$ gilt.

37. Man untersuche, ob die folgenden Matrizen regulär sind und bestimme gegebenenfalls die inverse Matrix:

(a)

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 5 \\ 0 & 1 & 1 & 2 \\ 1 & 0 & 1 & 4 \\ 1 & 1 & 0 & 3 \end{bmatrix},$$

(b)

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ 0 & 1 & 1 & \dots & 1 & 1 \\ 0 & 0 & 1 & \dots & 1 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

38. Man berechne:

(a)

$$\mathrm{Det} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 2 & 1 \end{bmatrix},$$

(b)

$$\mathrm{Det} \begin{bmatrix} 1+\cos x & 1+\sin x & 1 \\ 1-\sin x & 1+\cos x & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

39. Es sei

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 \\ 3 & 0 & 1 \\ 24 & -7 & 1 \end{bmatrix}.$$

Für welche λ ist die Matrix $\mathbf{A} - \lambda\mathbf{E}$ regulär?

40. Man berechne die Determinanten folgender Matrizen:

(a)

$$\mathbf{A} = \begin{bmatrix} 5 & 2 & 3 & 4 \\ 6 & 1 & 3 & 4 \\ 4 & 4 & 2 & 1 \\ 7 & 2 & 1 & 3 \end{bmatrix},$$

(b)

$$\mathbf{A} = \begin{bmatrix} 1 & a & -b \\ -a & 1 & c \\ b & -c & 1 \end{bmatrix},$$

(c)

$$\mathbf{A}_n = \begin{bmatrix} a & b & b & \dots & b & b \\ b & a & b & \dots & b & b \\ b & b & a & \dots & b & b \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ b & b & b & \dots & a & b \\ b & b & b & \dots & b & a \end{bmatrix}.$$

41. Man berechne die Determinante der quadratischen Matrix

$$\mathbf{A}_n = ((i+j-1)^2)_{n,n}.$$

42. Man löse $\det(\mathbf{A}) = 0$ für

$$\mathbf{A} = \begin{bmatrix} x & -1 & x \\ -1 & x & x \\ 1 & 2 & x \end{bmatrix}.$$

43. Es sei $(\cdot,\cdot)$ das Skalarprodukt und $||\cdot||$ die euklidische Norm im $\mathbb{R}^n$. Man beweise:

(a) $||\mathbf{x}+\mathbf{y}||^2 = ||\mathbf{x}||^2 + ||\mathbf{y}||^2 \Longleftrightarrow (\mathbf{x},\mathbf{y}) = 0$,
(b) $||\mathbf{x}+\mathbf{y}||^2 + ||\mathbf{x}-\mathbf{y}||^2 = 2(||\mathbf{x}||^2 + ||\mathbf{y}||^2)$.

44. Man gebe eine Matrizendarstellung für das Erhard-Schmidt-sche Orthogonalisierungsverfahren an und schreibe ein Programm für das Orthogonalisierungsverfahren.

45. Man finde eine orthonormale Basis von $\text{lin}\,(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3)$ mit

$$\mathbf{z}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{z}_2 = \begin{pmatrix} -1 \\ 0 \\ 2 \\ 3 \end{pmatrix}, \quad \mathbf{z}_3 = \begin{pmatrix} 3 \\ 1 \\ 0 \\ 2 \end{pmatrix}.$$

Wie kann man ein $\mathbf{z}_4 \in \mathbb{R}^4$ finden, das zu allen Elementen von $\text{lin}\,(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3)$ orthogonal ist?

46. Man finde eine orthonormale Basis von $\text{lin}\,(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3)$ mit

$$\mathbf{z}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{z}_2 = \begin{pmatrix} 7 \\ -4 \\ 2 \\ 4 \end{pmatrix}, \quad \mathbf{z}_3 = \begin{pmatrix} -18 \\ 9 \\ -3 \\ 0 \end{pmatrix}.$$

Weiterhin gebe man ein $\mathbf{z}_4 \in \mathbb{R}^4$ an, das zu allen Elementen von $\text{lin}\,(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3)$ orthogonal ist.

47. Man finde eine orthonormale Basis von $\text{lin}\,(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3)$ mit

$$\mathbf{z}_1 = \begin{pmatrix} 4 \\ 1 \\ -2 \\ 2 \end{pmatrix}, \quad \mathbf{z}_2 = \begin{pmatrix} 9 \\ 6 \\ -2 \\ 2 \end{pmatrix}, \quad \mathbf{z}_3 = \begin{pmatrix} 0 \\ 9 \\ 7 \\ 2 \end{pmatrix}.$$

Anschließend ergänze man die gefundene Basis zu einer orthonormalen Basis des $\mathbb{R}^4$.

48. Gegeben sei das lineare Gleichungssystem

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

mit

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 3 \\ 2 & 0 & 3 \\ -2 & 2 & -3 \end{bmatrix} \quad \text{und} \quad \mathbf{b} = \begin{pmatrix} 6 \\ 11 \\ -7 \end{pmatrix}.$$

Man löse das System auf folgende Weise:

- Man finde eine orthogonale Matrix $\mathbf{Q}$ und eine obere Dreiecksmatrix $\mathbf{R}$, so daß $\mathbf{A} = \mathbf{Q}\mathbf{R}$ gilt.
- Man berechne die Lösung des Systems $\mathbf{Q}\mathbf{y} = \mathbf{b}$, gemäß $\mathbf{y} = \mathbf{Q}^{\mathsf{T}}\mathbf{b}$.
- Man berechne die Lösung des Systems $\mathbf{R}\mathbf{x} = \mathbf{y}$.

49. Man berechne Eigenwerte und Eigenvektoren der Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

50. Man berechne die Inverse der Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

51. Man löse mit verschiedenen Verfahren das lineare Gleichungssystem $\mathbf{Ax} = \mathbf{y}$ mit

$$\mathbf{A} = (a_{ij})_{n,n} = (\frac{1}{i+j-1})_{n,n}, \quad y_j = 1, j = 1, \ldots, n, \quad n = 5, 6, 7, 12.$$

Dabei bedenke man, daß die Lösungen ganzzahlig sind.

52. Eine Matrix $\mathbf{A} = (a_{ij})_{n,n}$ heißt **streng diagonal dominant**, wenn

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n.$$

Man zeige, daß solche Matrizen regulär sind.

Kapitel 3

Graphentheorie

3.1. Gerichtete und ungerichtete Graphen

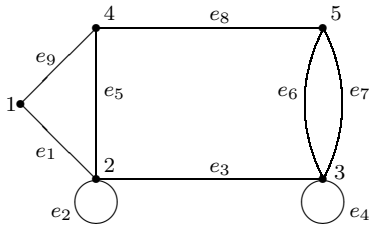
Die Darstellung und Untersuchung realer oder gedanklicher Strukturen ist ein wesentlicher Gegenstand der Informatik. Strukturen treten z. B. bei Rechnernetzen, Programmen, Datenbanken und elektrischen Netzwerken auf. Ihnen allen ist gemeinsam, daß zwischen den Objekten der Struktur gewisse Verbindungen existieren (oder auch nicht). Oft ist es zweckmäßig, von der Art der Verbindungen und den verbundenen Objekten zu abstrahieren und sich nur für die durch die Verbindungen definierte Struktur zu interessieren. In dieser Situation ist die Graphentheorie ein hervorragendes Hilfsmittel zur Strukturbeschreibung und zur Untersuchung von Struktureigenschaften. Wie in jedem mathematischen Gebiet ist auch hier ein gewisser grundlegender Begriffsapparat nötig, um die Sachverhalte in präziser Form aussprechen zu können.
Eine endliche Struktur $G = (V, R_u, R_g)$ heißt **Graph**, falls R_u endlich viele symmetrische Relationen auf V und R_g endlich viele asymmetrische Relationen auf V darstellen. Dabei heißt eine Relation R **asymmetrisch**, falls aus $(x, y) \in R$ mit $x \neq y$ stets $(y, x) \notin R$ folgt. Die Elemente der Trägermenge $V = \{ v_1, \ldots, v_l \}$ heißen **Knoten**. Die Zweiermengen $\{ (x, y), (y, x) \}$, wo die Paare (x, y) und (y, x) aus der gleichen definierenden symmetrischen Relation sind, heißen **ungerichtete Kanten**; alle Einermengen $\{ (x, y) \}$, wo (x, y) aus einer asymmetrischen Relation ist, heißen **gerichtete Kanten**. Auf diese Weise ist jedem Graphen seine wohlbestimmte Kantenmenge $E = \{ e_1, \ldots, e_r \}$ zugeordnet. Oft wird ein Graph auch durch seine Knoten- und Kantenmenge dargestellt: $G = (V, E)$. Kanten der Form (x, x) heißen **Schlingen**. Die obige Beschreibung eines Graphen ist nicht eindeutig; so kann man z. B. jede Kante durch eine Relation beschreiben. Die Beschreibung wird eindeutig, wenn wir zusätzlich fordern, daß je zwei definierende, elementfremde, symmetrische Relationen zu einer zusammenzufassen sind; analog für die asymmetrischen Relationen. Mit dieser Forderung erhalten wir eine Minimalbeschreibung für einen Graphen.
Sind alle definierenden Relationen symmetrisch, heißt der Graph **ungerichtet**; sind alle definierenden Relationen asymmetrisch, heißt der Graph **gerichtet**. Beispiele für Graphen sind:

- V : Menge von Städten, Kanten: Straßen,
- V : Menge von Relaisstationen, Kanten: Leitungen,
- V : Menge der Atome eines Moleküls, Kanten: Bindungen.

Es sei etwa $G = (\{ 1; 2; 3; 4; 5 \}, R_1, R_2)$ mit den symmetrischen Relationen

$$R_1 = \{ (1,2), (2,1), (2,2), (2,3), (3,2), (3,3), (2,4), (4,2), (3,5), (5,3), (4,5), (5,4), (1,4), (4,1) \},$$
$$R_2 = \{ (3,5), (5,3) \}$$

Eine mögliche graphische Darstellung zeigt das folgende Bild.



In Graphen dürfen Schlingen und Mehrfachkanten (d. h. zwei definierende Relationen sind nicht elementfremd) auftreten. Ist in einem Graphen dieses ausgeschlossen, heißt er **schlichter Graph**. Genauer: Ein Graph heißt

schlicht, wenn alle definierenden Relationen irreflexiv und je zwei von ihnen elementfremd sind. Für die Minimalbeschreibung eines Graphen bedeutet dies: Ein Graph heißt schlicht, wenn er durch höchstens eine asymmetrische und/oder höchstens eine symmetrische Relation definiert ist.
Bei unseren weiteren Überlegungen betrachten wir meist nur die reinen Fälle, d. h. die Graphen sollen entweder gerichtet oder ungerichtet sein. In der Graphentheorie wird versucht, eine möglichst anschauliche Sprechweise zu pflegen. Kanten haben Anfangs- und Endknoten. Bei einer ungerichteten Kante $e = \{ (x,y),(y,x) \}$ sind beide Knoten x und y sowohl Anfangs- als auch Endknoten. Bei einer gerichteten Kante $e = \{ (x,y) \}$ ist x der Anfangs- und y der Endknoten. Wir sagen: Die Kante e ist zu dem Knoten x **inzident**, wenn x Anfangsknoten von e ist. Ein Knoten y heißt **Nachbar** eines Knotens x (d. h. y ist **adjazent** zu x), wenn es eine Kante e gibt, so daß x Anfangs- und y Endknoten von e sind. Die Anzahl $d(x)$ aller zu einem Knoten x inzidenten Kanten nennt man **Grad** des Knotens x. Sollte zum Knoten x keine Kante inzident sein, d. h. $d(x) = 0$, so heißt der Knoten **isoliert**. Die **Endknoten** eines Graphen sind gerade jene, die zu genau einer Kante inzident sind. Da in ungerichteten Graphen selbst Schlingen zwei Endknoten haben, gilt

$$\sum_{x \in V} d(x) = 2|E|.$$

Daraus schließen wir

Satz 62. *Die Anzahl der Knoten mit ungeradem Grad ist in einem ungerichteten Graphen stets gerade.*

Ein Knoten x kann durch mehrere Kanten mit einem Knoten y verbunden sein. Dieser Sachverhalt äußert sich in der Graphdefinition darin, daß das Paar (x,y) in mehreren definierenden Relationen vorkommt. Deshalb sei $a_g(x,y)$ die Anzahl der gerichteten Kanten, die vom Knoten x zum Knoten y führen, d. h. die Anzahl der Paare (x,y) in den definierenden asymmetrischen Relationen; entsprechend sei $a_u(x,y)$ die Anzahl der Paare (x,y) in den definierenden symmetrischen Relationen. Wir nennen $a_u(x,y)$ den **ungerichteten Adjazenzgrad** des Knotenpaares (x,y) und $a_g(x,y)$ den **gerichteten Adjazenzgrad** des Knotenpaares (x,y). Die ungerichteten Adjazenzgrade fassen wir in einer Matrix, der **ungerichteten Adjazenzmatrix** $\mathbf{A}_u(G)$, zusammen: In ihr entsprechen jedem Knoten genau eine Zeile und Spalte; im Schnittpunkt der zum Knoten x gehörenden Zeile mit der zum Knoten y gehörenden Spalte steht der ungerichtete Adjazenzgrad $a_u(x,y)$. In analoger Weise bildet man die **gerichtete Adjazenzmatrix** $\mathbf{A}_g(G)$. Offensichtlich ist ein Graph durch seine beiden Adjazenzmatrizen vollständig beschrieben, da die Adjazenzmatrizen die definierenden Relationen charakterisieren. Damit ist die Adjazenz die strukturbestimmende Eigenschaft bei Graphen und wir dürfen einen Graph G auch in der Form $G = (V, \mathbf{A}_u, \mathbf{A}_g)$ darstellen. Wesentlich ist aber zu erwähnen, daß die Darstellung eines Graphen mittels seiner beiden Adjazenzmatrizen eine fixierte Numerierung seiner Knoten voraussetzt. Dies folgt daraus, daß jedem Knoten x genau eine natürliche Zahl i derart zuzuordnen ist, daß dem Knoten x die i-te Zeile und i-te Spalte in den Adjazenzmatrizen zugeordnet ist.
Liegt ein gerichteter bzw. ungerichteter Graph vor, so ist eine der beiden Adjazenzmatrizen die Nullmatrix; diese lassen wir weg und nennen die andere die dem Graphen zugeordnete Adjazenzmatrix $\mathbf{A}(G)$.
Die Adjazenzmatrix ist bei ungerichteten Graphen symmetrisch; bei schlichten, ungerichteten Graphen sind ihre Elemente gleich 0 oder 1 und auf der Hauptdiagonalen stehen nur Nullen. So lautet die Adjazenzmatrix für das obige Beispiel:

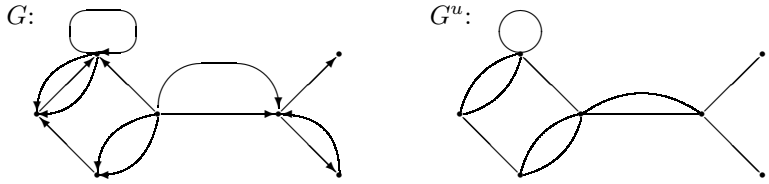
$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 2 & 1 & 1 & 0 \\ 0 & 1 & 2 & 0 & 2 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 2 & 1 & 0 \end{bmatrix}.$$

Beim Grad eines Knotens x in einem gerichteten Graphen müssen wir zwischen der Anzahl $d^+(x)$ der von x wegführenden Kanten – dem **Weggrad** – und der Anzahl $d^-(x)$ der zu x hinführenden Kanten – dem **Hingrad** – unterscheiden. Im ersten Falle ist der Knoten x Anfangspunkt und im zweiten Falle Endpunkt der betreffenden Kante. Aus $d(x) = d^+(x) + d^-(x)$ folgt $|E| = \sum_{x \in V} d^+(x) = \sum_{x \in V} d^-(x)$.
Jeder Graph G ist mit einem wohlbestimmten, ungerichteten Graphen G^u assoziiert, den man seinen **Schatten** nennt: G^u hat die gleiche Knotenmenge wie G, jedoch gilt für die Anzahl $a^u(x,y)$ der Kanten zwischen zwei beliebigen Knoten x, y:

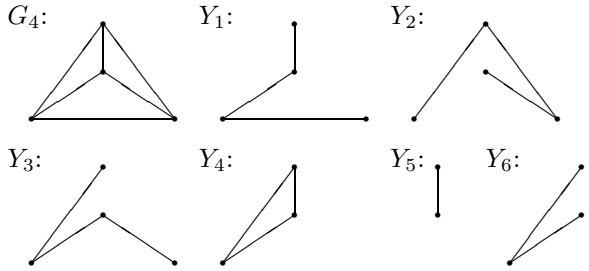
$$a^u(x,y) = a_u(x,y) + \max \{ a_g(x,y), a_g(y,x) \}.$$

Die Situation kann man sich leicht veranschaulichen:
Ein Graph $G' = (V', R'_u, R'_g)$ mit seinen Adjazenzmatrizen $\mathbf{A}_u(G')$ und $\mathbf{A}_g(G')$ heißt **Untergraph** des Graphen $G = (V, R_u, R_g)$ mit seinen Adjazenzmatrizen $\mathbf{A}_u(G)$ und $\mathbf{A}_g(G)$, wenn für je zwei Knoten $x, y \in V'$ die Ungleichungen

$$a'_g(x,y) \leqq a_g(x,y) \text{ und } a'_u(x,y) \leqq a_u(x,y)$$



gelten. Im Falle $V' = V$ heißt ein Untergraph **spannend** . Sind alle Knoten aus G', die in G adjazent sind, auch in G' adjazent, so heißt der Untergraph **gesättigt**. Ein schlichter Graph mit genau r Knoten heißt **abgeschlossen** oder auch **vollständig**, wenn zwischen je zwei Knoten genau eine Kante verläuft. Offenbar gibt es zu jedem r genau einen abgeschlossenen Graphen, den wir mit G_r bezeichnen. Beispielhaft betrachten wir den Graphen G_4 und die folgenden Untergraphen:



Hierin sind die Graphen Y_1, Y_2, Y_3 spannend, Y_4, Y_5 sind gesättigt, Y_6 ist nur ein einfacher Untergraph.

3.1.1. Isomorphie von Graphen

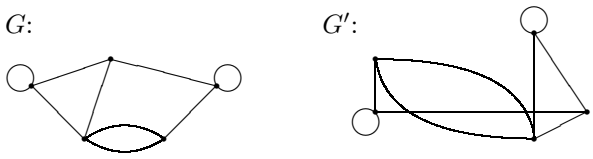
Im letzten Abschnitt haben wir insbesondere erkannt, daß die Adjazenzmatrix sowohl einen gerichteten als auch einen ungerichteten Graphen vollständig charakterisiert. Zwei Graphen $G = (V, R_u, R_g)$ und $G' = (V', R'_u R'_g)$ sind **isomorph**, wenn eine bijektive Abbildung

$$\psi : \quad V \longmapsto V'$$

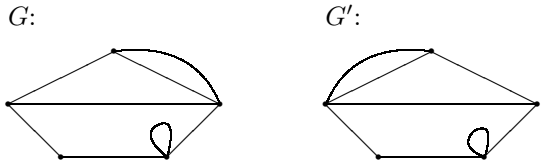
der Knoten von G auf die Knoten von G' derart existiert, daß sich die gerichteten und ungerichteten Adjazenzgrade jedes Knotenpaares nicht ändern:

$$a'_g(\psi(x), \psi(y)) = a_g(x, y)$$
$$a'_u(\psi(x), \psi(y)) = a_u(x, y) \quad \forall x, y \in V,$$

d. h. wenn von einem Knoten x zu einem Knoten y im Graphen G genau r Kanten führen, so muß dies auch für die Bildknoten im Graphen G' gelten. Bei der Graphenisomorphie bleibt somit die strukturbestimmende Eigenschaft, die Adjazenz, erhalten. So sind z. B. die beiden folgenden Graphen



isomorph, da man die Knoten und Kanten so numerieren kann, daß die Adjazenzen in beiden übereinstimmen. Generell kann man sagen: Zwei Graphen sind genau dann isomorph, wenn man die Knoten des einen Graphen so umnumerieren kann, daß die entsprechenden Adjazenzmatrizen mit denen des anderen Graphen übereinstimmen.
Zwei abgeschlossene Graphen mit gleicher Knotenzahl sind offenbar isomorph, so daß man von dem abgeschlossenen Graphen mit n Knoten sprechen kann. Das entscheidende Problem bei der Isomorphie ist hier, daß man aus der Darstellung von Graphen im allgemeinen nicht auf ihre Isomorphie schließen kann und der Isomorphienachweis algorithmisch sehr aufwendig ist. Stellt man Graphen mittels ihrer Adjazenzmatrizen dar, so wird die Isomorphie dadurch entschieden, daß man durch Zeilen- und Spaltenvertauschungen in den Adjazenzmatrizen diese als gleich zu identifizieren hat. Eine negative Entscheidung über die Isomorphie kann oft durch Vergleich gewisser charakteristischer Größen herbeigeführt werden. Solche Größen sind etwa die Knotenanzahl, die Kantenanzahl, die aufsteigende Gradfolge, maximaler Grad, minimaler Grad, Untergraphen eines gewissen Typs. Ist der Wert einer solchen Größe für zwei Graphen verschieden, so können diese nicht isomorph sein. Leider ist kein endliches System von charakteristischen Größen bekannt, aus deren Gleichheit man auf die Isomorphie schließen kann. So haben z. B. die beiden Graphen



gleiche Knoten- und Kantenzahlen, übereinstimmende aufsteigende Gradfolgen, minimaler und maximaler Grad sind gleich; trotzdem sind sie nicht isomorph, da die beiden Knoten mit der Schlinge aufeinander abgebildet werden müssen.
Für die graphentheoretische Beschreibung von Automaten benötigen wir den Begriff des bewerteten, gerichteten Graphen. Eine Struktur $G = (V, R_g, M, \sigma)$ heißt **bewerteter, gerichteter Graph**, wenn (V, R_g) ein gerichteter Graph ist und

$$\sigma : \quad E \longmapsto M$$

eine Abbildung der Kanten in die Menge M, die Bewertungsmenge des Graphen, darstellt, also jeder Kante zusätzlich eine Bewertung in Form eines Elementes aus der Menge M zugeordnet ist. Zwei bewertete, gerichtete Graphen G, G' sind isomorph, wenn sie zunächst als gerichtete Graphen isomorph sind und sich überdies die Bewertungen von sich einander entsprechenden Kanten bijektiv aufeinander abbilden lassen.

3.1.2. Zusammenhang

Sind zwei Knoten eines Graphen nicht durch eine Kante verbunden, fragt man nach einem Weg von einem Knoten zum anderen. Unter einem **Weg** zwischen zwei Knoten x, y eines ungerichteten Graphen $G = (V, R_u)$ versteht man eine endliche Folge $x_0, e_1, x_1, e_2, \ldots, e_n, x_n$ von Knoten und Kanten aus G, bei denen die auftretenden Kanten mit den rechts und links von ihnen stehenden Knoten inzident sind und $x_0 = x, x_n = y$ gilt. Falls ein Weg zwischen den Knoten x und y existiert, nennt man x und y durch einen Weg verbunden. Um eine innerhalb der Informatik typische Form der Definition, die **induktive Definition** zu verwenden, wollen wir den Wegbegriff induktiv definieren. Es sei $G = (V, R_u)$ ein ungerichteter Graph, die Menge $W(G)$ aller Wege in G ist dann durch folgende Regeln charakterisiert:

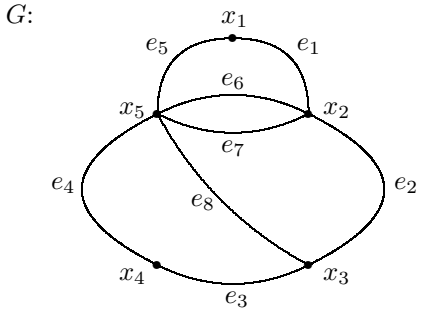
1. Jeder Knoten $x \in V$ ist ein Weg.
2. Es seien $w = x, \ldots, y$ und $w' = u, \ldots, v$ Wege.
   - Fall $y \neq u$: Gibt es eine Kante e mit y als Anfangs- und u als Endknoten (d. h. u ist adjazent zu y), so ist $x, \ldots, y, e, u, \ldots, v$ ein Weg.
   - Fall $y = u$: Es ist $x, \ldots, y, \ldots, v$ ein Weg.
3. Weitere Wege gibt es nicht.

Durch diese Regeln sind nicht nur alle Wege innerhalb eines ungerichteten Graphen definiert: Wir haben damit auch eine formale Entscheidungsgrundlage, die es uns gestattet, von einem vorgegebenen Objekt in endlich vielen Schritten zu entscheiden, ob das Objekt ein Weg ist oder nicht. Wege sind mittels der Regel 2 aus Knoten und adjazenten Kanten aufgebaut. Zu jedem Weg gibt es eine natürliche Zahl n, so daß man ihn durch n-malige Anwendung der Regel 2 aus den Knoten des Graphen gewinnen kann.
Ein Weg von x nach y heißt **einfach**, wenn seine Kanten alle verschieden sind. Ein Weg heißt **elementar**, wenn alle seine Knoten unterschiedlich sind mit eventueller Ausnahme von Anfangs- und Endknoten. Stimmen Anfangs- und Endknoten in einem Weg überein, so sprechen wir von einem **Kreis**; entsprechend von einfachen bzw. elementaren Kreisen. Im folgenden werden wir stets einfache Kreise meinen, wenn wir von Kreisen sprechen. Einfache oder elementare Wege und Kreise können in schlichten Graphen durch die Aufeinanderfolge ihrer Knoten beschrieben werden. Jeder Weg enthält einen elementaren Weg, den man beim Durchlaufen dadurch erhält, daß man aus dem Weg alle jene Knoten und Kanten streicht, die zum zweiten Male durchlaufen werden sollen. Analog enthält jeder Kreis einen elementaren. Unter der **Länge** eines Weges wird die Anzahl seiner Kanten verstanden. Betrachten wir z. B. den folgenden Graphen G:
Ein Weg hierin ist z. B.

$$x_1, e_1, x_2, e_6, x_5, e_6, x_2, e_7, x_5, e_8, x_3, e_3, x_4.$$

Die Folge

$$x_3, e_3, x_4, e_4, x_5, e_8, x_3, e_2, x_2, e_7, x_5$$



ist ein einfacher Weg, während

$$x_1, e_1, x_2, e_6, x_5, e_8, x_3, e_3, x_4$$

ein elementarer Weg ist. Ein Kreis der Länge 2 ist der Weg

$$x_5, e_6, x_2, e_7, x_5,$$

und ein Kreis der Länge 5 ist durch die Folge

$$x_5, x_4, x_3, x_2, x_1, x_5$$

gegeben.
Existiert zwischen zwei Knoten ein Weg, so gibt es auch einen kürzesten, d. h. einen Weg mit kleinster Länge. Der **Abstand** $d(x, y)$ zweier Knoten x und y des Graphen G ist die Länge des kürzesten Weges zwischen beiden Knoten; sollte kein Weg zwischen den betrachteten Knoten exisiteren, wird $d(x, y) = \infty$ gesetzt. Man überlegt sich leicht, daß der Abstand von Knoten Eigenschaften hat, die uns schon beim Abstand von Vektoren in einem euklidischen Vektorraum begegnet sind:

$$d(x, y) \geqq 0 \quad \text{und} \quad d(x, y) = 0 \Longleftrightarrow x = y,$$

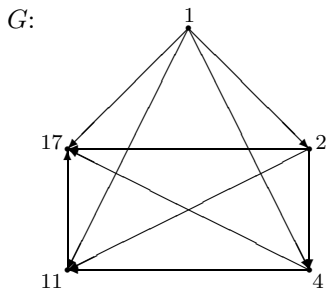
$$d(x, y) = d(y, x),$$

$$d(x, y) \leqq d(x, z) + d(z, y) \quad \text{Dreiecksungleichung.}$$

Wir nennen einen Graphen **zusammenhängend**, wenn es zwischen je zwei Knoten stets einen Weg gibt, d. h. wenn je zwei Knoten einen endlichen Abstand haben; andernfalls heißt er **unzusammenhängend**. Schließlich ist die **Komponente** $K(x)$ eines Knotens x die Menge aller jener Knoten, die durch einen Weg von x aus erreichbar sind, also einen endlichen Abstand von x haben:

$$K(x) = \{\, y \in V \mid d(x, y) < \infty \,\}.$$

Bei zusammenhängenden Graphen ist stets $K(x) = V$ für alle $x \in V$, während man bei unzusammenhängenden Graphen endlich viele Knoten $x_1, \ldots, x_n$ finden kann, so daß die Menge der Komponenten $K(x_1), \ldots, K(x_n)$ eine Zerlegung des Graphen in zusammenhängende Untergraphen bilden. Jeder beliebig gewählte Untergraph von G, der eine echte Komponente hat, kann offenbar nicht zusammenhängend sein.
In gerichteten Graphen sind die Begriffe analog. Man muß sich unter den Kanten stets nur gerichtete Kanten vorstellen. Bei gerichteten Graphen ist zwischen stark und schwach zusammenhängend zu unterscheiden. Ein gerichteter Graph heißt **stark zusammenhängend**, wenn je zwei Knoten durch einen gerichteten Weg verbunden sind und **schwach zusammenhängend**, wenn sein Schatten zusammenhängend ist. Als Beispiel sei G der Graph für die $<$-Relation auf der Menge $\{\, 1, 2, 4, 11, 17 \,\}$:



Ein gerichteter Graph, der keinen gerichteten Kreis enthält, heißt **azyklisch**. Diese Bezeichnung rührt daher, daß ein gerichteter Kreis oft auch Zyklus genannt wird. Ein Knoten ohne wegführende Kanten heißt **Senke** in einem gerichteten Graphen; ein Knoten ohne hinführende Kanten heißt **Quelle** des Graphen.

Satz 63. *Jeder azyklische, gerichtete Graph G hat sowohl eine Quelle als auch eine Senke.*

Beweis. In G gibt es einen Weg w maximaler Länge; ein solcher habe den Endpunkt x; im Falle $d^+(x) > 0$ gäbe es eine aus x herausführende, gerichtete Kante. Da der Weg w maximale Länge hat, muß diese Kante zu einem in w bereits vorkommenden Knoten führen, wodurch man einen gerichteten Kreis gewonnen hätte, was aber in einem azyklischen Graphen unmöglich ist. Folglich ist der Endpunkt jedes Weges maximaler Länge eine Senke. Analog zeigt man, daß der Anfangspunkt jedes Weges maximaler Länge eine Quelle des Graphen sein muß. □

Die Aussage des Satzes kann man ausnutzen, um zu entscheiden, ob ein gerichteter Graph azyklisch ist oder nicht: Man streiche alle Quellen einschließlich aller aus ihnen herausführenden Kanten. Dies wiederhole man solange, bis keine Kanten mehr existieren - in diesem Falle ist der Graph azyklisch - bzw. bis ein Untergraph entsteht, der keine Quellen hat. In gleicher Weise kann man mit den Senken verfahren; beide Vorgehensweisen dürfen auch gemischt werden.

In Rechnernetzen spielt u. a. die Frage nach solchen Knotenrechnern eine Rolle, von denen man jeden Rechner eines gewissen Unternetzes erreichen kann. Kennt man solche Rechner (oder gar alle), so braucht man nur an diese Informationen zu senden und ist sicher, daß alle Teilnehmer der Unternetze erreicht werden. Natürlich sollten in einem Netz möglichst wenig Knotenrechner installiert sein. Eine **Basis** eines gerichteten Graphen G ist daher eine minimale Untermenge B seiner Knotenmenge V derart, daß jeder Knoten aus V von einem Knoten aus B erreichbar ist, d. h. zu jedem Knoten $y \in V$ existiert ein Knoten $x \in B$, so daß ein gerichteter Weg von x nach y führt. Dabei setzt man zusätzlich fest, daß jeder Knoten x von x erreichbar ist. Offenbar hat jeder gerichtete Graph eine Basis. Jede Basis muß sicherlich alle Quellen des Graphen enthalten. Bei stark zusammenhängenden Graphen sind die Basen einelementig: Jeder Knoten bildet eine Basis. Allgemein ist eine Basis B eines gerichteten Graphen durch die folgenden beiden Bedingungen charakterisiert:

- Jeder Knoten ist von einem Knoten aus B erreichbar.
- Kein Knoten aus B ist von einem anderen aus B erreichbar.

Ein **Eulergraph** ist ein ungerichteter Graph, in dem es einen einfachen Kreis über alle Kanten gibt. Der betreffende Kreis heißt dann **Eulerkreis**. Eulergraphen sind gerade solche, deren graphische Darstellung in einem Zuge (ohne Absetzen) gezeichnet werden kann, wobei man zum Anfangsknoten zurückkehrt und jede Kante nur einmal durchlaufen wurde.

Satz 64. *Ein zusammenhängender Graph ist genau dann eulersch, wenn jeder seiner Knoten einen geraden Grad hat.*

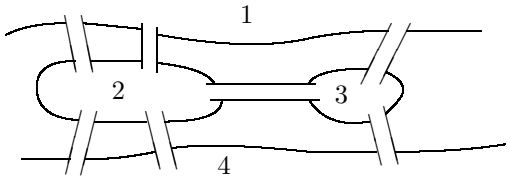
Beweis. Es sei im Graphen G ein Eulerkreis w gegeben. Dann muß es zu jedem Knoten von w eine Kante geben, auf der man zu ihm gelangt und eine weitere, auf der man ihn wieder verläßt. Tritt also ein Knoten x genau k-mal im Kreis w auf, so ist $d(x) = 2k$.

Es sei andererseits G ein zusammenhängender Graph mit n Kanten, jeder Knoten habe geraden Grad. Wir zeigen, wie man zu einem Eulerkreis kommt.

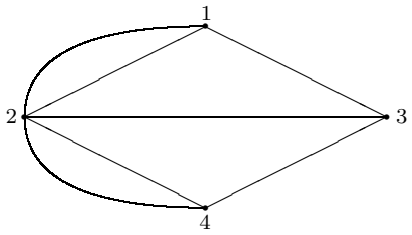
Von einem beliebigen Knoten x_1 aus starten wir das Durchlaufen eines einfachen Weges solange, wie noch eine nicht durchlaufene Kante vorhanden ist. Kann der Weg nicht fortgesetzt werden, muß er in x_1 enden, da in jedem anderen Knoten x ein Verlassen des Knotens möglich ist ($d(x)$ ist gerade). Enthält der Weg alle Kanten von G, sind wir fertig. Andernfalls streichen wir aus G alle durchlaufenen Kanten und alle danach isolierten Knoten. Nun ist der Graph in zusammenhängende Untergraphen zerfallen; jeder dieser Untergraphen hat weniger als n Kanten. Auf jeden einzelnen Untergraphen wenden wir das letzte Vorgehen an usw. bis wir nur noch Eulerkreise erhalten haben. Wir wollen nun alle entstandenen Eulerkreise von Untergraphen zu einem Eulerkreis von G zusammenfügen. Dazu nehmen wir an, daß die Methode nach r Schritten endet. Es sei V_i die Menge aller Knoten, die im i-ten Schritt durch Streichen von Kanten isoliert wurden und E_i die Menge aller Eulerkreise aus dem i-ten Schritt. So ist z. B. $x_1 \in V_1$. Alle Eulerkreise aus E_2 hängen an Knoten des einzigen Eulerkreises k_1 aus E_1; alle Eulerkreise aus E_i hängen an Knoten von Eulerkreisen aus E_{i-1} $(i = 1, \ldots, r)$. Durch folgendes Durchlaufen erhalten wir einen Eulerkreis von G: Wir beginnen in $x_{11} = x_1$ und laufen bis zum ersten Knoten $x_{12} \in V_2$; der Knoten x_{12} gehört auch zu einem Eulerkreis $k_2 \in E_2$ aus dem Schritt 2; diesen durchlaufen wir ab $x_{21} = x_{12}$ bis zum ersten Knoten, der zu einem Eulerkreis $k_3 \in E_3$ gehört usw. bis wir zu einem Eulerkreis $k_r \in E_r$ gekommen sind; an diesen hängen keine weiteren Eulerkreise. Nun steigen wir wieder suksessive bis zum Eulerkreis k_1 aus E_1 auf, indem k_{r-1} vollständig durchlaufen wird einschließlich aller an ihm hängenden Eulerkreis aus Schritt r; danach ist man in den Knoten $x_{r-1,1}$ zurückgekehrt und kann das Durchlaufen des Eulerkreises k_{r-2} fortsetzen usw. Auf diese Weise werden schließlich k_1 und alle nachgeordneten Eulerkreise durchlaufen. □

Mit diesen Überlegungen ist insbesondere das berühmte **Königsberger Brückenproblem** aus der Zeit Eulers (1736) gelöst worden. Die Aufgabe besteht in folgendem. Zu Eulers Zeiten gab es in Königsberg einige Brücken über die Pregel, wodurch auch zwei Inseln im Fluß mit dem Festland verbunden waren. Die Situation wird durch das folgende Bild dargestellt:

Die Aufgabe bestand nun darin zu entscheiden, ob es möglich ist, ausgehend von einem beliebigen Ort genau



einmal über alle Brücken wandernd zum Ausgangspunkt zurückzukehren. Offenbar entspricht der obigen Situation der folgende ungerichtete Graph:



Die Aufgabe bedeutet nun zu entscheiden, ob ein Eulergraph vorliegt oder nicht. Offensichtlich ist dies kein Eulergraph, da jeder Knoten ungeraden Grad hat.
Die getroffene Charakterisierung von Eulergraphen gibt uns die Möglichkeit zu entscheiden, wann ein Graph in einem Zuge gezeichnet werden kann, d. h. ob es einen einfachen Weg über alle Kanten des Graphen gibt. Dies ist genau dann der Fall, wenn er ein Eulergraph ist oder genau zwei Knoten x, y ungeraden Grades enthält. Im zweiten Falle startet man nämlich im Knoten x und durchläuft alle Kanten genau einmal, um im Knoten y zu enden. Erreicht man einen Knoten z mit geradem Grad, so kann man ihn stets auf einer anderen Kante verlassen. Kehrt man zum Ausgangspunkt zurück, kann man ihn ebenfalls auf einer noch nicht durchlaufenen Kante verlassen. Die einzige Ausnahme macht der Knoten y, da jedes Eintreffen in y und Verlasssen von y zwei durchlaufene Kanten ergibt, so daß man schließlich in y endet.
Eine erheblich schwierigere graphentheoretische Aufgabe ist die folgende. Ein Graph enthält einen **Hamiltonkreis**, wenn er einen elementaren Kreis über alle Knoten enthält. In vielen Anwendungen wird nach einem kürzesten Hamiltonkreis in einem gerichteten, bewerteten Graphen gefragt; so z. B. beim sog. **Rundreiseproblem**: Man möchte von einer Stadt ausgehend eine vorgegebene Anzahl von Städten bereisen und dabei minimale Reisekosten verursachen. Alle bisher bekannten Algorithmen zur exakten Lösung dieser Aufgabe haben ein exponentielles Aufwandsverhalten in Abhängigkeit von der Städtezahl und sind daher schon bei einer geringen Städtezahl (ca. 50) aus Zeitgründen praktisch undurchführbar.
Wir wollen hier auch kurz das wohl berühmteste Graphenproblem, das **Vierfarbenproblem** erwähnen. Es lautet wie folgt: Bekanntlich kann man jeder Landkarte einen Graphen zuordnen: Die Knoten bestehen aus den Ländern (zusammenhängende Wasserflächen bilden auch Länder). Zwei Knoten werden durch eine Kante verbunden, wenn sie eine gemeinsame Grenze haben, die sich nicht auf einen Punkt reduziert. Eine Färbung der Landkarte mit m Farben soll regulär heißen, wenn je zwei Länder mit einer gemeinsamen Grenze auch verschiedene Farben haben. Relativ einfach läßt sich zeigen, daß man mit 5 Farben jede gegebene Landkarte regulär färben kann. Andererseits ist es auch einfach, jede konkret vorgelegte Landkarte mit 4 Farben regulär zu färben. Alle Versuche, dies auch mathematisch zu beweisen, sind bisher fehlgeschlagen. Im Jahre 1976 wurde ein Beweis vorgelegt, der das Vierfarbenproblem auf die Untersuchung der regulären Färbung einer großen Anzahl spezieller Graphen reduziert (dies war schon seit etwa 1896 bekannt) und mittels Rechnerprogammen diese Frage für alle auftretenden Graphen positiv entscheidet. Leider gibt es keinen Beweis für die korrekte Arbeitsweise dieser Programme, weshalb dieser Beweis von Mathematikern auch nicht als vollwertig anerkannt ist. Der Korrektheitsnachweis ist in der Tat wesentlich, denn es werden immer noch Fehler in den Programmen gefunden (die bisher alle reparabel waren). Auch wenn über einen langen Zeitraum keine Fehler gefunden werden, bleibt die Situation unbefriedigend, wenngleich den Mathematikern die Show gestohlen wurde.

3.2. Relationen, Graphen und Automaten

Binäre Relationen lassen sich durch Matrizen darstellen. Ist r eine Relation über X mit $X = \{ x_1, \ldots, x_n \}$, so ist r die Matrix $\mathbf{R} = (r_{ij})_{n,n}$ mit $r_{ij} = 1$, falls $(x_i, x_j) \in r$ und $r_{ij} = 0$ sonst zugeordnet. Es sei s eine weitere Relation über X, der die Matrix $\mathbf{S} = (s_{jk})_{n,n}$ zugordnet ist. Dann gilt

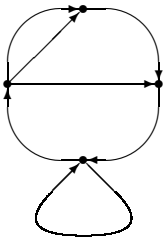
$$\mathbf{T} = \mathbf{R} \cdot \mathbf{S} = (t_{ik})_{n,n}, \quad t_{ik} = \sum_{j=1}^{n} r_{ij} s_{jk}.$$

Die Größe t_{ik} ist also die Anzahl der Werte j, für die $r_{ij} = s_{jk} = 1$ gilt, also die Anzahl der Möglichkeiten, von x_i nach x_k zu gelangen:

$$(x_i, x_k) \in r \circ s \Longleftrightarrow t_{ik} \neq 0.$$

Ebenso folgt für t Relationen $r_1, \ldots, r_t$ über X, daß der Relation $u = r_1 \circ r_2 \circ \ldots \circ r_t$ die Matrix $\mathbf{U} = \mathbf{R}_1 \mathbf{R}_2 \ldots \mathbf{R}_t = (u_{ik})_{n,n}$ zugeordnet ist und u_{ik} angibt, auf wieviele Arten man von x_i über eine Folge von Elementen aus X nach x_k gelangen kann.
Eine binäre Relation r über einer endlichen Menge $X = \{ x_1, \ldots, x_n \}$ kann auch durch einen gerichteten Graphen dargestellt werden. Dazu ordnen wir der Relation r einen gerichteten Graphen $G = (V, R_g)$ zu, wobei $V = X$ gilt und im Falle $(x, y) \in r$ vom Knoten x zum Knoten y eine Kante e führt. Es sei $\mathbf{A}(G) = (a_{ij})_{n,n}$ die Adjazenzmatrix des Graphen G; dann gibt a_{ij} die Anzahl der gerichteten Kanten von x_i nach x_j an, wie in dem folgenden Beispiel:

G:



$$\mathbf{A}(G) = \begin{bmatrix} 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Hat G keine Mehrfachkanten, so ist $a_{ij} = 1$ genau dann, wenn $(x_i, x_j) \in E$ und $a_{ij} = 0$ sonst. Also ist $\mathbf{A}(G)$ in diesem Falle die zur Relation r gehörende Matrix. Wir definieren die Relation $r^2 = r \circ r$ durch: $(x, y) \in r^2$ genau dann, wenn ein z existiert mit $(x, z) \in r$ und $(z, y) \in r$; entsprechend ist r^l definiert: $(y_0, y_l) \in r^l$ genau dann, wenn $y_1, \ldots, y_{l-1}$ existieren mit $(y_i, y_{i+1}) \in r, i = 0, \ldots, l-1$. Offenbar ist dann $\mathbf{A}^l(G)$ die zur Relation r^l gehörende Matrix. Im folgenden Satz sind einige Eigenschaften der Adjazenzmatrix zusammengestellt.

Satz 65. *Gegeben sei eine binäre Relation r über einer endlichen Menge $X = \{ x_1, \ldots, x_n \}$. Es sei ferner G der zugeordnete, gerichtete Graph und $\mathbf{A}(G)$ die zum Graphen gehörende Adjazenzmatrix. Dann gilt*

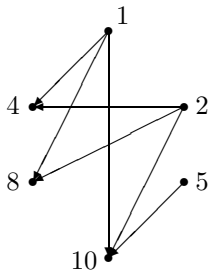
1. *Die Adjazenzmatrix ist genau dann symmetrisch, wenn die Relation r symmetrisch ist.*
2. *Die Zeilensumme ist gleich dem Weggrad, die Spaltensumme gleich dem Hingrad des zugeordneten Knotens:*
$$\sum_{j=1}^{n} a_{ij} = d^+(x_i), \quad \sum_{i=1}^{n} a_{ij} = d^-(x_j).$$
3. *Es sei G schlicht und $\mathbf{A}^k(G) = (a_{ij}^{(k)})_{n,n}$ das k-fache Produkt von $\mathbf{A}(G)$ mit sich. Dann ist $a_{ij}^{(k)}$ die Anzahl der gerichteten Wege der Länge k von x_i nach x_j.*
4. *Der Graph G ist genau dann azyklisch, wenn es eine Zahl l gibt mit $\mathbf{A}^l(G) \neq \mathbf{o}$ und $\mathbf{A}^{l+1}(G) = \mathbf{o}$, d. h. wenn es einen gerichteten Weg größter Länge gibt.*

Als Beispiel sei folgende Relation gegeben:

$$A = \{ 1, 2, 5 \}, B = \{ 4, 8, 10 \}, \quad (a, b) \in r \Longleftrightarrow a \text{ teilt } b.$$

Der zugehörende Graph sieht hier so aus:
Die zugeordnete Adjazenzmatrix lautet:

$$\mathbf{A}(G) = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$



Um den Zusammenhang zwischen Graphen und Automaten herauszuarbeiten, wollen wir zunächst den Begriff des Automaten mathematisch präzise einführen. Ein (endlicher) **Automat** $\mathcal{A}$ ist ein 5-Tupel

$$\mathcal{A} = (K, \Sigma, T, \sigma, \lambda),$$

wobei K, Σ, T Mengen und σ, λ Abbildungen sein mögen. Ein Element aus der Menge K nennen wir **Zustand** des Automaten $\mathcal{A}$, ein Element aus Σ heißt **Eingabe** für den Automaten und ein Element aus T heißt **Ausgabe** des Automaten $\mathcal{A}$. Man nennt die Menge Σ das **Eingabealphabet**, T das **Ausgabealphabet** und K die **Zustandsmenge** des Automaten. Mittels einer Eingabe wird der Automat von einem Zustand in einen weiteren überführt; daher heißt σ die **Überführungsfunktion** des Automaten:

$$\sigma: \quad K \times \Sigma \longmapsto K; (q, x) \in K \times \Sigma \longmapsto \sigma(q, x) \in K.$$

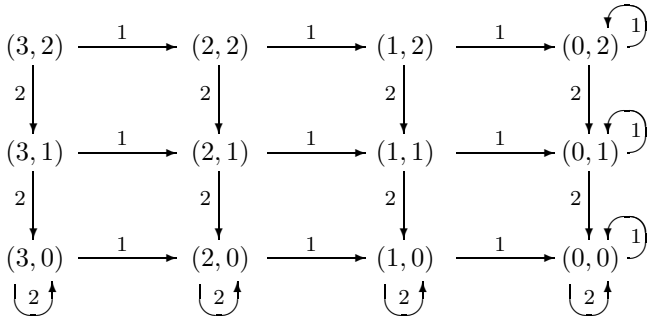
Die Befähigung des Automaten zur Ausgabe von Daten wird durch die **Ausgabefunktion** λ beschrieben:

$$\lambda: \quad K \times \Sigma \longmapsto T; (q, x) \in K \times \Sigma \longmapsto \lambda(q, x) \in T.$$

Zur Illustration kann man sich etwa einen Briefmarkenautomaten vorstellen, der bei Einwurf eines Markstückes eine Eine-Mark-Marke auswirft und bei Einwurf eines Zweimarkstückes eine Zwei-Mark-Marke. Als Zustand des Automaten sehen wir die Anzahl der noch vorhandenen Eine-Mark-Marken und die Anzahl der noch vorhandenen Zwei-Mark-Marken an, d. h. der Zustand wird durch ein Paar (x, y) von natürlichen Zahlen beschrieben. Jeder Einwurf eines Geldstückes ändert den Zustand. Jede Ausgabe ist eine Eine-Mark- oder eine Zwei-Mark-Marke. Zum richtigen Funktionieren des Briefmarkenautomaten gehört es natürlich, daß er im Zustand (0,y) bei Eingabe eines Markstückes das Geld wieder auswirft; entsprechend im Zustand (x,0) bei Einwurf eines Zweimarkstückes. Schließlich sollte nicht unerwähnt bleiben, daß unser Briefmarkenautomat bei Einwurf anderer Objekte diese ohne Änderung seines Zustandes wieder auswirft. Diese Eigenschaft wollen wir als selbstverständlich voraussetzen und nicht in unser Modell aufnehmen. Ein Automat arbeitet nach folgendem Prinzip: Auf Grund einer Eingabe ändert sich in definierter Weise der Zustand des Automaten. Nach diesem Grundprinzip arbeiten gegenwärtig auch alle Rechner.
Die inneren Verhältnisse eines Automaten können durch einen bewerteten, gerichteten Graphen $G = (V, \mathbf{A}, \Sigma, \sigma)$ beschrieben werden. Dabei ist die Knotenmenge V von G die Zustandsmenge K des Automaten; zwei Knoten q_1, q_2 sind durch so viele gerichtete Kanten verbunden, wie es Eingaben gibt, die den Automaten aus dem Zustand q_1 in den Zustand q_2 bringen, d. h. für die Adjazenzmatrix gilt

$$a(q_1, q_2) = |\{ x \in \Sigma \mid \sigma(q_1, x) = q_2 \}|.$$

Die einer Kante zugeordnete Eingabe ist die Bewertung der betreffenden Kante. Wir skizzieren von unserem Briefmarkenautomaten jenen Untergraphen G, der den Automaten vom Zustand $(3, 2)$ bis zum Zustand $(0, 0)$ beschreibt:



Graphentheoretisch können wir einige Eigenschaften von Automaten interpretieren. Wenn der Automat durch

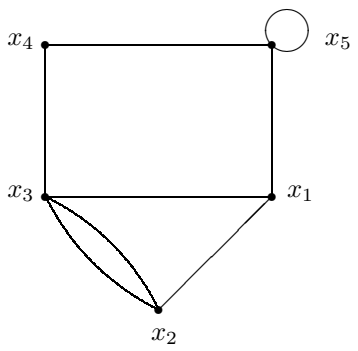
l Eingaben von einem Zustand $q \in K$ in einen Zustand $g' \in K$ übergehen kann, so gibt es in dem zugeordneten Graphen einen Weg der Länge l von q nach q'. Kann der Automat von einem gewissen Zustand in jeden anderen übergehen, ist der zugeordnete Graph schwach zusammenhängend. Sein Graph ist stark zusammenhängend, falls jeder Zustand des Automaten aus jedem anderen erzeugbar ist. Der Graph enthält einen Kreis, falls der entsprechende Automat von einem gewissen Zustand in diesen zurückgeführt werden kann. Die Isomorphie von Automaten wird so gefaßt, daß die zugeordneten Graphen isomorph sind. Zwei Automaten $\mathcal{A} = (K, \Sigma, T, \sigma, \lambda)$ und $\mathcal{A}' = (K', \Sigma', T', \sigma', \lambda')$ heißen **isomorph** , falls es bijektive Abbildungen $\varphi_K, \varphi_\Sigma, \varphi_T$ gibt, so daß für alle $q \in K$ und alle $x \in \Sigma$:

$$\varphi_K(\sigma(q,x)) = \sigma'(\varphi_K(q), \varphi_\Sigma(x)),$$
$$\varphi_T(\lambda(q,x)) = \lambda'(\varphi_K(q), \varphi_\Sigma(x)).$$

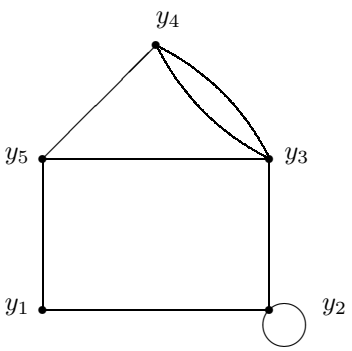
Bei isomorphen Automaten erreicht man also die gleiche Zustandsänderung bzw. Ausgabe unabhängig davon, ob man zuerst den ersten Automaten bedient und dann das isomorphe Bild des Ergebnisses im zweiten betrachtet oder ob man zuerst Zustand und Eingabe in den zweiten Automaten abbildet und diesen dann bedient.

3.3. Übungen

1. Man skizziere alle möglichen schlichten, gerichteten Graphen mit genau drei Knoten, wobei keine zwei Graphen einander isomorph sein sollen.
2. Man beweise die folgende Aussage für gerichtete und ungerichtete Graphen:
   Die Anzahl der Knoten mit ungeradem Grad ist gerade.
3. Die Knoten eines schlichten Graphen sollen so gefärbt werden, daß benachbarte Knoten verschiedene Farben erhalten. Man zeige: Wenn alle Knoten höchstens den Grad n haben, werden höchstens $n+1$ Farben benötigt.
4. Es sei Q_n ein Graph mit $n+1$ Knoten $a_0, \ldots, a_n$ und folgenden Eigenschaften: Genau ein Knoten (etwa a_0) ist zu allen anderen adjazent. Für die anderen Knoten $a_1, \ldots, a_n$ gilt: a_i und a_{i+1} $(i = 1, \ldots, n-1)$ sowie a_n und a_1 sind adjazent. Wie viele Farben werden höchstens benötigt, um Q_n so zu färben, daß je zwei benachbarte Knoten unterschiedlich gefärbt sind?
5. Man untersuche, für welche natürlichen Zahlen n es ungerichtete Graphen G_n mit genau n Knoten $x_1, x_2, \ldots, x_n$ derart gibt, daß für die Knotengrade $d(x_i) = i$, $i = 1, 2, \ldots, n$ gilt.
   - Für welche n gibt es solche Graphen nicht?
   - Man finde alle derartigen nichtisomorphen Graphen für die beiden kleinsten solcher natürlicher Zahlen n.
   - Man kennzeichne 4 Graphenpaare derart, daß jeweils der eine Graph isomorph zu einem echten Untergraphen des anderen ist.
6. Es sei G ein Graph mit 100 Knoten. Der Grad eines Knoten beträgt mindestens drei. Man untersuche, wieviel Kanten G mindestens besitzt.
7. Es sei $G = (V, R_g)$ ein endlicher ungerichteter Graph mit n Knoten $v_1 \ldots v_n$ mit den Knotengraden g_i und m Kanten. Man beweise:
$$\sum_{i=1}^{n} g_i = 2m\,.$$
8. Man untersuche die folgenden beiden Graphen auf Isomorphie:



und



.

9. Man zeige, daß in einem zusammenhängenden Graphen je zwei längste Wege einen Knoten gemeinsam haben.

10. Man beweise:
Wenn alle Knoten eines Graphen mit der der Knotenmenge V mindestens den Grad $\frac{|V|-1}{2}$ haben, ist er zusammenhängend.

11. Es sei G ein Graph ohne isolierte Knoten, der genau eine Kante weniger als Knoten hat. Man zeige: G enthält mindestens zwei Knoten mit dem Grad 1.

12. Man zeige: Jeder azyklische Graph mit n Knoten hat höchstens $n-1$ Kanten.

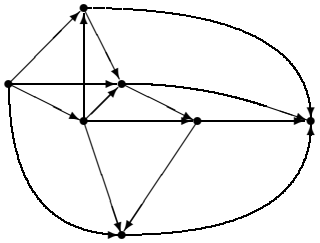
13. Man stelle sich einen dreidimensionalen Körper als Graphen vor: Die Kanten des Körpers entsprechen den Kanten im Graphen, die Eckpunkte den Knoten. Welche regelmäßigen Körper (d. h. mit nur kongruenten Seiten)

    (a) sind Eulergraphen?

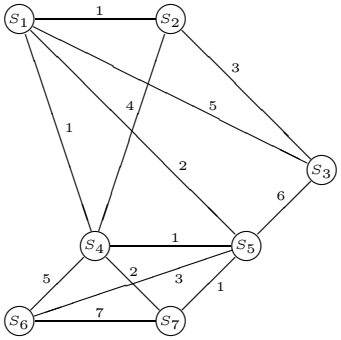
    (b) enthalten einen Hamilton-Kreis?

14. Ein Graph heißt n-**regulär**, wenn jeder Knoten den Grad n hat.
Man zeichne einen 5-regulären schlichten Graphen, dessen kürzester Kreis die Länge 3 und längster Kreis die Länge 8 hat.

15. Man untersuche nebenstehenden gerichteten Graphen auf Zyklenfreiheit:



16. Gegeben sei folgende Vereinfachung eines Ausschnitts einer Landkarte. Die Knoten $S_1, \ldots, S_7$ stellen Städte dar. Zwei Knoten S_i, S_j sind genau dann durch eine Kante verbunden, wenn S_i von S_j aus direkt erreichbar ist und umgekehrt. Die Reisekosten für eine direkte Verbindung stehen an der entsprechenden Kante.



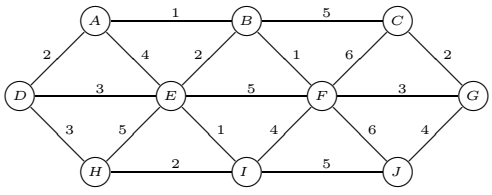
Gesucht ist ein Weg von S_1 nach S_1 über alle Städte $S_2, \ldots, S_7$ mit minimalen Gesamtkosten, wenn jede Stadt

(a) genau einmal,

(b) mindestens einmal

besucht werden soll.
Welche Auswirkung auf die Höhe der Ausgaben hat die Wahl des Start-Ziel-Ortes?

17. Es sei folgender bewertete, gerichtete Graph gegeben.



Man finde einen Weg von A nach J mit minimalen Kosten.

18. Man stelle sich einen unendlichen, gerichteten Graphen mit den Knoten a_i, $i = 1, 2, \ldots$ und den Kanten (a_i, a_j) vor, wobei für die Kanten gilt: Es existiert genau dann eine gerichtete Kante (a_i, a_j) von a_i nach a_j, wenn a_i ein Teiler von a_j ist. Was kann man über die Hin- und Weggrade der Knoten aussagen?

19. Ein Graph heißt **bipartit**, wenn eine Zerlegung der Knotenmenge V in zwei Mengen M, N so existiert, daß jede Kante zu je einem Knoten aus M und N inzident ist.
Man zeichne alle bis auf Isomorphie verschiedenen bipartiten Graphen mit $|M| = 2$, $|N| = 3$, die keinen isolierten Knoten enthalten.

Kapitel 4

Analysis

4.1. Erinnerung und Neues

Die Analysis ist nicht nur das umfangreichste mathematische Teilgebiet, sondern auch jenes mit den meisten außermathematischen Anwendungen. Wir können hier nur einige grundlegende Begriffe und Erkenntnisse studieren, um so in die analytische Denkweise einzuführen.
Für unsere Überlegungen verwenden wir den n-dimensionalen euklidischen Vektorraum $\mathbb{R}^n$ mit dem in Kap. 2 definierten Skalarprodukt $(\cdot,\cdot)$ und der euklidischen Norm

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x},\mathbf{x})} = \sqrt{\sum_{j=1}^{n} x_j x_j}.$$

Die Vektoren aus dem $\mathbb{R}^n$ werden wir auch **Punkte** nennen.
Für einen beliebigen Punkt $\mathbf{x} \in \mathbb{R}^n$ und eine beliebige Zahl $\varepsilon > 0$ heißt die Menge

$$U_\varepsilon(\mathbf{x}) = \{\, \mathbf{y} \mid \|\mathbf{x}-\mathbf{y}\| < \varepsilon \,\}$$

Umgebung von $\mathbf{x}$, genauer ε-Umgebung des Punktes $\mathbf{x}$. Eine ε-Umgebung um einen Punkt $\mathbf{x}$ ist offenbar eine Kugel um diesen mit dem Mittelpunkt in $\mathbf{x}$ und dem Radius ε. Im Falle $n = 1$, wenn wir also als Vektorraum die reellen Zahlen nehmen, stimmt die Norm mit dem Betrag überein; damit gilt hier

$$U_\varepsilon(\mathbf{x}) = \{\, \mathbf{y} \mid |\mathbf{x}-\mathbf{y}| < \varepsilon \,\} = \{\, \mathbf{y} \mid \mathbf{x}-\varepsilon < \mathbf{y} < \mathbf{x}+\varepsilon \,\},$$

d. h. in $\mathbb{R}$ sind die Umgebungen offene Intervalle.
Mit dem Zeichen $U_\varepsilon^\circ(\mathbf{x})$ bezeichnen wir eine punktierte Umgebung von $\mathbf{x}$, d. h. eine Umgebung von $\mathbf{x}$, wo der Punkt $\mathbf{x}$ herausgeschnitten wurde:

$$U_\varepsilon^\circ(\mathbf{x}) = U_\varepsilon(\mathbf{x}) \setminus \{\mathbf{x}\}.$$

Eine Menge $M \subseteqq \mathbb{R}^n$ heißt **beschränkt** , wenn sie in einer Umgebung des Nullpunktes liegt, d. h. wenn eine positive Zahl L existiert, so daß für alle $\mathbf{x} \in M$ die Ungleichung $\|\mathbf{x}\| < L$ gilt. Im Vektorraum $\mathbb{R}$ der reellen Zahlen kann man noch oben und unten unterscheiden: Eine Menge $M \subseteqq \mathbb{R}$ heißt **nach oben beschränkt**, wenn eine Zahl L existiert mit $\mathbf{x} \leqq L$ für alle $\mathbf{x} \in M$. Eine solche Zahl L heißt **obere Schranke** der Menge M. Die kleinste obere Schanke von M nennt man **obere Grenze** von M. Analog nennt man eine Menge $M \subseteqq \mathbb{R}$ **nach unten beschränkt**, wenn eine reelle Zahl l existiert, so daß $\mathbf{x} \geqq l$ für alle $\mathbf{x} \in M$ ausfällt. Jede dieser Zahlen heißt **untere Schranke** und die größte unter ihnen **untere Grenze** der Menge M. Hier sei bereits bemerkt, daß weder die obere noch die untere Grenze Element der Menge sein müssen. Offenbar ist im Bereich der reellen Zahlen eine Menge genau dann beschränkt, wenn sie nach oben und nach unten beschränkt ist.
Ein Punkt $\mathbf{x} \in M \subseteqq \mathbb{R}^n$ heißt **innerer Punkt** von M, wenn eine Umgebung von ihm in M liegt, d. h. wenn ein $\varepsilon > 0$ existert mit $U_\varepsilon(\mathbf{x}) \subseteqq M$. Mit $\operatorname{int}(M)$ bezeichnen wir die Menge aller inneren Punkte der Menge M. Besteht die Menge M nur aus inneren Punkten, d. h. gilt $\operatorname{int}(M) = M$, so heißt M **offen**. Andererseits nennen wir eine Menge $M \subseteqq \mathbb{R}^n$ **abgeschlossen**, wenn die Komplementmenge $\mathbb{R}^n \setminus M$ offen ist.
Ein Punkt $\mathbf{x} \in \mathbb{R}^n$ heißt **Randpunkt** der Menge M, wenn in jeder punktierten Umgebung von $\mathbf{x}$ sowohl Punkte aus M liegen als auch solche Punkte, die nicht zu M gehören, d. h. für jedes $\varepsilon > 0$ gilt

$$M \cap U_\varepsilon^\circ(\mathbf{x}) \neq \emptyset, \quad U_\varepsilon^\circ(\mathbf{x}) \setminus M \neq \emptyset.$$

Ein Punkt $\mathbf{x} \in M$ heißt **isoliert**, wenn es eine punktierte Umgebung von $\mathbf{x}$ gibt, die keinen Punkt aus M enthält, d. h. es gibt ein $\varepsilon > 0$, so daß $M \cap U_\varepsilon^\circ(\mathbf{x}) = \emptyset$ gilt.

Ein Punkt $\mathbf{x} \in \mathbb{R}^n$ heißt **Häufungspunkt** der Menge $M \subseteq \mathbb{R}^n$, wenn in jeder punktierten Umgebung von $\mathbf{x}$ auch Punkte aus M liegen, d. h. für jedes $\varepsilon > 0$ gilt $M \cap U_\varepsilon^\circ(\mathbf{x}) \neq \emptyset$. Es sei ausdrücklich erwähnt, daß Häufungspunkte einer Menge nicht automatisch auch zur Menge gehören müssen.

Satz 66. (Vereinigung von offenen Mengen)
Die Vereinigung von beliebig vielen offenen Mengen ist offen.

Beweis. Es sei

$$M = \bigcup_{\alpha \in I} M_\alpha$$

mit einer Indexmenge von beliebiger Mächtigkeit und $\mathbf{x} \in M$ beliebig ausgewählt. Dann muß $\mathbf{x}$ in mindestens einer der Mengen M_α, etwa M_{α_0} liegen. Da M_{α_0} eine offene Menge ist, enthält sie mit $\mathbf{x}$ auch eine Umgebung von $\mathbf{x}$, die folglich auch in der Vereinigung liegen muß, was uns sagt, daß M mit jedem Punkt $\mathbf{x}$ auch eine Umgebung von $\mathbf{x}$ enthält, d. h. M ist offen. □

Satz 67 (Durchschnitt abgeschlossener Mengen). *Der Durchschnitt von beliebig vielen abgeschlossenen Mengen ist abgeschlossen.*

Beweis. Es sei

$$M = \bigcap_{\alpha \in I} M_\alpha$$

und wir haben zu zeigen, daß die Menge $\mathbb{R}^n \setminus M$ offen ist. Dazu sei $\mathbf{x} \in \mathbb{R}^n \setminus M$ ein beliebiger Punkt; dann liegt $\mathbf{x}$ nicht im Durchschnitt M, d. h. es gibt unter den Mengen M_α mindestens eine, die den Punkt $\mathbf{x}$ nicht enthält; sei dies die abgeschlossene Menge M_{α_0}. Der Punkt $\mathbf{x}$ liegt dann aber in der offenen Menge $\mathbb{R}^n \setminus M_{\alpha_0}$ und mit ihm auch eine Umgebung $U_\varepsilon(\mathbf{x})$ von $\mathbf{x}$: $U_\varepsilon(\mathbf{x}) \subseteq \mathbb{R}^n \setminus M_{\alpha_0}$. Wegen $M \subseteq M_{\alpha_0}$ gilt $\mathbb{R}^n \setminus M_{\alpha_0} \subseteq \mathbb{R}^n \setminus M$, womit wir $U_\varepsilon(\mathbf{x}) \subseteq \mathbb{R}^n \setminus M$ schließen, was uns anzeigt, daß $\mathbb{R}^n \setminus M$ offen ist. □

Satz 68. (Abgeschlossenheitskriterium)
Eine Menge ist dann und nur dann abgeschlossen, wenn sie alle ihre Häufungspunkte enthält.

Beweis. Zunächst sei eine abgeschlossene Menge $M \subseteq \mathbb{R}^n$ gegeben. Wir zeigen, daß sie alle ihre Häufungspunkte enthält. Es sei $\mathbf{x}$ ein beliebiger Punkt aus der offenen Menge $\mathbb{R}^n \setminus M$. Dann existiert eine Umgebung $U_\varepsilon(\mathbf{x})$ des Punktes $\mathbf{x}$, die vollständig in $\mathbb{R}^n \setminus M$ liegt, woraus sich $U_\varepsilon(\mathbf{x}) \cap M = \emptyset$ ergibt. Dieser Schluß zeigt uns, daß außerhalb der Menge M keine Häufungspunkte von M liegen.
Nehmen wir nun umgekehrt an, daß die Menge M alle ihre Häufungspunkte enthält. Wir haben zu zeigen, daß $\mathbb{R}^n \setminus M$ offen ist. Es sei also $\mathbf{x} \in \mathbb{R}^n \setminus M$ beliebig ausgewählt. Da der Punkt $\mathbf{x}$ nicht Häufungspunkt von M sein kann, existiert ein $\varepsilon > 0$ und $U_\varepsilon(\mathbf{x}) \cap M = \emptyset$, also $U_\varepsilon(\mathbf{x}) \subseteq \mathbb{R}^n \setminus M$. □
Abschließend sei noch angemerkt, daß man Mengen oft auch durch ihre Indikatorfunktion darstellt. Genauer: Es sei $X \subseteq Y$ eine beliebige Teilmenge von Y. Eine auf einer Menge Y definierte reellwertige Funktion f_X, die nur die Werte 0 oder 1 annimmt (0-1-Funktion), heißt **Indikatorfunktion** von X bezüglich der Menge Y, falls

$$X = \{\, x \mid f_X(x) = 1 \,\}$$

gilt. Auf diese Weise ist jeder Menge eine wohlbestimmte Funktion, ihre Indikatorfunktion, zugeordnet. Ist umgekehrt f eine 0-1-Funktion, so definiert das Urbild von 1 eine Menge X derart, daß f die Indikatorfunktion dieser Menge darstellt. Die Indikatorfunktion L_i von $\{x_i\}$ bezüglich $\{\, x_1, \ldots, x_n \,\}$ lautet

$$L_i(x) = \begin{cases} 1 & x = x_i \\ 0 & x \neq x_i \end{cases} = \frac{(x - x_1) \cdot \ldots \cdot (x - x_{i-1})(x - x_{i+1}) \cdot \ldots \cdot (x - x_n)}{(x_i - x_1) \cdot \ldots \cdot (x_i - x_{i-1})(x_i - x_{i+1}) \cdot \ldots \cdot (x_i - x_n)}.$$

4.2. Folgen

Eine Funktion f, die jeder natürlichen Zahl n aus einer unendlichen Menge $N \subseteq \mathbb{N}$ ein Element $\mathbf{a}$ aus einer Menge M zuordnet, heißt **Folge**. Wir schreiben Folgen in der Form $(\mathbf{a}_k, k \in N)$ bzw. $(\mathbf{a}_k)$, falls $\mathbb{N}$ die Argumentmenge ist. Die Folgeglieder $\mathbf{a}_k$ müssen nicht notwendig verschiedene Elemente der Grundmenge M sein. Beispiele für Folgen reeller Zahlen sind folgende:

$$a_k = 2 \cdot (-1)^k \cdot k, \quad a_k = \frac{k}{k+1}, \quad a_k = \left(1 + \frac{1}{k}\right)^k, \quad k = 1, 2, \ldots.$$

Die Elemente $\mathbf{a}_k$ nennt man **Glieder** der Folge $(\mathbf{a}_k, k \in \mathbb{N})$. Eine Folge mit nur gleichen Gliedern heißt **stationär**. Wenn man aus einer Folge $(\mathbf{a}_k, k \in \mathbb{N})$ unendlich viele Folgeglieder herausgreift, erhält man eine **Unterfolge** oder auch **Teilfolge** $(\mathbf{a}_k, k \in K)$ mit

$$K = \{\, k_i, i = 1, 2, \ldots \,\}, \quad k_1 < k_2 < \cdots < k_i < \cdots.$$

Wir studieren hier Folgen, deren Glieder Vektoren aus dem $\mathbb{R}^n$ sind; Folgen von reellen Zahlen nennt man einfach **Zahlenfolgen**.
Unmittelbar einsichtig ist die Tatsache, daß man arithmetische Operationen mit Folgen ausführen darf: Sind $(\mathbf{a}_k), (\mathbf{b}_k)$ zwei Folgen, so sind auch $(\mathbf{a}_k \pm \mathbf{b}_k)$ Folgen. Zusätzlich sind bei Zahlenfolgen auch $(a_k \cdot b_k)$ und $(\frac{a_k}{b_k})$ Folgen, wobei im letzteren Falle gesichert sein muß, daß $b_k \neq 0$ für alle k gilt.
Eine Zahlenfolge (a_k) nennt man **monoton wachsend**, falls

$$a_k \leqq a_{k+1}, \quad k = 0, 1, 2, \ldots$$

und **monoton fallend**, falls

$$a_k \geqq a_{k+1}, \quad k = 0, 1, 2, \ldots$$

gilt. Sollten die Ungleichungen streng gelten, sprechen wir von streng monoton wachsend bzw. von streng monoton fallend.
Eine Folge $(\mathbf{a}_k)$ heißt **Nullfolge**, wenn in jeder Umgebung des Nullpunktes bis auf endlich viele Ausnahmen alle Folgeglieder liegen. Diese charakterisierende Eigenschaft läßt sich formal auf zwei Arten beschreiben:
Variante 1: Zu jedem $\varepsilon > 0$ existiert ein $k_0 = k_0(\varepsilon)$, so daß

$$\|\mathbf{a}_k\| < \varepsilon \quad \forall k \geqq k_0.$$

Variante 2: Für jedes $\varepsilon > 0$ enthält die Komplementmenge $\mathbb{R}^n \setminus U_\varepsilon(\mathbf{o})$ von $U_\varepsilon(\mathbf{o})$ höchstens endlich viele Folgeglieder:

$$|\left\{ k \;\middle|\; \|\mathbf{a}_k\| \geqq \varepsilon \right\}| < \infty.$$

Satz 69 (Nullfolgeneigenschaften). *Für Nullfolgen gelten die folgenden Aussagen.*

1. *Jede Unterfolge einer Nullfolge ist auch eine Nullfolge.*
2. *Eine Folge $(\mathbf{a}_k) \subseteqq \mathbb{R}^n$ ist genau dann eine Nullfolge, wenn $(\|\mathbf{a}_k\|) \subseteqq \mathbb{R}$ eine Nullfolge ist.*
3. *Jede Nullfolge ist beschränkt.*
4. *Sind $(\mathbf{a}_k), (\mathbf{b}_k)$ Nullfolgen, so sind auch $(\mathbf{a}_k + \mathbf{b}_k)$ und $(\mathbf{a}_k - \mathbf{b}_k)$ Nullfolgen. Die Menge aller Nullfolgen bildet einen Vektorraum über den reellen Zahlen.*
5. *(Majorantenkriterium.) Ist $(\mathbf{c}_k)$ eine Nullfolge und $(\mathbf{a}_k)$ eine Folge, zu der ein $L > 0$ existiert, so daß ab einem Index k_0*

   $$\|\mathbf{a}_k\| \leqq L \cdot \|\mathbf{c}_k\| \quad \forall k \geqq k_0$$

   gilt, dann ist auch $(\mathbf{a}_k)$ eine Nullfolge.
6. *Es sei $\mathbf{a}_k = (a_{1k}, a_{2k}, \ldots, a_{nk})$. Die Folge $(\mathbf{a}_k)$ ist genau dann Nullfolge, wenn alle Zahlenfolgen $(a_{jk}), j = 1, 2, \ldots, n$ Nullfolgen sind.*
7. *Ist $(\mathbf{a}_k)$ eine Nullfolge und (b_k) eine beschränkte Zahlenfolge, so ist $(b_k \cdot \mathbf{a}_k)$ eine Nullfolge.*

Beweis. Auf einen Beweis der ersten 4 Aussagen soll hier verzichtet werden. Für den Beweis des Majorantenkriteriums sei $\varepsilon > 0$ beliebig vorgegeben. Da wegen der Aussage 2 die Folge $(\|\mathbf{c}_k\|)$ eine Nullfolge ist, liegen außerhalb einer $\frac{\varepsilon}{L}$-Umgebung des Nullpunktes höchstens endlich viele Folgeglieder und wir können aus deren Indices den maximalen bilden:

$$k_0(\varepsilon) = \max\left\{ \{k_0\} \cup \left\{ k \;\middle|\; \|\mathbf{c}_k\| \geqq \frac{\varepsilon}{L} \right\} \right\}.$$

Für alle $k > k_0(\varepsilon)$ folgt daraus mit der Voraussetzung

$$\|\mathbf{a}_k\| \leqq L \cdot \|\mathbf{c}_k\| < L \cdot \frac{\varepsilon}{L} = \varepsilon,$$

was uns sagt, daß $(\mathbf{a}_k)$ eine Nullfolge ist.
Für die Aussage 6 zeigen wir zunächst: Wenn $(\mathbf{a}_k)$ eine Nullfolge darstellt, so ist auch jede Folge (a_{jk}) eine Nullfolge. Offensichtlich gilt

$$|a_{jk}| \leqq \|\mathbf{a}_k\|, \quad j = 1, 2, \ldots, n.$$

Mit dem Majorantenkriterium folgt hieraus, daß (a_{jk}) eine Nullfolge ist.
Nehmen wir nun umgekehrt an, daß alle Folgen $(a_{jk})(j = 1, 2, \ldots, n)$ Nullfolgen sind. Wegen Aussage 4 ist dann die Folge $(|a_{1k}| + |a_{2k}| + \cdots + |a_{nk}|)$ eine Nullfolge. Wegen der offensichtlichen Ungleichung

$$\|\mathbf{a}_k\| \leqq \sum_{j=1}^{n} |a_{jk}|$$

folgt aus dem Majorantenkriterium, daß $(\mathbf{a}_k)$ eine Nullfolge darstellt.
Wir kommen zum Beweis der Aussage 7. Die Zahlenfolge (b_k) ist nach Voraussetzung beschränkt; also existiert eine positive Zahl L mit

$$|b_k| \leqq L \quad \forall k.$$

Es sei nun $\varepsilon > 0$ beliebig vorgegeben und $k_0(\varepsilon)$ der maximale Index aller Folgeglieder $\mathbf{a}_k$, die außerhalb einer $\frac{\varepsilon}{L}$-Umgebung des Nullpunktes liegen:

$$k_0(\varepsilon) = \max \left\{ k \;\middle|\; \|\mathbf{a}_k\| \geqq \frac{\varepsilon}{L} \right\}.$$

Für $k > k_0(\varepsilon)$ folgt daraus:

$$\|\mathbf{a}_k \cdot b_k\| = \|\mathbf{a}_k\| \cdot |b_k| < L \cdot \frac{\varepsilon}{L} = \varepsilon,$$

was uns sagt, daß $(\mathbf{a}_k \cdot b_k)$ eine Nullfolge ist. □
Das folgende Beispiel soll zeigen, daß man unter Umständen die Glieder einer Nullfolge mit den Gliedern einer unbeschränkten Folge multiplizieren darf, ohne die Nullfolgeneigenschaft zu verlieren. Wir betrachten die beiden Zahlenfolgen $(q^k), (k)$ mit $0 < |q| < 1$ und bilden daraus die Folge $(k \cdot q^k)$. Indem wir $|q| = \frac{1}{1+x}$ setzen, folgt für $k \geqq 2$:

$$\begin{aligned} |k \cdot q^k| &= \frac{k}{(1+x)^k} = \frac{k}{1 + \binom{k}{1}x + \binom{k}{2}x^2 + \cdots x^k} \\ &< \frac{k}{\binom{k}{2}x^2} = \frac{2}{(k-1)x^2} = \frac{2}{x^2} \cdot \frac{1}{k-1}. \end{aligned}$$

Damit haben wir gezeigt, daß die Nullfolge $(\frac{1}{k-1})$ eine Majorante für die Folge $(k \cdot q^k)$ darstellt, und das Majorantenkriterium sagt uns, daß auch $(k \cdot q^k)$ eine Nullfolge ist. Die Folge $(\frac{1}{k-1})$ ist erst recht eine Majorante für die Folge (q^k), was uns das Nebenergebnis liefert, daß (q^k) für $|q| < 1$ eine Nullfolge darstellt.
Eine Folge $(\mathbf{a}_k)$ ist **konvergent**, falls ein $\mathbf{a}$ existiert, so daß $(\mathbf{a}_k - \mathbf{a})$ eine Nullfolge ist; andernfalls sagen wir, daß die betrachtete Folge **divergiert**. Bei divergenten Folgen unterscheidet man noch zwischen **bestimmt divergent** und **unbestimmt divergent**. Eine Folge $(\mathbf{a}_k)$ divergiert bestimmt, falls die Folge $(\frac{1}{\|\mathbf{a}_k\|})$ eine Nullfolge ist. Alle übrigen, nicht konvergenten Folgen nennt man unbestimmt divergent. Bei einer bestimmt divergenten Folge liegen in jeder Umgebung des Nullpunktes höchstens endlich viele Folgeglieder. Man sagt in einem solchen Falle, daß ∞ bzw. $-\infty$ der uneigentliche Grenzwert der Folge ist.
Zu einer Folge $(\mathbf{a}_k)$ gibt es höchstens einen Punkt $\mathbf{a}$, so daß $(\mathbf{a}_k - \mathbf{a})$ eine Nullfolge ist. Sind nämlich $(\mathbf{a}_k - \mathbf{a})$ und $(\mathbf{a}_k - \mathbf{b})$ Nullfolgen, so muß auch die Differenz eine Nullfolge sein, woraus sich $\mathbf{a} = \mathbf{b}$ ergibt. Zu einer konvergenten Folge $(\mathbf{a}_k)$ existiert daher genau ein $\mathbf{a}$, so daß $(\mathbf{a}_k - \mathbf{a})$ eine Nullfolge ist. Dieser eindeutig bestimmte Punkt $\mathbf{a}$ heißt **Grenzwert** oder **Limes** der Folge. Man sagt: Die Folge $(\mathbf{a}_k)$ konvergiert gegen den Punkt $\mathbf{a}$, in Zeichen:

$$\mathbf{a} = \lim_{k\to\infty} \mathbf{a}_k, \quad \mathbf{a}_k \overset{k\to\infty}{\longrightarrow} \mathbf{a} \quad \text{oder} \quad \mathbf{a}_k \longrightarrow \mathbf{a}.$$

Beispiele.

1. Wir betrachten für $|q| < 1$ die Folge $(\sum_{i=0}^{k} q^i)$. Aus der Gleichungskette

$$\begin{aligned} 1 - q^{k+1} &= 1 + q + q^2 + \cdots + q^k - (q + q^2 + \cdots + q^{k+1}) \\ &= (1-q)(1 + q + q^2 + \cdots + q^k) \\ &= (1-q)\sum_{i=0}^{k} q^i \end{aligned}$$

folgt durch Umstellen:

$$\sum_{i=0}^{k} q^i = \frac{1 - q^{k+1}}{1-q} = \frac{1}{1-q} - \frac{q^{k+1}}{1-q},$$

was uns zeigt, daß die Folge den Grenzwert $\frac{1}{1-q}$ hat, da der zweite Summand allgemeines Glied einer Nullfolge ist.

2. Als zweites Beispiel wählen wir die Folge $(\sqrt[k]{k})$. Indem wir $a_k = \sqrt[k]{k} - 1$ setzen, folgt

$$k = (1 + a_k)^k > \binom{k}{2} a_k^2 = \frac{k(k-1)}{2} a_k^2$$

und daraus durch Umstellen

$$a_k^2 < \frac{2}{k-1},$$

woraus wir mit dem Majorantenkriterium

$$\lim_{k\to\infty} \sqrt[k]{k} = 1$$

schließen.

Eine wichtige Abschwächung des Grenzwertbegriffes ist der Häufungspunkt einer Folge. Ein Punkt $\mathbf{a}$ heißt **Häufungspunkt** der Folge $(\mathbf{a}_k)$, wenn in jeder Umgebung von $\mathbf{a}$ unendlich viele Folgeglieder liegen. Bei Zahlenfolgen bezeichnen wir mit $\limsup_{k\to\infty} a_k$ den größten und mit $\liminf_{k\to\infty} a_k$ den kleinsten Häufungspunkt der Folge. Sollte kein größter Häufungspunkt existieren, setzen wir $\limsup_{k\to\infty} a_k = \infty$; sollte kein kleinster Häufungspunkt existieren, setzen wir $\liminf_{k\to\infty} a_k = -\infty$. Bei konvergenten Folgen stimmen beide überein: $\liminf_{k\to\infty} a_k = \limsup_{k\to\infty} a_k$.
Der Konvergenzbegriff hat den Nachteil, daß man zum Nachprüfen der Konvergenz eine Vermutung über den möglichen Grenzwert haben muß. Wir definieren daher: Eine Folge $(\mathbf{a}_k)$ heißt **Fundamentalfolge** oder auch **Cauchyfolge**, wenn es zu jedem $\varepsilon > 0$ ein $k_0(\varepsilon)$ gibt mit

$$\|\mathbf{a}_m - \mathbf{a}_k\| < \varepsilon \quad \forall m, k \geqq k_0(\varepsilon).$$

Satz 70 (Konvergenz von Folgen). *Konvergierende Folgen haben die folgenden Eigenschaften.*

1. *Jede Unterfolge einer konvergenten Folge ist konvergent.*

2. *Jeder Häufungspunkt einer Unterfolge ist auch Häufungspunkt der gesamten Folge.*

3. *Jede konvergente Folge hat genau einen Häufungspunkt.*

4. *Jede konvergente Folge ist beschränkt.*

5. (*Cauchysches Konvergenzkriterium.*) *Eine Folge ist genau dann eine Fundamentalfolge, wenn sie konvergiert.*

6. *Für konvergente Folgen gelten die folgenden Rechenregeln:*

$$\lim_{k\to\infty} (\mathbf{a}_k \pm \mathbf{b}_k) = \lim_{k\to\infty} \mathbf{a}_k \pm \lim_{k\to\infty} \mathbf{b}_k,$$
$$\lim_{k\to\infty} \|\mathbf{a}_k\| = \| \lim_{k\to\infty} \mathbf{a}_k \|$$

und zusätzlich bei Zahlenfolgen

$$\lim_{k\to\infty} (a_k \cdot b_k) = \lim_{k\to\infty} a_k \cdot \lim_{k\to\infty} b_k,$$
$$\lim_{k\to\infty} \frac{a_k}{b_k} = \frac{\lim_{k\to\infty} a_k}{\lim_{k\to\infty} b_k}.$$

Dabei müssen in der letzten Gleichung alle Folgeglieder b_k von 0 verschieden und (b_k) darf keine Nullfolge sein. Die Menge aller konvergenten Folgen des $\mathbb{R}^n$ bildet einen Vektorraum über den reellen Zahlen.

7. (*Satz von Bolzano-Weierstraß.*) *Jede beschränkte Folge hat einen Häufungspunkt (und damit eine konvergente Unterfolge).*

8. *Eine Folge konvergiert genau dann, wenn sie beschränkt ist und höchstens einen Häufungspunkt besitzt.*

9. *Für Zahlenfolgen gilt:*

(a) Aus

$$\lim_{k\to\infty} a_k = a, \quad \lim_{k\to\infty} b_k = b, \quad a_k \leqq b_k, \quad \forall k \geqq k_0$$

folgt $a \leqq b$.

(b) Aus

$$\lim_{k\to\infty} a_k = a, \quad \lim_{k\to\infty} b_k = a, \quad a_k \leqq c_k \leqq b_k, \quad \forall k \geqq k_0$$

folgt $\lim_{k\to\infty} c_k = a$.

(c) Eine monotone Folge konvergiert genau dann, wenn sie beschränkt ist.

(d) Jede Zahlenfolge enthält eine monotone Unterfolge.

Beweis. Zunächst soll die Aussage 9*d* bewiesen werden. Dazu sei eine Zahlenfolge (a_k) gegeben. Wir definieren

$$M = \left\{ l \;\middle|\; \exists k_0 :\ a_{k+l} \leqq a_k \ \forall k \geqq k_0 \right\} = \{ l_1 < l_2 < \cdots \}.$$

Es sei $l \in M$; dann gilt

$$a_{k_0+n\cdot l} \leqq a_{k_0+(n-1)\cdot l}, \quad n = 1, 2, \ldots,$$

also liegt eine monoton fallende Unterfolge (a_{k_0+kl}) vor. Wir haben daher nur noch den Fall $M = \emptyset$ zu untersuchen. In diesem Falle gibt es zu jedem Index k_0 einen Index k mit $a_k > a_{k_0}$. Folglich existiert zu $l_1 = 1$ ein kleinster Index $l_2 > l_1$ mit $a_{l_2} > a_{l_1}$. Dieses Vorgehen kann man iterieren: Es sei weiter $l_3 > l_2$ der erste Index mit $a_{l_3} > a_{l_2}$ usw.; die so entstehende Unterfolge (a_{l_i}) ist streng monoton wachsend. Damit ist in beiden Fällen die Aussage 9*d* nachgewiesen.
Für die Aussage 9*c* sei (a_k) eine monoton wachsende, gegen a konvergente Zahlenfolge:

$$a_1 \leqq a_2 \leqq \cdots \leqq a_k \leqq \cdots \leqq a,$$

also ist die Folge beschränkt. Entsprechend folgt die Beschränktheit bei einer monoton fallenden Zahlenfolge.
Es sei umgekehrt (a_k) nach oben beschränkt und monoton wachsend. Dann hat die Folge eine obere Grenze a, d. h. eine kleinste obere Schranke. Für jedes $\varepsilon > 0$ ist $a - \varepsilon$ nicht mehr obere Schranke; also existert zu ε ein Index $k_0(\varepsilon)$ mit $a_{k_0} > a - \varepsilon$; für alle $k \geqq k_0$ gilt dann

$$|a_k - a| = a - a_k = a - a_{k_0} - (a_k - a_{k_0}) \leqq a - a_{k_0} < \varepsilon,$$

womit wir gezeigt haben, daß $(a_k - a)$ eine Nullfolge ist, d. h. die Folge (a_k) konvergiert gegen a.
Der Satz von Bolzano-Weierstraß (Aussage 7) wird zunächst für Zahlenfolgen bewiesen. Es sei also (a_k) eine Zahlenfolge. Nach Aussage 9*d* enthält sie eine monotone Unterfolge, die nach Aussage 9*c* konvergiert; folglich hat (a_k) einen Häufungspunkt.
Es sei nun $(\mathbf{a}_k)$ eine beliebige beschränkte Punktfolge aus dem $\mathbb{R}^n$:

$$\|\mathbf{a}_k\| \leqq L \quad \forall k.$$

Wegen $\mathbf{a}_k = (a_{1k}, \ldots, a_{nk})$ und

$$|a_{jk}| \leqq \|\mathbf{a}_k\| \leqq L, \quad j = 1, \ldots, n$$

sind auch alle Zahlenfolgen (a_{jk}) beschränkt.
Daher enthält die Folge (a_{1k}) eine konvergente Unterfolge $(a_{1k}, k \in K_1), K_1 \subseteqq K_0$ mit $K_0 = \mathbb{N}$. Diese Folge enthält eine konvergente Unterfolge $(a_{2k}, k \in K_2), K_2 \subseteqq K_1$ usw. bis zu einer konvergenten Unterfolge $(a_{nk}, k \in K_n), K_n \subseteqq K_{n-1}$. Es sei a_j^* der Grenzwert der Folge $(a_{jk}, k \in K_j), K_j \subseteqq K_{j-1}$. Wegen

$$K_n \subseteqq K_{n-1} \subseteqq \cdots \subseteqq K_1 \subseteqq K_0$$

konvergiert die Folge $(\mathbf{a}_k, k \in K_n)$ gegen den Punkt $\mathbf{a}^* = (a_1^*, \ldots, a_n^*)$, folglich ist $\mathbf{a}^*$ ein Häufungspunkt der Folge $(\mathbf{a}_k)$, womit der Satz von Bolzano-Weierstraß bewiesen ist.
Wir kommen zum Beweis des Cauchyschen Konvergenzkriteriums.
Es sei $(\mathbf{a}_k)$ eine gegen $\mathbf{a}$ konvergente Folge. Wir haben zu beweisen, daß $(\mathbf{a}_k)$ Fundamentalfolge ist. Dazu geben wir uns ein $\varepsilon > 0$ beliebig vor und wählen $K > 0$ als obere Schranke für die Indexmenge

$$\left\{ k \;\middle|\; \|\mathbf{a}_k - \mathbf{a}\| \geqq \frac{\varepsilon}{2} \right\}.$$

Für alle $m, l \geqq K$ folgt damit:

$$\|\mathbf{a}_m - \mathbf{a}_l\| = \|(\mathbf{a}_m - \mathbf{a}) + (\mathbf{a} - \mathbf{a}_l)\| \leqq \|\mathbf{a}_m - \mathbf{a}\| + \|\mathbf{a}_l - \mathbf{a}\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

also ist $(\mathbf{a}_k)$ eine Fundamentalfolge.
Es sei umgekehrt $(\mathbf{a}_k)$ eine Fundamentalfolge. Wir zeigen zunächst, daß die Folge beschränkt ist. Nach der Definition einer Fundamentalfolge existert zu $\varepsilon = 1$ ein $k_0(1)$ mit

$$\|\mathbf{a}_m - \mathbf{a}_k\| < 1 \quad \forall m, k \geqq k_0(1).$$

Wir fixieren ein $m > k_0(1)$; für jedes $k \geqq m$ folgt dann:

$$\|\mathbf{a}_k - \mathbf{a}_1\| = \|(\mathbf{a}_m - \mathbf{a}_1) + (\mathbf{a}_k - \mathbf{a}_m)\| \leqq \|\mathbf{a}_m - \mathbf{a}_1\| + \|\mathbf{a}_k - \mathbf{a}_m\| < \|\mathbf{a}_m - \mathbf{a}_1\| + 1,$$

was uns sagt, daß die gegebene Fundamentalfolge beschränkt ist. Nach dem Satz von Bolzano-Weierstraß enthält sie eine konvergente Unterfolge $(\mathbf{a}_{k_i})$:

$$\lim_{i\to\infty} \mathbf{a}_{k_i} = \mathbf{a}.$$

Zu vorgegebenem $\varepsilon > 0$ existieren ein $k_1(\varepsilon), k_2(\varepsilon)$ mit

$$\|\mathbf{a}_{k_i} - \mathbf{a}\| < \frac{\varepsilon}{2} \quad \forall k_i \geqq k_1(\varepsilon), \quad \|\mathbf{a}_m - \mathbf{a}_l\| < \frac{\varepsilon}{2} \quad \forall m, l \geqq k_2(\varepsilon).$$

Damit können wir wie folgt abschätzen:

$$\|\mathbf{a}_k - \mathbf{a}\| \leqq \|\mathbf{a}_k - \mathbf{a}_{k_i}\| + \|\mathbf{a}_{k_i} - \mathbf{a}\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

womit nachgewiesen ist, daß die Folge konvergiert. □

Beispiele.
1. Wir wollen die beiden Folgen

$$(a_k) = \left(\left(1 + \frac{1}{k}\right)^k\right), \quad (b_k) = \left(\left(1 + \frac{1}{k}\right)^{k+1}\right)$$

untersuchen. Zunächst zeigen wir, daß die Folge (a_k) streng monoton wächst. Angenommen, dem ist nicht so; dann gibt es ein k mit $a_k \geqq a_{k+1}$, d. h.

$$\left(\frac{k+1}{k}\right)^k \geqq \left(\frac{k+2}{k+1}\right)^{k+1}.$$

Indem wir diese Ungleichung mit dem Faktor $(\frac{k}{k+1})^{k+1}$ multiplizieren, erhalten wir

$$\begin{aligned}\frac{k}{k+1} &\geqq \left(\frac{k+2}{k+1} \cdot \frac{k}{k+1}\right)^{k+1} = \left(1 - \frac{1}{(k+1)^2}\right)^{k+1} \\ &> 1 - (k+1) \cdot \frac{1}{(k+1)^2} = 1 - \frac{1}{k+1} = \frac{k}{k+1},\end{aligned}$$

was offenbar unmöglich ist. Analog zeigt man, daß auch die Folge

$$(c_k) = \left(\left(1 - \frac{1}{k}\right)^k\right)$$

streng monoton wächst. Wegen

$$b_k \cdot c_{k+1} = \left(1 + \frac{1}{k}\right)^{k+1} \cdot \left(1 - \frac{1}{k+1}\right)^{k+1} = \left(\frac{k+1}{k} \cdot \frac{k}{k+1}\right)^{k+1} = 1$$

ist die Folge (b_k) streng monoton fallend. Weiter folgt

$$\begin{aligned}b_k - a_k &= \left(1 + \frac{1}{k}\right)^{k+1} - \left(1 + \frac{1}{k}\right)^k = \left(1 + \frac{1}{k}\right)^k \left(\left(1 + \frac{1}{k}\right) - 1\right) \\ &= \left(1 + \frac{1}{k}\right)^k \cdot \frac{1}{k} = \frac{1}{k} \cdot a_k > 0.\end{aligned}$$

Wegen $a_k < b_k < b_1$ ist die streng monoton wachsende Folge (a_k) beschränkt und konvergiert daher nach Aussage 9*c*. Wegen $b_k - a_k = \frac{a_k}{k}$ bildet $(b_k - a_k)$ eine Nullfolge. Folglich haben beide Folgen einen gemeinsamen Grenzwert, den man mit e bezeichnet:

$$\lim_{k\to\infty} a_k = \lim_{k\to\infty} b_k = e = 2.71828182844590...$$

Es ist die **Eulersche Zahl**, die Basis des natürlichen Logarithmus.
2. Das folgende Beispiel soll zeigen, daß man mit dem Cauchyschen Konvergenzkriterium die Konvergenz einer Folge beweisen kann, ohne eine Vermutung über den Grenzwert zu verwenden. Es sei (a_k) eine beschränkte Zahlenfolge: $|a_k| \leqq L$ und $|q| < 1$. Wir bilden die Folge (s_k) mit

$$s_k = \sum_{i=0}^{k} a_i q^i$$

und zeigen, daß sie eine Cauchyfolge ist, woraus sich dann mit dem Cauchyschen Konvergenzkriterium die Konvergenz der Folge ergibt. Da (q^k) eine Nullfolge ist, gibt es zu beliebig vorgegebenem $\varepsilon > 0$ einen Index $k_0(\varepsilon)$ derart, daß

$$|q|^{k+1} < (1 - |q|)\frac{\varepsilon}{L} \quad \forall k \geqq k_0(\varepsilon).$$

Für solche k und $r \geqq 0$ gilt dann

$$\begin{aligned} |s_{k+r} - s_k| &= \left| \sum_{i=k+1}^{k+r} a_i q^i \right| \leqq \sum_{i=k+1}^{k+r} |a_i| \cdot |q|^i \\ &= L \cdot \sum_{i=k+1}^{k+r} |q|^i = L \cdot |q|^{k+1} \cdot \sum_{i=0}^{r-1} |q|^i \\ &= L \cdot |q|^{k+1} \cdot \frac{1 - |q|^r}{1 - |q|} \leqq L \cdot |q|^{k+1} \cdot \frac{1}{1 - |q|} < \varepsilon, \end{aligned}$$

d. h. (s_k) ist eine Cauchyfolge.

4.3. Unendliche Reihen

Im folgenden sei (a_n) eine Zahlenfolge. Wir betrachten den nachstehenden unendlichen Algorithmus:

$$\begin{cases} s_0 = a_0, \\ s_{n+1} = s_n + a_{n+1}, \quad n = 0, 1, 2, \ldots \end{cases}$$

Diesen Algorithmus nennt man **unendliche Reihe** oder einfach **Reihe** und schreibt abkürzend für ihn das Zeichen $\sum_{k=0}^{\infty} a_k$. Es sei ausdrücklich darauf hingewiesen, daß es sich hier nicht um so etwas wie eine „unendliche Summe“ handelt. Bei einer Summe gilt z. B. die Kommutativität der Addition, während hier über den Algorithmus genau vorgeschrieben ist, in welcher Reihenfolge die Folgeglieder zu addieren sind. Das Summenzeichen mag etwas irreführend sein; es ist jedoch insbesondere außerhalb der Mathematik sehr gebräuchlich. Wegen $s_n = \sum_{k=0}^{n} a_k$ heißt die Zahl s_n die **n-te Partialsumme** und die Folge (s_n) nennt man entsprechend **Partialsummenfolge**. Jeder Zahlenfolge ist eine Reihe und damit eine Partialsummenfolge zugeordnet. In der Partialsummenfolge widerspiegeln sich die Eigenschaften des obigen Algorithmus; sein Verhalten wird durch Eigenschaften der Partialsummenfolge beschrieben. Wir sagen daher, daß eine Reihe **konvergent** ist, **divergiert**, **bestimmt divergiert**, **unbestimmt divergiert**, falls die entsprechende Partialsummenfolge diese Eigenschaft hat. Falls die Reihe gegen den Wert s^* konvergiert, schreiben wir dies in der Form

$$\sum_{n=0}^{\infty} a_n = s^*$$

auf und nennen s^* den **Wert** der Reihe.
Im obigen Algorithmus werden die Folgeglieder in der aufgeführten Reihenfolge verarbeitet. Nun kann sich der Wert einer Reihe ändern oder auch nicht, falls man die Reihenfolge der Folgeglieder verändert. Es ist klar, daß höchstens dann eine Änderung im Wert der Reihe zu erwarten ist, wenn unendlich viele Folgeglieder einen anderen Platz in der Folge erhalten. Eine Reihe $\sum_{k=0}^{\infty} a_k$ **konvergiert bedingt**, falls sie gegen einen Wert s^* konvergiert, aber eine solche Umordnung der Folge (a_n) existiert, daß die daraus gebildete Reihe nicht gegen s^* konvergiert. Sie kann in einem solchen Falle also gegen einen anderen Wert konvergieren oder sogar divergieren. Eine Reihe ist **unbedingt konvergent**, wenn sie konvergiert und sich ihr Wert bei Umordnung der Folgeglieder nicht ändert. Eine Reihe $\sum_{k=0}^{\infty} a_k$ ist **absolut konvergent**, wenn die aus $(|a_n|)$ gebildete Reihe $\sum_{k=0}^{\infty} |a_k|$ konvergiert.

Satz 71. (Rechenregeln für Reihen)
Die Menge aller konvergenten Reihen bildet einen Vektorraum über dem Körper der reellen Zahlen.

1. *Wenn die Reihe $\sum_{n=0}^{\infty} a_n$ konvergiert, dann konvergiert für jede reelle Zahl α auch die Reihe $\sum_{n=0}^{\infty} \alpha a_n$ und es gilt*

$$\sum_{n=0}^{\infty} \alpha a_n = \alpha \cdot \sum_{n=0}^{\infty} a_n.$$

2. *Wenn die Reihen $\sum_{n=0}^{\infty} a_n, \sum_{n=0}^{\infty} b_n$ konvergieren, so auch die Reihen*

$$\sum_{n=0}^{\infty} (a_n \pm b_n)$$

und es gilt

$$\sum_{n=0}^{\infty} (a_n \pm b_n) = \sum_{n=0}^{\infty} a_n \pm \sum_{n=0}^{\infty} b_n.$$

Der Satz kann direkt durch Rückgang auf die Konvergenzdefinition bewiesen werden.

Satz 72. (Notwendiges Konvergenzkriterium)
Wenn eine Reihe $\sum_{n=0}^{\infty} a_n$ konvergiert, dann ist die Folge (a_n) eine Nullfolge.

Beweis. Die Behauptung des Satzes ergibt sich aus der folgenden Gleichungskette:

$$\lim_{n\to\infty} a_{n+1} = \lim_{n\to\infty} (s_{n+1} - s_n) = \lim_{n\to\infty} s_{n+1} - \lim_{n\to\infty} s_n = 0.$$

Die Tatsache, daß man höchstens aus einer Nullfolge eine konvergente Reihe erhalten kann, bedeutet nicht, daß aus **jeder** Nullfolge eine konvergente Reihe entsteht. Als Beispiel nehmen wir die Nullfolge $(\frac{1}{n})$ und die damit gebildete **harmonische Reihe**:

$$\sum_{n=1}^{\infty} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots .$$

Für die Partialsummen der harmonischen Reihe erhalten wir

$$s_{2n} - s_n = \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n} > n \cdot \frac{1}{2n} = \frac{1}{2},$$

was uns sagt, daß die Partialsummenfolge keine Fundamentalfolge ist. Daher divergiert die harmonische Reihe, obwohl die Reihenglieder eine Nullfolge bilden.

Satz 73 (Cauchysches Konvergenzkriterium). *Eine Reihe $\sum_{n=0}^{\infty} a_n$ konvergiert dann und nur dann, wenn es zu jedem $\varepsilon > 0$ ein $n_0(\varepsilon)$ gibt mit*

$$|a_{n+1} + a_{n+2} + \cdots + a_{n+m}| < \varepsilon \quad \forall n > n_0(\varepsilon),\ \forall m \geqq 1.$$

Das Cauchysche Konvergenzkriterium ist wegen

$$|s_{n+m} - s_n| = |a_{n+1} + \cdots + a_{n+m}|$$

zum Cauchyschen Konvergenzkriterium für Zahlenfolgen äquivalent.

Satz 74 (Reihen mit nichtnegativen Gliedern). *Eine Reihe, deren Glieder sämtlich nichtnegativ sind, konvergiert genau dann, wenn die zugeordnete Partialsummenfolge beschränkt ist.*

Beweis. Die Partialsummen von Reihen mit nichtnegativen Gliedern sind monoton wachsend; daher folgt die Behauptung aus der Aussage 9*c* für Folgen. □

Satz 75 (Leibniz-Kriterium). *Eine alternierende Reihe $\sum_{n=0}^{\infty} a_n$, d. h. bei aufeinander folgenden Gliedern wechselt das Vorzeichen, konvergiert, falls die Betragsfolge $(|a_n|)$ eine monotone Nullfolge ist.*

Beweis. Ohne Beschränkung der Allgemeinheit nehmen wir

$$a_0 \geqq |a_1| \geqq a_2 \geqq |a_3| \geqq \cdots$$

an. Die Partialsummenfolge spalten wir in 2 Folgen $(b_n), (c_n)$ mit

$$b_n = s_{2n+1}, \quad c_n = s_{2n}$$

auf. Es ist

$$\begin{aligned} b_{n+1} - b_n &= (s_{2n+3} - s_{2n+2}) + (s_{2n+2} - s_{2n+1}) \\ &= a_{2n+3} + a_{2n+2} = a_{2n+2} - |a_{2n+3}| \geqq 0, \\ c_{n+1} - c_n &= (s_{2n+2} - s_{2n+1}) + (s_{2n+1} - s_{2n}) \\ &= a_{2n+2} + a_{2n+1} = a_{2n+2} - |a_{2n+1}| \leqq 0, \\ c_n - b_n &= s_{2n} - s_{2n+1} = |a_{2n+1}| > 0. \end{aligned}$$

Daraus entnehmen wir, daß (b_n) monoton wächst, (c_n) monoton fällt, beide Folgen konvergieren und die Differenzfolge $(c_n - b_n)$ ist eine Nullfolge. Also haben beide Folgen den gleichen Grenzwert s. Da die beiden Partialsummen (s_{2n}) und (s_{2n+1}) die gesamte Folge (s_n) ausschöpfen, liegen in jeder Umgebung von s mit höchstens endlich vielen Ausnahmen alle Folgeglieder von (s_n), d. h. die Folge (s_n) konvergiert gegen s. □
Als Beispiel einer Leibnizreihe erwähnen wir die Reihe

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{n}.$$

Diese Reihe konvergiert nach dem Leibniz-Kriterium; sie konvergiert aber nicht absolut!

Satz 76. *Eine absolut konvergente Reihe ist konvergent.*

Beweis. Hier bemerken wir nur, daß

$$|a_{n+1} + a_{n+2} + \cdots + a_{n+m}| \leqq |a_{n+1}| + |a_{n+2}| + \cdots + |a_{n+m}|$$

gilt, so daß die Aussage direkt aus dem Cauchyschen Konvergenzkriterium folgt. □

Satz 77. *Eine Reihe $\sum_{n=0}^{\infty} a_n$ konvergiert genau dann absolut, wenn die Partialsummenfolge der Betragsfolge $(|a_n|)$ beschränkt ist.*

Beweis. Zunächst konvergiert eine Reihe $\sum_{n=0}^{\infty} a_n$ genau dann absolut, wenn die Reihe $\sum_{n=0}^{\infty} |a_n|$ konvergiert. Dies ist aber eine Reihe mit nichtnegativen Gliedern, worauf wir Satz 74 anwenden können und die Behauptung erhalten. □

Satz 78 (1. Majorantenkriterium). *Es sei (c_n) eine Folge mit nichtnegativen Gliedern.*
Wenn die Reihe $\sum_{n=0}^{\infty} c_n$ konvergiert und für eine Folge (a_n) ab einem gewissen Index n_0

$$|a_n| \leqq c_n \quad \forall n \geqq n_0$$

gilt, so konvergiert die Reihe $\sum_{n=0}^{\infty} a_n$ absolut.
Wenn die Reihe $\sum_{n=0}^{\infty} c_n$ divergiert und für eine Folge (a_n) ab einem gewissen Index n_0

$$|a_n| \geqq c_n \quad \forall n \geqq n_0$$

gilt so konvergiert die Reihe $\sum_{n=0}^{\infty} a_n$ nicht absolut.

Beweis. Für den 1. Teil sei $\sum_{n=0}^{\infty} c_n = c$. Dann erhalten wir

$$|a_{n_0}| + |a_{n_0+1}| + \cdots + |a_{n_0+m}| \leqq c_{n_0} + c_{n_0+1} + \cdots + c_{n_0+m} \leqq c,$$

woraus mit Satz 77 folgt, daß die Reihe $\sum_{n=0}^{\infty} a_n$ absolut konvergiert.
Nehmen wir andererseits an, daß

$$|a_n| \geqq c_n > 0 \quad \forall n \geqq n_0$$

gilt und die Reihe $\sum_{n=0}^{\infty} c_n$ divergiert. Dann folgt

$$|a_{n_0}| + |a_{n_0+1}| + \cdots + |a_{n_0+m}| \geqq c_{n_0} + c_{n_0+1} + \cdots + c_{n_0+m}.$$

Die rechte Seite dieser Ungleichung wird mit wachsendem m beliebig groß; also kann die Reihe $\sum_{n=0}^{\infty} a_n$ nicht absolut konvergieren. □

Satz 79 (2. Majorantenkriterium). *Es sei (c_n) eine Folge mit positiven Gliedern.*
Wenn die Reihe $\sum_{n=0}^{\infty} c_n$ konvergiert und für eine Folge (a_n), in der alle Glieder ungleich 0 sind, ab einem gewissen Index n_0

$$\frac{|a_{n+1}|}{|a_n|} \leqq \frac{c_{n+1}}{c_n} \quad \forall n \geqq n_0$$

gilt, so konvergiert die Reihe $\sum_{n=0}^{\infty} a_n$ absolut.
Wenn die Reihe $\sum_{n=0}^{\infty} d_n$ divergiert und für eine Folge (a_n), in der alle Glieder ungleich 0 sind, ab einem gewissen Index n_0

$$\frac{|a_{n+1}|}{|a_n|} \geqq \frac{d_{n+1}}{d_n} \quad \forall n \geqq n_0$$

gilt, so konvergiert die Reihe $\sum_{n=0}^{\infty} a_n$ nicht absolut.

Beweis. Wir schreiben die beiden Ungleichungen

$$\frac{d_{n+1}}{d_n} \leqq \frac{|a_{n+1}|}{|a_n|} \leqq \frac{c_{n+1}}{c_n}$$

für $n = n_0, \ldots, n_0 + m - 1$ auf:

$$\frac{d_{n_0+1}}{d_{n_0}} \leqq \frac{|a_{n_0+1}|}{|a_{n_0}|} \leqq \frac{c_{n_0+1}}{c_{n_0}}$$

$$\frac{d_{n_0+2}}{d_{n_0+1}} \leqq \frac{|a_{n_0+2}|}{|a_{n_0+1}|} \leqq \frac{c_{n_0+2}}{c_{n_0+1}}$$

usw. bis

$$\frac{d_{n_0+m}}{d_{n_0+m-1}} \leqq \frac{|a_{n_0+m}|}{|a_{n_0+m-1}|} \leqq \frac{c_{n_0+m}}{c_{n_0+m-1}}.$$

Wir multiplizieren nun – beginnend mit der letzten – die Ungleichungen suksessive miteinander und erhalten

$$\frac{d_{n_0+m}}{d_{n_0}} \leqq \frac{|a_{n_0+m}|}{|a_{n_0}|} \leqq \frac{c_{n_0+m}}{c_{n_0}},$$

d. h. mit $n = n_0 + m$:

$$\frac{|a_{n_0}|}{d_{n_0}} \cdot d_n \leqq |a_n| \leqq \frac{|a_{n_0}|}{c_{n_0}} \cdot c_n.$$

Mit dem 1. Majorantenkriterium folgen hieraus die behaupteten Eigenschaften. □

Satz 80 (Wurzelkriterium). *Wenn es zu einer Folge (a_n) eine positive Zahl $q < 1$ gibt, so daß ab einem Index n_0*

$$\sqrt[n]{|a_n|} \leqq q \quad \forall n \geqq n_0$$

gilt, so konvergiert die Reihe $\sum_{n=0}^{\infty} a_n$ absolut.
Falls

$$\sqrt[n]{|a_n|} \geqq 1 \quad \forall n \geqq n_0$$

gilt, divergiert die Reihe.

Beweis. Die Voraussetzung des Wurzelkriteriums schreiben wir in der Form

$$|a_n| \leqq q^n \quad \forall n \geqq n_0.$$

Wir wissen bereits, daß die Reihe $\sum_{n=0}^{\infty} q^n$ für $|q| < 1$ konvergiert; daher folgt die behauptete Konvergenz aus dem 1. Majorantenkriterium. Der 2. Teil ergibt sich dadurch, daß wegen der Voraussetzung die Folge (a_n) keine Nullfolge ist. □

Satz 81 (Quotientenkriterium). *Wenn es zu einer Folge (a_n) eine positive Zahl $q < 1$ gibt, so daß ab einem Index n_0*

$$\frac{|a_{n+1}|}{|a_n|} \leqq q \quad \forall n \geqq n_0$$

gilt, so konvergiert die Reihe $\sum_{n=0}^{\infty} a_n$ absolut.
Falls

$$\frac{|a_{n+1}|}{|a_n|} \geqq 1 \quad \forall n \geqq n_0$$

gilt, divergiert die Reihe.

Beweis. Die Voraussetzung des Quotientenkriteriums schreiben wir in der Form

$$\frac{|a_{n+1}|}{a_n} \leqq q = \frac{q^{n+1}}{q^n} < 1$$

und wenden das zweite Majorantenkriterium an. □

Satz 82 (Kleiner Umordnungssatz). *Eine Reihe konvergiert dann und nur dann absolut, wenn sie unbedingt konvergiert.*

Den Beweis dieses Satz übergehen wir hier, da er etwas länglich ist. Gleiches gilt für den folgenden Satz.

Satz 83 (Multiplikation von Reihen). *Wenn die Reihen $\sum_{n=0}^{\infty} a_n$, $\sum_{m=0}^{\infty} b_m$ absolut konvergieren mit den Werten s_a, s_b, so gilt*

$$s_a \cdot s_b = \sum_{n=0}^{\infty} \sum_{m=0}^{n} a_m b_{n-m}.$$

Beispiele. Die folgenden Beispiele sollen nicht nur die Aussagen illustrieren, sondern gleichzeitig neue, spezifische Aspekte beleuchten.
1. Für beliebiges x betrachten wir die Reihe

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^2}{3!} + \cdots .$$

Indem wir das Quotientenkriterium anwenden, erhalten wir

$$\frac{|a_{n+1}|}{|a_n|} = \frac{|x|}{n+1},$$

und das Quotientenkriterium sagt uns, daß die Reihe für alle x absolut konvergiert. Es sei erwähnt, daß die Reihe die Exponentialfunktion e^x darstellt. Die Reihe ist Grundlage für die Berechnung der Exponentialfunktion auf einem Rechner. Für die Auswertung der Reihe auf dem Rechner bei gegebenem x hat man so vorzugehen, daß man nur so viele Glieder der Reihe verwendet, wie zum Erreichen der Maschinengenauigkeit – im Rahmen einer gegebenen Mantissenlänge – nötig sind. Die Darstellung elementarer Funktionen mittels (geeigneter) unendlicher Reihen ist für die Standardsoftware ein wichtiges Problem, da auf dem Rechner nur die arithmetischen Operationen mehr oder weniger vollkommen nachgebildet sind.
2. Es sei die Reihe

$$\sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} \mp \cdots$$

vorgelegt. Da die absoluten Glieder der Reihe eine Teilfolge der Folge aus dem 1. Beispiel sind, konvergiert die Reihe absolut für alle x; sie stellt die trigonometrische Funktion $\cos x$ dar. Für Werte von x in der Nähe von 1 erkennt man die Näherungsformel

$$\cos x \approx 1 - \frac{x^2}{2}.$$

3. Mit dem gleichen Argument konvergiert auch die Reihe

$$\sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} \mp \cdots$$

für alle x absolut; ihr Wert ist gleich dem Wert der trigonometrischen Funktion $\sin x$. Für Werte von x in der Nähe von 0 ergibt sich die Näherungsformel

$$\sin x \approx x - \frac{x^3}{6}.$$

4. Auf die Reihe

$$\sum_{n=1}^{\infty} \frac{x^n}{n} = x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots$$

wenden wir das Quotientenkriterium an:

$$\frac{|a_{n+1}|}{|a_n|} = |x| \cdot \frac{n}{n+1}.$$

Für $|x| < 1$ ist das Quotientenkriterium mit $q = |x|$ erfüllt und die Reihe konvergiert in diesem Falle absolut. Ist $|x| > 1$, so gilt wegen $\frac{n}{n+1} \to 1$ ab einem gewissen n_0:

$$|x| \cdot \frac{n}{n+1} \geqq 1 \quad \forall n \geqq n_0,$$

womit nach dem zweiten Teil des Quotientenkriteriums die Reihe als divergent verifiziert ist.
Für $x = 1$ liegt die harmonische Reihe vor, von der wir bereits wissen, daß sie bestimmt divergiert.
Für $x = -1$ konvergiert die Reihe nach dem Leibniz-Kriterium.
5. Die sog. **geometrische Reihe**

$$\sum_{n=0}^{\infty} q^n$$

konvergiert für $|q| < 1$ und es gilt

$$\sum_{n=0}^{\infty} q^n = \frac{1}{1-q}.$$

Wir multiplizieren die Reihe mit sich und erhalten

$$\begin{aligned}\frac{1}{(1-q)^2} &= \sum_{n=0}^{\infty} q^n \cdot \sum_{n=0}^{\infty} q^n = \sum_{n=0}^{\infty} \sum_{k=0}^{n} q^k q^{n-k} \\ &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} q^n = \sum_{n=0}^{\infty} (n+1) q^n.\end{aligned}$$

4.4. Stetigkeit und Grenzwerte von Funktionen

Es sei f eine auf $X \subseteqq \mathbb{R}$ definierte, reellwertige Funktion. Die Funktion f heißt **stetig** im Punkte $a \in \operatorname{int}(X)$, wenn für jede gegen a konvergente Folge (a_n) die Folge $(f(a_n))$ der Funktionswerte konvergiert und den Grenzwert $f(a)$ hat:

$$\lim_{n\to\infty} f(a_n) = f(a) \quad \forall (a_n) : \lim_{n\to\infty} a_n = a$$

oder kürzer

$$\lim_{n\to\infty} f(a_n) = f(\lim_{n\to\infty} a_n).$$

Ist die Funktion f in jedem Punkte aus $\operatorname{int}(X)$ stetig, so heißt f stetig in X.

Satz 84 (Stetigkeitskriterium). *Eine auf $X \subseteqq \mathbb{R}$ definierte, reellwertige Funktion f ist genau dann im Punkte $a \in \operatorname{int}(X)$ stetig, wenn es zu jedem $\varepsilon > 0$ eine Zahl $\eta > 0$ derart gibt, daß für alle $x \in X$ mit $|x - a| < \eta$ die Ungleichung $|f(x) - f(a)| < \varepsilon$ erfüllt ist.*

Beweis. Wir zeigen indirekt, daß die angegebene Bedingung notwendig für die Stetigkeit ist. Es sei also $\varepsilon > 0$ eine solche Zahl, daß zu jedem $\eta > 0$ ein $x \in X$ mit $|x - a| < \eta$ existiert, für das aber $|f(x) - f(a)| \geqq \varepsilon$ gilt. Wir wählen $\eta = \frac{1}{n}$; dann gibt es zu jedem n ein $a_n \in X$ mit $|x - a_n| < \frac{1}{n}$ und $|f(x) - f(a_n)| \geqq \varepsilon$. Offenbar konvergiert die Folge (a_n) gegen a, aber die Folge der Funktionswerte konvergiert nicht gegen $f(a)$; folglich ist

f nicht stetig in a, was der Voraussetzung widerspricht.
Wir zeigen nun, daß die im Satz genannte Bedingung hinreichend für die Stetigkeit ist. Dazu sei $(a_n) \subseteqq X$ eine gegen $a \in X$ konvergente Folge, $\varepsilon > 0$ beliebig fixiert und $\eta > 0$ eine zu ε gehörende Zahl mit der Eigenschaft:

$$|f(x) - f(a)| < \varepsilon \quad \forall x \in X : |x - a| < \eta.$$

Da die Folge (a_n) gegen a konvergiert, existiert ein $n_0(\eta)$ mit

$$|a_n - a| < \eta \quad \forall n \geqq n_0(\eta)$$

und daher

$$|f(a_n) - f(a)| < \varepsilon \quad \forall n \geqq n_0(\eta),$$

was bedeutet, daß $f(a)$ der Grenzwert der Folge $(f(a_n))$ ist. □
Um auch Randpunkte der Menge X zu erfassen, benötigen wir noch den Grenzwert einer Funktionswertfolge für den Fall, daß der Grenzwert möglicherweise nicht zum Wertebereich der Funktion gehört. Wir sagen, daß eine Funktion f in $a \in X$ den **Grenzwert** b hat, wenn für jede gegen a konvergente Folge $(a_n) \subseteqq X$ die Funktionswerfolge $(f(a_n))$ konvergiert und den Grenzwert b hat:

$$\lim_{n\to\infty} f(a_n) = b \quad \forall (a_n) \subseteqq X : \lim_{n\to\infty} a_n = a.$$

Zur vereinfachenden Schreibweise: Mit einer Gleichung der Form

$$\lim_{x\to a} f(x) = b$$

ist folgendes gemeint: Für jede gegen a konvergente Folge konvergiert auch die entsprechende Funktionswertfolge und alle haben den gleichen Grenzwert, nämlich die Zahl b.
Beispiele.
1. Bei der Funktion f mit

$$f(x) = (\operatorname{sgn}(x))^2$$

gilt für alle $x \neq 0$:

$$|f(x) - 1| = \left|(\operatorname{sgn}(x))^2 - 1\right| = 0,$$

also

$$\lim_{x\to 0} f(x) = 1$$

aber

$$(\operatorname{sgn}(0))^2 = 0.$$

Insbesondere ist diese Funktion in $x = 0$ unstetig, hat aber dort einen endlichen Grenzwert.
2. Die Funktion f mit

$$f(x) = \frac{1}{x^2} \quad (x \neq 0)$$

hat in $x = 0$ den uneigentlichen Grenzwert ∞, denn für jede Nullfolge (a_n) ist die Folge $(\frac{1}{a_n^2})$ bestimmt divergent.
3. Es sei die Funktion f mit

$$f(x) = \frac{x^2}{1 + x^2}$$

gegeben und $(\frac{1}{a_n})$, $a_n \neq 0$ eine Nullfolge; dann divergiert die Folge (a_n) bestimmt und

$$\lim_{n\to\infty} f(a_n) = \lim_{n\to\infty} \frac{a_n^2}{1 + a_n^2} = \lim_{n\to\infty} \frac{1}{1 + \frac{1}{a_n^2}} = 1;$$

also hat die Funktion für jede unbedingt divergente Folge den Grenzwert 1.
Es sei X ein Intervall: $X = [a, b]$ und f eine auf X erklärte Funktion. Für einen Punkt $y \in X$ heißt die Funktion f **linksseitig stetig**, wenn für alle positiven Nullfolgen $(h_n), (h_n > 0)$ gilt:

$$\lim_{n\to\infty} f(y - h_n) = f(y).$$

Ganz analog nennt man die Funktion f in $y \in X$ **rechtsseitig stetig**, falls für jede positive Nullfolge gilt

$$\lim_{n\to\infty} f(y+h_n) = f(y).$$

Solche Grenzwerteigenschaft schreibt man meist kurz in der Form

$$\lim_{h\to 0-} f(y+h) = f(y) \text{ bzw. } \lim_{h\to 0+} f(y+h) = f(y).$$

Eine Funktion f heißt auf X stetig, wenn sie in X stetig und in den Randpunkten rechts- bzw. linksseitig stetig ist.
Beispiele.
1. Wir betrachten die Funktion

$$f(x) = \begin{cases} \sin\dfrac{\pi}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}.$$

Für $x = \frac{2}{4n+1}$ ist

$$\sin\frac{\pi}{x} = \sin\frac{\pi}{2}(4n+1) = \sin\left(\frac{\pi}{2} + 2n\pi\right) = \sin\frac{\pi}{2} = 1;$$

für $x = \frac{2}{4n+3}$

$$\sin\frac{\pi}{x} = \sin\left(\frac{3\pi}{2} + 2n\pi\right) = \sin\frac{3}{2}\pi = -1$$

und für $x = \frac{1}{n}$:

$$\sin\frac{\pi}{x} = \sin n\pi = 0.$$

Folglich ist f in $x = 0$ unstetig. Die Funktion nimmt in jeder noch so kleinen Umgebung vom Nullpunkt jeden Wert aus dem Intervall [-1,1] unendlich oft an. Das übersteigt die menschliche Vorstellungskraft.
2. Die Funktion

$$f(x) = \begin{cases} \frac{1}{n+1} & \frac{1}{n+1} < x \leqq \frac{1}{n} \\ 0 & x = 0 \end{cases}$$

ist in $x = 0$ stetig, da $|f(x)| \leqq |x|$.

Satz 85. *Die Menge $C(X)$ aller auf $X \subseteqq \mathbb{R}$ stetigen Funktionen bildet mit der Multiplikation eine Halbgruppe und ist ein Vektorraum über den reellen Zahlen.*

Den Beweis möge man als Übung selbst ausführen. Man hat nur zu zeigen: Sind f und g stetige Funktionen auf X, so auch $\alpha \cdot f, f+g, f \cdot g$.

Satz 86. *Ist die Funktion f stetig in $a \in \operatorname{int}(X)$, die Funktion g stetig in $f(a)$, so ist $g \circ f$ in a stetig.*

Auch der Beweis dieses Satzes sollte dem Leser leicht fallen.

Satz 87. *Das Bild $f(X)$ einer auf einer beschränkten, abgeschlossenen Menge X stetigen Funktion f ist abgeschlossen.*

Beweis. Es sei $(y_n) \subseteqq f(X)$ eine gegen y^* konvergente Folge. Zu jedem y_n existiert ein x_n mit $f(x_n) = y_n$. Die Folge $(x_n) \subseteqq X$ ist beschränkt, da X beschränkt ist und hat daher einen Häufungspunkt x^*, der wegen der Abgeschlossenheit von X auch in der Menge X liegen muß; mit der Stetigkeit von f folgt daraus:

$$f(x^*) = \lim_{n_i\to\infty} f(x_{n_i}) = \lim_{n_i\to\infty} y_{n_i} = y^*,$$

also gilt $y^* \in f(X)$. □

Satz 88 (Minimum-Maximum für stetige Funktionen). *Jede auf einer beschränkten, abgeschlossenen Menge stetige Funktion nimmt dort ihre untere und ihre obere Grenze an.*

Beweis. Wir beweisen den Satz nur für die obere Grenze; wegen

$$\inf f(x) = \sup -f(x)$$

gilt die Aussage dann auch für die untere Grenze.
Es sei $X \subseteqq \mathbb{R}$ eine beschränkte, abgeschlossene Menge und f eine auf X stetige Funktion; ferner sei M die obere Grenze von $f(X)$. Dann gibt es eine Folge $(y_n) \subseteqq f(X)$ mit $\lim_{n\to\infty} y_n = M$. Nach dem vorangegangenen Satz ist $f(X)$ eine abgeschlossene Menge, woraus $M \in f(X)$ folgt, d. h. es gibt ein $x^* \in X$ mit $f(x^*) = M$. □
Wenn die obere Grenze von einer Funktion angenommen wird, nennt man sie **Maximum** der Funktion; entsprechend spricht man von einem **Minimum**, wenn die Funktion ihre untere Grenze annimmt.

Satz 89 (Nullstelleneigenschaft). *Es sei f eine auf $[a, b]$ stetige Funktion. Haben die Funktionswerte $f(a)$ und $f(b)$ unterschiedliches Vorzeichen, dann hat f im Intervall $[a, b]$ eine Nullstelle.*

Beweis. Wir konstruieren eine Nullstelle nach dem sog. **Bisektionsverfahren**:

Eingabe:
a : untere Intervallgrenze,
b : obere Intervallgrenze,
f : stetige Funktion mit $f(a) < 0, f(b) > 0$,
Programm:
$x := a; y := b$
while *true* do
$z := \dfrac{x+y}{2}; u = f(z)$
if $u = 0$ do $out := z$ exit endif { z ist Nullstelle. }
if $u < 0$ $x := z$ else $y := z$ endif
endwhile.

Wenn der Algorithmus in endlicher Zeit endet, hat er offenbar eine Nullstelle von f gefunden. Andernfalls wird die Schleife unendlich oft durchlaufen und erzeugt so zwei Folgen $(x_n), (y_n)$, wobei die Folge (x_n) monoton wächst, die Folge (y_n) monoton fällt und

$$f(x_n) < 0, f(y_n) > 0, \quad y_n - x_n = \frac{b-a}{2^n} \quad \forall n$$

gilt. Die Intervallängen $y_n - x_n$ bilden also eine Nullfolge; daher haben beide Folgen einen gemeinsamen Grenzwert x^*; in einer Umgebung vom Grenzwert liegen links nur Punkte mit negativen Funktionswerten und rechts nur solche mit positiven Funktionswerten. Also muß $f(x^*) = 0$ sein. Wir bemerken noch, daß man den Test „$u = 0$“ durch einen Genauigkeitstest, etwa von der Form

$$\max\{\, |u|, y - x, f(y) + f(x) \,\} < \varepsilon$$

ersetzt. □

Satz 90 (Zwischenwerteigenschaft). *Jede auf einem gegebenen Intervall $[a, b]$ stetige Funktion f nimmt dort jeden zwischen $f(a)$ und $f(b)$ gelegenen Wert in mindestens einem Punkte an.*

Beweis. Es sei c ein beliebiger Wert zwischen $f(a)$ und $f(b)$; wir nehmen die Funktion

$$\varphi(x) = f(x) - c.$$

Diese Funktion ist stetig auf dem Intervall $[a, b]$ und nimmt in den Endpunkten des Intervalls Werte mit unterschiedlichem Vorzeichen an. Mit der Nullstelleneigenschaft schließen wir, daß es ein x^* mit $0 = \varphi(x^*) = f(x^*) - c$ gibt. □
Es sei erwähnt, daß auch die Menge aller auf einem Intervall definierten Funktionen, die die Zwischenwerteigenschaft haben, einen Vektorraum über den reellen Zahlen bildet.
Eine zentrale Bedeutung für Lösung vieler angewandter Aufgaben hat der nun folgende Fixpunktsatz, den wir im $\mathbb{R}^n$ formulieren wollen.
Dazu sei $X \subseteqq \mathbb{R}^n$ und f eine Abbildung von X in sich. Ein Punkt aus der Menge X, der bei der Abbildung f auf sich abgebildet wird, heißt **Fixpunkt** von f. Ein Fixpunkt ist also durch die Gleichung

$$f(\mathbf{x}^*) = \mathbf{x}^*, \quad \mathbf{x}^* \in X$$

charakterisiert. Eine Abbildung f von X in sich heißt **kontrahierend** auf X, wenn sich der Abstand von je zwei Punkten aus X bei der Abbildung gleichmäßig verkleinert, d. h. wenn es eine positive Zahl $q < 1$ gibt, so daß

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leqq q\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X$$

gilt. Aus dieser Ungleichung schließt man, daß jede kontrahierende Abbildung stetig sein muß. Wir überlegen uns sogleich, daß eine kontrahierende Abbildung höchstens einen Fixpunkt haben kann. Sind nämlich $\mathbf{x}, \mathbf{y}$ Fixpunkte von f, so folgt mit der Fixpunktgleichung und der Kontraktionsbedingung

$$\|\mathbf{x}-\mathbf{y}\| = \|f(\mathbf{x})-f(\mathbf{y})\| \leqq q\|\mathbf{x}-\mathbf{y}\|,$$

woraus wir wegen $0 < q < 1$ sofort $\mathbf{x} = \mathbf{y}$ schließen.

Satz 91 (Fixpunktsatz). *Jede auf einer abgeschlossenen Menge X kontrahierende Abbildung f mit einer Kontraktionskonstanten q hat genau einen Fixpunkt $\mathbf{x}^* \in X$. Dieser Fixpunkt ist Grenzwert der Folge $(\mathbf{x}_k)$, die gemäß*

$$\mathbf{x}_0 \in X, \quad \mathbf{x}_{k+1} = f(\mathbf{x}_k),\ k = 0, 1, 2, \ldots$$

konstruiert ist; außerdem gilt die Abschätzung

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leqq \frac{q^k}{1-q}\|\mathbf{x}_0 - \mathbf{x}_1\|.$$

Beweis. Wir schätzen den Abstand von zwei aufeinander folgenden Gliedern ab:

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \|f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})\| \leqq q\|\mathbf{x}_k - \mathbf{x}_{k-1}\| \leqq \cdots \leqq q^k\|\mathbf{x}_1 - \mathbf{x}_0\|$$

und daher

$$\|\mathbf{x}_{k+r+1} - \mathbf{x}_{k+r}\| \leqq q^r\|\mathbf{x}_{k+1} - \mathbf{x}_k\|,$$

womit wir erhalten:

$$\begin{aligned}\|\mathbf{x}_{k+r+1} - \mathbf{x}_k\| &\leqq \sum_{i=0}^{r}\|\mathbf{x}_{k+i+1} - \mathbf{x}_{k+i}\| \leqq \|\mathbf{x}_{k+1} - \mathbf{x}_k\|\sum_{i=0}^{r} q^i \\ &\leqq \|\mathbf{x}_1 - \mathbf{x}_0\|q^k\sum_{i=0}^{r} q^i \leqq \|\mathbf{x}_1 - \mathbf{x}_0\|\frac{q^k}{1-q};\end{aligned}$$

also ist $(\mathbf{x}_k)$ eine Fundamentalfolge, die nach dem Cauchyschen Konvergenzkriterium einen Grenzwert $\mathbf{x}^*$ hat, der in der abgeschlossenen Menge X liegen muß. Mit der Stetigkeit von f folgt

$$\mathbf{x}^* = \lim_{k\to\infty}\mathbf{x}_{k+1} = \lim_{k\to\infty} f(\mathbf{x}_k) = f(\mathbf{x}^*),$$

d. h. $\mathbf{x}^*$ ist ein Fixpunkt von f. Für den Abstand des k-ten Folgegliedes vom Fixpunkt berechnen wir

$$\begin{aligned}\|\mathbf{x}^* - \mathbf{x}_k\| &\leqq \|\mathbf{x}^* - \mathbf{x}_{k+r+1}\| + \|\mathbf{x}_{k+r+1} - \mathbf{x}_k\| \\ &\leqq \|\mathbf{x}^* - \mathbf{x}_{k+r+1}\| + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|\sum_{i=0}^{r} q^i\end{aligned}$$

und für $r \to \infty$:

$$\|\mathbf{x}^* - \mathbf{x}_k\| \leqq \|\mathbf{x}_{k+1} - \mathbf{x}_k\|\frac{1}{1-q} \leqq \|\mathbf{x}_k - \mathbf{x}_{k-1}\|\frac{q}{1-q} \leqq \cdots \leqq \|\mathbf{x}_1 - \mathbf{x}_0\|\frac{q^k}{1-q},$$

womit alles bewiesen ist. □

Die Bedeutung dieses Satzes liegt vor allem in seiner Konstruktivität: Er beinhaltet nicht nur eine qualitative Aussage, sondern liefert gleichzeitg eine Lösungsmethode nebst einer Genauigkeitsabschätzung über die erreichte Näherung bei Abbruch des Verfahrens.

4.5. Folgen und Reihen von Funktionen

Ein wichtiges Anliegen der Analysis ist es, komplizierte Funktionen durch möglichst einfache anzunähern. Eine solche Annäherung muß die Möglichkeit einer verbesserten Annäherung derart beinhalten, daß man eine beliebig genaue Annäherung erreichen kann, sofern man nur hinreichend lange rechnet. Für dieses Ziel ist es sachgemäß, Folgen von Funktionen zu untersuchen.

Wir betrachten eine Folge (f_n) von auf $X \subseteq \mathbb{R}$ definierten Funktionen und sagen, daß die Folge (f_n) auf X **konvergiert**, wenn eine auf X definierte Funktion f, die **Grenzfunktion**, existiert mit

$$f(x) = \lim_{n\to\infty} f_n(x) \quad \forall x \in X.$$

Beispiele. Die Folge (f_n) mit

$$f_n(x) = \left(1 + \frac{x}{n}\right)^n \quad (x \in \mathbb{R})$$

hat als Grenzfunktion die Exponentialfunktion e^x und $\ln x$ ist die Grenzfunktion der Funktionenfolge (φ_n) mit

$$\varphi_n(x) = n\left(\sqrt[n]{x} - 1\right) \quad (x > 0).$$

Sind die Glieder einer konvergenten Funktionenfolge sämtlich stetig, so braucht die Grenzfunktion $f(x)$ nicht stetig zu sein, wie das folgende Beispiel zeigt:

$$f_n(x) = \begin{cases} -nx + 1 & \text{für } 0 \leqq x \leqq \dfrac{1}{n} \\ 0 & \text{für } x > \dfrac{1}{n} \end{cases}.$$

Es ist

$$f(x) = \lim_{n\to\infty} f_n(x) = \begin{cases} 1 & \text{für } x = 0 \\ 0 & \text{für } x > 0 \end{cases}.$$

Aus diesem Grunde brauchen wir einen neuen Begriff, der uns sichert, daß die Grenzfunktion einer Folge stetiger Funktionen stetig ist. Eine Folge (f_n) von auf einer Menge $X \subseteq \mathbb{R}$ definierten Funktionen **konvergiert gleichmäßig** gegen eine Funktion f, wenn es zu jedem $\varepsilon > 0$ ein $n_0(\varepsilon)$ gibt, so daß

$$|f(x) - f_n(x)| < \varepsilon \quad \forall x \in X,\ \forall n \geqq n_0(\varepsilon)$$

ausfällt. Wesentlich an diesem Begriff ist es, daß die Zahl $n_0(\varepsilon)$ nur von ε und nicht noch von x abhängt. Inhaltlich besagt diese Definition, daß zu beliebig vorgegebenem $\varepsilon > 0$ ab einem gewissen Index n_0 alle Funktionen f_n in einem ε-Schlauch um die Grenzfunktion verlaufen. So konvergiert die obige Folge nicht gleichmäßig. Um dies einzusehen, setzen wir $\varepsilon = 1$ und nehmen die Folge $(x_n) = (\frac{1}{2n})$; es ist

$$|f(x_n) - f_n(x_n)| = |0 - \frac{1}{2}| = \frac{1}{2} \geqq \varepsilon.$$

Betrachten wir dagegen

$$f_n(x) = \sum_{k=1}^{n} \frac{\cos kx}{k^2}$$

und wählen $\varepsilon > 0$ beliebig; da die Reihe $\sum_{n=1}^{\infty} \frac{1}{n^2}$ konvergiert existiert ein $n_0(\varepsilon)$ mit

$$\frac{1}{(n+1)^2} + \frac{1}{(n+2)^2} + \cdots < \varepsilon \quad \forall n \geqq n_0(\varepsilon).$$

Damit folgt aber

$$|\sum_{k=1}^{\infty} \frac{\cos kx}{k^2} - f_n(x)| = |\sum_{k=n+1}^{\infty} \frac{\cos kx}{k^2}| \leqq \sum_{k=n+1}^{\infty} \frac{1}{k^2} < \varepsilon.$$

Satz 92. (Satz über die stetige Grenzfunktion)
Jede auf einer Menge X gleichmäßig konvergierende Folge (f_n) stetiger Funktionen hat eine stetige Grenzfunktion.

Beweis. Es seien (f_n) eine auf X gleichmäßig konvergente Folge stetiger Funktionen mit der Grenzfunktion f und $y \in X$; zu $\varepsilon > 0$ sei $n_0(\frac{\varepsilon}{3})$ so gewählt, daß

$$|f_n(x) - f(x)| < \frac{\varepsilon}{3} \quad \forall x \in X, \forall n \geqq n_0(\frac{\varepsilon}{3})$$

gilt. Wir fixieren ein beliebiges $n \geqq n_0(\frac{\varepsilon}{3})$. Zu $\frac{\varepsilon}{3}$ gibt es wegen der Stetigkeit von f_n ein $\eta > 0$ derart, daß

$$|f_n(x) - f_n(y)| < \frac{\varepsilon}{3} \quad \forall x \in X, |x - y| < \eta.$$

Für diese x folgt:

$$\begin{aligned} |f(x) - f(y)| &\leqq |f_n(x) - f(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon, \end{aligned}$$

was uns zeigt, daß die Grenzfunktion stetig ist. □
Wir sagen, daß eine Funktionenreihe s mit

$$s(x) = \sum_{n=0}^{\infty} f_n(x)$$

gleichmäßig konvergiert, wenn die zugeordnete Partialsummenfolge (s_n) gleichmäßig konvergiert.

Satz 93. (Kriterium von Weierstraß)
Wenn für eine Funktionenfolge (f_n) eine Abschätzung der Form

$$|f_n(x)| \leqq a_n \quad \forall x \in X$$

gilt und die Reihe $\sum_{n=0}^{\infty} a_n$ konvergiert, dann konvergiert die Funktionenreihe s mit

$$s(x) = \sum_{n=0}^{\infty} f_n(x)$$

gleichmäßig.

Beweis. Es sei

$$s_n(x) = \sum_{k=0}^{n} f_k(x);$$

dann ist

$$|s_{n+m}(x) - s_n(x)| \leqq a_{n+1} + a_{n+2} + \cdots + a_{n+m}.$$

Da die Reihe $\sum_{n=0}^{\infty} a_n$ konvergiert, ist die Folge $(s_n(x))$ eine Fundamentalfolge und nach dem Cauchyschen Konvergenzkriterium existert eine Grenzfunktion s mit

$$s(x) = \lim_{n\to\infty} s_n(x).$$

Bei beliebig fixiertem $\varepsilon > 0$ gibt es ein $n_0(\varepsilon)$ mit

$$a_{n+1} + a_{n+2} + \cdots < \varepsilon \quad \forall n \geqq n_0(\varepsilon),$$

also

$$|s(x) - s_n(x)| \leqq a_{n+1} + a_{n+2} + \cdots < \varepsilon \quad \forall n \geqq n_0(\varepsilon), \forall x \in X,$$

d. h. die Reihe $\sum_{n=0}^{\infty} f_n(x)$ konvergiert gleichmäßig. □
Beispiel. Für die Reihe

$$s(x) = \sum_{n=1}^{\infty} \frac{x^n}{n^2} \quad (|x| < 1)$$

folgt wegen

$$\left|\frac{x^n}{n^2}\right| \leqq \frac{1}{n^2}$$

mit dem Majorantenkriterium, daß die Reihe gleichmäßig konvergiert; da alle Glieder stetige Funktionen sind, folgt weiter, daß die Reihe eine stetige Funktion darstellt.

4.6. Eindimensionale Differentialrechnung

4.6.1. Differenzierbarkeit

Eine auf einer Menge $X \subseteq \mathbb{R}$ definierte reellwertige Funktion f heißt an einer Stelle $a \in \text{int}\,(X)$ **differenzierbar** (**ableitbar**), wenn die Funktion φ mit

$$\varphi(h) = \frac{f(a+h) - f(a)}{h}$$

an der Stelle 0 (d. h. für $h = 0$) einen endlichen Grenzwert hat. Dieser Grenzwert wird mit $f'(a)$ bezeichnet und heißt **Ableitung** (**Differentialquotient**) der Funktion f an der Stelle a. Andere übliche Schreibweisen für die Ableitung sind:

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

und mit $x = a + h$:

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}.$$

An dieser Stelle führen wir zwei sehr zweckmäßige Hilfsmittel der Analysis ein, die **Landau-Symbole**. Es seien φ, ψ zwei auf einer Menge X definierte Funktionen mit $\psi(x) \neq 0$ auf X. Falls es Zahlen $L > 0$ und $\eta > 0$ gibt, so daß die Ungleichung

$$\left| \frac{\varphi(x)}{\psi(x)} \right| \leqq L \quad \forall x \in X, |x - a| < \eta, x \neq a$$

gilt, nennt man φ eine $\mathcal{O}(\psi)$**-Funktion** für x gegen a und schreibt $\varphi(x) = \mathcal{O}(\psi(x))$. Sollte sogar zu jedem $L > 0$ ein $\eta > 0$ existieren, so daß die obige Ungleichung gilt, so schreibt man $\varphi(x) = o(\psi(x))$ und nennt φ eine $o(\psi)$**-Funktion** für x gegen a. Meist verwendet man Landau-Symbole, um das Verhalten einer Funktion φ für $x \to 0$ oder $x \to \infty$ abzuschätzen, so daß als Vergleichsfunktion ψ oft eine Funktion der Form $\psi(x) = x^r$ benutzt wird. Insbesondere bedeutet die Schreibweise $\varphi(x) = \mathcal{O}(1)$, daß die Funktion φ in einer Umgebung des Nullpunktes beschränkt ist. Leicht sieht man ein, daß die Summe zweier $\mathcal{O}$-Funktionen wieder eine $\mathcal{O}$-Funktion ist; gleiches gilt für die Summe zweier o-Funktionen. Wegen

$$\lim_{h \to 0} \frac{f(a+h) - f(a)}{h} - f'(a) = 0$$

gilt damit

$$\frac{f(a+h) - f(a)}{h} - f'(a) = \mathcal{O}(h)$$

und wir erhalten

$$f(a+h) = f(a) + h \cdot f'(a) + h \cdot \mathcal{O}(h)$$

mit $\lim_{h \to 0} \mathcal{O}(h) = 0$. Wegen $h \cdot \mathcal{O}(h) = o(h)$ ergibt sich die **Weierstraßsche Zerlegungsformel**:

$$f(a+h) = f(a) + h \cdot f'(a) + o(h)$$

mit

$$\lim_{h \to 0} \frac{o(h)}{h} = 0.$$

In erster Näherung gilt also

$$f(a+h) \approx f(a) + h \cdot f'(a).$$

Ist die Funktion f in jedem Punkte $x \in \operatorname{int}(X)$ differenzierbar, so heißt f differenzierbar in X; mit f' bezeichnet man die Ableitungsfunktion:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Sollte die Ableitung f' stetig in X sein, so heißt die Funktion f **stetig differenzierbar**.
Beispiele.
1. Für die Funktion f mit $f(x) = x^2$ erhalten wir:

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \lim_{x \to a} \frac{x^2 - a^2}{x - a} = \lim_{x \to a} x + a = 2a.$$

Also ist die Funktion in $\mathbb{R}$ differenzierbar und hat die Ableitung $2x$.
2. Für die Funktion f mit $f(x) = |x|$ ergibt sich mit $a \neq 0$ wegen $|x| = x \cdot \operatorname{sgn}(x)$:

$$\begin{aligned} \lim_{x \to a} \frac{|x| - |a|}{x - a} &= \lim_{x \to a} \frac{x \cdot \operatorname{sgn}(x) - a \cdot \operatorname{sgn}(a)}{x - a} \\ &= \lim_{x \to a} \operatorname{sgn}(a) \cdot \frac{x - a}{x - a} = \operatorname{sgn}(a), \end{aligned}$$

also

$$f'(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}.$$

Für $a = 0$ gilt für den Differenzenquotienten:

$$\frac{|x| - 0}{x - 0} = \operatorname{sgn}(x).$$

Somit ist die Funktion in $\mathbb{R} \setminus \{0\}$ differenzierbar, aber nicht in $x = 0$, dort aber stetig.
Wenn die einseitigen Grenzwerte

$$\lim_{\substack{h \to 0 \\ h > 0}} \frac{f(a+h) - f(a)}{h}, \quad \lim_{\substack{h \to 0 \\ h < 0}} \frac{f(a+h) - f(a)}{h}$$

existieren, so heißt f in a **rechtsseitig differenzierbar** bzw. **linksseitig differenzierbar** und die Grenzwerte sind die rechts- bzw. linksseitigen Ableitungen. So hat die Funktion f mit $f(x) = |x|$ an der Stelle 0 die rechtsseitige Ableitung $+1$ und die linksseitige Ableitung -1.
Wir erwähnen noch einige andere Schreibweisen für die Ableitung, die häufig in Anwendungen benutzt werden:

$$y', \quad \frac{dy}{dx}, \quad \frac{df}{dx}, \quad \frac{df(x)}{dx}$$

mit der Sprechweise „dy nach dx“ usw.
In früheren Zeiten mußte ein Anwender mathematischer Methoden insbesondere ein exellenter Handwerker im Ableiten mehr oder weniger komplizierter Funktionen sein. Heute gibt es selbst auf Kleinstrechnern Systeme, die bei Eingabe einer Funktion die Ableitung berechnen. Das enthebt uns aber nicht davon, die Ableitungen einiger elementarer Funktionen zu kennen, so wie wir auch für einfache arithmetische Operationen nicht erst einen Rechner bemühen sollten. Für einige elementare Funktionen wollen wir ihre Ableitungen herleiten.
1. Im Falle $f(x) = c$ folgt $f'(x) = 0$ für alle x, was man mit $f'(x) \equiv 0$ ausdrückt.
2. Um die Ableitung der Potenzfunktion

$$f(x) = x^n$$

zu berechnen, machen wir eine kleine Vorbemerkung. Es ist

$$\begin{aligned} y^n - x^n &= \sum_{k=0}^{n-1} y^{n-k} x^k - \sum_{k=1}^{n} y^{n-k} x^k \\ &= y \sum_{k=1}^{n} y^{n-k} x^{k-1} - \sum_{k=1}^{n} y^{n-k} x^k = (y - x) \sum_{k=1}^{n} y^{n-k} x^{k-1}. \end{aligned}$$

Die gefundene Endformel soll nun verwendet werden:

$$f'(y) = \lim_{x \to y} \frac{x^n - y^n}{x - y} = \lim_{x \to y} \sum_{k=1}^{n} x^{n-k} y^{k-1} = \sum_{k=1}^{n} y^{n-k} y^{k-1} = n \cdot y^{n-1},$$

also

$$(x^n)' = n \cdot x^{n-1}.$$

3. Es ist

$$(e^x)' = e^x.$$

Beweis. Wegen

$$\frac{e^{x+h} - e^x}{h} = e^x \cdot \frac{e^h - 1}{h}$$

genügt es zu zeigen:

$$\lim_{h \to 0} \frac{e^h - 1}{h} - 1 = 0,$$

was aber mit dem Majorantenkriterium aus der Abschätzung

$$0 < \left| \frac{1}{h}\left(e^h - 1\right) - 1 \right| \leqq |h| \cdot \left(\frac{1}{2!} + \frac{1}{3!} + \cdots \right) < |h| \cdot e$$

folgt. □

4. Es ist

$$(\ln x)' = \frac{1}{x} \quad (x > 0).$$

Beweis. Wir bemerken zunächst, daß mit

$$e^h = x$$

aus dem letzten Beispiel wegen

$$\left(\frac{e^h - 1}{h} \right) \cdot \frac{\ln x}{x - 1} = 1$$

der Grenzwert

$$\lim_{x \to 1} \frac{\ln x}{x - 1} = 1$$

folgt. Damit schließen wir für $a > 0$:

$$\lim_{x \to a} \frac{\ln x - \ln a}{x - a} = \frac{1}{a} \lim_{x \to a} \frac{\ln \frac{x}{a}}{\frac{x}{a} - 1} = \frac{1}{a} \lim_{x \to 1} \frac{\ln x}{x - 1} = \frac{1}{a},$$

womit die Behauptung bewiesen ist. □

5. Es ist

$$(\sin x)' = \cos x, \quad (\cos x)' = -\sin x.$$

Beweis. Wir verwenden das aus der Schule bekannte Additionstheorem

$$\sin x - \sin y = 2 \cos \frac{x + y}{2} \sin \frac{x - y}{2}$$

und die Grenzwertformel

$$\lim_{x \to 0} \frac{\sin x}{x} = 1.$$

die letzte Formel folgt mit dem Majorantenkriterium aus der Abschätzung

$$\left| \frac{\sin x}{x} - 1 \right| \leqq \frac{x^2}{3!} + \frac{x^5}{5!} + \cdots < x^2 \left(1 + \frac{1}{1!} + \frac{1}{2!} + \cdots \right) = x^2 e.$$

Damit erhält man

$$\lim_{h \to 0} \frac{\sin(x + h) - \sin x}{h} = \lim_{h \to 0} \cos\left(x + \frac{h}{2} \right) \cdot \frac{\sin \frac{h}{2}}{\frac{h}{2}} = \cos x.$$

Der Beweis für $\cos x$ verläuft analog. □

4.6.2. Eigenschaften differenzierbarer Funktionen

Aus der Definition stetiger Funktionen ergibt sich sofort, daß differenzierbare Funktionen stetig sind; aber nicht jede stetige Funktion ist auch diffenrenzierbar, wie wir bereits an einem Beispiel gesehen haben. Zunächst stellen wir einige wichtige Rechenregeln für differenzierbare Funktionen zusammen.

Satz 94 (Rechenregeln). *Die Menge $C^1(X)$ aller auf einer Menge X differenzierbaren Funktionen bildet einen Vektorraum über dem Körper der reellen Zahlen. Außerdem gelten die folgenden Regeln.*

Produktregel:

$$(f(x) \cdot g(x))' = f'(x)g(x) + f(x)g'(x) \quad \forall x \in X.$$

Quotientenregel:

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)} \quad (g(x) \neq 0) \quad \forall x \in X.$$

Umkehrregel *: Ist g die Umkehrfunktion von f und $(g' \circ f)(x) \neq 0$, so gilt*

$$f'(x) = \frac{1}{g'(f(x))}.$$

Kettenregel:

$$(f \circ g)'(x) = f'(g(x))g'(x)$$

oder als Merkregel mit $z = f(y), y = g(x)$:

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}.$$

Beweis. Die Produktregel erhält man aus der folgenden Grenzwertbetrachtung:

$$\begin{aligned}\lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h} &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} g(x+h) \\ &\quad + \lim_{h\to 0} f(x) \frac{g(x+h) - g(x)}{h} \\ &= f'(x)g(x) + f(x)g'(x).\end{aligned}$$

In der Quotientenregel ist $g(x) \neq 0$; also existert ein $\eta > 0$, so daß für alle h mit $|h| < \eta$ ebenfalls $g(x+h) \neq 0$ ausfällt. Für solche h schließen wir:

$$\begin{aligned}&\lim_{h\to 0} \frac{1}{h}\left(\frac{f(x+h)}{g(x+h)} - \frac{f(x)}{g(x)}\right) \\ &= \lim_{h\to 0} \frac{f(x+h)g(x) - f(x)g(x+h)}{hg(x+h)g(x)} \\ &= \lim_{h\to 0} \frac{1}{g(x+h)g(x)} \cdot \frac{f(x+h)g(x) - f(x)g(x+h)}{h} \\ &= \frac{1}{g^2(x)} \lim_{h\to 0} \left(\frac{f(x+h) - f(x)}{h} g(x) - f(x)\frac{g(x+h) - g(x)}{h}\right) \\ &= \frac{1}{g^2(x)} \left(f'(x)g(x) - f(x)g'(x)\right).\end{aligned}$$

Für die Umkehrregel verwenden wir die Weierstraßsche Zerlegungsformel:

$$\begin{aligned}x - a &= g(f(x)) - g(f(a)) \\ &= (f(x) - f(a))(g'(f(a)) + o(f(x) - f(a)))\end{aligned}$$

und erhalten

$$\lim_{x\to a} \frac{f(x) - f(a)}{x - a} = \lim_{x\to a} \frac{1}{g'(f(a)) + o(f(x) - f(a))} = \frac{1}{g'(f(a))}.$$

Für die Kettenregel verwenden wir die Weierstraßsche Zerlegungsformel in der Form

$$f(g(x)) - f(g(a)) = (g(x) - g(a)) \cdot (f'(g(a)) + o(g(x) - g(a)))$$

und folgern

$$\lim_{x\to a} \frac{f(g(x)) - f(g(a))}{x - a} = g'(a)f'(g(a)),$$

wobei die Ableitung von f an der Stelle $g(a)$ zu nehmen ist. □

Beispiele. Wir wollen einige Regeln auf bekannte elementare Funktionen anwenden:

$$\begin{aligned}(\tan x)' = \left(\frac{\sin x}{\cos x}\right)' &= \frac{\cos^2 x + \sin^2 x}{\cos^2 x} \\ &= \frac{1}{\cos^2 x} = 1 + \tan^2 x.\end{aligned}$$

Analog berechnet man

$$(\cot x)' = -\frac{1}{\sin^2 x} = -(1 + \cot^2 x).$$

Für die Umkehrfunktion $y = \arctan x$ ($x \in \mathbb{R}$) zu $x = \tan y$ mit $y \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ folgt daraus

$$(\arctan x)' = \frac{1}{1 + \tan^2(\arctan x)} = \frac{1}{1 + x^2}$$

und analog

$$(\operatorname{arccot} x)' = -\frac{1}{1 + x^2}.$$

Entsprechend ergibt sich für $|x| < 1$

$$(\arcsin x)' = \frac{1}{\cos(\arcsin x)} = \frac{1}{\sqrt{1 - x^2}}$$

und analog

$$(\arccos x)' = -\frac{1}{\sqrt{1 - x^2}}.$$

Als Ableitung für die Funktion $y = a^x$ folgt wegen

$$a^x = e^{x \cdot \ln a}$$

mit $f(y) = e^y, y = g(x) = x \cdot \ln a$:

$$(a^x)' = (f(g(x)))' = g'(x) \cdot f'(g(x)) = e^y \ln a = a^x \ln a.$$

Schließlich erhalten wir aus $x^\alpha = e^{\alpha \cdot \ln x}$ für $x > 0$:

$$(x^\alpha)' = \left(e^{\alpha \cdot \ln x}\right)' = \frac{\alpha}{x} x^\alpha = \alpha \cdot x^{\alpha - 1}.$$

Die wichtigsten Eigenschaften differenzierbarer Funktionen sollen nun bewiesen werden.

Satz 95 (Satz von Rolle). *Zwischen zwei Nullstellen einer gegebenen, differenzierbaren Funktion liegt eine Nullstelle der Ableitung.*

Beweis. Es sei f eine differenzierbare Funktion auf dem Intervall $[a, b], a < b$ und $f(a) = f(b) = 0$. Da die Funktion f stetig auf dem Intervall ist, nimmt sie ihren maximalen Wert in einem Punkte $x^* \in [a, b]$ an; dabei muß offenbar $f(x^*) \geqq 0$ sein. Wir unterscheiden nun zwei Fälle.
Fall 1: Es sei $f(x^*) > 0$; dann ist $a < x^* < b$ und für beliebige, aber kleine h gilt stets $f(x^* + h) - f(x^*) \leqq 0$. Damit wird das Vorzeichen des Differenzenquotienten

$$\frac{f(x^* + h) - f(x^*)}{h}$$

ausschließlich vom Vorzeichen von h bestimmt:

$$\frac{f(x^* + h) - f(x^*)}{h} \leqq 0 \quad \forall h > 0, a < x^* + h < b$$

und

$$\frac{f(x^* + h) - f(x^*)}{h} \geqq 0 \quad \forall h < 0, a < x^* + h < b.$$

Aus der ersten Ungleichung folgt

$$f'(x^*) = \lim_{h \to 0_+} \frac{f(x^* + h) - f(x^*)}{h} \leqq 0$$

und aus der zweiten

$$f'(x^*) = \lim_{h \to 0_-} \frac{f(x^* + h) - f(x^*)}{h} \geqq 0,$$

was zusammen $f'(x^*) = 0$ ergibt.
Fall 2: Es sei $f(x^*) = 0$. Ist auch der minimale Wert von f auf $[a, b]$ gleich Null, so ist $f(x) = 0$ für alle $x \in [a, b]$ und damit auch $f'(x) = 0$ für alle $x \in [a, b]$. Andernfalls wenden wir auf $-f$ den Fall 1 an. □
Der Beweis des Satzes ändert sich nicht, wenn wir die gegebene Funktion um eine Konstante abändern. Damit können wir den Satz auch so aussprechen:
Wenn eine differenzierbare Funktion in zwei Punkten den gleichen Funktionswert hat, so hat ihre Ableitung zwischen diesen Punkten eine Nullstelle.
In dieser Form werden wir den Satz auch anwenden.

Satz 96. (Mittelwertsatz der Differentialrechnung)
Ist die gegebene Funktion f im abgeschlossenen Intervall $[a, b]$ differenzierbar, so existiert zu je zwei Punkten $\alpha, \beta \in [a, b], \alpha < \beta$ ein $\xi \in (\alpha, \beta)$ mit

$$f'(\xi) = \frac{f(\beta) - f(\alpha)}{\beta - \alpha}.$$

Beweis. Wir verwenden die Hilfsfunktion

$$g(x) = f(x) - \left(f(\alpha) + \frac{f(\beta) - f(\alpha)}{\beta - \alpha}(x - \alpha) \right).$$

Es ist $g(\alpha) = 0 = g(\beta)$; nach dem Satz von Rolle existiert ein $\xi \in (\alpha, \beta)$ mit

$$0 = g'(\xi) = f'(\xi) - \frac{f(\beta) - f(\alpha)}{\beta - \alpha},$$

was gerade die Behauptung des Satzes darstellt. □
Der Mittelwertsatz wird oft in der folgenden Fassung angewendet: Ist die Funktion f im Intervall $[x - h, x + h]$ differenzierbar, so existiert eine Zahl $\varrho, \varrho \in [0, 1]$ mit

$$f(x + h) = f(x) + h \cdot f'(x + \varrho h).$$

Setzt man $h = y - x$, so kann man dies auch in der Form

$$f'(x + \varrho(y - x)) = \frac{f(y) - f(x)}{y - x}$$

schreiben, was anschaulich im Sinne der Schulmathematik bedeutet: Der Tangentenanstieg im Punkte $x+\varrho(y-x)$ an den Graphen der Funktion f ist gleich dem Anstieg der Sekante, die durch die Punkte x und y bestimmt ist.

Satz 97 (Verallgemeinerter Mittelwertsatz). *Sind die Funktionen f und g im Intervall (a, b) differenzierbar, auf dem abgeschlossenen Intervall $[a, b]$ stetig und gilt $g'(x) \neq 0$ für alle $x \in (a, b)$, so existiert ein $\xi \in (a, b)$ mit*

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}.$$

Beweis. Zunächst erkennen wir, daß die Funktion g in den Endpunkten a, b des Intervalls verschiedene Werte annehmen muß. Wären diese Werte gleich, so hätte nach dem Satz von Rolle die Ableitung im Intervall eine Nullstelle, was aber nach Voraussetzung ausgeschlossen ist. Wir nehmen die Hilfsfunktion

$$\varphi(x) = f(x) - \lambda g(x)$$

und wählen den Parameter λ so, daß $\varphi(a) = \varphi(b)$ ausfällt; dies führt uns zu

$$\lambda = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

Nach dem Satz von Rolle existiert dann ein $\xi \in (a, b)$ mit der Eigenschaft

$$0 = \varphi'(\xi) = f'(\xi) - \lambda g'(\xi) = f'(\xi) - \frac{f(b) - f(a)}{g(b) - g(a)} g'(\xi),$$

was nach Umstellung mit der Behauptung des Satzes übereinstimmt. □
Wir wissen bereits, daß die Ableitung einer Funktion nicht notwendigerweise stetig sein muß; wohl aber hat sie die Zwischenwerteigenschaft.

Satz 98 (Zwischenwertsatz). *Die Ableitung einer im abgeschlossenen Intervall $[a,b]$ differenzierbaren Funktion f nimmt jeden Wert zwischen $f'(a)$ und $f'(b)$ im Intervall an.*

Beweis. Für diesen Satz verwenden wir die beiden Hilfsfunktionen

$$\varphi(x) = \begin{cases} \dfrac{f(x)-f(a)}{x-a} & a < x \leqq b \\ f'(a) & x = a, \end{cases}$$

$$\psi(x) = \begin{cases} \dfrac{f(b)-f(x)}{x-a} & a \leqq x < b \\ f'(b) & x = b. \end{cases}$$

Beide Funktionen sind offenbar stetig auf dem Intervall $[a,b]$ und haben daher die Zwischenwerteigenschaft. Es sei nun α ein beliebiger, zwischen $f'(a)$ und $f'(b)$ gelegener Wert. Wegen

$$\varphi(a) = f'(a), \quad \varphi(b) = \psi(a), \quad \psi(b) = f'(b)$$

liegt der Wert α zwischen $\varphi(a)$ und $\varphi(b)$ oder zwischen $\psi(a)$ und $\psi(b)$. Es möge etwa der erste Fall eintreten:

$$\alpha = \varphi(x^*) = \frac{f(x^*)-f(a)}{x^*-a}, \quad a < x^* \leqq b.$$

Nach dem Mittelwertsatz existiert dann ein $\xi \in (a, x^*)$ mit

$$f'(\xi) = \frac{f(x^*)-f(a)}{x^*-a} = \alpha. \quad \square$$

Es sei Φ die Abbildung, die jeder auf dem Intervall $[a,b]$ differenzierbaren Funktion ihre Ableitung zuordnet. Dann ist Φ eine lineare Abbildung vom Vektorraum $C^1(a,b)$ in den Vektorraum aller auf $[a,b]$ definierten Funktionen mit Zwischenwerteigenschaft. Allgemein nennt man eine Abbildung eines Funktionenraumes in einen Funktionenraum **Operator**. So ist Φ ein linearer Operator vom Raum aller auf dem Intervall $[a,b]$ differenzierbaren Funktionen in den Raum aller auf $[a,b]$ definierten Funktionen, die die Zwischenwerteigenschaft besitzen.

Satz 99. *Wenn die Ableitung einer auf einem Intervall $[a,b]$ differenzierbaren Funktion f verschwindet, d. h.*

$$f'(x) = 0 \quad \forall x \in [a,b],$$

dann ist die Funktion konstant auf dem Intervall, d. h. es existiert eine Zahl c mit $f(x) = c$ für alle $x \in [a,b]$.

Beweis. Aus dem Mittelwertsatz erhalten wir unter den gemachten Voraussetzungen, daß

$$f(x+h) = f(x) + hf'(x+\varrho h) = f(x) \quad \forall x, x+h \in [a,b]$$

gilt, was bedeutet, daß die Funktion f konstant ist. $\square$

Mit diesem Satz folgt z. B. leicht die Beziehung $\sin^2 x + \cos^2 x = 1$. Für die Funktion

$$f(x) = \sin^2 x + \cos^2 x$$

gilt offenbar $f'(x) = 0$ für alle x; nach dem Satz folgt daraus, daß es eine Zahl c gibt mit $f(x) = c$ für alle x; wegen $c = f(0) = \sin^2 0 + \cos^2 0 = 1$ ist schon alles bewiesen.

Satz 100 (Monotoniesatz). *Ist die Ableitung einer auf einem Intervall $[a,b]$ differenzierbaren Funktion stets ungleich 0, so ist sie dort streng monoton.*

Beweis. Wir überlegen uns zunächst, daß wegen $f'(x) \neq 0$ auf $[a,b]$ die Ableitung entweder stets positiv oder stets negativ sein muß. Wäre $f'(x) < 0, f'(y) > 0$, so hätte die Ableitung nach dem Zwischenwertsatz eine Nullstelle, was aber der Voraussetzung widerspricht. Es sei etwa $f'(x) > 0$ für alle $x \in [a,b]$; mit dem Mittelwertsatz folgt daraus für $h > 0$:

$$f(x+h) - f(x) = h \cdot f'(x+\varrho h) > 0,$$

was bedeutet, daß die Funktion streng monoton wächst. $\square$

Als Übung kann man sich überlegen, ob auch die Umkehrung dieses Satzes gilt.

4.6.3. Taylor-Entwicklung

Ein wichtiges Problem für die Analysis ist es, für mehr oder weniger komplizierte Funktionen geeignete Näherungsformeln zu entwickeln. Die mittels dieser Näherungsformeln berechneten Werte sollen dann anstelle der Funktionswerte verwendet werden. Dieses Problem hat zwei Aspekte: Zum einen sollten die Näherungswerte „einfacher“ zu berechnen sein, was etwa durch eine geringere Anzahl von Rechenoperationen gemessen werden kann. Zum anderen sollten nur solche Rechenoperationen verwendet werden, die sich auch auf einem Rechner ausführen lassen. Für Funktionen, von denen man weiß, daß sie mehrfach differenzierbar sind und von denen die mehrfachen Ableitungen in einem vorgegebenen Punkt vorliegen, kann man leicht Näherungsformeln aufstellen. Die Annäherung einer Funktion durch ihre Ableitungswerte in einem Punkte gelingt exakt bei Polynomen n-ten Grades. Wenn wir nämlich ein Polynom n-ten Grades

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

wiederholt ableiten, erhalten wir

$$\begin{aligned}
P'(x) &= n \cdot a_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + 2a_2 x + a_1,\\
P''(x) &= n(n-1)a_n x^{n-2} + \cdots + 2 \cdot 1 \cdot a_2\\
&\vdots\\
P^{(n)}(x) &= n(n-1)(n-2)\cdots 2 \cdot 1 \cdot a_n = n!a_n,\\
P^{(k)}(x) &\equiv 0, \quad k > n.
\end{aligned}$$

Aus

$$P(0) = a_0 \quad P'(0) = 1 \cdot a_1, \quad P''(0) = 2!a_2, \ldots, P^{(n)}(0) = n!a_n$$

folgt mit $P^{(0)}(x) = P(x)$

$$a_j = \frac{P^{(j)}(0)}{j!}, \quad j = 0, 1, \ldots, n$$

und wir können das Polynom P formal wie folgt darstellen:

$$P(x) = P(0) + \frac{P'(0)}{1!}x + \frac{P''(0)}{2!}x^2 + \cdots + \frac{P^{(n)}(0)}{n!}x^n.$$

Analog folgt in einem beliebigen Punkt a:

$$P(x) = P(a) + \frac{P'(a)}{1!}(x-a) + \frac{P''(a)}{2!}(x-a)^2 + \cdots + \frac{P^{(n)}(a)}{n!}(x-a)^n.$$

Es sei die Funktion f in einer Umgebung eines Punktes a mindestens $(n+1)$-mal stetig differenzierbar und

$$P_n(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

Nach der obigen Darstellung eines Polynoms gilt dann

$$P^{(j)}(a) = f^{(j)}(a), \quad j = 0, 1, \ldots, n.$$

Das Polynom P_n kann als Näherungspolynom für die Funktion f an der Stelle a genommen werden. Die Güte der Näherung wird durch das **Restglied**

$$R_{n+1} = f(x) - P_n(x)$$

bestimmt, so daß

$$f(x) = P_n(x) + R_{n+1}(a, x)$$

gilt. Für das Restglied kann man verschiedene Darstellungen wählen. Das **Restglied nach Lagrange** erhält man aus dem Mittelwertsatz:

$$R_{n+1}(a, x) = \frac{f^{(n+1)}(a + \varrho(x-a))}{(n+1)!}(x-a)^{n+1}.$$

Setzt man $x - a = h$, so erhält man daraus

$$f(a+h) = f(a) + \frac{f'(a)}{1!}h + \frac{f''(a)}{2!}h^2 + \cdots + \frac{f^{(n)}(a)}{n!}h^n + \frac{f^{(n+1)}(a+\varrho h)}{(n+1)!}h^{n+1}.$$

Diese Formel nennt man **Taylor-Entwicklung** der Funktion f an der Stelle a. Das **Restglied nach Cauchy** lautet

$$R_{n+1}(a,x) = \frac{f^{(n+1)}(a+\varrho(x-a))}{n!}(x-a)^{n+1}(1-\varrho)^n.$$

Beispiel. Wir betrachten für $x > -1$ und reelles α die Funktion

$$f(x) = (1+x)^\alpha$$

und erhalten als n-te Ableitung:

$$f^{(n)}(x) = \alpha(\alpha-1)\cdots(\alpha-n+1)(1+x)^{\alpha-n} = n!\binom{\alpha}{n}(1+x)^{\alpha-n}.$$

Somit gilt die folgende Darstellung der Funktion in einer Umgebung von $x = 0$:

$$(1+x)^\alpha = 1 + \binom{\alpha}{1}x + \cdots + \binom{\alpha}{n}x^n + R_{n+1}(x).$$

Wie groß die Umgebung des Nullpunktes gewählt werden darf, zeigt eine genauere Untersuchung des Restgliedes nach Cauchy

$$R_{n+1}(x) = \binom{\alpha}{n+1}(n+1)x^{n+1}(1+\varrho x)^{\alpha-n-1}(1-\varrho)^n.$$

Die ersten drei Faktoren fassen wir zusammen

$$a_n = \binom{\alpha}{n}nx^n$$

und wenden das Quotientenkriterium an:

$$\left|\frac{a_{n+1}}{a_n}\right| = |x|\left|\frac{\alpha-n}{n+1}\right|\frac{n+1}{n} = |x|\left|\frac{\alpha-n}{n}\right|.$$

Dies zeigt uns, daß das Quotientenkriterium erfüllt ist, falls $|x| < 1$ gilt. Für den Restfaktor im Restglied erhalten wir wegen $\varrho \in (0,1)$ für $|x| < 1$:

$$\begin{aligned}(1+\varrho x)^{\alpha-n-1}(1-\varrho)^n &= \left(\frac{1-\varrho}{1+\varrho x}\right)^n(1+\varrho x)^{\alpha-1} < (1+\varrho x)^{\alpha-1} \\ &\leqq \begin{cases} 2^{\alpha-1} & \text{für} \quad \alpha \geqq 1 \\ (1-|x|)^{\alpha-1} & \text{für} \alpha < 1 \end{cases}.\end{aligned}$$

Damit haben wir gezeigt, daß die Reihe

$$\sum_{n=0}^{\infty}\binom{\alpha}{n}x^n$$

für alle x mit $|x| < 1$ absolut konvergiert und dort mit der Funktion $(1+x)^\alpha$ übereinstimmt:

$$(1+x)^\alpha = \sum_{n=0}^{\infty}\binom{\alpha}{n}x^n, \quad |x| < 1.$$

Nehmen wir nun den Spezialfall $\alpha = \frac{1}{2}$. Dann liefert die abgeleitete Formel für $|x| < 1$:

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{2\cdot 4}x^2 + \frac{1\cdot 3}{2\cdot 4\cdot 6}x^3 \mp \cdots + (-1)^{n-1}\frac{3\cdot\ldots\cdot(2n-3)}{2\cdot 4\cdot\ldots\cdot(2n)}x^n + \cdots.$$

Insbesondere wird die Funktion $\sqrt{1+x}$ für kleine $|x|$ durch die lineare Funktion $1 + \frac{1}{2}x$ oder durch die quadratische Funktion $1 + \frac{1}{2}x - \frac{1}{8}x^2$ angenähert.

4.6.4. Extremwerte

Wir sagen, daß eine auf einer gegebenen Menge X definierte Funktion im Punkte $a \in X$ ein **lokales Maximum** hat, wenn es eine ε-Umgebung $U_\varepsilon(a)$ von a gibt mit $f(x) \leqq f(a)$ für alle $x \in X \cap U_\varepsilon(a)$ gilt. Dabei sprechen wir im Falle $X \subseteqq U_\varepsilon(a)$ von einem **Maximum** schlechthin. Gilt die Gleichheit nur für $x = a$, so sprechen wir von einem **strengen** (lokalen) Maximum. Die Funktion f hat in $a \in X$ genau dann ein **lokales Minimum**, wenn $-f$ in a ein lokales Maximum hat. Als Oberbegriff für Maximum und Minimum verwendet man den Begriff **Extremum** bzw. **Extremwert**. Schließlich sprechen wir bei einer in a differenzierbaren Funktion f von einem **Wendepunkt**, wenn ein $\varepsilon > 0$ existiert mit

$$\text{entweder } \frac{f(x)-f(a)}{x-a} > f'(a) \text{ oder } \frac{f(x)-f(a)}{x-a} < f'(a) \quad \forall x: \ |x-a| < \varepsilon.$$

Für das Vorhandensein von Extrema und Wendepunkten kann man bei differenzierbaren Funktionen Bedingungen angeben.

Satz 101. *Hat die in X differenzierbare Funktion f in $a \in \operatorname{int}(X)$ ein lokales Extremum, so gilt $f'(a) = 0$.*

Beweis. Wir brauchen diese Bedingung nur für ein Minimum zu beweisen. Es sei

$$f(x) \geqq f(a) \quad \forall x \in U_\varepsilon(a).$$

Für kleine positive h gilt dann

$$\frac{f(a+h)-f(a)}{h} \geqq 0$$

und für kleine negative h

$$\frac{f(a+h)-f(a)}{h} \leqq 0,$$

woraus wir beim Grenzübergang im ersten Falle $f'(a) \geqq 0$ und im zweiten Falle $f'(a) \leqq 0$ erhalten, was zusammen die Behauptung liefert. □

Satz 102. *Wenn ein $\varepsilon > 0$ existiert mit*

$$f'(x) \cdot f'(y) < 0 \quad \forall x, y: \ a - \varepsilon \leqq x < a < y \leqq a + \varepsilon,$$

hat die Funktion f in a ein strenges lokales Extremum.

Beweis. Nach dem Mittelwertsatz gilt

$$f(a+h) = f(a) + hf'(a+\varrho h), \quad 0 < \varrho < 1, \quad |h| < \varepsilon.$$

Wegen der Voraussetzung haben die Funktionswerte $f'(x)$ und $f'(y)$ für x links von a und y rechts von a verschiedenes Vorzeichen; also wechselt $f'(a+\varrho h)$ mit h das Vorzeichen, d. h. $hf'(a+\varrho h)$ hat stets gleiches Vorzeichen. Ist das Produkt negativ, so folgt

$$f(a+h) < f(a),$$

ist es positiv, so erhalten wir

$$f(a+h) > f(a);$$

im ersten Falle liegt also ein strenges Maximum und im zweiten Falle ein strenges Minimum vor. □

Satz 103. *Hat die Ableitung f' einer Funktion f in $a \in X$ ein strenges lokales Extremum, so hat die Funktion selbst im Punkte a einen Wendepunkt.*

Beweis. Nehmen wir an, daß die Funktion f' in a ein strenges Maximum hat. Dann folgt mit dem Mittelwertsatz, daß für alle hinreichend kleinen $|h|$

$$\frac{f(a+h)-f(a)}{h} = f'(a+\varrho h) < f'(a) \quad (0 < \varrho < 1)$$

gilt; also liegt ein Wendepunkt vor. □

Satz 104. *Die Funktion f sei auf X m-mal stetig differenzierbar ($m > 1$), in einem inneren Punkte a aus X mögen die ersten $m-1$ Ableitungen verschwinden, nicht aber die m-te Ableitung, d. h.*

$$f'(a) = 0, f''(a) = 0, \ldots, f^{(m-1)}(a) = 0, f^{(m)}(a) \neq 0.$$

Für die Funktion f liegt im Punkte a genau dann ein Wendepunkt vor, wenn m ungerade ist. Ist die Zahl m gerade, so ist a im Falle $f^{(m)}(a) > 0$ eine lokale Minimumstelle und im Falle $f^{(m)}(a) < 0$ eine lokale Maximumstelle.

Beweis. Unter den gemachten Voraussetzungen lautet die Taylor-Entwicklung von f an der Stelle a:

$$f(x) = f(a) + \frac{f^{(m)}(a - \varrho(x-a))}{m!}(x-a)^m.$$

Da die Funktion $f^{(m)}$ stetig ist und $f^{(m)}(a) \neq 0$ gilt, existiert eine Umgebung $U_\varepsilon(a)$ mit

$$f^{(m)}(x) \neq 0 \quad \forall x \in U_\varepsilon(a);$$

also hat $f^{(m)}$ in $U_\varepsilon(a)$ einheitliches Vorzeichen. Damit folgt: Die Funktion f hat in a genau dann ein lokales Extremum, wenn $f'(a) = 0$ und m eine gerade Zahl ist; im Falle $f^{(m)}(a) > 0$ liegt ein lokales Minimum, im Falle $f^{(m)}(a) < 0$ ein lokales Maximum vor. Die Funktion f hat in a genau dann einen Wendepunkt, wenn $m > 1$ und ungerade ist. □

4.6.5. Grenzwertbestimmung

Mit Hilfe der Differentialrechnung lassen sich Grenzwerte von Quotienten differenzierbarer Funktionen berechnen. Wir geben hier zwei Möglichkeiten an; andere lassen sich auf diese zurückführen.

Satz 105 (Regel von de l'Hospital (1691)). *Falls*

$$\lim_{x \to a} f(x) = \lim_{x \to a} \varphi(x) = 0$$

gilt und der eigentliche oder uneigentliche Grenzwert

$$\lim_{x \to a} \frac{f'(x)}{\varphi'(x)}$$

existiert, dann gilt

$$\lim_{x \to a} \frac{f(x)}{\varphi(x)} = \lim_{x \to a} \frac{f'(x)}{\varphi'(x)}.$$

Beweis. Wir definieren zwei Hilfsfunktionen:

$$F(x) = \begin{cases} f(x) & x \neq a \\ 0 & x = a \end{cases}, \quad \Phi(x) = \begin{cases} \varphi(x) & x \neq a \\ 0 & x = a \end{cases}$$

und erhalten mit dem verallgemeinerten Mittelwertsatz

$$\frac{f(x)}{\varphi(x)} = \frac{F(x) - F(a)}{\Phi(x) - \Phi(a)} = \frac{F'(\xi)}{\Phi'(\xi)}, \quad a < \xi < x \text{ bzw. } x < \xi < a.$$

Durch Grenzübergang $x \to a$ folgt die Behauptung. □
Ganz ähnlich beweist man auch den nächsten Satz.

Satz 106. *Falls*

$$\lim_{x \to a} f(x) = \lim_{x \to a} \varphi(x) = \infty$$

gilt und der eigentliche oder uneigentliche Grenzwert

$$\lim_{x \to a} \frac{f'(x)}{\varphi'(x)} \quad (\varphi'(x) \neq 0)$$

existiert, dann gilt

$$\lim_{x \to a} \frac{f(x)}{\varphi(x)} = \lim_{x \to a} \frac{f'(x)}{\varphi'(x)}.$$

Beispiele. So folgt etwa

$$\lim_{x\to 0}\frac{\sin x}{x}=\lim_{x\to 0}\frac{\cos x}{1}=1.$$

Man darf die Regeln natürlich auch mehrfach hintereinander anwenden:

$$\begin{aligned}\lim_{x\to 0_+}\frac{\ln x}{\cot x}&=\lim_{x\to 0_+}\frac{\frac{1}{x}}{\frac{-1}{\sin^2 x}}=\lim_{x\to 0_+}-\frac{\sin^2 x}{x}\\&=\lim_{x\to 0_+}-\frac{2\sin x\cos x}{1}=0.\end{aligned}$$

Durch die Regel von de l'Hospital wird einem unbestimmten Ausdruck der Form $\frac{0}{0}$ bzw. $\frac{\infty}{\infty}$ mittels Ableiten in Zähler und Nenner ein Wert zugeordnet. Andere unbestimmte Ausdrücke sind $0\cdot\infty,\infty-\infty,0^0,1^\infty$, die beim Produkt, bei der Differenz bzw. beim Potenzieren auftreten können. Den Fall $f(x)\cdot\varphi(x)$ überführt man in einen der beiden obigen, indem man

$$\varphi(x)=\frac{1}{h(x)}$$

setzt. Mittels der Transformation

$$f(x)-\varphi(x)=\frac{\frac{1}{\varphi(x)}-\frac{1}{f(x)}}{\frac{1}{f(x)\cdot\varphi(x)}}$$

wird der Fall $\infty-\infty$ in den Fall $\frac{0}{0}$ überführt. Im Fall $f(x)>0$ kann man wegen

$$f(x)^{\varphi(x)}=e^{\varphi(x)\cdot\ln f(x)}$$

den Grenzwert der Funktion $\varphi(x)\cdot\ln f(x)$ berechnen. Zu diesem Fall sei ein Beispiel gegeben:

$$\begin{aligned}\lim_{x\to 0_+}\left(\frac{x}{x^2+1}\right)^x&=\exp\left(\lim_{x\to 0_+}x\cdot\ln\frac{x}{x^2+1}\right)=\exp\left(\lim_{x\to 0_+}\frac{\frac{1-x^2}{1+x^2}}{-\frac{1}{x^2}}\right)\\&=\exp\left(\lim_{x\to 0_+}-x^2\frac{1-x^2}{1+x^2}\right)=1.\end{aligned}$$

4.6.6. Potenzreihen

Eine unendliche Reihe der Form

$$\sum_{n=0}^{\infty}c_n(x-a)^n$$

heißt **Potenzreihe** an der Stelle a. Eine solche Reihe kann für gewisse Werte der Variablen x konvergieren, für andere nicht. Es sei X die Menge aller $x\in\mathbb{R}$, für die die Reihe konvergiert. Die Menge X ist offenbar nicht leer, denn es ist $a\in X$. Eine Funktion f, deren Funktionswerte in einer Umgebung von a durch eine Potenzreihe berechnet werden können, nennen wir **analytisch** in a; falls die Funktion diese Eigenschaft in jedem Punkte ihres Definitionsbereiches hat, heißt sie schlechthin analytisch. So sind $e^x,\sin x,\cos x$ analytische Funktionen. Die Funktionswerte von analytischen Funktionen können durch die Glieder der Partialsummenfolge beliebig genau angehnähert werden und sind daher durch elementare Operationen beliebig genau berechenbar. Dies ist der wesentliche Unterschied zur Taylor-Entwicklung: Bei der Taylor-Entwicklung bleibt ein Restglied, das wesentlich sein kann. Wir stellen die wichtigsten Eigenschaften zusammen.

Satz 107 (Konvergenzkreis). *Zu jeder Potenzreihe*

$$\sum_{n=0}^{\infty}c_n(x-a)^n$$

gibt es genau ein R, $0\leqq R\leqq\infty$ so, daß für alle x mit $|x-a|<R$ die Potenzreihe absolut konvergiert und für alle x mit $|x-a|>R$ divergiert.

Beweis. Es sei

$$L=\limsup_{n\to\infty}\sqrt[n]{|c_n|},\qquad R=\begin{cases}0 & L=\infty,\\ \frac{1}{L} & 0<L<\infty,\\ \infty & L=0\end{cases}.$$

Für $x = a$ konvergiert die Reihe; für $x \neq a$ konvergiert die Reihe nach dem Wurzelkriterium, wenn

$$\limsup \sqrt[n]{|c_n(x-a)^n|} = |x-a| \cdot L < 1$$

ausfällt und divergiert, falls

$$\limsup \sqrt[n]{|c_n(x-a)^n|} = |x-a| \cdot L > 1$$

gilt. Im Falle $R = 0$, d. h. $L = \infty$ divergiert die Reihe für $x \neq a$. Bei $R = \infty$, d. h. $L = 0$ konvergiert die Reihe absolut für alle x. Ist nun $0 < R < \infty$, so konvergiert die Reihe absolut, falls $|x-a| < R$; im Falle $|x-a| > R$ divergiert die Reihe. □
Die Menge $\{\, x \mid |x-a| < R \,\}$ heißt **Konvergenzkreis** der Potenzreihe; a ist der Mittelpunkt und R sein Radius.

Satz 108. *Eine Potenzreihe konvergiert gleichmäßig in jedem abgeschlossenen, beschränkten Bereich X, der vollständig im Konvergenzkreis liegt.*

Beweis. Es sei X eine beschränkte, abgeschlossene Menge, die ganz im Konvergenzkreis der Potenzreihe

$$\sum_{n=0}^{\infty} c_n(x-a)^n$$

liegen möge. Die Betragsfunktion nimmt als stetige Funktion ihren maximalen Wert auf X an:

$$r = \max\{\, |x-a| \mid x \in X \,\}.$$

Offenbar gilt $r < R$; also konvergiert die Reihe $\sum_{n=0}^{\infty} |c_n| r^n$. Wegen

$$|c_n(x-a)^n| \leqq |c_n| r^n \quad \forall x \in X$$

konvergiert die vorgegebene Reihe gleichmäßig. □

Satz 109. *Hat die Potenzreihe $\sum_{n=0}^{\infty} c_n(x-a)^n$ den Konvergenzkreis-Radius R, so stellt sie in $|x-a| < R$ eine Funktion dar, die dort beliebig oft differenzierbar ist und deren Ableitungen durch gliedweises Differenzieren gewonnen werden können. Aus*

$$f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n, \qquad |x-a| < r$$

folgt für die l-te Ableitung

$$f^{(l)}(x) = l! \sum_{n=l}^{\infty} \binom{n}{l} c_n(x-a)^{n-l}, \quad l = 0, 1, 2, \ldots.$$

Beweis. Wir brauchen offenbar nur die Formel für die l-te Ableitung zu beweisen. Dies soll durch vollständige Induktion über l geschehen. Für $l = 0$ ist nichts zu beweisen. Die Induktionsvoraussetzung lautet

$$f^{(l)}(x) = l! c_l + l! \sum_{n=l+1}^{\infty} \binom{n}{l} c_n(x-a)^{n-l}.$$

Für die $(l+1)$-te Ableitung erhalten wir daraus:

$$\begin{aligned} f^{(l+1)}(x) &= l! \sum_{n=l+1}^{\infty} (n-l) \binom{n}{l} c_n(x-a)^{n-l-l} \\ &= (l+1)! \sum_{n=l+1}^{\infty} \binom{n}{l+1} c_n(x-a)^{n-(l+l)}. \end{aligned}$$ □

Für $x = a$ folgt $f^{(l)}(a) = l! c_l$. Wenn wir c_l in die Potenzreihe einsetzen, erhalten wir die **Taylorreihe** einer analytischen Funktion:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n.$$

Satz 110. (Algebra der analytischen Funktionen)

- *Sind f und φ analytisch in a:*

$$f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n, \qquad \varphi(x) = \sum_{n=0}^{\infty} d_n (x-a)^n,$$

so sind auch $\lambda f, f+\varphi, f\cdot\varphi$ in a analytisch und es gilt

$$f(x)\cdot\varphi(x) = \sum_{l=0}^{\infty}\sum_{n=0}^{l} c_n d_{l-n} (x-a)^n.$$

- *Ist φ in a analytisch und f in $\varphi(a)$ analytisch, so ist auch die Funktion $f\circ\varphi$ in a analytisch.*
- *Ist die Funktion f in a analytisch und $f(a)\neq 0$, so ist auch die Funktion $\frac{1}{f}$ in a analytisch.*
- *Der Quotient $\frac{f}{\varphi}$ zweier in a analytischer Funktionen f und φ mit $\varphi(a)\neq 0$ ist in a analytisch.*

Beweis. Wir beweisen nur die vorletzte Eigenschaft. Es sei

$$f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n.$$

Wegen $f(a)\neq 0$, muß $c_0 \neq 0$ sein. Wir setzen

$$\varphi(x) = -\frac{1}{c_0}\sum_{n=1}^{\infty} c_n (x-a)^n, \qquad h(y) = \frac{1}{1-y} = \sum_{n=0}^{\infty} y^n \quad (|y|<1).$$

Die so definierte Funktion φ ist in a analytisch, die Funktion h ist in 0 analytisch und

$$\frac{1}{f(x)} = \frac{1}{c_0}\frac{1}{1-\varphi(x)} = \frac{1}{c_0} h(\varphi(x)).$$

Damit haben wir $\frac{1}{f}$ als Verkettung zweier analytischer Funktionen dargestellt. □

4.7. Integralrechnung

4.7.1. Das bestimmte Integral

Die Integralrechnung geht von der klassischen Aufgabenstellung aus, daß man bei einer auf einem Intervall $[a,b]$ gegebenen Funktion f den Flächeninhalt der ebenen Menge

$$I_0^f = \left\{ (x,y) \mid a \leqq x \leqq b, \quad 0 \leqq y \leqq f(x) \right\}$$

berechnen möchte. Diese Aufgabe führt unmittelbar zum Riemannschen Integralbegriff.
Es sei eine auf einem Intervall $I=[a,b]$ definierte Funktion f gegeben. Einem beliebigen Teilintervall $I' \subseteqq I$ ordnen wir die obere und die untere Grenze der Funktionswerte von f auf diesem Teilintervall zu:

$$\underline{f}(I') = \inf\{ f(x) \mid x\in I' \}, \quad \overline{f}(I') = \sup\{ f(x) \mid x\in I' \}.$$

Wir benutzen ferner beliebige, endliche Zerlegungen des Intervalls I:

$$\mathcal{Z} = \{ I_1, I_2, \ldots, I_m \}, \quad I_j = (x_{j-1}, x_j), \quad (j=1,\ldots,m),$$

$$a = x_0 < x_1 < \ldots < x_m = b.$$

Einem Paar $(f,\mathcal{Z})$ ordnen wir die **Untersumme**

$$\underline{S}(f,\mathcal{Z}) = \sum_{j=1}^{m} \underline{f}(I_j)(x_j - x_{j-1})$$

und die **Obersumme**

$$\overline{S}(f,\mathcal{Z}) = \sum_{j=1}^{m} \overline{f}(I_j)(x_j - x_{j-1})$$

zu. Geometrisch ist die Untersumme gerade die Summe der Flächeninhalte aller „einbeschriebenen“ Rechtecke mit den Seitenlängen $\underline{f}(I_j)$ und $x_j - x_{j-1}$; analog läßt sich die Obersumme deuten. Nach Definition folgt

$$\overline{S}(f, \mathcal{Z}) - \underline{S}(f, \mathcal{Z}) \geqq 0.$$

Wegen

$$\underline{f}(I_j) \geqq \underline{f}(I), \quad \overline{f}(I_j) \leqq \overline{f}(I), \quad j = 1, \ldots, m$$

ist

$$\underline{S}(f, \mathcal{Z}) = \sum_{j=1}^{m} \underline{f}(I_j)(x_j - x_{j-1}) \geqq \underline{f}(I) \sum_{j=1}^{m} (x_j - x_{j-1}) = \underline{f}(I)(b - a)$$

und entsprechend für die Obersumme

$$\overline{S}(f, \mathcal{Z}) \leqq \overline{f}(I)(b - a),$$

was zusammen

$$(b - a)\underline{f}(I) \leqq \underline{S}(f, \mathcal{Z}) \leqq \overline{S}(f, \mathcal{Z}) \leqq \overline{f}(I)(b - a)$$

für jede Zerlegung $\mathcal{Z}$ des Intervalls $[a, b]$ liefert. Insbesondere sind die Obersummen nach unten und die Untersummen nach oben beschränkt.
Die untere Grenze $\overline{J}(f, a, b)$ aller Obersummen nennt man **Oberintegral** der Funktion f über dem Intervall $[a, b]$:

$$\overline{J}(f, a, b) = \inf_{\mathcal{Z}} \overline{S}(f, \mathcal{Z}).$$

Entsprechend heißt die obere Grenze $\underline{J}(f, a, b)$ aller Untersummen **Unterintegral** der Funktion f über dem Intervall $[a, b]$:

$$\underline{J}(f, a, b) = \sup_{\mathcal{Z}} \underline{S}(f, \mathcal{Z}).$$

Für alle Zerlegungen $\mathcal{Z}$ gilt offenbar

$$\underline{S}(f, \mathcal{Z}) \leqq \underline{J}(f, a, b) \leqq \overline{J}(f, a, b) \leqq \overline{S}(f, \mathcal{Z}).$$

Diese Konstruktion kann man für jede Funktion durchführen. Nun müssen Ober- und Unterintegral durchaus nicht übereinstimmen. Bei der Funktion

$$f(x) = \begin{cases} 1 & x \text{ rational} \\ 0 & x \text{ irrational} \end{cases}$$

auf dem Intervall $I = [0, 1]$ gilt offenbar $\underline{f}(I) = 0$ und $\overline{f}(I) = 1$. In jedem Teilintervall von I liegen eine rationale und eine irrationale Zahl; also ist stets $\underline{f}(I') = 0$ und $\overline{f}(I') = 1$ für alle $I' \subseteqq I$, was $\underline{J}(f, I) = 0$ und $\overline{J}(f, I)$=1 liefert.
Eine auf dem Intervall $[a, b]$ beschränkte Funktion f heißt **integrierbar** (nach B. Riemann), wenn Ober- und Unterintegral übereinstimmen; den gemeinsamen Wert nennt man **bestimmtes Integral** der Funktion f über dem Intervall $[a, b]$ und schreibt es nach Leibniz (1675) in der Form

$$\int_a^b f(x)dx,$$

wobei a und b als **Integrationsgrenzen** bezeichnet werden.
Eine notwendige und hinreichende Bedingung für die Integrierbarkeit einer Funktion liefert der folgende Satz.

Satz 111 (Riemannsches Integrabilitätskriterium). *Eine Funktion f ist genau dann über dem Intervall $[a, b]$ integriebar, wenn es zu jedem $\varepsilon > 0$ eine Zerlegung $\mathcal{Z}$ gibt mit*

$$\overline{S}(f, \mathcal{Z}) - \underline{S}(f, \mathcal{Z}) < \varepsilon.$$

Beweis. Die Hinlänglichkeit der Bedingung ist offensichtlich. Wir haben nur zu zeigen, daß die Bedingung auch notwendig ist. Es sei also f über $[a,b]$ integrierbar und $\varepsilon > 0$ beliebig vorgegeben. Dann existieren zwei Zerlegungen $\mathcal{Z}', \mathcal{Z}''$ mit

$$\int_a^b f(x)dx - \underline{S}(f,\mathcal{Z}') < \frac{\varepsilon}{2}, \quad \overline{S}(f,\mathcal{Z}'') - \int_a^b f(x)dx < \frac{\varepsilon}{2}.$$

An dieser Stelle sei eine Zwischenbemerkung eingeschoben. Wir nennen eine Zerlegung $\tilde{\mathcal{Z}}$ **Verfeinerung** der Zerlegung $\mathcal{Z}$, wenn jedes Intervall aus $\tilde{\mathcal{Z}}$ in einem Intervall aus $\mathcal{Z}$ liegt. Gilt etwa $\tilde{I}' \subseteqq I'$, so ist

$$\overline{f}(\tilde{I}') \leqq \overline{f}(I'), \quad \underline{f}(\tilde{I}') \geqq \underline{f}(I').$$

Bei einer Verfeinerung kann sich die Obersumme höchstens verkleinern und die Untersumme höchstens vergrößern:

$$\overline{S}(f,\tilde{\mathcal{Z}}) \leqq \overline{S}(f,\mathcal{Z}), \quad \underline{S}(f,\tilde{\mathcal{Z}}) \geqq \underline{S}(f,\mathcal{Z}).$$

Kehren wir sogleich zum Beweis zurück und bilden eine Überlagerung $\mathcal{Z}$ der beiden Zerlegungen $\mathcal{Z}'$ und $\mathcal{Z}''$, d. h. eine Verfeinerung, die sowohl Verfeinerung von $\mathcal{Z}'$ als auch von $\mathcal{Z}''$ ist. Damit erhalten wir

$$\overline{S}(f,\mathcal{Z}) - \underline{S}(f,\mathcal{Z}) \leqq \overline{S}(f,\tilde{\mathcal{Z}}'') - \underline{S}(f,\tilde{\mathcal{Z}}') < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \quad \square$$

4.7.2. Eigenschaften integrierbarer Funktionen

In letzten Abschnitt haben wir bei der Einführung des bestimmten Integrals einer Funktion f über einem Intervall $[a,b]$ vorausgesetzt, daß $a < b$ gilt. Würde man die Einführung für den Fall $a > b$ wiederholen, erhielte man, daß

$$\int_a^b f(x)dx = -\int_b^a f(x)dx$$

gilt, was wir damit als gegeben annehmen wollen. Speziell ist

$$\int_a^a f(x)dx = 0.$$

Satz 112. *Jede auf einem Intervall stetige Funktion ist dort integrierbar.*

Beweis. Aus technischen Gründen beweisen wir diese Aussage nur für stetig differenzierbare Funktionen. Es sei f eine auf $[a,b]$ stetig differenzierbare Funktion. In jedem Teilintervall $I \subseteqq [a,b]$ gibt es Punkte $\underline{x}, \overline{x}$ mit $f(\underline{x}) = \underline{f}(I)$ und $f(\overline{x}) = \overline{f}(I)$; außerdem gibt es eine Zahl $M > 0$ mit $|f'(x)| \leqq M$ für alle $x \in [a,b]$. Aus dem Mittelwertsatz folgt

$$\overline{f}(I) - \underline{f}(I) = (\overline{x} - \underline{x})f'(\xi) \leqq M|\overline{x} - \underline{x}| \leqq M(b-a).$$

Damit schließen wir für jede Zerlegung $\mathcal{Z} = \{ I_1, \ldots, I_m \}$:

$$\begin{aligned} 0 \leqq \overline{S}(f,\mathcal{Z}) - \underline{S}(f,\mathcal{Z}) &= \sum_{j=1}^m (\overline{f}(I_j) - \underline{f}(I_j))(x_j - x_{j-1}) \\ &\leqq M\varrho(\mathcal{Z}) \sum_{j=1}^m (x_j - x_{j-1}) \\ &= M\varrho(\mathcal{Z})(b-a). \end{aligned}$$

Hierin bezeichnet $\varrho(\mathcal{Z})$ die Größe

$$\varrho(\mathcal{Z}) = \max_j |x_j - x_{j-1}|,$$

d. h. $\varrho(\mathcal{Z})$ ist die maximale Intervallänge der Intervalle aus $\mathcal{Z}$; die Größe nennt man **Durchmesser** der Zerlegung $\mathcal{Z}$. Ist nun $\varepsilon > 0$ beliebig vorgegeben, so existiert dazu eine Zerlegung $\mathcal{Z}$ mit

$$\varrho(\mathcal{Z}) < \frac{\varepsilon}{M(b-a)},$$

womit aus dem Riemannschen Integrabilitätskriterium folgt, daß die Funktion f integrierbar ist. $\square$

Satz 113. *Sind m, M eine untere bzw. obere Schranke einer auf $[a,b]$ integrierbaren Funktion f, so gelten die Abschätzungen:*

$$m(b-a) \leqq \int_a^b f(x)dx \leqq M(b-a).$$

Beweis. Die Aussage des Satzes folgt mit $I = [a,b]$ aus der folgenden Ungleichungskette, die für jede Zerlegung $\mathcal{Z}$ gilt:

$$\begin{aligned} m(b-a) &\leqq (b-a)\underline{f}(I) \leqq \underline{S}(f,\mathcal{Z}) \\ &\leqq \int_a^b f(x)dx \\ &\leqq \overline{S}(f,\mathcal{Z}) \leqq (b-a)\overline{f}(I) \\ &\leqq M(b-a). \quad \square \end{aligned}$$

Die untere und obere Abschätzung in diesem Satz sind dann besonders gut, wenn sie mit der unteren bzw. oberen Grenze der zu integrierenden Funktion ausgeführt werden.

Satz 114. *Die Menge aller auf einem Intervall $[a,b]$ integrierbaren Funktionen bildet einen Vektorraum über den reellen Zahlen. Das bestimmte Integral ist eine lineare Abbildung dieses Vektorraumes in die reellen Zahlen.*

Beweis. Für die Linearität haben wir zu zeigen:

$$\int_a^b (\lambda f(x) + \mu g(x))dx = \lambda \int_a^b f(x)dx + \mu \int_a^b g(x)dx \qquad \forall \lambda, \mu \in \mathbb{R}.$$

Es sei $h(x) = \lambda f(x) + \mu g(x)$; für jedes Teilintervall $I \subseteqq [a,b]$ gilt dann

$$\lambda \underline{f}(I) + \mu \underline{g}(I) \leqq \underline{h}(I) \leqq \overline{h}(I) \leqq \lambda \overline{f}(I) + \mu \overline{g}(I),$$

woraus sich für jede Zerlegung $\mathcal{Z}$

$$\lambda \underline{S}(f,\mathcal{Z}) + \mu \underline{S}(g,\mathcal{Z}) \leqq \underline{S}(h,\mathcal{Z}) \leqq \overline{S}(h,\mathcal{Z}) \leqq \lambda \overline{S}(f,\mathcal{Z}) + \mu \overline{S}(g,\mathcal{Z})$$

ergibt, und daraus für $\varrho(\mathcal{Z}) \to 0$ die Behauptung folgt. $\square$

Wir erwähnen, daß man eine lineare Abbildung eines Funktionenraumes in die reellen Zahlen **Funktional** nennt. Damit ist die bestimmte Integration ein lineares Funktional auf dem Raum aller über dem Intervall $[a,b]$ integrierbaren Funktionen. Wir haben schon andere Funktionale betrachtet. So ist z. B. die Abbildung, die einer konvergenten Folge ihren Grenzwert zuordnet, ein lineares Funktional. Wenn wir im Raum aller konvergenten Zahlenfolgen den Unterraum betrachten, der durch alle jene konvergenten Nullfolgen gebildet wird, deren zugeordnete Reihe konvergiert, so ist die Abbildung, die einer solchen Nullfolge den Wert der entsprechenden Reihe zuordnet, ein lineares Funktional.

Satz 115. *Es sei $a < c < b$; eine Funktion f ist genau dann über dem Intervall $[a,b]$ integrierbar, wenn sie über den Intervallen $[a,c]$ und $[c,b]$ integrierbar ist; außerdem gilt*

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx.$$

Beweis. Es sei $I' = [a,c]$, $I'' = [c,b]$, $\mathcal{Z}'$ eine Zerlegung von I' und $\mathcal{Z}''$ eine Zerlegung von I''. Dann ist $\mathcal{Z}' \cup \mathcal{Z}''$ eine Zerlegung von $[a,b]$ und es gilt daher

$$\underline{S}(f,\mathcal{Z}') + \underline{S}(f,\mathcal{Z}'') = \underline{S}(f,\mathcal{Z}) \leqq \underline{J}(f,a,b),$$

also

$$\underline{J}(f,a,c) + \underline{J}(f,c,b) \leqq \underline{J}(f,a,b).$$

Es sei nun $\mathcal{Z}$ eine Zerlegung von $[a,b]$, die c als Randpunkt eines Teilintervalls enthält; dann zerfällt $\mathcal{Z}$ in eine Zerlegung $\mathcal{Z}'$ von I' und eine Zerlegung $\mathcal{Z}''$ von I''. Daher haben wir

$$\underline{S}(f,\mathcal{Z}) = \underline{S}(f,\mathcal{Z}') + \underline{S}(f,\mathcal{Z}'') \leqq \underline{J}(f,a,c) + \underline{J}(f,c,b)$$

und damit

$$\underline{J}(f,a,b) \leqq \underline{J}(f,a,c) + \underline{J}(f,c,b),$$

was zusammen

$$\underline{J}(f,a,b) = \underline{J}(f,a,c) + \underline{J}(f,c,b)$$

liefert. Analog ergibt sich für die Oberintegrale

$$\overline{J}(f,a,b) = \overline{J}(f,a,c) + \overline{J}(f,c,b),$$

also zusammen

$$\overline{J}(f,a,b) - \underline{J}(f,a,b) = [\overline{J}(f,a,c) - \underline{J}(f,a,c)] + [\overline{J}(f,c,b) - \underline{J}(f,c,b)].$$

Die linke Seite ist nichtnegativ, ebenso sind rechts die beiden Summanden nichtnegativ. Somit ist die linke Seite genau dann gleich 0, wenn beide rechts stehenden Summanden gleich 0 sind. Dies ist gleichwertig zum Satz. □

Satz 116. (Monotoniesatz)
Die bestimmte Integration ist eine monotone Operation: Sind f und g integrierbar über dem Intervall $[a,b]$ und gilt

$$f(x) \leqq g(x) \qquad \forall x \in [a,b],$$

dann ist

$$\int_a^b f(x)dx \leqq \int_a^b g(x)dx.$$

Beweis. Wir setzen $h(x) = g(x) - f(x)$; die Funktion h nimmt nur nichtnegative Werte an und ist integrierbar. Nach Satz 113 folgt mit $m = 0$:

$$0 \leqq \int_a^b h(x)dx = \int_a^b g(x)dx - \int_a^b f(x)dx,$$

womit bereits alles bewiesen ist. □

Satz 117. *Für eine über dem Intervall $[a,b]$ integrierbare Funktion f gilt*

$$\left|\int_a^b f(x)dx\right| \leqq \int_a^b |f(x)|dx.$$

Beweis. Es sei $\varphi(x) = |f(x)|$. Für jedes Intervall $I \subseteqq [a,b]$ gilt dann mit $x, y \in I$:

$$|\varphi(x) - \varphi(y)| = ||f(x)| - |f(y)|| \leqq |f(x) - f(y)| \leqq \overline{f}(I) - \underline{f}(I)$$

und daher $\overline{\varphi}(I) - \underline{\varphi}(I) \leqq \overline{f}(I) - \underline{f}(I)$. Für jede Zerlegung $\mathcal{Z}$ folgt

$$\overline{S}(\varphi, \mathcal{Z}) - \underline{S}(\varphi, \mathcal{Z}) \leqq \overline{S}(f, \mathcal{Z}) - \underline{S}(f, \mathcal{Z}),$$

woraus sich die Integrierbarkeit mit dem Riemannschen Integrabilitätskriterium ergibt. Wegen

$$f(x) \leqq |f(x)|, -f(x) \leqq |f(x)|$$

liefert die Monotonie der Integration

$$\int_a^b f(x)dx \leqq \int_a^b |f(x)|dx, \qquad -\int_a^b f(x)dx \leqq \int_a^b |f(x)|dx \qquad (a \leqq b),$$

was gerade im Satz behauptet wird. □

Satz 118 (Mittelwertsatz der Integralrechnung). *Zu jeder auf $[a,b]$ stetigen Funktion f gibt es ein $\xi \in [a,b]$ mit*

$$\int_a^b f(x)dx = f(\xi)(b-a).$$

Beweis. Es sei m der minimale und M der maximale Wert von f auf dem Intervall $[a,b]$; beide Werte existieren, da f stetig ist. Wir setzen

$$\eta = \frac{1}{b-a}\int_a^b f(x)dx$$

und erhalten mit Satz 113: $m \leqq \eta \leqq M$. Da die stetige Funktion f jeden Zwischenwert annimmt, existiert ein $\xi \in [a,b]$ mit $f(\xi) = \eta$, womit der Satz bewiesen ist. □
Eine andere Formulierung der Aussage des Mittelwertsatzes lautet:

$$\int_a^{a+h} f(x)dx = h \cdot f(a+\varrho h) \text{ mit } \varrho \in [0,1].$$

Satz 119 (Verallgemeinerter Mittelwertsatz). *Es seien f, φ stetige Funktionen auf dem Intervall $[a,b]$ und $\varphi(x) \geqq 0$ für alle $x \in [a,b]$. Dann gibt es ein $\xi \in [a,b]$ mit*

$$\int_a^b f(x)\varphi(x)dx = f(\xi)\int_a^b \varphi(x)dx.$$

Beweis. Es seien m, M wie beim Mittelwertsatz; dann gilt zunächst

$$m \cdot \varphi(x) \leqq f(x)\varphi(x) \leqq M \cdot \varphi(x)$$

und wegen der Monotonie der Integration

$$m \cdot \int_a^b \varphi(x)dx \leqq \int_a^b f(x)\varphi(x)dx \leqq M \cdot \int_a^b \varphi(x)dx.$$

Im Falle

$$\int_a^b \varphi(x)dx = 0$$

ist die Behauptung klar. Andernfalls sei

$$\eta = \frac{\int_a^b f(x)\varphi(x)dx}{\int_a^b \varphi(x)dx}.$$

Es ist $m \leqq \eta \leqq M$ und mit dem Zwischenwertsatz folgt, daß es ein $\xi \in [a,b]$ gibt mit $f(\xi) = \eta$. □
Für den Hauptsatz der Differential- und Integralrechnung brauchen wir einen neuen, wichtigen Begriff. Wir nennen eine auf einem Intervall $[a,b]$ differenzierbare Funktion F **Stammfunktion** einer dort definierten Funktion f, wenn

$$F'(x) = f(x) \qquad \forall x \in [a,b].$$

Eine Funktion hat unendlich viele Stammfunktionen. Nach der Definition können sich zwei Stammfunktionen zu einer festen Funktion nur um eine Konstante unterscheiden.

Satz 120. (Hauptsatz der Differential- und Integralrechnung)
Es sei f eine auf dem Intervall $[a,b]$ stetige Funktion. Dann ist die Funktion F mit

$$F(x) = \int_a^x f(t)dt, \qquad a \leqq x \leqq b$$

eine Stammfunktion von f.

Beweis. Der Hauptsatz ist eine unmittelbare Folge des Mittelwertsatzes:

$$\begin{aligned}\frac{F(x+h)-F(x)}{h} &= \frac{1}{h}\left[\int_a^{x+h} f(t)dt - \int_a^x f(t)dt\right] \\ &= \frac{1}{h}\int_x^{x+h} f(t)dt = f(x+\varrho h) \qquad (0 \leqq \varrho \leqq 1).\end{aligned}$$

Für $h \to 0$ folgt die Behauptung. □
Nach diesem Satz ist bei stetigen Funktionen die Integration die Umkehrung der Differentiation.

Satz 121. *Ist F eine Stammfunktion einer auf $[a,b]$ stetigen Funktion f, so gilt*

$$\int_a^b f(x)dx = F(b) - F(a).$$

Beweis. Es sei F eine beliebige Stammfunktion von f. Nach dem Hauptsatz ist auch

$$\Phi(x) = \int_a^x f(t)dt, \qquad (a \leqq x \leqq b)$$

eine Stammfunktion; also können sich beide nur um eine Konstante c unterscheiden:

$$\Phi(x) = F(x) + c.$$

Speziell erhalten wir für $x = a$:

$$0 = \Phi(a) = F(a) + c,$$

also $c = -F(a)$. Damit gilt $\Phi(x) = F(x) - F(a)$, woraus

$$F(b) - F(a) = \Phi(b) = \int_a^b f(t)dt$$

folgt, was zu zeigen war. □
Oft verwendet man die Schreibweise

$$[F(x)]_a^b = F(b) - F(a) \text{ bzw. } F(x)|_a^b = F(b) - F(a).$$

Bei einer stetig differenzierbaren Funktion f ist die Funktion f eine Stammfunktion von f' und daher

$$\int_a^b f'(x)dx = [f(x)]_a^b.$$

4.7.3. Integrationsmethoden

Unter dem **unbestimmten Integral** einer auf einem Intervall I integrierbaren Funktion f versteht man die Menge aller Stammfunktionen F von f auf I; meist schreibt man dafür

$$\int f(x)dx + C$$

und nennt C Integrationskonstante. In dieser Darstellung steht der erste Summand für eine beliebig gewählte Stammfunktion von f. Die übliche Schreibweise ist nicht eindeutig. Manchmal meint man mit $\int f(x)dx$ schon die Menge aller Stammfunktionen von f:

$$\int xdx = \frac{1}{2}x^2 + C.$$

Hinzu kommt noch die Problematik, daß das Zeichen x auf beiden Seiten der Gleichung völlig anders interpretiert werden muß, damit die Gleichung als sinnvoll angesehen werden kann. Mit der Gleichung ist gemeint: Die Funktion F mit $F(x) = \frac{1}{2}x^2$ ist eine Stammfunktion der Funktion f mit $f(x) = x$. Allgemein bedeutet damit die Schreibweise

$$\int f(x)dx = F(x) + C,$$

daß F eine Stammfunktion von f ist. In anderen Fällen steht $\int f(x)dx$ für eine geeignet zu wählende Stammfunktion, z. B. in

$$u(x)v(x) = \int u'(x)v(x)dx + \int u(x)v'(x)dx.$$

Was im konkreten Falle gemeint ist, muß man aus dem Zusammenhang entnehmen. Für die Mathematik ist es leicht, eine eindeutige Notation festzulegen. Es bleibt aber sehr zweifelhaft, ob eine solche Notation von Nichtmathematikern akzeptiert und angewendet wird. Daher lebt die Mathematik schon seit Jahrhunderten

mit diesem ungelösten Konflikt.
Aus den Rechenregeln der Differentialrechnung ergeben sich sofort einige unbestimmte Integrale:

$$\int x^\alpha dx = \frac{1}{\alpha+1}x^{\alpha+1} + C \qquad (\alpha \neq -1),$$
$$\int \frac{dx}{x} = \ln|x| + C,$$
$$\int e^x dx = e^x + C,$$
$$\int \cos x dx = \sin x + C,$$
$$\int \sin x dx = -\cos x + C,$$
$$\int \frac{dx}{\sqrt{1-x^2}} = \arcsin x + C, \quad |x| < 1,$$
$$\int \frac{dx}{\sqrt{1-x^2}} = -\arccos x + C, \quad |x| < 1,$$
$$\int \frac{dx}{1+x^2} = \arctan x + C.$$

Eine Stammfunktion heißt **elementar**, wenn sie durch endlich viele Verknüpfungen $+, -, \cdot, /, \circ$ aus den bekannten elementare Funktionen dargestellt werden kann. Man nennt daher eine Funktion **elementar integrierbar**, wenn sie eine elementare Stammfunktion besitzt. Es gibt elementare Funktionen, die nicht elementar integrierbar sind, so z. B. die Funktion

$$f(x) = \frac{\sin x}{x}.$$

Für die unbestimmte Integration (d. h. die Bestimmung des unbestimmten Integral) gibt es viele Regeln, von denen wir hier nur die wichtigsten kurz diskutieren werden. Es gibt heute effziente Programmsysteme, die uns das mühselige Handwerk des Integrierens abnehmen.

Satz 122 (Partielle Integration). *Sind die Funktionen f, g stetig differenzierbar auf einem Intervall I, so gilt*

$$\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx + C.$$

Beweis. Nach der Produktregel für die Differentiation gilt

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x),$$

also

$$\begin{aligned} f(x)g(x) &= \int [f'(x)g(x) + f(x)g'(x)]\, dx \\ &= \int f'(x)g(x)dx + \int f(x)g'(x)dx + C. \quad \square \end{aligned}$$

Beispiel:

$$\begin{aligned} \int \cos^2 x dx &= \int \cos x \cos x dx = \cos x \sin x + \int \sin^2 x dx \\ &= \cos x \sin x + \int (1 - \cos^2 x)dx = \cos x \sin x + x - \int \cos^2 x dx \\ &= \frac{1}{2}(\cos x \sin x + x) + C. \end{aligned}$$

Für bestimmte Integrale lautet die partielle Integration:

$$\int_a^b f(x)g'(x)dx = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)dx.$$

Beispiel. Wir wollen $\int_0^x t^2 e^{-t} dt$ berechnen. Es ist

$$\begin{aligned} \int t^2 e^{-t} dt &= -t^2 e^{-t} + 2\int te^{-t} dt = -t^2 e^{-t} - 2\left[te^{-t} - \int e^{-t} dt\right] \\ &= -e^{-t}\left(t^2 + 2t + 2\right) + C. \end{aligned}$$

Damit erhalten wir

$$\int_0^x t^2 e^{-t} dt = \left[-e^{-t}(t^2+2t+2)\right]_0^x = -e^{-x}(x^2+2x+2)+2.$$

Für $x \to \infty$ folgt

$$\lim_{x\to\infty} x^2 e^{-x} = \lim_{x\to\infty} \frac{x^2}{e^x} = 2 \cdot \lim_{x\to\infty} \frac{x}{e^x} = 2 \cdot \lim_{x\to\infty} \frac{1}{e^x} = 0,$$

also

$$\begin{aligned}\int_0^\infty t^2 e^{-t} dt &= \lim_{x\to\infty} \int_0^x t^2 e^{-t} dt \\ &= \lim_{x\to\infty} \left(-e^{-x}(x^2+2x+2)+2\right) = 2.\end{aligned}$$

Satz 123 (Substitutionsregel). *Ist die Funktion f stetig auf I, die Funktion g stetig differenzierbar mit Werten in I, so gilt*

$$\int f(g(x))g'(x)dx = \int f(t)dt \qquad (t = g(x)).$$

Beweis. Die Funktion g im Satz heißt **Substitutionsfunktion**. Die Formel folgt direkt aus der Kettenregel für die Differentiation. □

Beispiele. Mit $t = f(x), f(x) \neq 0$ erhält man

$$\int \frac{f'(x)}{f(x)} dx = \int \frac{dt}{t} dt = \ln|t| + C = \ln|f(x)| + C.$$

Entsprechend ergibt sich mit $t = f(x), f(x) > 0, \alpha \neq -1$:

$$\int (f(x))^\alpha f'(x)dx = \int t^\alpha dt = \frac{t^{\alpha+1}}{\alpha+1} + C = \frac{1}{\alpha+1}(f(x))^{\alpha+1} + C.$$

Für den Fall, daß die Substitutionsfunktion umkehrbar eindeutig ist, kann man die Subtitutionsregel auch von rechts nach links lesen:

$$\int f(x)dx = \int f(g(t))g'(t)dt \qquad (t = g^{-1}(x)).$$

Beispiel. Für $n > 1$ sei $t = n \cdot x - 1$; dann gilt $\frac{dx}{dt} = \frac{1}{n}$ und

$$\begin{aligned}\int \sin(nx-1)dx &= \int \frac{1}{n}\sin t dt = -\frac{1}{n}\cos t + C \\ &= -\frac{1}{n}\cos(nx-1) + C.\end{aligned}$$

Satz 124 (1. Substitutionsregel für bestimmte Integrale). *Ist φ stetig differenzierbar auf $[a,b]$ und f stetig auf $\varphi([a,b])$, so gilt*

$$\int_a^b f(\varphi(x))\varphi'(x)dx = \int_{\varphi(a)}^{\varphi(b)} f(t)dt.$$

Beweis. Mit einer Stammfunktion F von f gilt

$$F'(\varphi(x)) = f(\varphi(x))\varphi'(x),$$

also ist $F(\varphi(\cdot))$ eine Stammfunktion von $f(\varphi(\cdot))\varphi'(\cdot)$ und

$$\int_a^b f(\varphi(x))\varphi'(x)dx = F(\varphi(b)) - F(\varphi(a)) = \int_{\varphi(a)}^{\varphi(b)} f(t)dt. \quad \square$$

Satz 125. (2. Substitutionsregel für bestimmte Integrale)
Es sei f eine auf dem Intervall $[a,b]$ stetige Funktion; φ sei stetig differenzierbar und bilde ein Intervall umkehrbar eindeutig auf das Intervall $[a,b]$ ab. Dann gilt

$$\int_a^b f(x)dx = \int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} f(\varphi(t))\varphi'(t)dt.$$

Der Beweis folgt sofort mit dem letzten Satz.

Satz 126. *Jede rationale Funktion ist elementar integrierbar.*

Diesen Satz beweist man durch eine sog. Partialbruchzerlegung einer rationalen Funktion, was eine rein technische Angelegenheit ist und daher auch wegen seiner Länge übergangen werden soll.

Satz 127. *Ist (f_n) eine gleichmäßig konvergente Folge von auf dem Intervall $[a,b]$ stetigen Funktionen, dann gilt*

$$\int_a^b \lim_{n\to\infty} f_n(x)dx = \lim_{n\to\infty} \int_a^b f_n(x)dx.$$

Beweis. Die Grenzfunktion

$$f(x) = \lim_{n\to\infty} f_n(x)$$

ist stetig; wegen der gleichmäßigen Konvergenz gibt es zu jedem $\varepsilon > 0$ eine Zahl $N = N(\varepsilon)$, so daß

$$|f(x) - f_n(x)| < \frac{\varepsilon}{b-a} \qquad \forall n \geqq N, x \in [a,b].$$

Damit folgt

$$|\int_a^b f(x)dx - \int_a^b f_n(x)dx| \leqq \int_a^b |f(x) - f_n(x)|dx \leqq \frac{\varepsilon}{b-a}\int_a^b dx = \varepsilon. \quad \square$$

Für gleichmäßig konvergente Reihen gilt entsprechend:

$$\int_a^b \left(\sum_{n=0}^{\infty} f_n(x)\right) dx = \sum_{n=0}^{\infty} \int_a^b f_n(x)dx.$$

Beispiel.

$$\begin{aligned}
\int_0^{\frac{\pi}{4}} \sum_{n=0}^{\infty} \frac{\sin(10^{2n}x)}{10^n}dx &= \sum_{n=0}^{\infty} \frac{1}{10^n} \int_0^{\frac{\pi}{4}} \sin(10^{2n}x)dx \\
&= -\sum_{n=0}^{\infty} \frac{1}{10^{3n}} \left[\cos(10^{2n}x)\right]_0^{\frac{\pi}{4}} \\
&= -\sum_{n=0}^{\infty} \frac{1}{10^{3n}} \left(\cos(10^{2n}\frac{\pi}{4}) - 1\right) \\
&\approx 1,002 - 0,707.
\end{aligned}$$

4.7.4. Uneigentliche Integrale

Es sei f eine im Intervall $[a,\infty)$ integrierbare Funktion. Für jedes $b \geqq a$ existiert dann $\int_a^b f(x)dx$ und

$$F(t) = \int_a^t f(x)dx.$$

ist eine stetige Funktion. Falls F für $t \to \infty$ einen endlichen Grenzwert hat, setzen wir

$$\int_a^{\infty} f(x)dx = \lim_{t\to\infty} F(t) = \lim_{t\to\infty} \int_a^t f(x)dx$$

und nennen den Grenzwert **uneigentliches Integral** der Funktion f über dem Intervall $[a,\infty)$. Existiert dieser Grenzwert nicht, so sagen wir, daß das Integral divergiert. Analog zu unendlichen Reihen konvergiert das uneigentliche Integral absolut, wenn das uneigenliche Integral der Funktion $|f|$ über $[a,\infty)$ existiert. Analog denken wir uns das uneigentliche Integral einer über $(-\infty,a]$ integrierbaren Funktion eingeführt. Falls die Funktion f über dem Intervall $(-\infty,\infty)$ integrierbar ist, setzt man

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{a} f(x)dx + \int_a^{\infty} f(x)dx.$$

Für eine im Intervall $[a,\infty)$ stetige Funktion f mit einer Stammfunktion F folgt

$$\int_a^\infty f(x)dx = \lim_{b\to\infty}\int_a^b f(x)dx = \lim_{b\to\infty}(F(b)-F(a)) = [F(x)]_a^\infty;$$

analog im Intervall $(-\infty,a]$:

$$\int_{-\infty}^a f(x)dx = F(a) - \lim_{c\to-\infty} F(c) = [F(x)]_{-\infty}^a$$

und zusammen

$$\int_{-\infty}^\infty f(x)dx = \lim_{b\to\infty} F(b) - \lim_{c\to\infty} F(c) = [F(x)]_{-\infty}^\infty.$$

Beispiele.

$$\int_1^\infty \frac{dx}{x^2} = \lim_{b\to\infty}\int_1^b \frac{dx}{x^2} = \lim_{b\to\infty}\left[-\frac{1}{x}\right]_1^b = 1 - \lim_{b\to\infty}\frac{1}{b} = 1.$$

$$\int_1^\infty \frac{dx}{x} = \lim_{b\to\infty}\int_1^b \frac{dx}{x} = \lim_{b\to\infty}(\ln b - \ln 1) = \lim_{b\to\infty}\ln b.$$

Damit haben wir insbesondere, daß das letzte uneigentliche Integral divergiert. Dagegen folgt

$$\begin{aligned}\int_{-\infty}^\infty \frac{dx}{1+x^2} &= \lim_{b\to\infty}\int_0^b \frac{dx}{1+x^2} + \lim_{c\to-\infty}\int_c^0 \frac{dx}{1+x^2}\\ &= \lim_{b\to\infty}[\arctan x]_0^b + \lim_{c\to-\infty}[\arctan x]_c^0\\ &= \lim_{b\to\infty}\arctan b - \lim_{c\to-\infty}\arctan c\\ &= \frac{\pi}{2}+\frac{\pi}{2} = \pi.\end{aligned}$$

Es sei nun die Funktion f in jedem offenen Intervall (a,c) mit $a<c<b$ beschränkt und integrierbar. Wir setzen

$$\int_a^b f(x)dx = \lim_{\substack{c\to b\\ c<b}}\int_a^c f(x)dx$$

und sagen, daß das Integral konvergiert, wenn dieser Grenzwert existiert; andernfalls divergiert das Integral. Analog setzen wir

$$\int_a^b f(x)dx = \lim_{\substack{c\to a\\ c>a}}\int_c^b f(x)dx.$$

Ist die Funktion f auf jedem abgeschlossenen Teilintervall aus $[a,c)$ und $(c,b]$ integrierbar, so setzt man

$$\int_a^b f(x)dx = \lim_{\substack{t\to c\\ t<c}}\int_a^t f(x)dx + \lim_{\substack{t\to c\\ t>c}}\int_t^b f(x)dx,$$

falls beide Integrale konvergieren.
Beispiele.

$$\int_0^2 \frac{dx}{\sqrt{x}} = \lim_{\substack{t\to 0\\ t>0}}\int_t^2 \frac{dx}{\sqrt{x}} = \lim_{\substack{t\to 0\\ t>0}}\left[2\sqrt{x}\right]_t^2 \approx 2\cdot 1,414\ldots,$$

$$\int_0^1 \frac{dx}{x} = \ln 1 - \lim_{\substack{t\to 0\\ t>0}}\ln t = \infty.$$

Wegen

$$\frac{1}{1-x^2} = \frac{1}{2}\cdot\frac{1}{1-x} + \frac{1}{2}\cdot\frac{1}{1+x}$$

folgt

$$\int_0^1 \frac{dx}{1-x^2} = \frac{1}{2}\int_0^1 \frac{dx}{1-x} + \frac{1}{2}\int_0^1 \frac{dx}{1+x},$$

woraus wir ersehen, daß das Integral divergiert, da der erste Summand divergiert. Insbesondere divergiert damit jedes Integral der Funktion

$$\frac{1}{1-x^2},$$

wenn die Integration über ein Intervall I mit $1\in I$ erstreckt wird.

4.8. Übungen

1. Man gebe $N(\varepsilon) \in \mathbb{R}$ an, so daß gilt: $\mid x_n \mid < \varepsilon \quad \forall n > N(\varepsilon)$.

   (a)

   $$x_n = \frac{(-1)^{n^2+1}}{4n^3}\,,$$

   (b)

   $$x_n = \frac{2n}{n^2-2}.$$

2. Es sei (x_n) die Ziffernfolge der Zahl π ($x_0 = 3$, $x_1 = 1$, $x_2 = 4$, ...).

   (a) Besitzt die Folge (x_n) Häufigkeitspunkte?

   (b) Besitzt die Folge einen Grenzwert?

3. Man untersuche die Folgen (q_n) auf Monotonie, Beschränktheit und Häufigkeitspunkte.

   (a)

   $$q_n = \frac{(-2)^{n+1} + 3^n}{3^{n+1} + (-2)^n}\,,$$

   (b)

   $$q_n = \cos\left(\frac{n\pi}{4}\right)\,,$$

   (c)

   $$q_{n+1} = \frac{2}{q_n},\ q_0 \in (1,2)\,,$$

   (d)

   $$q_{n+1} = \sqrt{2+q_n},\ q_0 = \sqrt{2}\,.$$

4. Man gebe $n_0(\varepsilon) \in \mathbb{R}$ an, so daß gilt: $\mid a_n - a \mid < \varepsilon \quad \forall n > n_0(\varepsilon)$.

   (a)

   $$a_n = \frac{1-\sqrt{n}}{1+\sqrt{n}} \quad ,\ a = -1\,,$$

   (b)

   $$a_n = \frac{n^4}{n!} \quad ,\ a = 0\,.$$

5. Man bestimme den Grenzwert $\lim\limits_{n\to\infty} a_n$.

   (a)

   $$a_n = q^n\,,$$

   (b)

   $$a_n = \frac{2n^3 + 6^n}{n!}\,,$$

   (c)

   $$a_n = \left(1 - \frac{1}{n^2}\right)^n\,,$$

(d)

$$a_n = \frac{n!}{n^n}\,.$$

6. Es sei (a_n) eine Folge nichtnegativer reeller Zahlen. Man zeige:

$$\sum_{n=0}^{\infty} a_n \quad \text{konv.} \Rightarrow \sum_{n=0}^{\infty} a_n^2 \quad \text{konv.}$$

Gilt auch die Umkehrung?

7. Man untersuche folgende Reihen auf Konvergenz:

(a)

$$\sum_{k=0}^{\infty} (-1)^k\,,$$

(b)

$$\sum_{n=2}^{\infty} \sqrt[n]{a} \quad , \quad 0 < a < 1\,,$$

(c)

$$\sum_{k=1}^{\infty} \left(\frac{1}{4^k} - \frac{3}{2^k}\right)\,,$$

(d)

$$\sum_{n=1}^{\infty} \left(1 - \frac{1}{n}\right)^n\,,$$

(e)

$$\sum_{n=1}^{\infty} \frac{(-1)^{\frac{n(n-1)}{2}}}{3^n}\,.$$

8. Man untersuche folgende Reihen mit Hilfe des Wurzel- bzw. Quotientenkriteriums auf Konvergenz:

(a)

$$\sum_{n=1}^{\infty} \frac{(n!)^2 5^n}{(2n)!}\,,$$

(b)

$$\sum_{k=1}^{\infty} \frac{k^2}{\left(2 - \frac{1}{k}\right)^k}\,,$$

(c)

$$\sum_{n=1}^{\infty} \frac{n^3}{2^n}\,,$$

(d)

$$\sum_{n=1}^{\infty} \frac{n!}{n^n}\,,$$

(e)

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

9. Konvergiert die Reihe

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n^2-10n+30}}\,?$$

10. Man gebe alle $x \in \mathbb{R}$ an, für die die Reihe

$$\sum_{n=1}^{\infty} \frac{x^n}{n2^{n-1}}$$

konvergiert.

11. Man zeige: Für

$$P(x) = a_n x^n + \cdots + a_1 x + a_0 \qquad (a_n \neq 0),$$

$$Q(x) = b_m x^m + \cdots + b_1 x + b_0 \qquad (b_m \neq 0)$$

gilt:

$$\lim_{x\to\infty} \frac{P(x)}{Q(x)} = \lim_{x\to\infty} \frac{a_n x^{n-m}}{b_m}.$$

12. Man berechne:

(a)

$$\lim_{k\to\infty} \frac{k+19k^{21}+21k^{19}-100}{100k+19k^{19}+21k^{21}+1}\,,$$

(b)

$$\lim_{k\to\infty} \frac{\sqrt[10]{k^{15}}}{\sqrt[5]{k^{10}+k}}\,,$$

(c)

$$\lim_{\varphi\to\infty} \frac{\varphi \sin \varphi^2}{\varphi^2+\sin\varphi}\,.$$

13. Mit Hilfe des Stetigkeitskriteriums zeige man die Stetigkeit der folgenden Funktionen im $\mathbb{R}^1$:

(a) $f(x) = \sin 3x\,,$

(b) $f(x) = x^n \ (n \in \mathbb{N})\,.$

14. Man berechne die Grenzwerte der folgenden Funktionen:

(a)

$$\lim_{x\to 0-} e^{\frac{1}{x}}\,,$$

(b)

$$\lim_{x\to 0+} e^{\frac{1}{x}}\,,$$

(c)

$$\lim_{x\to 0} \frac{\sin x}{x}\,,$$

(d)

$$\lim_{x\to 0} \frac{\sin 5x}{x}\,,$$

(e)

$$\lim_{x\to 0} \frac{1-\cos x}{x^2}\,,$$

(f)

$$\lim_{x\to a} \frac{\cos x - \cos a}{x-a}\,.$$

15. Der Umfang eines regelmäßigen n-Ecks, das einem Kreis vom Radius R einbeschrieben ist, beträgt

$$U_n = 2Rn \sin\frac{\pi}{n}.$$

Man bestimme

$$u = \lim_{n\to\infty} U_n.$$

16. Es gilt der Satz:
Für n stetige Funktionen $f_1 \dots f_n$ in $\mathbb{R}$ sind auch die Funktionen

$$F_{min}(x) = \min_{1 \leqq k \leqq n} f_k(x) \qquad F_{max}(x) = \max_{1 \leqq k \leqq n} f_k(x)$$

stetige Funktionen.
Man zeige, daß die Funktion

$$g_c(x) = \begin{cases} -c & f(x) < -c \\ f(x) & -c \leqq f(x) \leqq c \\ c & \text{sonst} \end{cases}$$

für jede stetige Funktion in $\mathbb{R}$ ebenfalls stetig in $\mathbb{R}$ ist.

17. Man untersuche die Folge von Funktionen (f_n) auf gleichmäßige Konvergenz und bestimme die zugehörige Grenzfunktion:

(a)

$$f_n(x) = \frac{1}{1+e^{n(a-x)}} \qquad x \in (a,\infty)\,,$$

(b)

$$f_n(x) = \frac{1}{1+e^{nx}} \qquad x \in (1,\infty)\,,$$

(c)

$$f_n(x) = \sqrt[n]{x} \qquad x > 0\,.$$

18. Man bestimme $a, b \in \mathbb{R}$ so, daß die Funktion

$$f(x) = \begin{cases} x & \text{für} \quad x \leqq a \\ 2 + bx^2 & \text{für} \quad x > a \end{cases}$$

stetig differenzierbar in $\mathbb{R}$ ist.

19. Man bestimme die 1. Ableitung folgender Funktionen:

(a)

$$p(y) = \frac{b+ay}{(b-ay)^c} \qquad y \neq \frac{b}{a}\,,$$

(b)

$$q(x) = \frac{1}{\log_2(x^2)} \qquad x \neq 0\,.$$

20. Es gilt die Regel

$$\frac{d}{dx}(\ln f(x)) = \frac{f'(x)}{f(x)}$$

für differenzierbare Funktionen f mit positiven Funktionswerten. Man berechne damit die 1. Ableitung von:

(a)

$$f(x) = (x^2)^{2x} \qquad (x > 0),$$

(b)

$$f(x) = x^{\sin x} \qquad (x \in (0, \pi)),$$

(c)

$$f(x) = (\ln x)^{\ln x} \qquad (x > 1).$$

21. Man zeige mit Hilfe des Mittelwertsatzes :

$$|\arctan x - \arctan y| \leqq \frac{1}{2}\,|x - y| \text{ für alle } x, y \geqq 1.$$

22. Man ermittle die ersten sechs Glieder der Taylor-Entwicklung folgender Funktionen in $x = 0$.
Wo konvergieren die Reihen?

(a) $f(x) = \cos^2 x - \sin^2 x\,,$

(b) $g(x) = \tan x - x\,,$

(c) $h(x) = \ln(\cos x)$ für $(|x| < \frac{\pi}{2})\,.$

23. Man berechne $\sin 2^\circ$ so, daß der absolute Fehler kleiner als $5 \cdot 10^{-4}$ ist. Wieviele Reihenglieder sind nötig?

24. Kann man $p, q \in \mathbb{R}$ so wählen, daß $|x + p\sin x + q \sin 2x| \leqq C \cdot |x^5|$ für ein $C > 0$ und genügend kleines x gilt?
Hinweis: Man ermittle die Reihenentwicklung des linken Ausdrucks!

25. Man ermittle Extrema und Wendepunkte der Funktion $f(x) = e^{-x^2}$.

26. Mit der Regel von l'Hospital ermittle man

(a)

$$\lim_{x\to 0} \frac{\ln(\cos ax)}{\ln(\cos bx)}\,,$$

(b)

$$\lim_{x\to \infty} x^2 \cdot e^{\frac{-x}{1000}}\,,$$

(c)

$$\lim_{x\to \infty} \frac{\ln x}{x^p} \qquad (p \in \mathbb{R})\,,$$

(d)

$$\lim_{x\to 0} x^x\,,$$

(e)

$$\lim_{x\to 1}\left(\frac{1}{\ln x}-\frac{1}{x-1}\right),$$

(f)

$$\lim_{x\to 1} x^{\frac{1}{1-x}},$$

(g)

$$\lim_{x\to 0}\left(\sin^2 x\right)^{\frac{1}{\ln x^2}}.$$

27. Man ermittle die Taylor-Reihe im Punkt $x_0 = 0$ sowie ihren Konvergenzradius:

(a)

$$f(x) = a^x \qquad (a > 0),$$

(b)

$$f(x) = \frac{x^{10}}{(1-x)^2} \qquad (x \neq 1).$$

28. Wie groß ist der Konvergenzradius folgender Potenzreihen?
Was kann man über die Konvergenz an den Grenzen des Konvergenzbereiches aussagen?

(a)

$$\sum_{n=1}^{\infty}\frac{x^n}{n^p} \qquad (p \in \mathbb{R}),$$

(b)

$$\sum_{n=1}^{\infty}\frac{3^n+(-2)^n}{n}(x+1)^n,$$

(c)

$$\sum_{n=1}^{\infty}\frac{x^n}{a^n+b^n} \qquad (a > b \geqq 0).$$

29. Man bestimme unter Zurückführung auf Grundintegrale:

(a)

$$\int (1+x)(1-2x)(1+3x)\,dx,$$

(b)

$$\int\left[\left(\frac{1+x}{x}\right)^2-\left(\frac{1-x}{x}\right)^2\right]dx,$$

(c)

$$\int\frac{e^{-2x}+2e^x+5}{e^x}\,dx,$$

(d)

$$\int\frac{x^2+3\sqrt[5]{x}}{\sqrt[5]{x^6}}\,dx.$$

30. Man bestimme mittels linearer Substitution:

(a)

$$\int \frac{dx}{\sin^2(\frac{x+\pi}{4})}\,,$$

(b)

$$\int_1^2 \sqrt[9]{27x-26}\,dx\,,$$

(c)

$$\int_0^1 \frac{x^2}{\sqrt{2-x}}\,dx\,,$$

(d)

$$\int \frac{2}{1+(x-1)^2}\,dx\,.$$

31. Man berechne mit Hilfe der partiellen Integration:

(a)

$$\int z^2 \sin z\,dz\,,$$

(b)

$$\int (x^2+x)e^x\,dx\,.$$

Kapitel 5

Stochastik

5.1. Wahrscheinlichkeit

Der Zufall tritt in der Tätigkeit des Informatikers bei zahlreichen Gelegenheiten auf. So spricht man z. B. von einer zufälligen Laufzeit eines Programms innerhalb eines Mehraufgabensystems. Auch die Simulation realer Vorgänge anhand zufällig gewählter Daten ist hier zu nennen. Die mathematische Wahrscheinlichkeitstheorie ist der Versuch, mittels mathematischer Modelle den Zufall in einer Form zu beschreiben, daß daraus praktische Schlußfolgerungen gezogen werden können. Da es sich hierbei um eine Abstraktion handelt, kann man nicht erwarten, daß durch ein mathematisches Modell alle auftretenden Formen des Zufalls beschrieben werden können. Wir beschränken uns vielmehr auf jene Form des Zufalls, wie er uns in sog. zufälligen Versuchen begegnet. Ein zufälliger Versuch kann beliebig oft wiederholt werden, d. h. die Versuchsbedingungen sind beliebig oft wiedereinstellbar. Der Versuchsausgang wird jedoch nicht vollständig durch die Versuchsparameter festgelegt und ist daher in einem gewissen Rahmen ungewiß. Einen möglichen Ausgang eines zufälligen Versuchs nennt man **zufälliges Ereignis**. Als wohl bekanntestes Beispiel kann man ein Lotteriespiel anführen, z. B. 6 aus 49. In diesem „zufälligen“ Versuch gibt es 13983816 mögliche Versuchsausgänge, nämlich so viele Möglichkeiten, 6 Zahlen ohne Rücklegen zu ziehen.

5.1.1. Wahrscheinlichkeit zufälliger Ereignisse

Zufällige Ereignisse werden im Rahmen der Mengenlehre modelliert. Dabei ist es wesentlich zu wissen, welches die sog. Elementarereignisse eines zufälligen Versuches sind. In einer konkreten Situation kann man oft die Elementarereignisse verifizieren: Jedes **Elementarereignis** ist ein möglicher Versuchsausgang. So sind die Elementarereignisse beim Würfeln das Auftreten der Augenzahlen 1, 2, 3, 4, 5, 6, und andere gibt es nicht. Wichtig ist hier, daß man alle Elementarereignisse in die Bertrachtungen einbezieht.
Es sei Ω eine beliebige, nichtleere Menge (d. h. die Menge aller Elementarereignisse) und $\mathcal{A} \subseteqq \mathcal{P}(\Omega)$ eine Untermenge der Potenzmenge von Ω. Die Menge $\mathcal{A}$ heißt **σ-Algebra** über Ω, wenn Ω in dem Mengensystem $\mathcal{A}$ liegt und $\mathcal{A}$ abgeschlossen ist bezüglich der Komplementbildung und der Vereinigung von abzählbar vielen Elementen aus $\mathcal{P}(\Omega)$:

- $\Omega \in \mathcal{A}$,
- $A \in \mathcal{A} \Longrightarrow \overline{A} \in \mathcal{A}$,
- $(A_n) \subseteqq \mathcal{A} \Longrightarrow \bigcup\limits_{n=1}^{\infty} A_n \in \mathcal{A}$.

Die Potenzmenge $\mathcal{P}(\Omega)$ ist die feinste (größte) σ-Algebra über Ω und $\mathcal{A} = \{\emptyset, \Omega\}$ die gröbste (kleinste). Ein Element einer σ-Algebra nennt man **zufälliges Ereignis** oder einfach **Ereignis**. Mit den de Morganschen Regeln folgt sofort, daß die leere Menge zu jeder σ-Algebra gehört und jede σ-Algebra auch abgeschlossen gegenüber einer abzählbaren Durchschnittsbildung ist:

$$(A_n) \subseteqq \mathcal{A} \Longrightarrow \bigcap_{n=1}^{\infty} A_n \in \mathcal{A}.$$

Wenn man bei einer abzählbaren Vereinigungsbildung ab einem gewissen Index nur noch die leere Menge nimmt und bei einer abzählbaren Durchschnittsbildung nur noch die Menge Ω, so sieht man, daß eine σ-Algebra abgeschlossen ist gegenüber der Vereinigung und dem Durchschnitt. Damit ist $\mathcal{A}(\cap, \cup, \overline{\ })$ eine algebraische Struktur im üblichen Sinne. Formal gesehen sind die Elemente eines minimalen Erzeugendensystems einer σ-Algebra die Elementarereignisse. Sehr anschaulich wird dieser algebraische Modellierungsansatz in dem Falle,

daß es nur endlich viele Elementarereignisse gibt. Dann enthält jede Menge nur endlich viele Elemente. Ist nun $A = \{ a_1, \ldots, a_m \}$ ein Element der σ-Algebra $\mathcal{A}$, dann charakterisiert A das Ereignis

„a_1 oder a_2 oder $\cdots$ oder a_m“.

Es ist klar, daß das sichere Ereignis durch die Menge Ω und das unmögliche Ereignis durch die leere Menge $\emptyset$ repräsentiert sind. Wir suchen nun nach einem quantitativen Maß für die Zufälligkeit. Dieses Maß soll ausdrücken, wie wahrscheinlich das Eintreten eines Ereignisses ist. Wenn wir einen zufälligen Versuch n-mal wiederholen und dabei das Ereignis A genau $H_n(A)$-mal eintritt, ist

$$h_n(A) = \frac{H_n(A)}{n}$$

die relative Häufigkeit für das Eintreten des Ereignisses A. Die relative Häufigkeit hat folgende, unmittelbar einsichtige Eigenschaften:

$$0 \leqq h_n(A) \leqq 1,$$

$$h_n(\emptyset) = 0, \qquad h_n(\Omega) = 1,$$

$$h_n(\overline{A}) = 1 - h_n(A),$$

$$h_n(A \cup B) = h_n(A) + h_n(B) - h_n(A \cap B),$$

$$h_n(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} h_n(A_i), \quad \text{falls } A_i \cap A_j = \emptyset \quad (i \neq j).$$

Dabei heißen zwei Ereignisse **unvereinbar**, wenn ihr Durchschnitt leer ist.
Aus der Erfahrung weiß man nun, daß mit wachsendem n die relativen Häufigkeiten $h_n(A)$ immer weniger stark um einen gewissen Wert schwanken. Diesen Wert nennt man die **empirische Wahrscheinlichkeit** $P(A)$ für das Ereignis A. Allgemein versteht man unter einem **Wahrscheinlichkeitsmaß** P eine auf einer σ-Algebra $\mathcal{A}$ (über einer nichtleeren Menge Ω) definierte reellwertige Funktion mit folgenden charakterisierenden Eigenschaften:

$$0 \leqq P(A) \leqq 1, \quad \forall A \in \mathcal{A}, \quad P(\Omega) = 1,$$

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n) \quad \forall (A_n) \subseteq \mathcal{A}, \quad A_n \cap A_m = \emptyset \quad (n \neq m).$$

Die letzte Eigenschaft nennt man auch **σ-Additivität** . Der Funktionswert $P(A)$ heißt die **Wahrscheinlichkeit** für das Ereignis $A \in \mathcal{A}$.

Satz 128. (Grundeigenschaften eines Wahrscheinlichkeitsmaßes)
Es sei P ein Wahrscheinlichkeitsmaß auf einer σ-Algebra $\mathcal{A}$: Dann gelten die folgenden Regeln.

1. $P(\emptyset) = 0$,
2. $P(\overline{A}) = 1 - P(A) \quad \forall A \in \mathcal{A}$.
3. **Monotonie**: *Aus $A \subseteq B$ folgt $P(A) \leqq P(B)$.*
4. **Subtraktivität**: *Aus $A \subseteq B$ folgt $P(B \setminus A) = P(B) - P(A)$.*
5. **Unterhalbstetigkeit**: *Für jede monoton wachsende Ereignisfolge, d. h.*

   $$(A_n) \subseteq \mathcal{A}, \quad A_n \subseteq A_{n+1}$$

   gilt

   $$P(\bigcup_{n=1}^{\infty} A_n) = \lim_{n \to \infty} P(A_n).$$

6. **Oberhalbstetigkeit**: *Für jede monoton fallende Ereignisfolge, d. h.*

   $$(A_n) \subseteq \mathcal{A}, \quad A_{n+1} \subseteq A_n$$

   gilt

   $$P(\bigcap_{n=1}^{\infty} A_n) = \lim_{n \to \infty} P(A_n).$$

7. **Subadditivität**: *Für alle Ereignisfolgen* $(A_n) \subseteqq \mathcal{A}$ *ist stets*

$$P(\bigcup_{n=1}^{\infty} A_n) \leqq \sum_{n=1}^{\infty} P(A_n).$$

8. **Siebformel**:

$$P(\bigcup_{k=1}^{n} A_k) = \sum_{k=1}^{n} P(A_k) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - + \\ \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n).$$

9. **Bonferoni-Ungleichung**:

$$P(\bigcup_{k=1}^{n} A_k) \geqq \sum_{k=1}^{n} P(A_k) - \sum_{i<j} P(A_i \cap A_j).$$

Beweis. Die Regel 2 folgt, wenn man die σ-Additivität des Wahrscheinlichkeitsmaßes P auf die Folge

$$(A, \overline{A}, \emptyset, \emptyset, \ldots)$$

anwendet. Setzen wir in 2. speziell $A = \Omega$, so erhalten wir die Regel 1. Für die Monotonie beachten wir, daß im Falle $A \subseteqq B$ offenbar

$$B = A \cup (B \setminus A)$$

gilt, die Ereignisse A und $B \setminus A$ unvereinbar sind und damit aus der σ-Additivität folgt:

$$P(B) = P(A \cup (B \setminus A)) = P(A) + P(B \setminus A) \geqq P(A).$$

Gleichzeitig folgt daraus auch die Subtraktivität.
Für den Beweis der Unterhalbstetigkeit konstruieren wir eine neue Ereignisfolge:

$$B_1 = A_1, \quad B_{n+1} = A_{n+1} \setminus A_n, \quad n = 1, 2, \ldots.$$

Die Ereignisse B_n sind paarweise unvereinbar und es gilt

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n.$$

Die σ-Additivität und die Subtraktivität liefern nun:

$$\begin{aligned} P(\bigcup_{n=1}^{\infty} A_n) = P(\bigcup_{n=1}^{\infty} B_n) &= \sum_{n=1}^{\infty} P(B_n) \\ &= P(A_1) + \sum_{n=2}^{\infty} (P(A_n) - P(A_{n-1}) \\ &= P(A_1) + \lim_{m\to\infty} \sum_{n=2}^{m} (P(A_n) - P(A_{n-1})) \\ &= P(A_1) + \lim_{m\to\infty} (P(A_m) - P(A_1)) = \lim_{m\to\infty} P(A_m). \end{aligned}$$

Die Oberhalbstetigkeit folgt aus der Unterhalbstetigkeit durch Komplementbildung. Die Subadditivität erhält man aus dem Beweis der Unterhalbstetigkeit, da dort wegen der Monotonie gilt:

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(B_n) \leqq \sum_{n=1}^{\infty} P(A_n).$$

Die Siebformel beweisen wir induktiv. Für $n = 1$ ist nichts zu beweisen. Die Siebformel möge also für eine natürliche Zahl n gelten. Es sei

$$A = \bigcup_{k=1}^{n} A_k, \quad B = A_{n+1}.$$

Wir erhalten mittels Subtraktivität die folgende Gleichungskette:

$$\begin{aligned} P(\bigcup_{k=1}^{n+1} A_k) &= P(A \cup B) = P((A \setminus B) \cup (B \setminus A) \cup (A \cap B)) \\ &= P(A \setminus B) + P(B \setminus A) + P(A \cap B) \\ &= P(A \setminus (A \cap B)) + P(B \setminus (A \cap B)) + P(A \cap B) \\ &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

Mit der Induktionsvoraussetzung schließen wir:

$$\begin{aligned} P(A \cap B) &= P(\bigcup_{k=1}^{n} A_k \cap A_{n+1}) \\ &= \sum_{k=1}^{n} P(A_k \cap A_{n+1}) - \sum_{i<j} P(A_i \cap A_j \cap A_{n+1}) \\ &+ \sum_{i<j<k} P(A_i \cap A_j \cap A_k \cap A_{n+1}) - + \cdots \\ &\cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n \cap A_{n+1}). \end{aligned}$$

Wenn wir nun die Induktionsvoraussetzung noch auf $P(A)$ anwenden und alle Teile geeignet zusammenfügen, erhalten wir die Siebformel für $n+1$.
Die Bonferoni-Ungleichung ergibt sich analog durch Induktion über n mittels der Subtraktivität. □
Die Monotonie-Regel bedeutet, daß im Falle $A \subseteqq B$ das Ereignis B wahrscheinlicher ist als das Ereignis A. Die Siebformel erlaubt es, die Wahrscheinlichkeit für das Eintreten des Ereignisses „A_1 oder A_2 oder ... oder A_n“ auch dann zu berechnen, wenn die Einzelereignisse nicht paarweise unvereinbar sind. Abschließend sei noch erwähnt, daß man zwei Ereignisse A, B **unabhängig** nennt, wenn

$$P(A \cap B) = P(A) \cdot P(B)$$

gilt. Unabhängige Ereignisse haben nichts mit unvereinbaren Ereignissen zu tun. Es sind solche Ereignisse, die gleichzeitig eintreten können, das Eintreten des einen Ereignisses aber nicht durch das Eintreten des anderen Ereignisses beeinflußt wird.

5.1.2. Zufallsgrößen und Verteilungsfunktionen

Es sei Ω die Menge aller Elementarereignisse, $\mathcal{A}$ die von Ω erzeugte σ-Algebra und P ein Wahrscheinlichkeitsmaß auf $\mathcal{A}$. Eine auf Ω erklärte reellwertige Funktion X heißt **Zufallsgröße** (**zufällige Veränderliche**, **Zufallsvariable**), wenn das Urbild jedes reellen offenen Intervalls I der Form $(-\infty, x)$ ein zufälliges Ereignis ist: $X^{-1}(I) \in \mathcal{A}$. Auf den Intervallen $(-\infty, x)$ führen wir ein Wahrscheinlichkeitsmaß P^* ein, indem wir $P^*(I) = P(A)$ setzen, falls A das Urbild des Intervalls I ist. Damit können wir $P^*(I)$ als die Wahrscheinlichkeit dafür interpretieren, daß die Zufallsgröße X einen Wert aus dem Intervall I annimmt. Jedes halboffene Intervall $I = [a, b)$ mit $a < b$ kann man als Differenz der Intervalle $(-\infty, b)$ und $(-\infty, a)$ darstellen, so daß auch für solche Intervalle die Wahrscheinlichkeit $P^*(I)$ erklärt ist. Anstelle von $P^*(I)$ mit $I = (-\infty, x)$ schreiben wir $P(X < x)$ und bei $I = [a, b)$ analog $P(a \leqq X < b)$. Entsprechend bedeutet $P(X = a)$ die Wahrscheinlichkeit, daß die Zufallsgröße X den Wert a annimmt. Betrachten wir das Würfeln. Die Elementarereignisse sind hier durch die Zahlen $1, 2, 3, 4, 5, 6$ repräsentiert. Alle Ereignisse haben die gleiche Wahrscheinlichkeit; da wir alle möglichen Versuchsausgänge erfaßt haben, hat jede Zahl die Wahrscheinlichkeit $\frac{1}{6}$. Unsere Zufallsgröße X kann hier die 6 Werte $x_i = i \quad (i = 1, \ldots, 6)$ annehmen und es gilt $P(X = x_i) = \frac{1}{6}$. Offenbar ist $P(X < 1) = 0$. Für $1 < x \leqq 2$ ist

$$P(X < x) = P(X = 1) = \frac{1}{6}$$

und im Falle $2 < x \leqq 3$ erhält man

$$P(X < x) = P(X = 1) + P(X = 2) = \frac{1}{3}.$$

Schließlich folgt für $5 < x \leqq 6$:

$$P(X < x) = \sum_{i=1}^{5} P(X = i) = \frac{5}{6}$$

und für $x > 6$:

$$P(X < x) = P(X \leqq 6) = \sum_{i=1}^{6} P(X = i) = 1.$$

Wir sehen an diesem Beispiel, daß die Wahrscheinlichkeit $P(X < x)$ mit x monoton wächst. Die durch die Formel

$$F(x) = P(X < x)$$

definierte Funktion heißt **Verteilungsfunktion** der Zufallsgröße X. Kennt man die Verteilungsfunktion, so kann man alle Wahrscheinlichkeiten berechnen. So ist z. B.

$$P(a \leqq X < b) = F(b) - F(a).$$

Satz 129. *Die Verteilungsfunktion F einer Zufallsgröße X hat folgende Eigenschaften:*

1. $$0 \leqq F(x) \leqq 1 \quad \forall x \in \mathbb{R}.$$

2. *F ist monoton wachsend: Aus $x < y$ folgt $F(x) \leqq F(y)$.*

3. *F ist linksseitig stetig:*

$$\lim_{\substack{x \to a \\ x < a}} F(x) = F(a).$$

4. $$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1.$$

5. $$P(X = a) = \lim_{\substack{x \to a \\ x < a}} F(x) - F(a) = F(a^+) - F(a).$$

Beweis. Wir zeigen, daß die Verteilungsfunktion monoton wächst. Es sei $x < y$; dann liegt das Intervall $(-\infty, x)$ im Intervall $(-\infty, y)$, so daß $P(X < x) \leqq P(X < y)$ gilt, also $F(x) \leqq F(y)$.
Um die linksseitige Stetigkeit zu beweisen, sei (x_n) eine beliebige monoton wachsende, gegen a konvergente Folge; mit A_n bezeichnen wir das Ereignis, das die Zufallsgröße X einen Wert aus dem Intervall $[x_n, a)$ annimmt. Der Grenzwert a gehört zu keinem der betrachteten Intervalle. Also ist es unmöglich, daß die Zufallsgröße X einen zu allen Intervallen gehörenden Wert annimmt. Daher ist

$$A = \bigcap_{n=1}^{\infty} A_n$$

ein unmögliches Ereignis, also $P(A) = 0$. Die Ereignisfolge (A_n) ist monoton fallend:

$$A_{n+1} \subseteq A_n, \quad n = 1, 2, \ldots,$$

wodurch wir mit der Oberhalbstetigkeit von P erhalten

$$\begin{aligned} 0 = P(A) &= \lim_{n \to \infty} P(A_n) = \lim_{n \to \infty} P(x_n \leqq X < a) = \lim_{n \to \infty} (F(a) - F(x_n)) \\ &= F(a) - \lim_{n \to \infty} F(x_n). \end{aligned}$$

Die übrigen Eigenschaften sind offensichtlich. □
Ist der Wertebereich einer Zufallsgröße X höchstens abzählbar, so nennt man die Zufallsgröße **diskret**. Die Funktionswerte einer diskreten Zufallsgröße lassen sich indizieren: $x_1, x_2, \ldots$. Für die vollständige Beschreibung einer diskreten Zufallsgröße X braucht man noch für jedes n die Wahrscheinlichkeit p_n, mit der der Wert x_n angenommen wird: $p_n = P(X = x_n)$. Die Größe p_n nennt man **Einzelwahrscheinlichkeit**. Die zugehörige Verteilungsfunktion ist dann eine Treppenfunktion und hat die Form

$$F(x) = \sum_{x_n < x} p_n.$$

Eine Zufallsgröße X heißt **stetige Zufallsgröße**, wenn eine nichtnegative, stückweise stetige Funktion f existiert, so daß sich die Verteilungsfunktion F von X in der Form

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

darstellen läßt. Der Integrand f heißt dann **Dichtefunktion** oder einfach **Dichte** der Zufallsgröße X. Mit dem Hauptsatz der Differential- und Integralrechnung folgt sofort

Satz 130. *Sind f die Dichtefunktion und F die Verteilungsfunktion einer stetigen Zufallsgröße X, so gilt*

- $$\int_{-\infty}^{+\infty} f(x)dx = 1,$$
- $$P(a \leqq X < b) = F(b) - F(a) = \int_{a}^{b} f(x)dx,$$
- $$F'(x) = f(x) \quad \forall x.$$

Es sei bemerkt, daß jede nichtnegative, stückweise stetige Funktion f, die für alle reellen Zahlen erklärt ist und die die erste Eigenschaft des letzten Satzes hat, als Dichtefunktion einer stetigen Zufallsfröße fungieren kann.
Beispiel. Es sei f wie folgt definiert:

$$f(x) = \begin{cases} 0 & \text{für } x < 0, \\ \dfrac{x}{2} & \text{für } 0 \leqq x \leqq 2, \\ 0 & \text{für } x > 2. \end{cases} .$$

Die Verteilungsfunktion F der Zufallsgröße X mit dieser Dichtefunktion lautet offenbar:

$$F(x) = \begin{cases} 0 & \text{für } x < 0, \\ \dfrac{x^2}{4} & \text{für } 0 \leqq x \leqq 2, \\ 1 & \text{für } x > 2. \end{cases} .$$

Bei einer stetigen Zufallsgröße X gilt offenbar

$$P(X = a) = \int_{a}^{a} f(x)dx = 0,$$

und jede reelle Zahl ist ein Elementarereignis. Für das Ereignis $\mathbb{R} \setminus \{a\}$ folgt

$$P(\mathbb{R} \setminus \{a\}) = 1,$$

obwohl das Ereignis, das der Zufallsgröße X einen Wert aus $\mathbb{R} \setminus \{a\}$ zuweist, nicht das sichere Ereignis ist. Allgemein muß man daraus schlußfolgern: Wenn bei einer stetigen Zufallsgröße die Wahrscheinlichkeit eines gewissen Ereignisses gleich 0 ist, so kann man dieses nicht als unmögliches Ereignis ansehen, sondern muß es als ein Ereignis betrachten, dessen Eintreten sehr wenig wahrscheinlich ist. Ist andererseits bei einer stetigen Zufallsgröße die Wahrscheinlichkeit eines Ereignisses gleich 1, so kann man es als sehr wahrscheinlich ansehen, jedoch nicht als sicher.
In vielen praktischen Anwendungen ist man nur daran interessiert, gewisse prinzipielle Aussagen über die Verteilung einer Zufallsgröße zu machen. Dies geschieht durch verschiedene quantitative Kenngrößen der Verteilungsfunktion. Eine solche Kenngröße ist ihr **Mittelwert**. Dazu ein Beispiel. Hat man keine Ahnung vom Fußballspiel, so tippt man beim Fußballtoto jeden Spielausgang der 12 Spiele mit der Wahrscheinlichkeit $\frac{1}{3}$ richtig, so daß man etwa mit $12 \cdot \frac{1}{3} = 4$ richtigen Tips rechnen kann und jedes andere Ergebnis ist weniger wahrscheinlich: 4

richtige Tips ist der Erwartungswert beim ahnungslosen Totospieler. Dieser Wert verschiebt sich sofort, wenn der Spieler z. B. weiß, daß Heimspiele häufiger als Auswärtsspiele gewonnen werden. Allgemein sagen wir: Bei einer diskreten Zufallsgröße X mit den Werten x_n und den Einzelwahrscheinlichkeiten p_n nennt man die Größe

$$\mu = E(X) = \sum_n p_n x_n$$

den **Erwartungswert** bzw. **Mittelwert** der Zufallsgröße X. Dabei muß im Falle abzählbar vieler Werte x_n gefordert werden, daß die Reihe absolut konvergiert; andernfalls existiert der Erwartungswert nicht. Diese Einschränkung folgt aus folgender Überlegung: Die Numerierung der Einzelereignisse x_n ist willkürlich; also darf sich der Erwartungswert beim Umnumerieren der Reihenglieder nicht ändern, d. h. die Reihe muß unbedingt konvergieren. In der Analysis haben wir gelernt, daß diese Forderung gleichwertig mit der absoluten Konvergenz der Reihe ist.
Analog definiert man bei einer stetigen Zufallsgröße X mit der Dichte f den Erwartungswert als

$$\mu = E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx,$$

sofern das uneigentliche Integral absolut konvergiert, d. h.

$$\int_{-\infty}^{+\infty} |x| \cdot f(x)dx < \infty.$$

Beispiele. Ist X jene Zufallsgröße, die jedem Würfeln die gewürfelte Augenzahl zuordnet, so gilt $p_n = \frac{1}{6}, x_n = n, n = 1, \ldots, 6$; also erhalten wir

$$\mu = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3,5.$$

Dieses Beispiel zeigt uns zusätzlich, daß der Erwartungswert im allgemeinen kein Wert ist, der von der Zufallsgröße angenommen werden kann.
Die Zufallsgröße X möge die Werte

$$x_n = \frac{(-2)^n}{n}, \quad n = 1, 2 \ldots$$

mit den Wahrscheinlichkeiten

$$p_n = \frac{1}{2^n}, \quad n = 1, 2 \ldots$$

annehmen. Dann folgt

$$\sum_{n=1}^{\infty} p_n x_n = \sum_{n=1}^{\infty} \frac{(-1)^n}{n} = \ln 2;$$

jedoch existiert der Erwartungswert nicht, da die Reihe nicht absolut konvergiert.
Eine kleine Rechnung zeigt, daß man jeden existierenden Erwartungswert durch Transformieren der Zufallsgröße auf den Wert 0 einstellen kann. Ist nämlich X eine Zufallsgröße, so auch $Y = aX + b$, wo a, b reelle Zahlen sind. Im diskreten Fall folgt

$$\begin{aligned} E(Y) &= \sum_n p_n y_n = \sum_n p_n (a x_n + b) = a \sum_n p_n x_n + b \sum_n p_n \\ &= aE(X) + b \end{aligned}$$

und im stetigen Falle zeigen wir

$$E(Y) = \int_{-\infty}^{+\infty} (ax + b) f(x)dx = aE(X) + b.$$

Dazu seien F_Y, f_Y die Verteilungs- und die Dichtefunktionen der Zufallsgröße Y sowie F, f die Verteilungs-und Dichtefunktionen der Zufallsgröße X und $a > 0$. Wir erhalten

$$F_Y(t) = P(Y < t) = P(aX + b < t) = P\left(X < \frac{t-b}{a}\right) = F\left(\frac{t-b}{a}\right)$$

woraus für die Dichtefunktionen

$$f_Y(t) = f\left(\frac{t-b}{a}\right)\frac{1}{a}$$

folgt. Damit schließen wir

$$\begin{aligned} E(Y) &= \int_{-\infty}^{+\infty} y f_Y(y)dy = \int_{-\infty}^{+\infty} y f\left(\frac{y-b}{a}\right)\frac{1}{a}dy = \int_{-\infty}^{+\infty}(ax+b)f(x)dx \\ &= a\int_{-\infty}^{+\infty} x f(x)dx + b\int_{-\infty}^{+\infty} f(x)dx = aE(X)+b \end{aligned}$$

Setzen wir speziell $a = 1, b = -E(X)$, so folgt

$$E(Y) = E(X - E(X)) = 0$$

und man nennt den Übergang von der Zufallsgröße X zur Zufallsgröße $X - E(X)$ **Zentrieren** der Zufallsgröße X. Insbesondere lernen wir hieraus, daß verschiedene Zufallsgrößen den gleichen Erwartungswert haben können; man benötigt also ein Maß, das die Abweichung der Werte von X vom Erwartungswert ausdrückt. Natürlich sollen alle Abweichungen gleichbehandelt werden. Hierfür kann man die quadratische Abweichung verwenden. Es sei X eine Zufallsgröße mit dem Erwartungswert $\mu = E(X)$. Dann heißt im Falle einer diskreten Zufallsgröße die Zahl

$$\sigma^2 = V(X) = \sum_n (x_n - \mu)^2 p_n$$

Streuung oder **Varianz** von X. Bei einer stetigen Zufallsgröße X mit der Dichte f und dem Erwartungswert μ lautet die Streuung

$$\sigma^2 = V(X) = \int_{-\infty}^{+\infty}(x-\mu)^2 f(x)dx.$$

Die Wurzel σ aus der Streuung nennt man **Standardabweichung** von X. Aus den Rechenregeln für unendliche Reihen schließen wir bei einer diskreten Zufallsgröße

$$\begin{aligned} \sigma^2 &= \sum_n (x_n-\mu)^2 p_n = \sum_n x_n^2 p_n - 2\mu\sum_n x_n p_n + \mu^2\sum_n p_n \\ &= E(X^2) - \mu^2 \end{aligned}$$

und bei einer stetigen Zufallsgröße X mit der Dichte f:

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{+\infty}(x-\mu)^2 f(x)dx \\ &= \int_{-\infty}^{+\infty} x^2 f(x)dx - 2\mu\int_{-\infty}^{+\infty} x f(x)dx + \mu^2\int_{-\infty}^{+\infty} f(x)dx \\ &= \int_{-\infty}^{+\infty} x^2 f(x)dx - \mu^2 = E(X^2) - E(X)^2, \end{aligned}$$

womit wir zusammen den folgenden Satz gewonnen haben.

Satz 131. *Für jede Zufallsgröße X mit dem Erwartungswert $E(X)$ und der Varianz $V(X)$ gilt:*

$$V(X) = E(X^2) - E(X)^2.$$

Untersuchen wir weiter, wie sich die Varianz gegenüber einer linearen Transformation der Zufallsgröße verhält:

$$\begin{aligned} V(aX+b) &= \sum_n \left(ax_n + b - E(aX+b)\right)^2 p_n \\ &= \sum_n \left(ax_n + b - aE(X) - b\right)^2 p_n \\ &= a^2\sum_n \left(x_n - E(X)\right)^2 p_n \\ &= a^2 V(X). \end{aligned}$$

Analoges rechnet man für eine stetige Zufallsgröße aus. Folglich gilt der nächste Satz.

Satz 132. *Ist X eine Zufallsgröße mit der Varianz $V(X)$, so gilt für beliebige reelle Zahlen a, b:*

$$V(aX + b) = a^2 V(X).$$

Insbesondere ist also $V(-X) = V(X)$ und $V(X + b) = V(X)$. Die Streuung ist somit symmetrisch und unempfindlich gegenüber einer Parallelverschiebung. Außerdem folgt

$$V(\frac{X}{\sigma}) = 1.$$

Den Übergang von der Zufallsgröße X zur Zufallsgröße Y:

$$X \quad \Longrightarrow \quad Y = \frac{X}{\sigma} \text{ mit } V(Y) = 1$$

nennt man **Normierung** der Zufallsgröße X. Wenn wir das Zentrieren hinzunehmen, nennt man den Übergang

$$X \quad \Longrightarrow \quad \frac{X - \mu}{\sigma}$$

Standardisierung; die neue Zufallsgröße heißt **standardisierte Zufallsgröße**; sie hat den Erwartungswert 0 und die Streuung 1.

Satz 133 (Tschebyscheff-Ungleichung). *Für jede Zufallsgröße X mit dem Erwartungswert μ und der Sreuung σ^2 gilt bei beliebig gewähltem $\varepsilon > 0$ die Ungleichung*

$$P(|X - \mu| \geqq \varepsilon) \leqq \frac{\sigma^2}{\varepsilon^2}.$$

Beweis. Zunächst wird die folgende Aussage bewiesen: Wenn die Zufallsgröße Y mit dem Erwartungswert $E(Y)$ nur nichtnegative Werte annimmt, so gilt für jedes $\alpha > 0$ die Ungleichung

$$P(Y \geqq \alpha) \leqq \frac{E(Y)}{\alpha}.$$

Für diskretes Y mit den Werten y_n und den Einzelwahrscheinlichkeiten p_n folgt die behauptete Ungleichung aus

$$E(Y) = \sum_n y_n p_n \geqq \sum_{n: y_n \geqq \alpha} y_n p_n \geqq \alpha \sum_{n: y_n \geqq \alpha} p_n = \alpha P(Y \geqq \alpha).$$

Für stetiges Y mit der Dichte f ergibt sich:

$$E(Y) = \int_{-\infty}^{+\infty} y \cdot f(y)dy \geqq \int_{\alpha}^{+\infty} y \cdot f(y)dy \geqq \alpha \int_{\alpha}^{+\infty} f(y)dy = \alpha P(Y \geqq \alpha).$$

Wir setzen nun $\alpha = \varepsilon^2$ und $Y = (X - E(X))^2$; dann ist

$$E(Y) = E\left((X - E(X))\right) = V(X)$$

und die obige Ungleichung liefert

$$P\left((X - E(X))^2 \geqq \varepsilon^2\right) \leqq \frac{V(X)}{\varepsilon^2},$$

was mit der Behauptung übereinstimmt, da die beiden Ereignisse

$$(X - E(X))^2 \geqq \varepsilon^2 \text{ und } |X - E(X)| \geqq \varepsilon$$

die gleichen sind. □

Setzt man in der Tschebyscheff-Ungleichung $\varepsilon = n\sigma$, so erhält man die Form

$$P(|X - E(X)| \geqq n\sigma) \leqq \frac{1}{n^2}.$$

Für n=4 folgt daraus z. B.

$$P(|X - \mu| < 4\sigma) \geqq 1 - \frac{1}{16} = \frac{15}{16} = 0,9375,$$

was man im Falle $\sigma = 1$ so lesen kann: Jede Zufallsgröße X nimmt mit mindestens der Wahrscheinlichkeit 0,9375 nur Werte an, deren Abstände vom Erwartungswert kleiner als 4 sind.

Abschließend soll noch eingeführt werden, was man unter unabhängigen Zufallsgrößen versteht. Zwei Zufallsgrößen X, Y heißen **unabhängig**, wenn die sie repräsentierenden zufälligen Ereignisse unabhängig sind. Ist daher A das Urbild eines Intervalls I bei der Zufallsgröße X und B das Urbild bei der Zufallsgröße Y, so gilt bei unabhängigen Zufallssgrößen stets $P(A \cap B) = P(A)P(B)$.

5.1.3. Einige diskrete Verteilungen

Die wohl einfachste diskrete Verteilung ist die diskrete **Gleichverteilung** oder **gleichmäßige Verteilung**. Bei dieser Verteilung nimmt die Zufallsgröße X nur endlich viele Werte $x_1, x_2, \ldots, x_n$ an und jeden mit der gleichen Wahrscheinlichkeit:

$$p_i = P(X = x_i) = \frac{1}{n}, \quad i = 1, \ldots, n.$$

Bei den meisten Glücksspielen liegt eine solche Verteilung vor. Für den Erwartungswert und die Varianz folgt hier:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2.$$

Der Erwartungswert einer gleichmäßig verteilten Zufallsgröße ist also das arithmetische Mittel der möglichen Werte.
Eine weitere diskrete Verteilung erhalten wir bei der Betrachtung des folgenden Urnenmodells. Wir betrachten einen zufälligen Versuch, bei dem der Versuchsausgang für jede Wiederholung unabhängig von den bereits durchgeführten Versuchen ist. Also etwa das Ziehen einer gewissen Kugelanzahl aus einer Urne. Ein gewisses Ereignis A möge mit der Wahrscheinlichkeit p als Versuchsausgang eintreten: $P(A) = p$. Dann tritt das Ereignis $\overline{A}$ mit der Wahrscheinlichkeit $1 - p$ ein. Die n-malige Wiederholung des Versuches liefert uns ein n-Tupel aus den Ereignissen A und $\overline{A}$; jedes solche n-Tupel repräsentiert eine Versuchsserie aus n Wiederholungen. Alle möglichen, aus n Wiederholungen bestehenden Versuchsserien werden also durch alle n-Tupel, die aus A und $\overline{A}$ bestehen, charakterisiert. Von diesen n-Tupeln gibt es genau $\binom{n}{r}$, in denen das Ereignis A genau r-mal auftritt. Jedes n-Tupel hat die gleiche Wahrscheinlichkeit, als Resultat einer Versuchsserie aufzutreten. Enthält ein n-Tupel genau r-mal das Ereignis A, so enthält es genau $(n - r)$mal das Ereignis $\overline{A}$. Das Auftreten von A und $\overline{A}$ bei n-maliger Wiederholung sind unabhängige Ereignisse, so daß sich die Wahrscheinlichkeiten multiplizieren. Also hat ein n-Tupel von Ereignissen A und $\overline{A}$, in dem r-mal das Ereignis A auftritt, die Wahrscheinlichkeit $p^r(1-p)^{n-r}$, um als Resultat einer Versuchsserie aus n Wiederholungen aufzutreten. Es sei nun X die Anzahl der Ereignisse A in einem n-Tupel, also die absolute Häufigkeit des Eintretens von A bei einer Versuchsserie von n Wiederholungen; X ist dann eine Zufallsgröße und kann die Werte $0, 1, 2, \ldots, n$ annehmen. Nach den obigen Überlegungen ist

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r}.$$

Als konkretes Beispiel nehmen wir wie angekündigt das Ziehen von Kugeln aus einer Urne. In der Urne mögen N Kugeln liegen, R davon seien rot und nach dem Ziehen wird die Kugel zurückgelegt. Ist X die Anzahl der roten Kugeln unter n zufällig gezogenen, so sei A das Ereignis, eine rote Kugel zu ziehen. Dieses Ereignis hat offenbar die Wahrscheinlichkeit $p = \frac{R}{N}$.
Allgemein sagen wir, daß eine diskrete Zufallsgröße X, die die Werte $0, 1, 2, \ldots, n$ annehmen kann, einer **Binomialverteilung** genügt, wenn

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r}, \quad r = 0, 1, 2, \ldots, n$$

gilt. Die Binomialverteilung hängt von den beiden Parametern n und p ab. Aus

$$\sum_{r=0}^{n} P(X = r) = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} = (p + 1 - p)^n = 1$$

ergibt sich, daß tatsächlich eine Verteilung vorliegt.
Eine typische Anwendung für die Binomialverteilung ist die folgende. Für eine Ware sei bekannt, daß sich in einem hinreichend großen Warenposten ungefähr $p \cdot 100\%$ Ausschuß befindet. Die Anzahl der zum Ausschuß gehörenden Einzelstücke bei einer zufällig entnommenen Stichprobe vom Umfang n ist dann eine Zufallsgröße mit einer Binomialverteilung und den Parametern n und p. Um die Unabhängigkeit einer Warenentnahme von den vorangegangenen Entnahmen zu sichern, muß der Warenposten sehr groß gegenüber dem Stichprobenumfang n sein oder aber man legt jedes entnommene Stück nach der Prüfung zurück.
Erwartungswert und Varianz lassen sich hier leicht berechnen.

Satz 134. *Für eine binomialverteilte Zufallsgröße X mit den Parametern n und p gilt:*

$$E(X) = n \cdot p, \quad V(X) = n \cdot p(1-p).$$

*Beweis.*Der Beweis erfolgt durch Ausrechnen:

$$\begin{aligned}E(X) &= \sum_{r=0}^{n} r \binom{n}{r} p^r (1-p)^{n-r} = \sum_{r=1}^{n} r \binom{n}{r} p^r (1-p)^{n-r} \\ &= \sum_{r=1}^{n} n \binom{n-1}{r-1} p^r (1-p)^{n-r} \\ &= np \sum_{r=1}^{n} \binom{n-1}{r-1} p^{r-1} (1-p)^{n-1-(r-1)} \\ &= np\,(p+(1-p))^{n-1} = np.\end{aligned}$$

Analog berechnet man die Varianz. □

Eine binomialverteilte Zufallsgröße X mit den Parametern n und p kann nach den obigen Überlegungen als absolute Häufigkeit interpretiert werden; also ist $Y = \frac{1}{n}X$ die relative Häufigkeit und ebenfalls eine Zufallsgröße. Nach den Rechenregeln für Erwartungswert und Varianz bei einer linearen Transformation der Zufallsgröße folgt:

$$E(Y) = p, \quad V(Y) = \frac{1}{n}p(1-p).$$

Der Erwartungswert der relativen Häufigkeit ist somit die Wahrscheinlichkeit p selbst; außerdem folgt aus dem Wert der Varianz, daß die Abweichung vom Erwartungswert mit wachsendem n beliebig klein und sehr selten wird.

Als dritte diskrete Verteilung betrachten wir die **Poissonverteilung**. Eine diskrete X, die jede natürliche Zahl als Wert annehmen kann, heißt **poissonverteilt** mit dem Parameter λ, $\lambda > 0$, wenn

$$P(X = r) = \frac{\lambda^r}{r!} e^{-\lambda}, \quad r = 0, 1, 2, \ldots$$

gilt. Durch Bildung der entsprechenden unendlichen Reihe überzeugen wir uns davon, daß wirklich eine Verteilung vorliegt:

$$\sum_{r=0}^{\infty} \frac{\lambda^r}{r!} e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1.$$

Diese Verteilung hat große praktische Bedeutung, da man bei vielen Zufallsgrößen eine Poissonverteilung annehmen kann. Dazu einige

Beispiele.

- Die Anzahl der Anrufe, die in einem gegebenen Zeitintervall in einer Zentrale eintreffen. Allgemein bei Bedienungssystemen: Die Anzahl der Kunden, die in einer gegebenen Zeiteinheit vor einem Bedienungssytem auf eine Bedienung warten.
- Die Anzahl des Eintretens eines Ereignisses A mit kleiner Wahrscheinlichkeit p bei einer sehr großen Zahl von Wiederholungen des entsprechenden Versuches.
- Die Anzahl der Atome eines radioaktiven Materials, die in einer gegebenen Zeiteinheit zerfallen.
- In einer Telefonzentrale mögen durchschnittlich 10 Anrufe pro Minute eintreffen. Dann ist $\lambda = 10$ und für die Wahrscheinlichkeit, daß in einer Minute mehr als 2 Anrufe eingehen, ergibt sich

$$\begin{aligned}P(X > 2) &= 1 - P(X \leqq 2) \\ &= 1 - (P(X=0) + P(X=1) + P(X=2)) \\ &= 1 - \frac{10^0}{0!}e^{-10} - \frac{10^1}{1!}e^{-10} - \frac{10^2}{2!}e^{-10} \\ &\approx 0,997.\end{aligned}$$

Satz 135. *Für eine poissonverteilte Zufallsgröße X mit dem Parameter λ gilt*

$$E(X) = \lambda, \quad V(X) = \lambda.$$

Beweis. Wir berechnen nur den Erwartungswert, da sich die Varianz analog ausrechnen läßt:

$$\begin{aligned}E(X) &= \sum_{r=0}^{\infty} P(X=r) r = e^{-\lambda} \sum_{r=0}^{\infty} \frac{\lambda^r}{r!} r \\ &= \lambda e^{-\lambda} \sum_{r=1}^{\infty} \frac{\lambda^{r-1}}{(r-1)!} = \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda.\end{aligned}$$

□

Wir beweisen nun einen wichtigen Zusammenhang zwischen der Binomial- und der Poissonverteilung.

Satz 136 (Grenzwertsatz von Poisson). *Für alle r ($r = 0, 1, 2, \ldots$) und ein beliebig fixiertes $\lambda > 0$ gilt:*

$$\lim_{n\to\infty} \binom{n}{r} \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r} = \frac{\lambda^r}{r!} e^{-\lambda}.$$

Beweis. Das links stehende Folgeglied schreiben wir in der Form

$$\frac{n(n-1)\cdots(n-r+1)}{n^r} \frac{\lambda^r}{r!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-r}.$$

Der erste Faktor strebt für $n \to \infty$ gegen 1, der dritte gegen $e^{-\lambda}$ und der vierte gegen 1, so daß die Behauptung schon bewiesen ist. □

Der Inhalt dieses Satzes soll nun interpretiert werden. Die Glieder der Folge

$$\binom{n}{r} \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r}$$

sind für fixiertes r bei $n \geqq r$ erklärt. Ist nun $a(r, n, p)$ die r-te Einzelwahrscheinlichkeit einer binomialverteilten Zufallsgröße mit den Parametern n und p, $b(r, \lambda)$ die r-te Einzelwahrscheinlichkeit einer poissonverteilten Zufallsgröße mit dem Parameter $\lambda = np$, so folgt aus dem Grenzwertsatz, daß für große n beide näherungsweise übereinstimmen:

$$a(r, n, p) \approx b(r, \lambda).$$

Die Annäherung ist bereits für $n > 10$ und kleine Zahlen p für praktische Zwecke völlig ausreichend. Diese Tatsache ist praktisch wichtig, da die Werte $b(r, \lambda)$ in Tabellen vorliegen, während $a(r, n, p)$ für große n schlecht berechnet werden kann.
Aus der Definition der Poissonverteilung ergeben sich die beiden folgenden Rekursionformeln:

$$b(r+1, \lambda) = \frac{\lambda}{r+1} b(r, \lambda), \quad r \geqq 0,$$

$$b(r-1, \lambda) = \frac{r}{\lambda} b(r, \lambda), \quad r \geqq 1,$$

die man vorteilhaft für nicht zu große r verwenden kann.
Abschließend wollen wir noch zusammenstellen, wie sich diskrete Zufallsgrößen bei Addition verhalten.

Satz 137. *Die diskreten, unabhängigen Zufallsgrößen X, Y seien binomialverteilt mit den Parametern n, p bzw. m, p. Dann ist die Summe $X + Y$ binomialverteilt mit den Parametern $n + m$ und p.*

Anschaulich kann man diese Aussage so interpretieren. Es sei X eine Zufallsgröße, die das Eintreten eines Ereignisses A mit $P(A) = p$ bei n-maliger Wiederholung beschreibt; entsprechend Y bei m-maliger Wiederholung. Dann gehört $X + Y$ offenbar zur $(n + m)$-maligen Wiederholung.

Satz 138. *Die diskreten, unabhängigen Zufallsgrößen X, Y seien poissonverteilt mit den Parametern λ, ϱ. Dann ist die Summe poissonverteilt mit dem Parameter $\lambda + \varrho$.*

Die beiden letzten Aussagen können durch Ausrechnen verifiziert werden.

5.1.4. Einige stetige Verteilungen

Die einfachste stetige Verteilung ist die stetige **Gleichverteilung** oder **Rechteckverteilung** . Eine stetige Zufallsgröße X heißt **gleichverteilt** mit den Parametern a und h ($h > 0$) – kurz $R(a-h, a+h)$-verteilt –, wenn ihre Dichtefunktion f die folgende Form hat:

$$f(x) = \begin{cases} \dfrac{1}{2h} & a - h \leqq x \leqq a + h \\ 0 & \text{sonst.} \end{cases}$$

Wegen

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{2h} \int_{a-h}^{a+h} dx = 1$$

liegt eine Verteilung vor. Wir berechnen die Verteilungsfunktion. Für $x < a - h$ ist offenbar $F(x) = 0$ und für $x > a + h$ ist $F(x) = 1$. Für $a - h \leqq x \leqq a + h$ folgt

$$F(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{2h} \int_{a-h}^{x} dt = \frac{x - (a-h)}{2h},$$

also zusammen

$$F(x) = \begin{cases} 0 & x < a - h \\ \dfrac{x-(a-h)}{2h} & a - h \leqq x \leqq a + h \\ 1 & x > a + h. \end{cases}$$

Der Erwartungswert ergibt sich zu

$$\mu = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{2h} \int_{a-h}^{a+h} xdx = \frac{1}{2h} \frac{(a+h)^2 - (a-h)^2}{2} = a$$

und

$$E(X^2) = \frac{1}{2h} \int_{a-h}^{a+h} x^2 dx = \frac{1}{2h} \frac{(a+h)^3 - (a-h)^3}{3} = \frac{3a^2 + h^2}{3},$$

woraus für die Varianz folgt:

$$\sigma^2 = E(X^2) - (E(X))^2 = \frac{3a^2 + h^2}{3} - a^2 = \frac{h^2}{3}.$$

Wir fassen alles in einem Satz zusammen.

Satz 139. *Eine rechteckverteilte stetige Zufallsgröße X mit den Parametern a und h hat den Erwartungswert a und die Varianz $\frac{h^2}{3}$. Die transformierte Zufallsgröße*

$$Y = \frac{X - (a-h)}{2h}$$

ist $R(0,1)$-verteilt mit der Dichte

$$f(y) = \begin{cases} 1 & 0 \leqq y \leqq 1 \\ 0 & \textit{sonst,} \end{cases}$$

dem Erwartungswert $\frac{1}{2}$ und der Varianz $\frac{1}{12}$.

Eine wichtige Bedeutung erhält die $R(0,1)$-Verteilung durch den folgenden Umstand.

Satz 140. *Es seien X eine stetige Zufallsgröße mit der Verteilungsfunktion F und Y jene stetige Zufallsgröße, die den Wert $F(x)$ annimmt, wenn X den Wert x annimmt, kurz als $Y = F(X)$ geschrieben. Dann ist Y eine $R(0,1)$-verteilte Zufallsgröße.*

Beweis. Jedem Werteintervall $(-\infty, x)$ der Zufallsgröße X entspricht eine Wertemenge der Zufallsgröße Y, die im Intervall $[0, F(x)]$ mit $F(x) \leqq 1$ liegt. Andererseits entspricht jedem $y \in [0,1]$ ein Wert x, der die Beziehung $y = F(x) = P(X < x)$ erfüllt. Diese Transformation ist umkehrbar eindeutig, wenn F streng monoton wächst. Im allgemeinen wird $F^{-1}(y)$ für gewisse y ein Intervall sein, in dem die Verteilungsfunktion F konstant ist. Ist nun F_1 die Verteilungsfunktion von Y, so erhalten wir

$$F_1(y) = P(Y < y) = P(F(X) < y) = P(X < F^{-1}(y)) = F(F^{-1}(y)) = y$$

für $y \in [0,1]$ und $F_1(y) = 0$ für $y < 0$, sowie $F_1(y) = 1$ für $y > 1$; damit

$$F_1'(y) = f_1(y) = \begin{cases} 1 & 0 \leqq y \leqq 1 \\ 0 & \text{sonst.} \end{cases} \qquad \square$$

Dieser Satz zeigt uns, daß man prinzipiell aus $R(0,1)$-verteilten Zufallsgrößen mittels geeigneter Transformationen Zufallszahlen mit anderen Verteilungen berechnen kann.

Eine stetige Zufallsgröße X unterliegt einer **Exponentialverteilung** mit dem Parameter α ($\alpha > 0$), wenn ihre Dichtefunktion f die Form

$$f(x) = \begin{cases} 0 & x \leqq 0 \\ \alpha \cdot e^{-\alpha x} & x > 0 \end{cases}$$

hat. Durch Integration überzeugt man sich sofort, daß eine Verteilung vorliegt. Für die Verteilungsfunktion F folgt:

$$F(x) = \begin{cases} 0 & x \leqq 0 \\ 1 - e^{-\alpha x} & x > 0. \end{cases}$$

Satz 141. *Eine exponentialverteilte Zufallsgöße mit dem Parameter α hat den Erwartungswert $\frac{1}{\alpha}$ und die Varianz $\frac{1}{\alpha^2}$.*

Beweis. Der Beweis erfolgt durch direktes Ausrechnen:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx = \alpha \int_{0}^{\infty} x e^{-\alpha x} dx = [-x e^{-\alpha x}]_0^{\infty} + \int_{0}^{\infty} e^{-\alpha x} dx \\ &= [0 - \frac{1}{\alpha} e^{-\alpha x}]_0^{\infty} = \frac{1}{\alpha}. \end{aligned}$$

Analog berechnet man die Varianz. □

Die Erfahrung zeigt, daß viele zufallsabhängige Zeiten einer Exponentialverteilung unterliegen. Folgende Größen sind meist exponentialverteilt:

- die Dauer eines Telefonanrufes,
- die Dauer einer Reparatur, einer Bedienung,
- Zeitdifferenzen, die keinen vorhersagbaren Wert haben, wie z. B. die Zeit zwischen zwei Ausfällen eines Rechners oder die Zeit zwischen zwei ankommenden Nachrichten.

Die einfache Formel für den Erwartungswert erlaubt es, empirisch eine Näherung für den Parameter α einer exponentialverteilten Zufallsgröße zu ermitteln. Ist etwa X die zufällige Zeit zwischen zwei Rechnerstörungen, so mißt man diese hinreichend oft und bildet über die Meßwerte den Mittelwert. Als Parameter α kann man dann näherungsweise den reziproken Wert davon nehmen.
Beispiel. Die Zufallsgröße X beschreibe die Laufzeit eines Rechners zwischen zwei Störungen. Aus einer längeren Meßreihe möge man wissen, daß der Rechner durchschnittlich 2 Stunden störungsfrei läuft; daraus erhält man $\alpha = \frac{1}{2}$. Die Wahrscheinlichkeit, daß der Rechner mehr als 3 Stunden störungsfrei läuft, beträgt dann

$$P(X > 3) = 1 - P(X \leqq 3) = 1 - (1 - e^{-0,5 \cdot 3}) \approx 0,3232.$$

Natürlich ist dieser Wert unrealistisch, wenn ein Eingriff in die Funktionsweise des Rechners vorgenommen wurde.
Zwischen der Poisson- und der Exponentialverteilung besteht in den Anwendungen oft ein inniger Zusammenhang: So ist die Anzahl der Programme, die in einer Stapelmaschine auf ihren Start warten, meist poisson- und die Abarbeitungszeit exponentialverteilt. Zusammen ergibt sich die Gesamtbearbeitungszeit für ein Programm.
Die wohl wichtigste Verteilung ist die **Normalverteilung**. Eine stetige Zufallsgröße X nennt man **normalverteilt** mit den positiven Parametern μ, σ – kurz $N(\mu, \sigma)$-verteilt –, wenn die Dichte φ von X die folgende Form hat:

$$\varphi(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

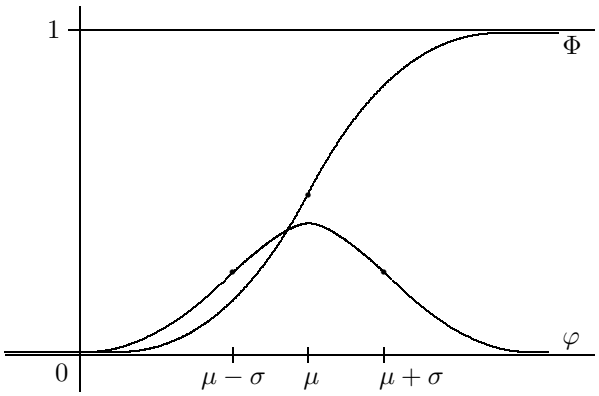
Ohne Beweis wollen wir hinnehmen, daß eine Verteilung vorliegt, also

$$\int_{-\infty}^{+\infty} \varphi(x, \mu, \sigma) dx = 1$$

gilt. Die zugehörige Verteilungsfunktion lautet dann:

$$\Phi(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt.$$

Empirisch hat man festgestellt, daß alle jene Zufallsgrößen näherungsweise als normalverteilt angesehen werden können, die durch additive Überlagerung vieler, voneinander unabhängiger, kleiner zufälliger Einflüsse entstehen, bei denen keiner besonders ausgezeichnet ist; so z. B. Meß- und Beobachtungsfehler, bei denen insbesondere kein systematischer Fehlereinfluß vorliegt, Normabweichungen eines Werkstückes, die insbesondere nicht auf einer falschen Maschineneinrichtung beruhen usw. Bei oftmaliger Wiederholung eines Versuches passiert es häufig, daß man sog. Ausreißer im Versuchsergebnis erhält, die dann aus der Versuchsserie herausgelassen werden, um zum einen das Ergebnis „zu schönen“ und zum anderen Normalverteilung annehmen zu dürfen. Gelegentlich zeigen dann Versuchswiederholungen durch andere Experimentatoren, daß gerade die Ausreißer näher an der Wahrheit waren als das publizierte statistische Ergebnis.
Die folgende Abbildung zeigt einen typischen Verlauf von Dichte und Verteilungsfunktion ($\sigma = 1, \mu = 3$).



Die Dichte φ hat an der Stelle $x = \mu$ ein absolutes Maximum mit dem Funktionswert $1/(\sigma\sqrt{2\pi})$ und verläuft symmetrisch zur Maximumstelle; außerdem hat die Funktion in $\mu - \sigma$ und $\mu + \sigma$ je einen Wendepunkt. Je kleiner σ ist, um so höher ist der Maximalwert und umso stärker konzentriert sich der gesamte Flächeninhalt zwischen dem Graphen der Funktion und der x-Achse im Intervall $(\mu - \sigma, \mu + \sigma)$. Ohne Beweis sei der nächste Satz angegeben.

Satz 142. *Eine normalverteilte Zufallsgröße mit den Parametern μ und σ hat den Erwartungswert μ und die Varianz σ^2.*

Mittels der Standardisierung

$$Y = \frac{X - \mu}{\sigma}$$

erhält man aus einer $N(\mu, \sigma)$-verteilten Zufallsgröße X eine $N(0, 1)$-verteilte Zufallsgröße Y, d. h. eine Zufallsgröße mit dem Erwartungswert 0 und der Varianz 1; diese Verteilung nennt man **standardisierte Normalverteilung** mit der Dichte und der Verteilungsfunktion

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2}), \quad \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left(-\frac{t^2}{2}\right) dt.$$

Wegen

$$\varphi(x, \mu, \sigma) = \frac{1}{\sigma}\varphi\left(\frac{x - \mu}{\sigma}\right), \quad \Phi(x, \mu, \sigma) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

genügt es, Dichte und Verteilungsfunktion der standardisierten Normalverteilung zu kennen, die in Tabellen vorliegt. Wegen der Symmetrie

$$\varphi(-x) = \varphi(x), \quad \Phi(-x) = 1 - \Phi(x)$$

kann man sich auf die nichtnegativen Werte von x beschränken. Ist nun X eine $N(\mu, \sigma)$-verteilte Zufallsgröße, so folgt

$$\begin{aligned} P(a < X \leqq b) &= P(a \leqq X \leqq b) = \Phi(b, \mu, \sigma) - \Phi(a, \mu, \sigma) \\ &= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right). \end{aligned}$$

Beispiel. Das Gewicht von geschlachteten Hähnchen sei normalverteilt mit $\mu = 1000g, \sigma = 20g$. Die Wahrscheinlichkeit, daß ein Hähnchen zwischen $960g$ und $1040g$ wiegt, ist dann

$$P(960 \leqq X \leqq 1040) = \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 \approx 0,954.$$

Allgemein ergibt sich für Intervalle, die symmetrisch zum Erwartungswert μ liegen:

$$\begin{aligned} P(|X-\mu| \leqq r\sigma) &= P(\mu - r\sigma \leqq X \leqq \mu + r\sigma) = \Phi(r) - \Phi(-r) \\ &= 2\Phi(r) - 1, \end{aligned}$$

also z. B.

$$P(|X-\mu| < \sigma) \approx 0,683,$$

$$P(|X-\mu| < 2\sigma) \approx 0,955,$$

$$P(|X-\mu| < 3\sigma) \approx 0,997.$$

Der letzte Wert besagt insbesondere, daß es im Falle einer Normalverteilung eine $99,7\%$-ige Sicherheit dafür gibt, daß die Realisierungen der Werte von X im Intervall $(\mu - 3\sigma, \mu + 3\sigma)$ liegt; dies ist die sog. **3σ-Regel**.
Für die nächste Verteilung benötigen wir die **Gammafunktion** oder auch **Fakultätsfunktion**, die für $x > 0$ definiert ist:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

Das Integral konvergiert gleichmäßig; daher ist Γ eine stetige Funktion und mittels partieller Integration folgt

$$\begin{aligned} \Gamma(x+1) &= \int_0^\infty t^x e^{-t} dt = [-e^{-t} t^x]_0^\infty + x \int_0^\infty t^{x-1} e^{-t} dt \\ &= x\Gamma(x). \end{aligned}$$

Wegen

$$\Gamma(1) = [-e^{-x}]_0^\infty = 1$$

ergibt sich für jede natürliche Zahl n:

$$\Gamma(n+1) = n\Gamma(n) = n(n-1)\Gamma(n-1) = n(n-1)\cdots 2\Gamma(1) = n!.$$

Die Gammafunktion ist somit die reelle Erweiterung der Fakultät, die wir für natürliche Zahlen kennen.
Wir sagen, daß eine stetige Zufallsgröße X einer **χ^2-Verteilung** mit m Freiheitsgraden unterliegt, wenn ihre Dichte die folgende Form hat:

$$f(x) = \begin{cases} 0 & x \leqq 0 \\ \dfrac{x^{\frac{m}{2}-1} e^{-\frac{m}{2}}}{2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)} & x > 0. \end{cases}$$

Diese Verteilung wird bei statistischen Untersuchungen verwendet. Ohne Beweis vermerken wir den nächsten Satz.

Satz 143. *Eine χ^2-verteilte Zufallsgröße mit m Freiheitsgraden hat den Erwartungswert m und die Varianz $2m$.*

Als letzte Verteilung erwähnen wir die **Studentverteilung**. Eine stetige Zufallsgröße X unterliegt der Studentverteilung mit n Freiheitsgraden, wenn ihre Dichte die Form

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \cdot \Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

hat.

Satz 144. *Eine Studentverteilung mit $n \geqq 2$ Freiheitsgraden hat den Erwartungswert 0 und für $n \geqq 3$ die Varianz $\frac{n}{n-2}$.*

Viele praktisch auftretende Verteilungen sind Mischverteilungen. Darum wollen wir zusammenstellen, wie sich Zufallsgrößen verhalten, wenn man sie elementaren Operationen unterzieht.

Satz 145. *Die unabhängigen Zufallsgrößen X, Y seien normalverteilt mit den Parametern μ_x, σ_x bzw. μ_y, σ_y. Dann ist $X + Y$ normalverteilt mit den Parametern $\mu_x + \mu_y$ und $\sqrt{\sigma_x^2 + \sigma_y^2}$.*

Allgemeiner gilt

Satz 146. *Sind die unabhängigen Zufallsgrößen $X_1, X_2 \ldots, X_n$ normalverteilt mit den gleichen Parametern μ, σ, so ist ihr arithmetisches Mittel*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

eine normalverteilte Zufallsgröße mit den Parametern μ und $\frac{\sigma}{\sqrt{n}}$.

Diese Eigenschaft folgt durch vollständige Induktion aus dem vorletzten Satz. Eine mögliche Interpretation ist die folgende: Bei einem Versuch möge ein Merkmal normalverteilt mit den Parametern μ, σ auftreten. Es sei X_i die dem Merkmal entsprechende Zufallsgröße bei der i-ten Wiederholung des Versuches, wobei die Versuche unabhängig voneinander ausgeführt werden. Der Satz gibt dann Auskunft über das mittlere Auftreten des betreffenden Merkmals nach n Versuchen. Die folgenden Sätze zeigen Zusammenhänge zwischen verschiedenen Verteilungen auf.

Satz 147. *Sind die unabhängigen Zufallsgrößen $X_1, \ldots, X_n$ alle $N(0,1)$-verteilt, dann ist die Zufallsgröße*

$$X = X_1^2 + X_2^2 + \cdots + X_n^2$$

χ^2-verteilt mit n Freiheitsgraden.

Satz 148. *Sind die unabhängigen Zufallsgrößen $X_1, \ldots, X_n$ normalverteilt mit den einheitlichen Parametern μ, σ, so hat die quadratische Abweichung*

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

vom arithmetischen Mittel $\overline{X}$ eine χ^2-Verteilung mit $n-1$ Freiheitsgraden.

Satz 149. *Es seien X, Y unabhängige Zufallsgrößen; X sei $N(0,1)$-verteilt und Y χ^2-verteilt mit n Freiheitsgraden. Dann hat*

$$Z = \frac{\sqrt{n} X}{\sqrt{Y}}$$

eine Studentverteilung mit n Freiheitsgraden.

Eine mögliche Anwendung dieser Aussagen ist die folgende. Es seien $X_1, \ldots, X_n$ unabhängige und normalverteilte Zufallsgrößen mit den einheitlichen Parametern μ, σ. Dann ist ihr arithmetisches Mittel $\overline{X}$ normalverteilt mit den Paramertern μ, $\frac{\sigma}{\sqrt{n}}$, so daß die standardisierte Zufallsgröße

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$N(0,1)$-verteilt ist. Die quadratische Abweichung

$$Y = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

ist χ^2-verteilt mit $n-1$ Freiheitsgraden. Setzen wir alles ineinander ein, so folgt mit dem letzten Satz: Sind die Zufallsgrößen $X_1, \ldots, X_n$ unabhängig und normalverteilt mit den einheitlichen Parametern μ, σ, so hat die Zufallsgröße

$$\sqrt{n} \frac{\overline{X} - \mu}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

eine Studentverteilung mit $n-1$ Freiheitsgraden.

5.1.5. Grenzwertsätze

Grenzwertsätze haben grundlegende Bedeutung für die Anwendungen. Es werden Folgen von Zufallsgrößen untersucht; dabei interessiert die sich ergebende Verteilungsfunktion beim Grenzübergang. Dadurch erhalten wir einerseits eine theoretische Begründung für empirisch gefundene Verteilungen und andererseits die Möglichkeit, Grenzverteilungen zu approximieren. Aus den zahlreich vorhandenen Grenzwertsätzen wählen wir nur drei aus.

Satz 150 (Gesetz der großen Zahlen). *Es sei $h_n(A)$ die relative Häufigkeit für das Eintreten eines Ereignisses A bei n-maliger, unabhängiger Wiederholung des zufälligen Versuches; das Ereignis A habe die Wahrscheinlichkeit p. Dann gilt für jedes $\varepsilon > 0$:*

$$\lim_{n\to\infty} P(|h_n(A) - p| < \varepsilon) = 1.$$

Beweis. Wie wir bereits wissen, hat das Ereignis $h_n(A)$ den Erwartungswert p und die Varianz $p(1-p)/n$. Aus der Tschebyscheff-Ungleichung folgt damit

$$0 \leqq P(|h_n(A) - p| \geqq \varepsilon) \leqq \frac{p(1-p)}{n\varepsilon^2}$$

bzw.

$$1 \geqq P(|h_n(A) - p| < \varepsilon) \geqq 1 - \frac{p(1-p)}{n\varepsilon^2}.$$

Für $n \to \infty$ folgt daraus die Behauptung. □

Für große n kommt es nach diesem Satz sehr selten vor, daß die relative Häufigkeit des Ereignisses A bei n unabhängigen Wiederholungen des Versuches sich wesentlich von der Wahrscheinlichkeit p unterscheidet. Auch sehr unwahrscheinliche Ereignisse treten mit großer Wahrscheinlichkeit ein, sofern der Versuch nur hinreichend oft wiederholt wird. Der Satz ist daher ein mathematische Pendant zu Volksweisheiten wie z. B.

- Was lange währt, wird endlich gut.
- Der Krug geht so lange zu Wasser, bis er bricht.

Betrachten wir nun die absolute Häufigkeit $H_n(A)$ als Zufallsgröße; sie hat den Erwartungswert np und die Varianz $np(1-p)$. Die Verteilungsfunktionen der standardisierten, absoluten Häufigkeiten streben für $n \to \infty$ gegen die Verteilungsfunktion der Normalverteilung.

Satz 151. (Grenzwertsatz von deMoivre-Laplace)
Es seien $H_n, n = 1, 2, \ldots$ binomialverteilte Zufallsgrößen mit den Parametern n und p, $0 < p < 1$; X_n seien die standardisierten Zufallsgrößen

$$X_n = \frac{H_n - np}{\sqrt{np(1-p)}}.$$

Dann gilt für die Verteilungsfunktionen F_n:

$$\lim_{n\to\infty} F_n(x) = \Phi(x).$$

Nach diesem Satz ist eine Binomialverteilung mit den Parametern n und p für große n näherungsweise eine Normalverteilung mit den Parametern $\mu = np$ und $\sigma = \sqrt{np(1-p)}$. Für eine binomialverteilte Zufallsgröße X mit den Parametern n und p gilt also für große n näherungsweise:

$$\begin{aligned} P(a \leqq X \leqq b) &= P\left(\frac{a-np}{\sqrt{np(1-p)}} \leqq \frac{X-np}{\sqrt{np(1-p)}} \leqq \frac{b-np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right). \end{aligned}$$

Diese Werte kann man aus den bekannten Tabellen entnehmen. Meist erhält man schon für $np(1-p) > 9$ gute Näherungswerte.

Satz 152. (Zentraler Grenzwertsatz)
Es sei (X_i) eine Folge unabhängiger Zufallsgrößen mit dem gemeinsamen Erwartungswert μ und der gemeinsamen Varianz σ^2. Dann konvergiert die Folge (F_n) der Verteilungsfunktionen für die standardisierten Zufallsgrößen

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

gegen die Verteilungsfunktion der standardisierten Normalverteilung:

$$\lim_{n\to\infty} F_n(x) = \Phi(x).$$

Nach diesem Satz hat die Summe von n unabhängigen Zufallsgrößen X_i, die alle den gleichen Erwartungswert μ und die gleiche Standardabweichung σ haben, näherungsweise eine Normalverteilung mit den Parametern $n\mu$ und $\sqrt{n\sigma}$. Das arithmetische Mittel von n unabhängigen Zufallsgrößen mit Erwartungswert μ und Standardabweichung σ ist annähernd normalverteilt mit dem Erwartungswert μ der Varianz σ^2/n. Dies ist die theoretische Begründung dafür, daß eine zufällige Erscheinung, die durch additive Überlagerung vieler unabhängiger Einflußgrößen entsteht, bei denen keine sonderlich bevorteilt ist, näherungsweise normalverteilt ist.
Beispiel. Die Zufallsgrößen $X_i, i = 1, 2, \ldots$ seien unabhängig und mögen nur die Werte $r = 0, 1, \ldots, 9$ mit der einheitlichen Wahrscheinlichkeit $0,1$ annehmen:

$$P(X_i = r) = 0,1.$$

Es ist dann

$$\mu = E(X_i) = \frac{1}{10}\sum_{r=0}^{9} r = 4,5,$$

$$\sigma^2 = \frac{1}{10}\sum_{r=0}^{9} r^2 - \mu^2 = 28,50 - 20,25 = 8,25,$$

also $\sigma \approx 2,87$. Wir fragen nun danach, wie groß die Wahrscheinlichkeit dafür ist, daß die Zufallsgröße

$$Y_{100} = \frac{1}{100}(X_1 + X_2 + \cdots + X_{100})$$

einen Wert annimmt, der größer als 5 ist. Nach dem zentralen Grenzwertsatz ist die Zufallsgröße Y_{100} annähernd normalverteilt mit dem Erwartungswert $\mu = 4,5$ und der Standardabweichung

$$\sigma = \frac{2,87}{10} = 0,287.$$

Wir erhalten also

$$\begin{aligned} P(Y_{100} > 5) &= P\left(\frac{Y_{100} - 4,5}{0,287} > \frac{5 - 4,5}{0,287}\right) = P\left(\frac{Y_{100} - 4,5}{0,287} > 1,74\right) \\ &\approx 1 - \Phi(1,74) \approx 0,041. \end{aligned}$$

5.2. Anwendungen in Simulation und Statistik

5.2.1. Erzeugung von Pseudozufallszahlen

Ein großes Gebiet der Informatik ist die Simulation realer Prozesse auf einem Rechner. Hier ist man insbesondere daran interessiert, eine Vielzahl von Daten in kürzester Zeit verfügbar zu machen, die dann ausreichen, die betrachtete Situation hinreichend genau darzustellen. Auch bei Laufzeituntersuchungen von Algorithmen benötigt man oft Eingabedaten, die 'zufällig' erzeugt sind und einer gewissen Verteilung genügen. Natürlich ist kein Rechner in der Lage, wirklich Zufall zu erzeugen. Daher stellt sich besser die Frage, wie man Daten erzeugen kann, die für einen neutralen Beobachter 'zufällig' aussehen und deren Zufälligkeit man wegen gewisser Untersuchungen nicht ablehnen kann. Solche Zahlen nennt man **Pseudozufallszahlen**.
Wie wir gezeigt haben, können wir uns zunächst auf die Erzeugung von $R(0,1)$-verteilten Zufallszahlen beschränken, da man daraus mittels geeigneter Transformationen andere Verteilungen berechnen kann.
Als leicht zu realisierende Methode hat sich die **multiplikative Kongruenzmethode** durchgesetzt. Bei dieser Methode wird eine Folge von Zahlen $x_1, x_2, \ldots$ aus einer Menge

$$M = \{\, 1, 2, \ldots, m-1 \,\}$$

nach der Vorschrift

$$x_{i+1} = a \cdot x_i \pmod m$$

erzeugt, wobei der Faktor a, der Modul m und der Startwert x_1 geeignet gewählt werden müssen. Als Zufallszahlen verwendet man dann

$$z_i = \frac{x_i}{m}, \quad i = 1, 2, \ldots.$$

Auf Grund unserer algebraischen Kenntnisse wissen wir, daß sich die nach dieser Methode erzeugten Zahlen nach einer gewissen Vorlaufphase periodisch wiederholen müssen. Man kann zeigen, daß für $m = 2^n$ mit $n \geqq 3$ die maximale Periodenlänge $m/4$ beträgt. Diese Schranke wird angenommen, wenn der Startwert x_1 ungerade ist und der Faktor a der Bedingung

$$a = 3 \pmod 8 \text{ oder } a = 5 \pmod 8)$$

genügt. Alle erzeugten Zahlen haben den Abstand $\frac{1}{m}$; daher sollte man, um näherungsweise eine $R(0,1)$-Verteilung zu sichern, den Modul m möglichst groß wählen, etwa $m = 2^{35}$, wodurch die maximale Periodenlänge

$$2^{33} = 8589934592$$

beträgt. Für die Wahl des Faktors a ist zu beachten, daß das Produkt $a \cdot m$ noch auf dem Rechner ausführbar sein muß. Bei einem 64-bit-Rechner darf a nicht größer als

$$2^{28} = 268435456$$

sein. Andererseits darf man a auch nicht zu klein wählen, da sonst die produzierten Zahlen nicht mehr unabhängig sind. Ein Kompromiß ist etwa

$$a = 8^9 + 5 = 134217733.$$

Bezeichnet man mit $[x]$ den Nachkomma-Anteil einer reellen Zahl x, so kann die obige Methode auch als

$$x_{i+1} = [a \cdot x_i], \quad i = 1, 2, \ldots$$

geschrieben werden.
Wenn wir nun die $R(0,1)$-verteilte Zufallsgröße X in der Form $X = F(Y)$ darstellen, wobei F die Verteilungsfunktion der Zufallsgröße Y sein soll, so können wir mittels $Y = F^{-1}(X)$ weitere Verteilungen berechnen. Sucht man etwa eine exponentialverteilte Zufallsgröße mit dem Parameter α, also

$$F(y) = 1 - e^{\alpha y} \quad (y > 0)$$

und setzt man

$$\mathrm{Ln}x = \begin{cases} \ln x & x > 0 \\ 0 & \text{sonst,} \end{cases}$$

so erfüllt

$$Y = -\frac{1}{\alpha}\mathrm{Ln}X$$

diese Forderung.
$N(0,1)$-verteilte Zufallsgrößen erhält man durch die sog. **Polarmethode**:
Sind X, Y unabhängige, $R(0,1)$-verteilte Zufallsgrößen, so kann man zeigen, daß

$$U = \sqrt{-2\ln X}\sin(2\pi Y), \quad V = \sqrt{-2\ln X}\cos(2\pi Y)$$

normalverteilt sind mit dem Erwartungswert 0 und der Varianz 1.

5.2.2. Monte-Carlo-Methoden

Wir wollen hier nur die sog. rohe Monte-Carlo-Methode anhand einer konkreten Aufgabe besprechen. Daraus wird das grundlegende Prinzip dieser Methoden klar hervortreten. Es sei ein Gebiet G in der Ebene gegeben, das vollständig im Einheitsquadrat liegt:

$$G \subseteqq \left\{ (x,y) \mid 0 \leqq x \leqq 1, 0 \leqq y \leqq 1 \right\}.$$

Auf G sei eine stetige Funktion h erklärt:

$$h : G \longmapsto \mathbb{R}.$$

Berechnet werden soll das bestimmte Integral von h über G:

$$I = \iint\limits_G h(x,y)dxdy,$$

d. h. das Volumen zwischen der durch h beschriebenen Fläche und dem Gebiet G. Ist ein Gebiet gegeben, das nicht im Einheitsquadrat liegt, bildet man es zunächst mittels einer geeigneten Transformation in das Einheitsquadrat ab. Es seien nun X, Y unabhängige, $R(0,1)$-verteilte Zufallsgrößen und I_G die Indikatorfunktion von G, d. h.

$$I_G(x,y) = \begin{cases} 1 & (x,y) \in G, \\ 0 & \text{sonst}, \end{cases}$$

so läßt sich das Volumenintegral offenbar auch in der Form

$$I = E(h(X,Y)I_G(X,Y))$$

darstellen. Die Idee besteht nun darin, diesen Erwartungswert nach dem Gesetz der großen Zahlen zu approximieren. Sind (X_n) und (Y_n) unabhängige Folgen von $R(0,1)$-verteilten Zufallsgrößen, dann gilt

$$I = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} h(X_k, Y_k) I_G(X_k, Y_k) = \lim_{n\to\infty} \frac{1}{n} \sum_{k} h(X_k, Y_k),$$

wobei in der letzten Summe über alle jene k zu summieren ist, für die $(x_k, y_k) \in G$ gilt. Diese Grundmethode läßt sich noch wesentlich verfeinern.

5.2.3. Vertrauensintervalle

Bisher hatten wir angenommen, daß die Verteilung und die Parameter einer Zufallsgröße bekannt sind. In der Praxis stellt sich aber die Frage, ob die ursprünglichen Wahrscheinlichkeitsannahmen gerechtfertigt sind bzw. wie genau die wirkliche Situation erfaßt worden ist. Dazu kann man den folgenden Weg einschlagen: Man führe einen zufälligen Versuch hinreichend oft durch und schließe von den Versuchsergebnissen auf die Verteilung und die Parameter der Zufallsgröße. So kann man etwa Glühlampen auf ihre Lebenszeit untersuchen und Leute nach ihrer Einstellung zu politischen Parteien befragen, um damit Rückschlüsse auf die Gesamtheit aller Glühlampen bzw. der Bevölkerung eines Landes zu ziehen. Allgemein wird von einer **Stichprobe** durch Hochrechnung auf die Grundgesamtheit geschlossen. Wichtig dabei ist, daß durch die Stichprobe ein repräsentativer Querschnitt erreicht wird. So hat man z. B. in den USA vor dem zweiten Weltkrieg per Telefon eine repräsentative Umfrage nach dem Namen des nächsten Präsidenten gemacht. Es ergab sich eine überwältigende Mehrheit für einen, der es schließlich doch nicht wurde. Die Stichprobe war allein schon dadurch nicht repräsentativ, daß nur wenige Menschen über ein Telefon verfügten und jene, die telefonisch erreichbar waren, einer ausgewählten Bevölkerungsschicht angehörten. Bei Meinungs-Umfragen entsteht ein weiteres, wichtiges Problem: Durch die Art der Frage, wird die Antwort wesentlich beeinflußt. Beispiel: 1. Frage: Wollen Sie, daß in Ihrem Garten eine atomare Mittelstrecken-Rakete der NATO aufgestellt wird? 2. Frage: Glauben Sie, daß der NATO-Doppelbeschluß Ihrer und damit unserer Sicherheit dient? Jeder aufmerksame Wahlbeobachter kann über einige Wahlen hinweg selbst erkennen, daß ein Meinungsforschungs-Institut in seinen Prognosen mehr rechts von der eingetretenen Situation und ein anderes mehr links davon liegt. Diese Tatsache wechselt nicht zwischen den Instituten.

Wir wollen zunächst ein Vertrauensintervall für eine Wahrscheinlichkeit p konstruieren.

Beispiel: Angenommen, bei der letzten Wahl haben 43% der Wähler die Partei A gewählt. Dann ist die Wahrscheinlichkeit p dafür, daß auf einem zufällig ausgewählten Stimmzettel die Partei A angekreuzt ist, gleich 0,43. Bei 1000 zufällig ausgewählten Stimmzetteln wird man ca. 430 Stimmen für die Partei A erwarten: $\mu = 0,43 \cdot 1000 = 430$. Aber weder 410 noch 450 Stimmen für A werden uns überraschen, denn die absolute Häufigkeit $H_n(A)$ ist binomialverteilt mit den Parametern $n = 1000$ und $p = 0,43$; $H_n(A)$ ist näherungsweise normalverteilt; wegen

$$np(1-p) = 430 \cdot 0,57 > 9$$

folgt

$$P(405 \leqq H_n(A) \leqq 455) \approx 0,9.$$

Erst bei weniger als 400 oder mehr als 460 Stimmen wäre man stutzig, denn

$$P(H_n(A) \notin [400, 460]) < 0,05.$$

Nehmen wir umgekehrt an, daß das Wahlergebnis nicht bekannt ist, 1000 zufällig ausgewählte Stimmzettel bereits ausgezählt sind und dabei ein Anteil $h_n(A) = 0,43$ auf die Partei A entfällt. Selbst wenn sich später herausstellen sollte, daß der wahre Anteil nur 42% oder aber gar 44% beträgt, würden wir unser Stichprobenergebnis akzeptieren, für wahrscheinlich halten. Für welche Stimmenanteile p in der Gesamtwählerschaft wird

nun das Stichprobenergebnis unwahrscheinlich? Die Frage kann man auch so formulieren: Für welche Werte von p liegt $h_n(A)$ noch nicht in einem Bereich mit geringer Wahrscheinlichkeit?
Wir wissen: Für eine normalverteilte Zufallsgröße X gilt

$$P(\mu - k\sigma \leqq X \leqq \mu + k\sigma) = 2\Phi(k) - 1.$$

Die absolute Häufigkeit $H_n(A)$ ist näherungsweise normalverteilt mit

$$\mu = np, \quad \sigma = \sqrt{np(1-p)},$$

also folgt

$$P\left(np - k\sqrt{np(1-p)} \leqq H_n(A) \leqq np + k\sqrt{np(1-p)}\right) \approx 2\Phi(k) - 1,$$

d. h.

$$P\left(h_n(A) - k\sqrt{\frac{p(1-p)}{n}} \leqq p \leqq h_n(A) + k\sqrt{\frac{p(1-p)}{n}}\right) \approx 2\Phi(k) - 1,$$

bzw.

$$P\left(-k \leqq \sqrt{n}\frac{h_n(A) - p}{\sqrt{p(1-p)}} \leqq k\right) \approx 2\Phi(k) - 1.$$

Daraus folgt für eine vorgegebene Wahrscheinlichkeit $\varrho > 0$, $\varrho = 2\Phi(k) - 1$ ein Intervall

$$\left[h_n(A) - k\sqrt{\frac{p(1-p)}{n}}, h_n(A) + k\sqrt{\frac{p(1-p)}{n}}\right]$$

mit folgender Eigenschaft: In $100 \cdot \varrho\%$ aller Stichproben wird das Intervall den Wert p enthalten. Also liegt p mit der Wahrscheinlichkeit ϱ in diesem Intervall; man nennt es ϱ-**Vertrauensintervall** für die gesuchte Wahrscheinlichkeit p. Die Größe $\alpha = 1 - \varrho$ heißt **Irrtumswahrscheinlichkeit**. Das obige Intervall ist eine Zufallsgröße, da seine Lage noch von $h_n(A)$ abhängt. Zur Ermittlung des Intervalls ist die quadratische Ungleichung

$$|h_n(A) - p| \leqq k \cdot \sqrt{\frac{p(1-p)}{n}}$$

zu lösen.
In unserem obigen Beispiel wählen wir $\alpha = 0,05$ als Irrtumswahrscheinlichkeit; dann folgt $\varrho = 0,95$ und aus $\varrho = 2\Phi(k) - 1$ ergibt sich $k = 1,96$, womit die fragliche Ungleichung lautet:

$$|0,43 - p| \leqq 1,96\sqrt{\frac{p(1-p)}{1000}}$$

mit den Lösungen $p_1 \approx 0,4000, p_2 \approx 0,460$; also können wir sagen, daß mit einer 95%-igen Sicherheit der tatsächliche Stimmenanteil für die Partei A zwischen 40% und 46% liegen wird.
Wir wollen nun ein Vertrauensintervall für den Erwartungswert ermitteln.
Dazu seien $X_1, \ldots, X_n$ identisch normalverteilte Zufallsgrößen mit den Parametern μ, σ; dann ist der Mittelwert

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

normalverteilt mit den Parametern $\mu, \frac{\sigma}{\sqrt{n}}$ und daher die standardisierte Größe

$$T_{\mu,\sigma} = \sqrt{n}\frac{\overline{X} - \mu}{\sigma}$$

$N(0,1)$-verteilt. Wie oben schließen wir

$$P\left(-k \leqq \sqrt{n}\frac{\overline{X} - \mu}{\sigma} \leqq k\right) = 2\Phi(k) - 1 = \varrho$$

bzw.

$$P\left(\overline{X} - k\frac{\sigma}{\sqrt{n}} \leqq \mu \leqq \overline{X} + k\frac{\sigma}{\sqrt{n}}\right) = 2\Phi(k) - 1 = \varrho.$$

Folglich ist

$$\left[\overline{X} - k\frac{\sigma}{\sqrt{n}}, \overline{X} + k\frac{\sigma}{\sqrt{n}}\right]$$

ein ϱ-Vertrauensintervall für den Erwartungswert μ einer Normalverteilung, falls die Standardabweichung σ bekannt ist. Sollte die Standardabweichung unbekannt sein, ersetzt man sie durch den Schätzwert s mit

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

und erhält als Testgröße

$$T_{\mu,s} = \sqrt{n}\frac{\overline{X} - \mu}{s};$$

diese ist studentverteilt mit $n-1$ Freiheitsgraden. Damit lautet die Bedingung

$$P(-t \leqq T_{\mu,s} \leqq t) = \varrho.$$

Bei gegebenem ϱ entnehmen wir den Wert für t der Tabelle für die Studentverteilung mit $n-1$ Freiheitsgraden. Wegen der Symmetrie dieser Verteilung gilt

$$P(-t \leqq T_{\mu,s} \leqq t) = 2 \cdot P(-t \leqq T_{\mu,s}) - 1.$$

Daher haben wir in der Tabelle bei $n-1$ und $p = \frac{1+\varrho}{2}$ nachzusehen. Die obige Bedingung stellen wir nun nach dem Erwartungswert um:

$$P\left(\overline{X} - t\frac{s}{\sqrt{n}} \leqq \mu \leqq \overline{X} + t\frac{s}{\sqrt{n}}\right) = \varrho$$

und erhalten

$$\left[\overline{X} - t\frac{s}{\sqrt{n}}, \overline{X} + t\frac{s}{\sqrt{n}}\right]$$

als ein ϱ-Vertrauensintervall für den Erwartungswert μ einer Normalverteilung bei unbekannter Varianz.
Beispiel. Holzbretter werden auf Länge gesägt; die letzten 10 hatten eine mittlere Länge von $201,5$ cm mit einer Standardabweichung von 2,4 cm. Die Schnittlängen seien normalverteilt mit dem Erwartungswert μ und der unbekannten Varianz σ^2. Für μ berechnen wir ein 95%-iges Vertrauensintervall:

$$\begin{aligned}\left[\overline{X} - t\frac{s}{\sqrt{n}}, \overline{X} + t\frac{s}{\sqrt{n}}\right] &= \left[201,5 - 2,262\frac{2,4}{\sqrt{10}}; 201,5 + 2,262\frac{2,4}{\sqrt{10}}\right] \\ &\approx [199,8; 203,2].\end{aligned}$$

Der t-Wert ist der Tabelle für die Studentverteilung bei $n = 9$ und $\alpha = \frac{1+0,95}{2}$ zu entnehmen.
Angenommen, die Varianz σ^2 ist bekannt, z. B. $\sigma = 2,4$; dann kann man mit der Normalverteilung rechnen und erhält als 95%-iges Vertrauensintervall:

$$\begin{aligned}\left[\overline{X} - k\frac{\sigma}{\sqrt{n}}, \overline{X} + k\frac{\sigma}{\sqrt{n}}\right] &= \left[201,5 - 1,96\frac{2,4}{\sqrt{10}}; 201,5 + 1,96\frac{2,4}{\sqrt{10}}\right] \\ &\approx [200,0; 203,0].\end{aligned}$$

Abschließend soll ein Vertrauensintervall für die Varianz bestimmt werden.
Es seien $X_1, \ldots, X_n$ unabhängige, normalverteilte Zufallsgrößen mit den gleichen Parametern μ, σ. Dann ist die Zufallsgröße

$$T_{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

χ^2-verteilt mit $n-1$ Freiheitsgraden. Damit folgt aus

$$P(c_1 \leqq T_{\sigma^2} \leqq c_2) = \varrho,$$

d. h.

$$P\left(c_1 \leqq \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \leqq c_2\right) = \varrho,$$

daß c_1, c_2 aus der Tabelle für die χ^2-Verteilung mit $n-1$ Freiheitsgraden zu ermitteln ist (es liegt eine unsymmetrische Verteilung vor!). Sind nun c_1 und c_2 bestimmt, so können wir die Ungleichung mit der Varianzschätzung

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

umformen:

$$P\left(\frac{(n-1)s^2}{c_2} \leqq \sigma^2 \leqq \frac{(n-1)s^2}{c_1}\right) = \varrho.$$

Folglich ist

$$\left[\frac{(n-1)s^2}{c_2}, \frac{(n-1)s^2}{c_1}\right]$$

ein ϱ-Vertrauensintervall für die Varianz einer Normalverteilung. Im Zusammenhang mit der Bestimmung von c_1, c_2 erwähnen wir noch, daß man wegen

$$\begin{aligned}\varrho = P(c_1 \leqq T_{\sigma^2} \leqq c_2) &= P(T_{\sigma^2} \leqq c_2) - P(T_{\sigma^2} \geqq c_1)\\ &= \frac{1+\varrho}{2} - \frac{1-\varrho}{2}\end{aligned}$$

den Wert für c_1 für $p = \frac{1-\varrho}{2}$ und c_2 für $p = \frac{1+\varrho}{2}$ zu ermitteln hat.
Im obigen Beispiel war $s = 2,4$. Als 95%-iges Vertrauensintervall für σ^2 folgt

$$\left[\frac{(n-1)s^2}{c_2}, \frac{(n-1)s^2}{c_1}\right] = \left[\frac{9 \cdot 2,4^2}{19,02}; \frac{9 \cdot 2,4^2}{2,70}\right] \approx [2,73; 19,2],$$

also

$$1,65 \leqq \sigma \leqq 4,38.$$

Natürlich ist dies nur eine grobe Schätzung, die sich aber mit einer größeren Stichprobe verbessern läßt.

5.2.4. Testen von Hypothesen

Die prinzipielle Vorgehensweise soll an einem Beispiel erläutert werden: Die Partei A behauptet am Wahltag, daß sie die absolute Mehrheit der abgegebenen Stimmen erringen wird. Mit den ersten 1000 zufällig ausgewählten Stimmzetteln soll die Behauptung $p > 0,5$ überprüft werden. Wir lehnen die Behauptung ab, wenn für das Stichprobenergebnis $H_n(A)$ mit einem gewissen a gilt:

$$P(H_n(A) \leqq a) \leqq \alpha \ll 1, \text{ z. B. } \alpha = 0,01$$

unter der Annahme $p > 0,5$.
Es sind

$$E(H_n(A)) = n \cdot p = 500, \quad \sigma^2 = V(H_n(A)) = np(1-p) = 250,$$

und $H_n(A)$ ist annähernd normalverteilt, also

$$P(H_n(A) \leqq \mu - k\sigma) \approx \Phi(-k) = 1 - \Phi(k)$$

und mit $\alpha = 1 - \Phi(k) = 0,01$:

$$P(H_n(A) \leqq 500 - 2,33\sqrt{250}) \approx 0,01$$

oder

$$P(H_n(A) \leqq 463) < 0,01.$$

Also kann man so argumentieren: Erhält die Partei A wirklich einen Stimmenanteil von 50%, so ist es sehr unwahrscheinlich, daß unter den 1000 zufällig ausgewählten Stimmzetteln höchstens 463 Stimmen für A sind. Sollte dies trotzdem eintreten, werden wir die Behauptung $p > 0,5$ ablehnen, wobei wir uns im ersten Falle mit der Wahrscheinlichkeit $\alpha = 0,01$ irren; daher heißt α **Irrtumswahrscheinlichkeit**. Die Hypothese $p >$

$0,5$ werden wir nur dann annehmen, wenn das Stichprobenergebnis $H_n(A)$ unter der Annahme $p \leqq 0,5$ sehr unwahrscheinlich wird, also mit einem gewissen a gilt:

$$P(H_n(A) \geqq a) \leqq \alpha.$$

Dafür folgt $(p = 0,5)$:

$$P(H_n(A) \geqq \mu + k\sigma) \approx 1 - \Phi(k) = \alpha = 0,01$$

bzw. $(k = 2,33, \sigma = \sqrt{250})$

$$P(H_n(A) \geqq 537) < 0,01,$$

was man so interpretieren kann: Unter der Annahme, daß der wahre Wähleranteil unter 50% liegen wird, entfallen höchstens mit der Wahrscheinlichkeit 0,01 mehr als 536 Stimmen aus der Stichprobe auf die Partei A. Werden aber mehr als 536 Stimmen für A gezählt, wird man die Hypothese $H_0 : p \leqq 0,5$ verwerfen und $p > 0,5$ annehmen. Im Falle

$$463 < H_n(A) < 537$$

(das sind $7,45\%$ Abweichung von 500) ist mit einer Irrtumswahrscheinlichkeit von $\alpha = 0,01$ die Hypothese $H_0 : p > 0,5$ weder anzunehmen noch abzulehnen. Durch Erhöhung der Irrtumswahrscheinlichkeit oder des Stichprobenumfangs kann man das Intervall verkleinern. Bei einer Stichprobe von $n = 2000$ beträgt die Abweichung nur noch $5,2\%$.
Allgemein: Wir betrachten die Testgröße

$$T_h = \frac{h_n(A) - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

wobei p_0 eine angenommene Wahrscheinlichkeit in der Hypothese $H_0 : p \geqq p_0$ ist. Die Hypothese wird abgelehnt, wenn $T_h < -k$ bei einer Irrtumswahrscheinlichkeit $\alpha \ll 1$ und die Gegenhypothese $H_1 : p < p_0$ angenommen. Die Hypothese $H_0 : p \leqq p_0$ wird bei $T_h > k$ abgelehnt. Beides sind einseitige Tests. Ein zweiseitiger Test ist z. B. $H_0 :\ p = p_0$. Dieser Test wird abgelehnt, wenn $T_h > k$ oder $T_h < -k$ ausfällt; dabei muß α aufgeteilt werden:

$$P(T_h > k) = 1 - \Phi(k) \leqq \frac{\alpha}{2}.$$

Mit den Testgrößen

$$T_{\mu,\sigma} = \sqrt{n}\frac{\overline{X} - \mu}{\sigma}, \quad T_{\mu,s} = \sqrt{n}\frac{\overline{X} - \mu}{s}, \quad T_{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$$

kann man μ bzw. σ testen. Dabei wird die Hypothese $H_0 :\ \mu = \mu_0$ für $T_{\mu,\sigma} < -k$ oder $T_{\mu,\sigma} > k$ abgelehnt; ebenso für $T_{\mu,s} < -t$ oder $T_{\mu,s} > t$. Die Hypothese $H_0 :\ \sigma^2 = \sigma_0^2$ wird für $T_{\sigma^2} < c_1$ oder $T_{\sigma^2} > c_2$ abgelehnt und jeweils die Gegenhypothese angenommen.
Beispiel. Wir nehmen das Holzsägen mit verschiedenen Hypothesen:
$H_0 :\ \mu = 202,5$ cm, $\alpha = 0,05,\ \sigma = 2,4$; die Hypothese kann nicht abgelehnt werden, da

$$-1,96 < \sqrt{10}\frac{201,5 - 202,5}{2,4} < 1,96;$$

$H_0 :\ \mu \leqq 200,\ \alpha = 0,05$ und unbekannte Varianz; die Hypothese wird abgelehnt mit $T_{\mu,s}$:

$$\sqrt{10}\frac{201,5 - 200,0}{2,4} > 1,833;$$

$H_0 :\ \sigma^2 \geqq 16,\ \alpha = 0,05$ wird abgelehnt, da

$$\frac{(n-1)s^2}{\sigma_0^2} = \frac{9 \cdot 2,4^2}{16} < c_1 = 3,3251.$$

Man hat zwei Fehlerarten bei Testentscheidungen:
1. Die Hypothese H_0 wird abgelehnt, obwohl sie richtig ist.
2. Die Hypothese wird angenommen, obwohl sie falsch ist.
Bei fixiertem Stichprobenumfang bewirkt eine Verringerung des ersten Fehlers eine Vergrößerung des zweiten. Nur eine Vergrößerung des Stichprobenumfangs verringert beide Fehlerrisken gleichzeitig. Ein kleiner Stichprobenumfang verlangt eine nicht zu kleine Irrtumswahrscheinlichkeit. Welcher Fehler folgenschwerer ist, kann mathematisch nicht entschieden werden. Nehmen wir nur die beiden Hypothesen: „Das Medikament ist wirksam“ und „Es treten Nebenwirkungen auf“. Bei der ersten Hypothese ist der zweite Fehler bedeutungsvoller; bei der zweiten Hypothese ist sicherlich der erste Fehler folgenreicher.

5.2.5. Tabellen von Verteilungen

1. **Die Poisson-Verteilung mit dem Parameter λ**

r	λ							
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8
0	0,904837	0,818731	0,740818	0,670320	0,606531	0,548812	0,496585	0,449329
1	0,090484	0,163746	0,222245	0,268128	0,303265	0,329287	0,347610	0,359463
2	0,004524	0,016375	0,033337	0,053626	0,075816	0,098786	0,121663	0,143785
3	0,000151	0,001092	0,003334	0,007150	0,012636	0,019757	0,028388	0,038343
4	0,000004	0,000055	0,000250	0,000715	0,001580	0,002964	0,004968	0,007669
5	—	0,000002	0,000015	0,000057	0,000158	0,000356	0,000696	0,001227
6	—	—	0,000001	0,000004	0,000013	0,000036	0,000081	0,000164
7	—	—	—	—	0,000001	0,000003	0,000008	0,000019
8	—	—	—	—	—	—	0,000001	0,000002

r	λ							
	0,9	1,0	1,5	2,0	2,5	3,0	3,5	4,0
0	0,406570	0,367879	0,223130	0,135335	0,082085	0,049787	0,030197	0,018316
1	0,365913	0,367879	0,334695	0,270671	0,205212	0,149361	0,150691	0,073263
2	0,164661	0,183940	0,251021	0,270671	0,256516	0,224042	0,184959	0,146525
3	0,049398	0,061313	0,125510	0,180447	0,213763	0,224042	0,215785	0,195367
4	0,011115	0,015328	0,047067	0,090224	0,133602	0,168031	0,188812	0,195367
5	0,002001	0,003066	0,014120	0,036089	0,066801	0,100819	0,132169	0,156293
6	0,000300	0,000511	0,003530	0,012030	0,027834	0,050409	0,077098	0,104196
7	0,000039	0,000073	0,000756	0,003437	0,009941	0,021604	0,038549	0,059540
8	0,000004	0,000009	0,000142	0,000859	0,003106	0,008102	0,016865	0,029770
9	—	0,000001	0,000024	0,000191	0,000863	0,002701	0,006559	0,013231
10	—	—	0,000004	0,000038	0,000216	0,000810	0,002296	0,005292
11	—	—	—	0,000007	0,000049	0,000221	0,000730	0,001925
12	—	—	—	0,000001	0,000010	0,000055	0,000213	0,000642
13	—	—	—	—	0,000002	0,000013	0,000057	0,000197
14	—	—	—	—	—	0,000003	0,000014	0,000056
15	—	—	—	—	—	0,000001	0,000003	0,000015
16	—	—	—	—	—	—	0,000001	0,000004
17	—	—	—	—	—	—	—	0,000001

r	λ						
	4,5	5,0	6,0	7,0	8,0	9,0	10
0	0,011109	0,006738	0,002479	0,000912	0,000335	0,000123	0,000045
1	0,049990	0,033690	0,014873	0,006383	0,002684	0,001111	0,000454
2	0,112479	0,084224	0,044618	0,022341	0,010735	0,004998	0,002270
3	0,168718	0,140374	0,089235	0,052129	0,028626	0,014994	0,007567
4	0,189808	0,175467	0,133853	0,091226	0,057252	0,033737	0,018917
5	0,170827	0,175467	0,160623	0,127717	0,091604	0,060727	0,037833
6	0,128120	0,146223	0,160623	0,149003	0,122138	0,091090	0,063055
7	0,082363	0,104445	0,137677	0,149003	0,139587	0,117116	0,090079
8	0,046329	0,065278	0,103258	0,130377	0,139587	0,131756	0,112599
9	0,023165	0,036266	0,068838	0,101405	0,124077	0,131756	0,125110
10	0,010424	0,018133	0,041303	0,070983	0,099262	0,118580	0,125110
11	0,004264	0,008242	0,022629	0,045171	0,072190	0,097020	0,113736
12	0,001599	0,003434	0,011264	0,026350	0,048127	0,072765	0,094780
13	0,000554	0,001321	0,005199	0,014188	0,029616	0,050376	0,0,2908
14	0,000178	0,000472	0,002228	0,007094	0,016924	0,032384	0,052077
15	0,000053	0,000157	0,000891	0,003311	0,009026	0,019431	0,034718
16	0,000015	0,000049	0,000334	0,001448	0,004513	0,010930	0,021699
17	0,000004	0,000014	0,000118	0,000596	0,002124	0,005786	0,012764
18	0,000001	0,000004	0,000039	0,000232	0,000944	0,002893	0.007091
19	—	0,000001	0,000012	0,000085	0,000397	0,001370	0,003732
20	—	—	0,000004	0,000030	0,000159	0,000617	0,001866
21	—	—	0,000001	0,000010	0,000061	0,000264	0,000889
22	—	—	—	0,000003	0,000022	0,000108	0,000404
23	—	—	—	0,000001	0,000008	0,000042	0,000176
24	—	—	—	—	0,000003	0,000016	0,000073
25	—	—	—	—	0,000001	0,000006	0,000029
26	—	—	—	—	-	0,000002	0,000011
27	—	—	—	—	—	0,000001	0,000004
28	—	—	—	—	—	—	0,000001
29	—	—	—	—	—	—	0,000001

2. **Die Normalverteilung mit Erwartungswert 0 und Varianz 1**

x	$\varphi(x)$	x	$\varphi(x)$	x	$\varphi(x)$	x	$\varphi(x)$	x	$\varphi(x)$
0,00	0,3989	0,60	0,3332	1,20	0,1942	1,80	0,0790	2,40	0,0224
0,05	0,3984	0,65	0,3230	1,25	0,1826	1,85	0,0721	2,45	0,0198
0,10	0,3970	0,70	0,3123	1,30	0,1714	1,90	0,0656	2,50	0,0176
0,15	0,3945	0,75	0,3011	1,35	0,1604	1,95	0,0596	2,55	0,0154
0,20	0,3910	0,80	0,2897	1,40	0,1497	2,00	0,0040	2,60	0,0136
0,25	0,3867	0,85	0,2780	1,45	0,1394	2,05	0,0488	2,65	0,0119
0,30	0,3814	0,90	0,2661	1,50	0,1295	2,10	0,0440	2,70	0,0104
0,35	0,3752	0,95	0,2541	1,55	0,1200	2,15	0,0396	2,75	0,0091
0,40	0,3683	1,00	0,2420	1,60	0,1109	2,20	0,0355	2,80	0,0079
0,45	0,3605	1,05	0,2299	1,65	0,1023	2,25	0,0317	2,85	0,0069
0,50	0,3521	1,10	0,2179	1,70	0,0940	2,30	0,0283	2,90	0,0060
0,55	0,3429	1,15	0,2059	1,75	0,0863	2,35	0,0252	2,95	0,0051
								3,00	0,0044

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,500000	0,75	0,773373	1,50	0,933193	2,25	0,987776
0,05	0,519939	0,80	0,788145	1,55	0,939429	2,30	0,989276
0,10	0,539828	0,85	0,802338	1,60	0,945201	2,35	0,990613
0,15	0,559618	0,90	0,815940	1,65	0,950528	2,40	0,991802
0,20	0,579260	0,95	0,828944	1,70	0,955434	2,45	0,992857
0,25	0,598706	1,00	0,841345	1,75	0,959941	2,50	0,993790
0,30	0,617911	1,05	0,853141	1,80	0,964070	2,55	0,994614
0,35	0,636831	1,10	0,864334	1,85	0,967843	2,60	0,995339
0,40	0,655422	1,15	0,874928	1,90	0,971283	2,65	0,995975
0,45	0,673645	1,20	0,884930	1,95	0,974412	2,70	0,996533
0,50	0,691463	1,25	0,894350	2,00	0,977250	2,75	0,997020
0,55	0,708840	1,30	0,903200	2,05	0,979818	2,80	0,997445
0,60	0,725747	1,35	0,911492	2,10	0,982136	2,85	0,997814
0,65	0,742154	1,40	0,919243	2,15	0,984222	2,90	0,998134
0,70	0,758036	1,45	0,926471	2,20	0,986097	2,95	0,998411
						3,00	0,998650

3. **Die χ^2-Verteilung**

Die Tabelle gibt die Werte von χ_α für einige Werte α an. Dabei ist χ^2_α so bestimmt, daß die Wahrscheinlichkeit dafür, daß die Zufallsgröße χ^2 mit n Freiheitsgraden nicht kleiner als χ^2_α ist, gleich α ist:

$$P(\chi^2 \geqq \chi^2_\alpha) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \int_{\chi^2_\alpha}^{\infty} e^{-\frac{x}{2}} x^{\frac{n}{2}-1} dx = \alpha$$

n	α								
	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01
1	0,064	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345
4	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277
5	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	3,070	3,828	6,348	7,231	8,558	10,645	12,592	15,033	16,812
7	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475
8	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	6,989	8,148	10,341	1,899	14,631	17,275	19,675	22,618	24,725
12	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217
13	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688
14	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578
16	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000
17	12,002	13,531	16,338	19,511	21,615	24,769	27,687	30,995	33,409
18	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805
19	13,716	16,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191
20	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566
21	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932
22	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289
23	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638
24	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980
25	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314
26	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642
27	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963
28	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278
29	22,475	24,577	28,336	32,461	35,139	39,087	42,657	46,693	49,588
30	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892

4. **Die Student-Verteilung**

Die Tabelle enthält die Werte von t_α für einige Werte α. Dabei ist t_α derart gewählt, daß die Wahrscheinlichkeit dafür, daß die studentverteilte Zufallsgröße t mit n Freiheitsgraden absolut genommen nicht kleiner als t_α ist, gleich α ist:

$$P(|t| \geqq t_\alpha) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \cdot \Gamma(\frac{n}{2})} \int_{t_\alpha}^{\infty} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx = \alpha$$

n	α							
	0,80	0,60	0,40	0,20	0,10	0,05	0,02	0,01
1	0,325	0,727	1,376	3,078	6,314	12,706	31,821	63,657
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660
120	0,254	0,526	0,845	1,289	1,658	1,980	2,358	2,617
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576

5.3. Übungen

1. Eine Reederei besitzt n Schiffe, A_i sei das zufällige Ereignis „Das i-te Schiff sinkt.“ $(i = 1, \ldots, n)$. Man beschreibe die folgenden Ereignisse durch A_i und die üblichen Operationen mit zufälligen Ereignissen:

   B: „Mindestens ein Schiff sinkt.“,
   C: „Keines der n Schiffe sinkt.“,
   D: „Genau ein Schiff sinkt.“,
   E: „Höchstens ein Schiff sinkt.“.

2. Zwei Schützen A und B schießen unabhängig voneinander 5 Schuß auf eine Zielscheibe. Die Trefferwahr-

scheinlichkeit von A beträgt 0,7, die von B 0,6. Man bestimme die Wahrscheinlichkeit für folgende Ereignisse:

(a) A hat mindestens einen Treffer,

(b) B hat höchstens 2 Treffer,

(c) A hat genau 3 Treffer,

(d) A und B haben zusammen mindestens 2 Treffer.

3. Wieviele Teilnehmer muß man höchstens zu einem fünftägigen Kongreß einladen, damit mit der Wahrscheinlichkeit 0,95 mindestens einer während dieser 5 Tage Geburtstag hat?

4. In einer Urne sind 4 Kugeln mit den Zahlen 1 bis 4. Bei einem zufälligen Versuch werden nacheinander 2 Kugeln gezogen (ohne Zurücklegen der 1. Kugel). Die Zufallsgröße X sei die Differenz zwischen der 1. und der 2. gezogenen Zahl.

(a) Man ermittle die Einzelwahrscheinlichkeiten $p_k = P(X = k)$.

(b) Man skizziere die Verteilungsfunktion F.

(c) Man ermittle $P(-1 < X < 3)$.

5. Es sei $F_X(x) = a + b \arctan x$ $(x \in \mathbb{R})$ die Verteilungsfunktion einer Zufallsgröße X.

(a) Man bestimme die Konstanten a und b.

(b) Wie lautet die Dichtefunktion?

(c) Man bestimme den Erwartungswert und die Varianz von X.

6. Es sei f eine durch

$$f(x) = \begin{cases} \alpha x^2(1-x) & 0 \leqq x \leqq 1 \\ 0 & \text{sonst} \end{cases}$$

gegebene Funktion.

(a) Man bestimme α so, daß f die Dichtefunktion einer stetigen Zufallsgröße X ist.

(b) Man ermittle Verteilungsfunktion, Erwartungswert und Varianz.

(c) Man berechne $P(X < 1/2)$ und $P(X < E(X))$.

7. In einem Meßgerät seien 4 unabhängig voneinander arbeitende Transistoren gleicher Bauart installiert. Die zufällige Zeit T bis zum Ausfall unterliege einer Exponentialverteilung.

$$f(t) = \begin{cases} 0,15e^{-0,15t} & t > 0 \\ 0 & \text{sonst.} \end{cases}$$

(a) Man berechne die Wahrscheinlichkeit dafür, daß ein solcher Transistor mindestens 10 Zeiteinheiten arbeitet.

(b) Man berechne die Wahrscheinlichkeit dafür, daß mindestens einer der 4 Transistoren länger als 10 Zeiteinheiten arbeitet.

(c) Man berechne die mittlere Anzahl der Transistoren, die länger als 10 Zeiteinheiten arbeiten.

8. Gegeben sei die Funktion f mit

$$f(x) = \begin{cases} 0 & x < 1 \\ a \ln x & 1 \leqq x \leqq e \\ 0 & x > e \end{cases}$$

(a) Man bestimme die Konstante a derart, daß f Dichtefunktion einer Zufallsgröße X ist.

(b) Man ermittle die Verteilungsfunktion F.

9. Es sei X eine diskrete Zufallsgröße mit der Verteilungstabelle:

x_i	-1	0	1	2	3
p_i	1/5	1/5	1/5	1/5	1/5

Man berechne für $Y = \mid X - E(X) \mid$

(a) die Verteilungsfunktion F_Y und die Einzelwahrscheinlichkeiten p_{y_i},

(b) $E(Y)$,

(c) $P(Y > 0)$.

10. Gegeben sei eine Funktion f mit

$$f(x) = \begin{cases} 0 & x < -1 \quad \text{und} \quad x \geqq 1 \\ a & -1 \leqq x < 0 \\ b & 0 \leqq x < 1 \end{cases}$$

Welche Bedingungen müssen a und b erfüllen, damit f Dichtefunktion einer stetigen Zufallsgröße ist? Man ermittle unter diesen Bedingungen Erwartungswert und Varianz.

11. Man zeige, daß beim Würfelspiel mit drei Würfeln die Wahrscheinlichkeit für die Augensumme 11 größer als die Wahrscheinlichkeit für die Augensumme 12 ist.

12. Wie oft muß man einen Spielwürfel werfen, damit mit Wahrscheinlichkeit 0,3 zu erwarten ist, daß keine 6 gewürfelt wird?

13. An einer Tankstelle kommen zwischen 16^{00} und 18^{00} Uhr durchschnittlich 2,5 Fahrzeuge pro Minute an. Man bestimme die Wahrscheinlichkeit, daß während einer Minute

(a) kein Fahrzeug,

(b) genau ein Fahrzeug,

(c) genau 2 Fahrzeuge,

(d) mehr als 3 Fahrzeuge,

(e) weniger als 6 Fahrzeuge

eintreffen. Die Anzahl der eintreffenden Fahrzeuge sei dabei poissonverteilt.

14. Die Zerfallszeit T für Polonium ist eine exponentialverteilte Zufallsgröße. Mittels der Halbwertzeit, die für dieses radioaktive Element 140 Tage beträgt, bestimme man

(a) den Parameter λ der Exponentialverteilung,

(b) die Zeitdauer t_0, so daß mit einer Wahrscheinlichkeit $p = 0,95$ ein Zerfall erfolgt.

(Unter Halbwertzeit versteht man diejenige Zeit, in deren Verlauf die Wahrscheinlichkeit eines Zerfalls gleich 0,5 ist.)

15. Bei der Abfüllung von 0,5l-Flaschen wird das Füllvolumen F als normalverteilt mit $\mu = 500$, $\sigma = 5$ (Maßeinheit cm^3) angenommen.

(a) Wie groß ist die Wahrscheinlichkeit, daß eine Flasche weniger als $490\,\text{cm}^3$ enthält?

(b) Wie groß ist die Wahrscheinlichkeit, daß die Flasche bei der Abfüllung überläuft, wenn

i. das Flaschenvolumen $510\,\text{cm}^3$ beträgt,

ii. das Flaschenvolumen (unabhängig vom Füllvolumen) normalverteilt mit $\mu = 500$ und $\sigma = 2$ ist.

16. Aus der Produktion von Kugellagern werden 150 Stück zufällig entnommen. In dieser Stichprobe sind 6 unbrauchbare. Der Ausschußprozentsatz $p \cdot 100\%$ der Gesamtproduktion ist unbekannt. Mit Hilfe der Stichprobe ist ein konkretes Vertrauensintervall für p mit $\alpha = 0,05$ zu berechnen.

17. Bei 10 Messungen der Streckgrenze S des Stahls ST70 ergeben sich folgende Werte:
332, 354, 338, 340, 345, 360, 366, 352, 346, 342.
Unter der Annahme, daß die Werte $S_1, \ldots, S_{10}$ eine Stichprobe aus einer Grundgesamtheit darstellen, in der die Streckgrenze eine normalverteilte Zufallsgröße ist, ermittle man Vertrauensintervalle mit $\alpha = 0,05$ für

(a) den Erwartungswert $\mu = E(S)$ bei bekannter Varianz $\sigma^2 = V(S) = 105$,

(b) den Erwartungswert μ bei unbekannter Varianz,

(c) die Varianz $\sigma^2 = V(S)$.

(d) Wie groß müßte der Stichprobenumfang im Falle von (a) mindestens gewählt werden, damit bei gleichem $\alpha = 0,05$ die Länge des Vertrauensintervalls 8 beträgt?

18. Aus einem Sortiment haben 20 Schrauben die Längen [mm]:
    10, 11, 13, 11, 12, 13, 14, 10, 9, 10, 10, 11, 12, 14, 14, 10, 11, 10, 16, 9.
    Unter der Voraussetzung, daß die Stichprobe aus einer Grundgesamtheit ist, in der die Schraubenlänge eine normalverteilte Zufallsgröße mit $\sigma = 2$ [mm] ist, prüfe man die Hypothese: $\mu = 11$ [mm] mit einer Irrtumswahrscheinlichkeit von $\alpha = 0,01$.

19. Man zeige, daß die Binomialverteilung für $n \to \infty$ gegen die Poissonverteilung konvergiert.

Kapitel 6

Numerische Mathematik

6.1. Einführung

Die numerische Mathematik hat insbesondere die Aufgabe, die Genauigkeit eines Rechenergebnisses zu beurteilen, das mittels eines Rechenprogramms erzielt wurde. Das Rechenprogramm sehen wir als rechnerinterne Realisierung eines numerischen Algorithmus an. Das rechnerinterne Abbild eines numerischen Algorithmus muß leider nicht notwendigerweise die gewünschten Ergebnisse liefern.
Beispiel. 1 Wir wollen die Größe

$$w = 9x^4 - y^4 + 2y^2$$

für $x = 10864.0, y = 18817.0$ berechnen. Dazu können wir 4 mathematisch gleichwertige Formeln verwenden:

$$w1 = 9 \cdot x \cdot x \cdot x \cdot x - y^4 + 2 \cdot y \cdot y,$$
$$w_2 = (3 \cdot x \cdot x - y \cdot y) \cdot (3 \cdot x \cdot x + y \cdot y) + 2 \cdot y \cdot y,$$
$$w_3 = 9 \cdot x^4 + (2 \cdot y^2 - y \cdot y \cdot y \cdot y),$$
$$w_4 = (9 \cdot x^4 + 2 \cdot y^2) - y^4.$$

Bei 7-stelliger Rechnung erhalten wir auf einem gewissen Rechner die Werte

$$w_1 = 236052992.0, \quad w_2 = 708158976.0, \quad w_3 = 0.0, \quad w_4 = 0.0$$

und bei 16-stelliger Rechnung

$$w_1 = -320.0, \quad w_2 = 1.0, \quad w_3 = 160.0, \quad w_4 = -160.0.$$

Wenn man die Werte von x und y in der Eingabe vertauscht, liefern alle Formeln sowohl bei 7- als auch bei 16-stelliger Rechnung das gleiche Ergebnis, nämlich $w = 1.114420 \cdot 10^{19}$. Was ist hier richtig? Das Beispiel lehrt uns zunächst, daß mathematisch gleichwertige Formeln in einem Rechenprogramm verschiedene Resultate liefern können. Vergleichen wir die Berechnungen miteinander, so müssen wir berechtigt vermuten, daß bekannte Rechengestze für die reellen Zahlen, wie die Assoziativität der Addition auf dem Rechner nicht gelten. Um die verschiedenen Ergebnisse bei der Berechnung der gleichen mathematischen Formel erklären zu können, müssen wir genauer untersuchen, was unsere Vorstellung bezüglich der Ausführung eines numerischen Algorithmus von der wirklichen Ausführung im Rahmen eines Rechenprogramms auf dem Rechner unterscheidet.
Beantworten wir zunächst die Frage nach dem Weg, auf dem man nach der Formulierung einer Aufgabe zu einem Maschinenergebnis gelangt. Aus einer konkreten Aufgabe wird ein Modell für die Aufgabe entworfen. In dieses Modell gehen Daten ein, die wir aus unterschiedlichen Gründen nicht genau kennen. Wegen dieser Datenunsicherheit wird das Modell viele mögliche Ergebnisse haben. Aus diesem Modell muß man durch geeignete Idealisierungen – Anwendung von Theorien und Gesetzmäßigkeiten, Vernachläßigung von Einflüssen und Abhängigkeiten, die man als unwesentlich ansieht – eine mathematische Aufgabenstellung herausarbeiten. Diese Aufgabe ist einer mathematischen Analyse zugängig. Insbesondere müssen die Existenz und Eindeutigkeit von Lösungen untersucht werden. Dabei ist die Untersuchung der Existenz einer Lösung besonders wichtig. Dazu ein Beispiel: Nehmen wir an, es gibt eine größte natürliche Zahl. Wenn wir uns nicht um die Richtigkeit dieser These kümmern würden, zeigt uns eine kleine mathematische Analyse (selbst durchführen!), daß es überhaupt nur die natürlichen Zahlen 0 und 1 gibt. Aber auch die Untersuchung der Eindeutigkeit der Lösung ist wichtig, zumal man oft weiß, daß es aus nicht modellierten Beziehungen nur eine Lösung geben darf.
Eine mathematische Aufgabe, die einer numerischen Behandlung zugängig ist, besteht i. a. aus Gleichungen, Ungleichungen oder ähnlichen Beziehungen zwischen bekannten Größen und Funktionen, den sog. Daten und unbekannten Größen und Funktionen. Durch die numerische Rechnung kann lediglich über eine endliche Folge von

Operationen aus den gegebenen Größen eine Reihe von Zahlen gewonnen werden; sie sind entweder Näherungswerte für die gesuchte Lösung oder bestimmen als Parameter eine Näherungsfunktion, die Näherungslösung. Die Spezifikation einer endlichen Folge von Rechenschritten, die aus den Daten diese Ergebniswerte erzeugen, heißt **numerischer Algorithmus**. Damit auf einem konkreten Rechner aus konkret vorgegebenen Daten nach einem spezifizierten Algorithmus Zahlenwerte für die Ergebnisgrößen berechnet werden können, muß der Algorithmus in ein Rechenprogramm umgesetzt werden. Dies geschieht meist durch Formulierung des Algorithmus in einer Programmiersprache. Das Rechenprogramm wird sodann durch Komponenten des Betriebssystems in eine Folge von Anweisungen übersetzt, die der Rechner ausführen kann, das sog. Maschinenprogramm. Die Umwandlung eines numerischen Algorithmus in ein auf dem Rechner ausführbares Rechenprogramm bezeichnet man als **Implementierung** des numerischen Algorithmus. Das Rechenprogramm operiert dabei stets nur mit solchen Zahlen, die auf dem Rechner dargestellt werden können; das sind aber nur endlich viele. Auf diesem Wege der Aufgabentransformation liegt kein Isomorphismus vor; bei jedem Schritt gehen Informationen verloren. Damit ist klar, daß das vom Rechner präsentierte Ergebnis praktisch nie mit dem gesuchten übereinstimmen kann. Die Abweichung eines nach einer bestimmten Vorschrift sich ergebenden Resultates von dem gewünschten nennt man in der numerischen Mathematik **Fehler**. Die drei wesentlichen Fehlerarten sind folgende.

1. **Rechenfehler**: Sie sind Folge der Implementierung eines numerischen Algorithmus; anstelle des gewünschten Resultates berechnet das Programm ein Maschinenresultat. Bei jeder arithmetischen Operation muß das Ergebnis auf eine Rechnerzahl abgebildet, gerundet werden.

2. **Verfahrensfehler**: Nur bei wenigen mathematischen Aufgaben kann ein numerischer Algorithmus angegeben werden, dessen Ergebis mit dem Ergebnis der mathematischen Aufgabe übereinstimmt. Meist sind unendliche Algorithmen in geeigneter Weise durch endliche zu ersetzen; ein unendlicher Algorithmus muß nach endlicher Zeit abgebrochen werden. Das erzielte Ergebnis wird daher vom wahren abweichen; diese Abweichung nennt man Verfahrensfehler.

3. **Eingabefehler**: Wegen der idealisierten Annahmen im mathematischen Modell sind einige Daten i. a. mit beträchtlichen Unsicherheiten behaftet. Solche Unsicherheiten bewirken natürlich auch Unsicherheiten im Ergebnis der mathematischen Aufgabe. Kleine Änderungen in den Eingabedaten des Problems bewirken u. U. große Änderungen in den Ergebnissen. Diese Fehlerart hat einen prinzipiell anderen Charakter als Rechen- und Verfahrensfehler, weil man hier nicht genau sagen kann, was eine Lösung der Aufgabe sein soll. Die Auswirkung dieser Unsicherheit auf das Ergebnis des mathematischen Modells begrenzt unmittelbar die Genauigkeit, mit der die numerische Lösung des entsprechenden mathematischen Modells sinnvollerweise anzugeben ist. Der Eingabefehler begrenzt damit auch die Genauigkeit, die bei einem Maschinenergebnis anzustreben ist. Andererseits kann der Eingabefehler auch so schwerwiegend sein, daß die im Rechner vorhandene Aufgabe mit jedem numerischen Algorithmus eine Lösung liefert, die mit der wahren nichts zu tun hat.

Die theoretischen Überlegungen sollen nun an weiteren Beispielen illustriert werden.
Beispiel 2. Wir wollen das folgende lineare Gleichungssystem in 4 Variablen lösen:

$$\begin{aligned} x + \frac{1}{2}y + \frac{1}{3}u + \frac{1}{4}v &= 1 \\ \frac{1}{2}x + \frac{1}{3}y + \frac{1}{4}u + \frac{1}{5}v &= 1 \\ \frac{1}{3}x + \frac{1}{4}y + \frac{1}{5}u + \frac{1}{6}v &= 1 \\ \frac{1}{4}x + \frac{1}{5}y + \frac{1}{6}u + \frac{1}{7}v &= 1. \end{aligned}$$

Durch Einsetzen überzeugen wir uns, daß $x = -4, y = 60, u = -180, v = 140$ die exakte Lösung ist. Einige der Koeffizienten sind nicht exakt im Rechner darstellbar; ihnen müssen bei der Eingabe Maschinenzahlen zugeordnet werden. Wenn wir diese Zahlen der Reihe nach mit 4, 5, 6 und 8 Ziffern eingeben, erhalten wir die folgenden Maschinenergebnisse:

	x	y	u	v
4 :	−5.8999	80.5437	−228.5033	171.1528
5 :	−4.1814	61.9951	−184.7562	143.0748
6 :	−4.0262	60.2963	−180.7181	140.4694
8 :	−4.0003	60.0033	−180.0080	140.0052.

Wir erkennen insbesondere, daß erst ab einer Eingabegenauigkeit von 5 Ziffern sich die Maschinenlösung der wahren Lösung annähert. Außerdem erkennen wir, daß die Lösung sehr empfindlich auf eine Änderung der Eingabedaten reagiert. Die gefundenen Lösungen sind in dem Sinne exakt, daß sie Lösungen jener Aufgaben sind,

die sich im Rechner befinden. Die sensible Änderung der Lösung bei Änderung der Eingabedaten ist nicht etwa Folge des verwendeten Algorithmus, sondern eine Eigenschaft der Aufgabe selbst. Man nennt eine numerische Aufgabe **stabil**, wenn der Fehler in der Lösung in der Größenordnung des Eingabefehlers liegt. In diesem Sinne ist obige Aufgabe instabil. Die Verstärkung des Eingabefehlers in der Lösung kann kein noch so ausgefeilter Algorithmus verhindern; er ist unvermeidbar. Instabile Aufgaben verlangen zu ihrer numerischen Behandlung solche Algorithmen, die den Eingabefehler möglichst nicht noch zusätzlich verstärken. Instabile Aufgaben treten oft in den Anwendungen auf. Die meistens Jugendlichen haben schon einmal versucht, einen flachen Stein über das Wasser springen zu lassen. Sie wissen aus Erfahrung, daß man nicht nur einen geeigneten Stein finden muß, sondern auch ziemlich genau auf den richtigen Anstellwinkel und eine hohe Anfangsgeschwindigkeit zu achten hat. Schon eine kleine Änderung dieser beiden Parameter führt zur Erfolglosigkeit.

Beispiel 3. Wir wollen die Quadratwurzel aus einer positiven Zahl ziehen. Dies ist eine Aufgabe, die sich auf einem Rechner nicht exakt ausführen läßt. Also muß man eine Näherungsmethode anwenden. Ist a eine positive Zahl, aus der die Quadratwurzel gezogen werden soll, so suchen wir eine Zahl $x > 0$ mit $x^2 = a$ oder besser $x = \frac{a}{x}$. Wählt man nun x, so wird i. a. die eine Seite der Gleichung größer als die andere sein; daher dürfte der Mittelwert von beiden Seiten ein besserer Näherungswert für die gesuchte Wurzel sein: $x := \frac{1}{2}(x + \frac{a}{x})$; mit diesem neuen x vergleicht man wieder beide Seiten usw., d. h. wir können die Gleichung iterieren:

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right), \quad n = 0, 1, 2, \ldots$$

In der folgenden Tabelle sind die berechneten Näherungen mit den exakten vergleichen. Dabei wurden jeweils 4 Iterationen ausgeführt und als Startwert a genommen:

Wert	Wurzel	berechnet	Fehler
0.000001	0.001000	0.062505	6150.53%
0.000010	0.003162	0.062553	1878.10%
0.000100	0.010000	0.063030	530.31%
0.001000	0.031623	0.067725	114.66%
0.010000	0.100000	0.108404	8.40%
0.100000	0.316228	0.316246	0.0056%
1.000000	1.000000	1.000000	0.0000%
10.00000	3.162278	3.162456	0.0056%
100.0000	10.00000	10.84044	8.40%
1000.000	31.62278	67.72538	114.66%
10000.00	100.0000	630.3036	530.31%
100000.0	316.2278	6255.312	1878.10%

Eine oberflächliche Betrachtung dieser Tabelle zeigt uns folgendes: Bei 4 Iterationen ist der Verfahrensfehler für Zahlen außerhalb des Intervalls $[\frac{1}{10}, 10]$ so groß, daß die Methode unbrauchbar erscheint, falls man nicht eine erheblich höhere Anzahl von Iterationen verwendet. Andererseits sollte uns aber die besonders einfache Iterationsformel ermutigen darüber nachzudenken, wie man sie eventuell trotzdem anwenden kann. Eine weitere Überlegung führt uns darauf, daß das Wurzelziehen nur für Zahlen aus dem Intervall $[\frac{1}{10}, 10]$ numerisch ausgeführt werden muß. Jede Zahl a kann man nämlich als Produkt einer geradzahligen Potenz von 10 und einer Zahl aus diesem Intervall darstellen:

$$a = 10^{2n} \cdot b, \quad 0.1 \leqq b \leqq 10.$$

Für den ersten Faktor läßt sich die Wurzel sofort angeben, während die Wurzel für den zweiten Faktor nach der obigen Methode berechnet werden kann. Die folgende Tabelle zeigt die Ergebnisse nach dieser Änderung und mit 3 Iterationen, wobei als Startwert stets $(1 + a)/2$ genommen wurde:

Wert	Wurzel	berechnet	Fehler
0.000001	0.001000	0.001000	0.0000%
0.000010	0.003162	0.003162	0.0056%
0.000100	0.010000	0.010000	0.0000%
0.001000	0.031623	0.031625	0.0056%
0.010000	0.100000	0.100000	0.0000%
0.100000	0.316228	0.316246	0.0056%
1.000000	1.000000	1.000000	0.0000%
10.00000	3.162278	3.162456	0.0056%
100.0000	10.00000	10.00000	0.0000%
1000.000	31.62278	31.62456	0.0056%
10000.00	100.0000	100.0000	0.0000%
100000.0	316.2278	316.2456	0.0056%

Wir erkennen. daß sich bei Ausnutzung einer einfachen Tatsache der Verfahrensfehler drastisch reduziert und dadurch die Maschinenergebnisse brauchbar werden. Durch eine geeignete Wahl des Startwertes konnte der Operationsaufwand noch reduziert werden. In anderen Fällen wird man weit kompliziertere mathematische Sachverhalte oder Theorien heranziehen müssen, um solche Arbeitsbereiche für numerische Algorithmen zu finden, die den Verfahrensfehler möglichst gering halten. Oft ist es auch schwierig oder gar unmöglich, realistische Abschätzungen für den Verfahrensfehler zu gewinnen. Es gibt aber auch Verfahren, bei denen jede Mühe vergebens ist, weil der Rechenfehler in entscheidender Weise dominiert, wie das folgende Beispiel zeigt.
Beispiel 4. Es soll das Integral

$$I_n = \int_0^1 x^n e^{1-x} dx$$

berechnet werden. Dazu kann man den folgenden Zusammenhang ausnutzen:

$$I_0 = e - 1 = 1.7182818284590\ldots$$
$$I_n = -1 + n \cdot I_{n-1}, \quad n = 1, 2, 3, \ldots$$
$$0 < I_{n+1} < I_n, \quad \lim_{n\to\infty} I_n = 0.$$

Die folgende Tabelle zeigt die Werte von I_n bis $n = 22$ bei Rechnung mit 7 bzw. 16 Ziffern:

n	$I_n(7-\text{stellig})$	$I_n(16-\text{stellig})$
0	1.71828	1.71828
2	0.43656	0.43656
4	0.23876	0.23876
6	0.16294	0.16292
8	0.12480	0.12332
10	0.2315	0.09911
12	17.6434	0.08281
14	3196.10	0.07110
16	767048	0.06506
18	Überlauf	0.90685
20		323.604
22		149482.0

In Übereinstimmung mit unserer mathematischen Erwartung werden die berechneten Integralwerte zunächst mit jedem Schritt kleiner. Doch urplötzlich wachsen sie von Schritt zu Schritt immer rascher. Haben wir vielleicht einen Fehler gemacht? Um dies festzustellen, wenden wir die Formel rückwärts an, indem wir annehmen, daß für ein großes N bereits $I_N = 0$ gilt. Die Rückwärtsformel lautet

$$I_{n-1} = \frac{1 + I_n}{n}, \quad n = N, N-1, \ldots, 1.$$

Die folgende Tabelle zeigt die erhaltenen Werte:

n	$I_n(7-\text{stellig})$	$I_n(16-\text{stellig})$
7	0.14028	0.14028
6	0.16290	0.16290
5	0.19382	0.19382
4	0.23876	0.23976
3	0.30369	0.30369
2	0.43656	0.43656
1	0.71828	0.71828
0	1.71828	1.71828

Wir erkennen, daß unsere mathematischen Voraussetzungen an die Methode richtig waren. Als Ursache für das äußerst fehlerhafte Maschinenergebnis müssen wir die sich schnell aufschaukelnden Rechenfehler ansehen. Die obige Vorwärtsformel zeigt, daß der Rechenfehler aus dem Schritt $n-1$ im Schritt n um den Faktor n verstärkt werd; also verstärkt sich der Eingabefehler aus dem Schritt 0 im n-ten Schritt um den Faktor $1 \cdot 2 \cdot \ldots \cdot n = n!$, so daß schon der unvermeidbare Eingabefehler zum Versagen der Methode führen muß. Um diese Aussage noch

zu stützen, wollen wir den Startwert mit nur 4 Ziffern eingeben. Die erhaltenen Werte zeigt die nächste Tabelle:

n	$I_n(7-\text{stellig})$	$I_n(16-\text{stellig})$
0	1.71800	1.71800
1	0.71800	0.71800
2	0.44300	0.43600
3	0.30800	0.30800
4	0.23200	0.23200
5	0.16001	0.16000
6	−0.03999	−0.40000
7	−1.27973	−1.28000
8	−11.2378	−11.2400
9	−102.140	−102.160
10	−1022.40	−1022.60
11	−11247.4	−11249.6
12	Überlauf	−134996

Im Gegensatz zu der obigen Rechnung strebt hier die Folge der berechneten Integralwerte sogar gegen $-\infty$! Die Diskussion dieses Beispiels zeigt insbesondere, daß es numerische Methoden gibt, bei denen ein winziger Fehler in den Eingabedaten zu katastrophalen Fehlern im Maschinenergebnis führt. Solche Methoden nennt man **instabil**.

6.2. Rechnerzahlen und Rundung

Aus technischen Gründen stützen sich die elektronischen Rechner auf das Dualsystem, indem die Koeffizienten α_i der Dualdarstellung einer reellen Zahl

$$x = \pm(\alpha_n 2^n + \alpha_{n-1}2^{n-1} + \cdots + \alpha_0 2^0 + \alpha_{-1}2^{-1} + \cdots), \quad \alpha_i \in \{0,1\}, \alpha_n = 1$$

benutzt werden. Um Verwechslungen zu vermeiden, bezeichnet man oft in der Dualdarstellung die Zahlen 0 und 1 mit **O** und **L**. So gilt z. B.

$$19.5 = \textbf{LOOLL.L}.$$

Die Dezimaldarstellung einer reellen Zahl ist nicht eindeutig, z. B. gilt

$$1.9999\cdots = 2.$$

Wir wählen hier stets die endliche Form, falls eine solche existiert. Konstruktionsbedingt gibt es für die Zahlendarstellung im Rechner nur eine feste Anzahl n – die **Wortlänge** – von Dualstellen. Meist läßt sich die Wortlänge nur in Vielfachen ändern. Um die Wortänge voll ausnutzen zu können, werden im allgemeinen Zahlen in normalisierter Form dargestellt:

$$x = a_x \cdot 2^{b_x}, \quad b \in \mathbb{Z}, \quad a \in [0.5, 1),$$

z. B. für $x = 19.5$:

$$x = 0.\textbf{LOOLLL}_2\textbf{LOL}.$$

Für die **Mantisse** a_x gibt es dabei m und für den Exponenten e Stellen mit $n = m + e$. Das Tripel $(2, e, m)$ charakterisiert vollständig die Menge A aller Zahlen, die in einem Rechner exakt dargestellt werden können; ihre Elemente heißen **Rechnerzahl**en. Meist wird anstelle der Basis 2 die Basis 8 oder 16 genommen; im letzteren Falle liegt die Mantisse stets zwischen 0.0625 (einschließlich) und 1 (ausschließlich).
Die Anzahl der Rechnerzahlen ist endlich. Daher entsteht die Frage, wie man eine reelle Zahl $x \notin A$ durch eine Rechnerzahl approximieren sollte. Dieses Problem steht sowohl bei der Eingabe von Zahlen als auch bei arithmetischen Operationen, die i. a. keine Rechnerzahlen liefern werden. Von einer sinnvolen Approximation einer Zahl x durch eine Rechnerzahl $\overline{x}$ wird man verlangen, daß

$$|x - \overline{x}| \leqq |x - y| \quad \forall y \in A$$

gilt. Eine solche Zahl erhält man gewöhnlich durch Rundung. Allgemein wird bei einer m-stelligen Dezimalmaschine eine reelle Zahl x wie folgt gerundet. Es sei

$$x = a \cdot 10^b, \quad |a| \geqq 10^{-1},$$
$$|a| = 0.\alpha_1\alpha_2\ldots\alpha_m\alpha_{m+1}\ldots, \quad 0 \leqq \alpha_i \leqq 9, \quad \alpha_1 \neq 0.$$

Man bildet

$$a' = \begin{cases} 0.\alpha_1\alpha_2\ldots\alpha_m & \alpha_{m+1} \leqq 4 \\ 0.\alpha_1\alpha_2\ldots\alpha_m + 10^{-m} & \alpha_{m+1} \geqq 5 \end{cases}$$

und danach

$$\overline{x} = \text{sign}(x) \cdot a' \cdot 10^b.$$

Dann ergibt sich der relative Fehler von $\overline{x}$ zu

$$\left|\frac{\overline{x}-x}{x}\right| = \left|\frac{a'-|a|}{a}\right| \leqq \begin{cases} \dfrac{0.\alpha_{m+1}\ldots\cdot 10^{-m}}{0.\alpha_1\alpha_2\ldots} \\ \dfrac{10^{-m} - 0.\alpha_{m+1}\alpha_{m+2}\ldots\cdot 10^{-m}}{0.\alpha_1\alpha_2\ldots} \end{cases} \leqq 5\cdot 10^{-m},$$

also mit der Abkürzung $eps = 5 \cdot 10^{-m}$:

$$\overline{x} = x(1+\varepsilon), \quad |\varepsilon| \leqq eps.$$

Auf einem konkreten Rechner bestimmt man *eps* als kleinste positive Rechnerzahl, für die der Test „if $1.+eps > 1.$“ positiv ausfällt.
Wegen des hohen konstruktiven Aufwandes vollzieht sich die Rechner-Rundung nach anderen Prizipien, jedoch meist so, daß sie bis auf einen konstanten Faktor mit der obigen übereinstimmt.
Für den Exponenten einer Rechnerzahl ist nur eine beschränkte Stellenzahl reserviert; daher kann es während der Rechnung zu Exponenten-Unterlauf bzw. zu Exponenten-Überlauf kommen. Der erste Fall wird meist ohne Fehlermeldung übergangen, während bei Exponenten-Überlauf das Programm mit einem Laufzeitfehler abbricht. Wir wollen hier die Stellenzahl e für den Exponenten als hinreichend groß annehmen.
Da arithmetische Operationen mit Rechnerzahlen i. a. keine Rechnerzahlen liefern, sind sie als Ersatzoperationen (Real-Operationen) $(+), (-), (\cdot), (/)$ realisiert, etwa in der Form

$$x(\circ)y = \overline{x \circ y}, \quad \circ \in \{+,-,\cdot,/\},$$

so daß

$$x(\circ)y = (x \circ y)(1+\varepsilon), \quad |\varepsilon| \leqq eps$$

gilt. Wegen der meist etwas anderen Rundung wird der Fehler etwas größer sein, jedoch so, daß noch $|\varepsilon| \leqq \nu \cdot eps$ mit $\nu \geqq 1$ gilt.
Interessant und wichtig ist der Fall der **Auslöschung**, der bei der Subtraktion zweier Rechnerzahlen x, y mit gleichen Vorzeichen, Exponenten und übereinstimmenden führenden Mantissenstellen eintritt, z. B. bei

$$x = 0.315876 \cdot 10^1,$$
$$y = 0.315289 \cdot 10^1.$$

Die Differenz $x - y$ ist wieder eine Rechnerzahl, so daß die Operation exakt ausgeführt wird:

$$x(-)y = x - y = 0.587000 \cdot 10^{-2}.$$

Jedoch geraten wegen der Normalisierung von $x - y$ alte Rundungsgehler in höhere Mantissenstellen. Waren etwa bei x und y noch die ersten 3 Ziffern richtig, so ist bei der Differenz keine Ziffer mehr sicher. Also werden jene Fehler, die bei der Berechnung von x und y vor der Subtraktion entstanden, verstärkt. Eine rechnerunabhängige Näherungsmethode zur Auslöschungsmessung ist im Programm **AUSL** implementiert. Dabei wird die Auslöschung bei Zahlen, die kleiner als 1 sind, als absoluter, und bei Zahlen größer 1 als relativer Fehler gemessen. Für das Ergebnis von Real-Operationen hat sich die Schreibweise

$$\text{gl}(x \circ y) = x(\circ)y$$

eingebürgert, die wir auch verwenden wollen.
Eine numerische Aufgabe besteht darin, aus gewissen Zahlen $x_1, x_2, \ldots, x_n$ (Input) gewisse andere Zahlen $y_1, y_2, \ldots, y_m$ (Output) zu berechnen.
Ein Problem dieser Art zu lösen bedeutet, den Wert $\mathbf{y}$ einer gewissen Vektorfunktion $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_m)$ im Punkte $\mathbf{x}$ zu bestimmen:

$$y_i = \varphi_i(x_1, x_2, \ldots, x_n), \quad i = 1, 2, \ldots, m.$$

Ein Algorithmus ist eine endliche Rechenvorschrift zur Berechnung von $\varphi(\mathbf{x})$. Die Abbildung φ sei Verknüpfung von elementaren Operationen:

$$\varphi = \varphi^{(r)} \circ \varphi^{(r-1)} \circ \ldots \circ \varphi^{(1)} \circ \varphi^{(0)}.$$

Unter den elementaren Operationen kann man etwa die arithmetischen Operationen und die üblichen Standardfunktionen verstehen, wie sie über einen Sprachübersetzer verfügbar sind. Auf einem Rechner sind für die elementaren Operationen $\varphi^{(i)}$ Ersatzabbildungen $\mathrm{gl}(\varphi^{(i)})$ implementiert. Entsprechend ist

$$\mathrm{gl}(\varphi^{(i)})(\mathbf{x}^{(i)}) - \varphi^{(i)}(\mathbf{x}^{(i)})$$

der Rundungsfehler, der bei der Berechnung von $\varphi^{(i)}(\mathbf{x}^{(i)})$ auf dem Rechner entsteht.
Beispiel 1. Es sei $\varphi(a,b,c) = a+b+c$. Wir haben zwei Algorithmen:

ALG1:	ALG2:
$\eta = a+b$	$\eta = b+c$
$y = c+\eta$	$y = a+\eta$
$\varphi^{(0)}(a,b,c) = \begin{pmatrix} a+b \\ c \end{pmatrix}$	$\varphi^{(0)}(a,b,c) = \begin{pmatrix} a \\ b+c \end{pmatrix}$

$$\varphi^{(1)}(u,v) = u+v$$
$$\varphi = \varphi^{(1)} \circ \varphi^{(0)}.$$

Beispiel 2. Es sei $\varphi(a,b) = a^2 - b^2$. Auch hier haben wir zwei Algorithmen:

ALG1:	ALG2:
$\eta_1 = a \cdot a$	$\eta_1 = a+b$
$\eta_2 = b \cdot b$	$\eta_2 = a-b$
$y = \eta_1 - \eta_2$	$y = \eta_1 \cdot \eta_2$
$\varphi^{(0)}(a,b) = \begin{pmatrix} a^2 \\ b^2 \end{pmatrix}$	$\varphi^{(0)}(a,b) = \begin{pmatrix} a+b \\ a-b \end{pmatrix}$
$\varphi^{(1)}(u,v) = u-v$	$\varphi^{(1)}(u,v) = u \cdot v$

Am ersten Beispiel soll gezeigt werden, daß verschiedene Algorithmen zur Lösung eines Problems verschiedene Resultate liefern. In ALG 1 erhält man für

$$y = a+b+c$$

einen Näherungswert

$$\tilde{y} = \mathrm{gl}(\mathrm{gl}(a+b)+c)$$

mit

$$\begin{aligned}
\eta &= \mathrm{gl}(a+b) = (a+b)(1+\varepsilon_1) \\
\tilde{y} &= \mathrm{gl}(\eta + c) = (\eta + c)(1+\varepsilon_2) \\
&= [(a+b)(1+\varepsilon_1)+c](1+\varepsilon_2) \\
&= (a+b+c)\left[1 + \frac{a+b}{a+b+c}\varepsilon_1(1+\varepsilon_2) + \varepsilon_2\right].
\end{aligned}$$

Für den relativen Fehler ε_y von $\tilde{y}$ folgt

$$\varepsilon_y = \frac{\tilde{y}-y}{y} = \frac{a+b}{a+b+c}\varepsilon_1(1+\varepsilon_2) + \varepsilon_2$$

und in erster Näherung

$$\varepsilon_y \doteq \frac{a+b}{a+b+c}\varepsilon_1 + 1 \cdot \varepsilon_2.$$

Die beiden Faktoren vor ε_1 und ε_2 geben an, wie sich die Rundungsfehler $\varepsilon_1, \varepsilon_2$ im relativen Fehler des Ergebnisses verstärken. Der kritische Faktor ist dabei jener vor ε_1; je nachdem, welcher der beiden Faktoren $|a+b|, |b+c|$ kleiner ist, wird es numerisch günstiger, den ersten bzw. den zweiten Algorithmus anzuwenden.
Man nennt einen Algorithmus zur Berechnung von $\varphi(\mathbf{x})$ **numerisch stabiler** als einen zweiten, falls der Gesamtfehler beim ersten Algorithmus kleiner als beim zweiten ist.

6.3. Interpolation

Das Interpolationsproblem ist ein grundlegendes innerhalb der numerischen Mathematik. Wir formulieren es in folgender Form. Es seien eine Funktion Φ:

$$y = \Phi(x; a_0, a_1, \ldots, a_n)$$

und $n+1$ Paare $(x_i, y_i), i = 0, 1, \ldots, n, x_i \neq x_k$ für $i \neq k$ gegeben. Die Funktion Φ hänge von $n+1$ unbekannten Parametern $a_0, a_1, \ldots, a_n$ ab. Die Paare nennt man **Stützstellen** oder auch **Stützpunkte**. Die unbekannten Parameter sind so zu bestimmen, daß

$$\Phi(x_i; a_0, a_1, \ldots, a_n) = y_i, \quad i = 0, 1, \ldots, n$$

gilt. Ein Interpolationsproblem heißt **linear** , wenn die Funktion Φ linear von den Parametern abhängt, also die Form

$$\Phi(x; a_0, a_1, \ldots, a_n) = \sum_{i=0}^{n} a_i \Phi_i(x)$$

hat. Zu den linearen Interpolationsproblemen gehören die **Polynom-Interpolation** mit

$$\Phi(x; a_0, a_1, \ldots, a_n) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

die **trigonometrische Interpolation**

$$\Phi(x; a_0, a_1, \ldots, a_n) = a_0 + a_1 e^{ix} + a_2 e^{2ix} + \cdots + a_n e^{nix} \quad (i^2 = -1)$$

und die **Spline-Interpolation**, bei der im Falle kubischer Splines eine Funktion Φ benutzt wird, die zweimal steitg differenzierbar ist und in jedem Teilintervall $[x_i, x_{i+1}]$ mit einem Polynom 3. Grades übereinstimmt.
Interpolationsaufgaben treten sehr vielfältig auf. Polynom-Interpolation verwendet man zur näherungsweisen Berechnung von Werten einer Funktion, die nur an diskreten Stellen gegeben ist. Auch bei Näherungsformeln für die numerische Integration tritt Polynom-Interpolation auf. Die trigonometrische Interpolation wird meist für die numerische Auswertung von Meßreihen verwendet. Die Spline-Interpolation benutzt man zum Zeichnen von Kurven, die möglichst glatt durch vorgegebene Punkte verlaufen sollen.
Zu den nichtlinearen Interpolationsaufgaben gehören die Interpolation durch rationale Funktionen

$$\Phi(x; a_0, \ldots, a_n, b_0, \ldots, b_m) = \frac{a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0}{b_m x^m + a_{m-1} x^{m-1} + \cdots + b_1 x + b_0}$$

und die Interpolation durch Exponentialsummen

$$\Phi(x; a_0, \ldots, a_n, \lambda_0, \ldots, \lambda_n) = a_0 e^{\lambda_0 x} + a_1 e^{\lambda_1 x} + \ldots + a_n e^{\lambda_n x}.$$

Rationale Interpolation verwendet man zur Konvergenzbeschleunigung von Algorithmen; die Interpolation durch Exponentialreihen wird in Physik und Chemie bei der Analyse von Zerfallsreihen eingesetzt.
Wir besprechen hier die Polynominterpolation und die Interpolation mittels natürlicher kubischer Splinefunktionen.
Es sei Π_n die Menge aller Polynome P vom Grade höchstens n:

$$P(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

Satz 153. (Existenz- und Eindeutigkeitssatz)
Zu beliebigen $n+1$ Stützstellen $(x_i, y_i), i = 0, 1, \ldots, n, x_i \neq x_k (i \neq k)$ gibt es genau ein Polynom $P \in \Pi_n$ mit $P(x_i) = y_i, i = 0, 1, \ldots, n$.

Beweis. Zunächst zeigen wir, daß es höchstens ein solches Polynom gibt. Angenommen, die Polynome P, Q erfüllen die Bedingungen des Satzes, also

$$P(x_i) = Q(x_i) = y_i, \quad i = 0, 1, \ldots, n.$$

Dann hat das Polynom $R = P - Q$ vom Grade höchstens n mindestens $n+1$ Nullstellen:

$$R(x_i) = 0, \quad i = 0, 1, \ldots, n.$$

Ein Nichtnull-Polynom vom Grade n kann aber nur n Nullstellen haben; also folgt $P = Q$.
Die im Satz behauptete Existenz des Polynoms beweisen wir direkt. Es sei L_i die Indikatorfunktion von $\{x_i\}$ bezüglich der Menge $\{ x_0, x_1, \ldots, x_n \}$, d. h.

$$L_i(x_k) = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases} .$$

Offenbar gilt

$$\begin{aligned} L_i(x) &= \frac{(x-x_0)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_0)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)} \\ &= \frac{\omega(x)}{(x-x_i)\omega'(x_i)} \end{aligned}$$

mit

$$\omega(x) = (x-x_0)(x-x_1)\dots(x-x_n).$$

Wir setzen

$$P(x) = \sum_{i=0}^{n} y_i L_i(x) = \sum_{i=0}^{n} y_i \prod_{\substack{k=0 \\ k\neq i}}^{n} \frac{x-x_k}{x_i-x_k}.$$

Dieses Polynom leistet das Verlangte; es heißt **Langrange-sches Interpolationspolynom**. □
Die Indikatorfunktionen $L_i(x)$ hängen nicht von den y_i ab; daher erhalten wir aus $P(x) = 1$ also $y_i = 1, i = 1, \dots, n$ die Bedingung

$$\sum_{i=0}^{n} L_i(x) = 1.$$

Für die algorithmische Berechnung eines Funktionswertes für das Lagrange-sche Interpolationspolynom eignet sich besonders der **Neville-Algorithmus**.

Satz 154 (Neville-Algorithmus). *Es sei P_{ij} $(i \geqq 0)$ das Interpolationspolynom höchstens j-ten Grades $(i \geqq j)$ mit*

$$P_{ij}(x_k) = y_k, \quad k = i-j, \dots, i.$$

Dann gilt für fixiertes x:

$$\begin{aligned} &P_{i0}(x) = y_i, \\ &P_{ij}(x) = \frac{(x-x_{i-j})P_{i,j-1}(x) - (x-x_i)P_{i-1,j-1}(x)}{x_i - x_{i-j}}, \quad j = 1, \dots, i. \end{aligned}$$

Beweis. Die Richtigkeit dieser Formel sieht man wie folgt ein. Es ist

$$\begin{aligned} P_{i,j-1}(x_k) &= y_k, \quad k = i-j+1, \dots, i-1, i, \\ P_{i-1,j-1}(x_k) &= y_k, \quad k = i-j, \dots, i-1, \end{aligned}$$

also folgt für die rechte Seite der Formel, die mit $P(x)$ bezeichnet werden soll:

$$\begin{aligned} P(x_{i-j}) &= P_{i-1,j-1}(x_{i-j}) = y_{i-j}, \\ P(x_k) &= \frac{(x_k - x_{i-j})y_k - (x_k - x_i)y_k}{x_i - x_{i-j}} = y_k, \quad k = i-j+1, \dots, i-1, \\ P(x_i) &= P_{i,j-1}(x_i) = y_i, \end{aligned}$$

d. h. P ist ein interpolierendes Polynom für die Stützstellen (x_k, y_k), $k = i-j, \dots, i$. Wegen der Eindeutigkeit dieses Polynoms muß $P = P_{ij}$ sein. □
Der Neville-Algorithmus berechnet somit nach der obigen Formel folgendes Schema, das die Werte der interpolierenden Polynome P_{ij} an der Stelle x enthält:

$$\begin{array}{c|cccc} x_0 & P_{00} & & & \\ x_1 & P_{10} & P_{11} & & \\ x_2 & P_{20} & P_{21} & P_{22} & \\ x_3 & P_{30} & P_{31} & P_{32} & P_{33} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

mit z. B.

$$P_{32} = \frac{(x-x_1)P_{31} - (x-x_3)P_{21}}{x_3 - x_1}.$$

Eine weitere Möglichkeit zur Berechnung des Lagrange-schen Interpolationspolynoms ergibt sich durch folgende Betrachtung. Der Nenner der Indikatorfunktion L_i hängt nicht von x ab; wir setzen also

$$a_i = \frac{1}{\prod\limits_{\substack{k=0\\k\neq i}}^{n}(x_i - x_k)}$$

und erhalten

$$P(x) = \sum_{i=0}^{n} y_i a_i \prod_{\substack{k=0\\k\neq i}}^{n}(x - x_k).$$

Wegen

$$1 = \sum_{i=0}^{n} L_i(x) = \sum_{i=0}^{n} a_i \prod_{\substack{k=0\\k\neq i}}^{n}(x - x_k) \quad \text{und} \quad \prod_{\substack{k=0\\k\neq i}}^{n}(x - x_k) = \frac{\prod\limits_{k=0}^{n}(x - x_k)}{x - x_i}$$

erhalten wir schließlich

$$P(x) = \frac{\sum\limits_{i=0}^{n} \frac{a_i}{x - x_i} y_i}{\sum\limits_{i=0}^{n} \frac{a_i}{x - x_i}}.$$

Diese Darstellung ist für $x \neq x_i$ definiert. Zusammen gilt somit

$$P(x) = \begin{cases} y_i & x = x_i \ (i = 0, \ldots, n) \\ \dfrac{\sum\limits_{i=0}^{n} \frac{a_i}{x - x_i} y_i}{\sum\limits_{i=0}^{n} \frac{a_i}{x - x_i}} & x \neq x_i \ (i = 0, \ldots, n) \end{cases}.$$

Dies nennt man **baryzentrische Darstellung** des Polynoms $P(x)$; sie läßt sich gut numerisch auswerten und wird auch im Programm LPOLYNOM angewendet.

```
//=============================================================================
//  Polynomwertberechnung mittels Langrange-schem Interpolationspolynom
//                 und seiner baryzentrischen Darstellung
// Rückkehrwert: Polynomwert
//=============================================================================
#define REAL double
#include<math.h>
#include<stdlib.h>
REAL lpolynom(ushort n,    // Stützstellen-Anzahl
              REAL *x,  // Feld mit den x-Werten der Stützstellen
              REAL *y,  // Feld mit den y-Werten der Stützstellen
              REAL t)   // Argument fuer den Polynomwert
{ static REAL *a=NULL, s, ss, z, epslpolynom=1.e-10;
  ushort nn=0,i,j;
  if(n!=nn)
  { if(a) delete []a; if(!n) return(0);
    a=new REAL[n]; nn=n;
    for(i=0; i<n; a[i++]=1/s)
      for(s=1, ss=x[i], j=0; j<n;j++) if(j!=i) s*=(ss-x[j]);
  }
  for(i=0; i<n; i++) if(fabs(t-x[i])<epslpolynom) return y[i];
  for(i=0, s=ss=0; i<n; z=a[i]/(t-x[i]), s+=y[i++]*z, ss+=z);
  return s/ss;
}
```

Es gibt Aufgaben, bei denen man nicht nur den Wert des interpolierenden Polynoms an einer gewissen Stelle haben möchte, sondern die Koeffizienten des Polynoms benötigt. Ein anderer Ansatz ist die algorithmische Frage, ob man nicht bei oftmaligem Aufruf des Neville-Algorithmus in einem Vorspann alle jene Rechenoperationen ausführen kann, die vom Eingabeparameter x unabhängig sind. Beide Ausgangsfragen führen zum gleichen Ziel. Wir stellen das gesuchte Polynom P in der Form des sog. **Newton-schen Interpolationspolynoms** dar:

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

In dieser Form kann man es nach einem Horner-artigen Schema auswerten:

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

Prinzipiell kann man die Koeffizienten a_i nacheinander aus den Beziehungen

$$\begin{aligned} f_0 &= P(x_0) = a_0, \\ f_1 &= P(x_1) = a_0 + a_1(x_1 - x_0), \\ f_2 &= P(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1), \\ &\ldots \end{aligned}$$

ermitteln. Für die Abschnittspolynome

$$Q_i(x) = a_0 + a_1(x - x_0) + \ldots + a_i(x - x_0) \cdot \ldots \cdot (x - x_{i-1}), \quad (i = 0, \ldots, n)$$

folgt sofort

1. $Q_i(x) = P_{ii}(x)$,
2. $Q_{i+1} = Q_i(x) + a_{i+1}(x - x_0) \cdot \ldots \cdot (x - x_i)$,
3. a_i ist der Koeffizient von x^i im Polynom Q_i.

Betrachten wir nun die folgenden Größen:

$$\begin{aligned} f_{i0} &= y_i, \quad i = 0, \ldots, n \\ f_{ij} &= \frac{f_{i,j-1} - f_{i-1,j-1}}{x_i - x_j}, \quad i = 1, \ldots, n; j = 1, \ldots, i. \end{aligned}$$

Man nennt die Größe f_{ij} die **j-te dividierte Differenz** .

Satz 155 (Newton-Interpolation). *Die Koeffizienten a_i des Newton-schen Interpolationspolynoms*

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

sind gleich der i-ten dividierten Differenz

$$a_i = f_{ii}, \quad i = 0, 1, \ldots, n.$$

Beweis. Wir zeigen durch Induktion, daß

$$P_{ij}(x) = f_{i0} + f_{i1}(x - x_{i-j}) + \ldots + f_{ij}(x - x_{i-j+1}) \cdot \ldots \cdot (x - x_{i-1})$$

gilt. Für $j = 0$ gilt diese Formel offenbar. Nehmen wir an, sie ist für $j - 1$ richtig. Aus den obigen Eigenschaften der Abschnittspolynome folgt

$$P_{ij}(x) = P_{i-1,j-1}(x) + a(x - x_{i-j+1})(x - x_{i-j+2}) \cdot \ldots \cdot (x - x_i),$$

wobei der unbekannte Faktor a gerade der Koeffizient von x^j des Polynoms P_{ij} darstellt. Für den Induktionsschritt ist somit $a = f_{ij}$ zu zeigen. Nach Induktionsvoraussetzung gilt:

$$\begin{aligned} P_{i-1,j-1}(x) &= \ldots + f_{i-1,j-1}x^{j-1}, \\ P_{i,j-1}(x) &= \ldots + f_{i,j-1}x^{j-1}. \end{aligned}$$

Die Nevillesche Interpolationsformel liefert

$$P_{ij}(x) = \frac{(x - x_{i-j})P_{i,j-1}(x) - (x - x_i)P_{i-1,j-1}(x)}{x_i - x_{i-j}}.$$

Der Koeffizient von x^j ergibt sich daraus zu

$$\frac{f_{i,j-1} - f_{i-1,j-1}}{x_i - x_{i-j}},$$

was mit der obigen Rekursionsformel übereinstimmt. □
Das Differenzenschema für die Newton-Interpolation lautet also

$$\begin{array}{c|cccc} x_0 & f_{00} & & & \\ x_1 & f_{10} & f_{11} & & \\ x_2 & f_{20} & f_{21} & f_{22} & \\ x_3 & f_{30} & f_{31} & f_{32} & f_{33} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

mit z. B.

$$f_{32} = \frac{f_{31} - f_{21}}{x_3 - x_1}.$$

Das Programm NPOLYNOM berechnet einen Polynomwert mittels des Newtonschen Interpolytionspolynoms.

```
//=============================================================================
//     Polynomwertberechnung mittels Newtonschem Interpolationspolynom
// Rückkehrwert: Polynomwert
//=============================================================================
#define REAL double
#include<stdlib.h>
REAL npolynom(ushort n,    // Stützstellen-Anzahl
              REAL *x,  // Feld mit den x-Werten der Stützstellen
              REAL *y,  // Feld mit den y-Werten der Stützstellen
              REAL t)   // Argument für den Polynomwert
{ static REAL *a=NULL, s;
  static ushort nn=0, i, j;
  if(n!=nn)
  { if(a) delete []a; if(!n) return 0;
    a=new REAL[n]; for(i=0; i<n; a[i++]=y[i]); nn=n;
    for(i=n-1; i; i-=2)
      for(j=i, s=x[i-1]; j<n ;a[j++]=(a[j]-a[j-1])/(x[j]-s));
  }
  for(i=n-1, s=a[i]; i--; s=s*(t-x[i])+a[i]);
  return(s);
}
```

Wir wollen nun untersuchen, wie genau die Polynominterpolation arbeitet, falls die Stützstellen von einer auf einem Intervall $[a, b]$ definierten Funktion f stammen.

Satz 156 (Restgliedsatz für die Polynominterpolation). *Ist f eine auf dem Intervall $[a, b]$ $(n+1)$-mal stetig differenzierbare Funktion mit*

$$f(x_i) = y_i, \quad i = 0, \ldots, n \quad a = x_0 < x_1 < \ldots < x_n = b,$$

so gibt es zu jedem $\overline{x}$ ein ξ aus dem kleinsten Intervall, das die Punkte $\overline{x}, a, b$ enthält mit

$$f(\overline{x}) - P_{nn}(\overline{x}) = \frac{\omega(\overline{x}) f^{(n+1)}(\xi)}{(n+1)!}.$$

Beweis. Es sei $\overline{x} \neq x_i$, $i = 0, \ldots, n$. Wir verwenden die Hilfsfunktion

$$F(x) = f(x) - P(x) - K \cdot \omega(x)$$

und wählen den Parameter K so, daß $F(\overline{x}) = 0$ gilt.
Dann hat die Funktion F im Intervall die $n+2$ Nullstellen $x_0, x_1, \ldots, x_n, \overline{x}$. Nach dem Satz von Rolle hat die Ableitung F' dort $n+1$ Nullstellen; F'' hat n Nullstellen usw.; die $(n+1)$-te Ableitung $F^{(n+1)}$ hat dort eine Nulstelle ξ. Wegen $P^{(n+1)} \equiv 0$ folgt

$$0 = F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - K \cdot (n+1)!,$$

d. h.

$$K = \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

womit der Satz bereits bewiesen ist. □

Als letztes Beispiel für die Interpolation wollen wir die Spline-Interpolation studieren. Gegeben seien ein Intervall $[a,b]$ und Stützstellen $(x_i,y_i), i=0,\ldots,n$ mit $a=x_0<x_1<\ldots<x_n=b$. Eine **kubische Spline-Funktion** S ist im Intervall $[x_i,x_{i+1}]$ ein Polynom 3. Grades $(i=0,1,\ldots,n)$, wobei die Ableitungen in den Randpunkten stetig anschließen mögen. Im Falle $S''(a)=S''(b)=0$ spricht man von **natürlichen kubischen Spline-Funktionen**, die hier untersucht werden sollen.

Satz 157. (Existenz- und Eindeutigkeit für natürliche, kubische Spline-Funktionen.) *Zu jedem System von Stützstellen $(x_i,y_i), i=0,\ldots,n$ existiert genau eine natürliche kubische Spline-Funktion S mit $S(x_i)=y_i, i=0,\ldots,n$.*

Beweis. Der Beweis des Satzes gibt uns gleichzeitig eine Methode zur Konstruktion einer natürlichen kubischen Spline-Funktion.
Es sei

$$\begin{aligned} h_{i+1} &= x_{i+1}-x_i, \quad i=0,\ldots,n-1,\\ M_i &= S''(x_i), \quad i=0,\ldots,n, \quad M_0=M_n=0. \end{aligned}$$

Die Größen M_i nennt man **Momente**. Da die gesuchte Funktion S auf dem Intervall $[x_i,x_{i+1}]$ ein Polynom 3. Grades sein soll, muß die zweite Ableitung dort linear sein:

$$S''(x)=M_i\frac{x_{i+1}-x}{h_{i+1}}+M_{i+1}\frac{x-x_i}{h_{i+1}}, \quad x\in[x_i,x_{i+1}].$$

Diese Funktion integrieren wir zweimal unbestimmt:

$$S'(x)=-M_i\frac{(x_{i+1}-x)^2}{2h_{i+1}}+M_{i+1}\frac{(x-x_i)^2}{2h_{i+1}}+A_i,$$

$$S(x)=M_i\frac{(x_{i+1}-x)^3}{6h_{i+1}}+M_{i+1}\frac{(x-x_i)^3}{6h_{i+1}}+A_i(x-x_i)+B_i,\ i=0,\ldots,n-1,$$

wobei A_i,B_i Integrationskonstanten darstellen, die wir in Abhängigkeit von den Momenten berechnen werden. In der Darstellung von S in Abhängigkeit von den Momenten setzen wir einmal $x=x_i$ und dann $x=x_{i+1}$, woraus folgt:

$$\begin{aligned} M_i\frac{h_{i+1}^2}{6}+B_i &= S(x_i)=y_i,\\ M_{i+1}\frac{h_{i+1}^2}{6}+A_ih_{i+1}+B_i &= S(x_{i+1})=y_{i+1}. \end{aligned}$$

Aus diesen Gleichungen erhalten wir die gewünschten Darstellungen für die Integrationskonstanten:

$$B_i=y_i-M_i\frac{h_{i+1}^2}{6},$$

$$A_i=\frac{y_{i+1}-y_i}{h_{i+1}}-\frac{h_{i+1}}{6}(M_{i+1}-M_i),\ i=0,\ldots,n-1.$$

Die Größen A_i setzen wir in die Darstellung von S' ein. Im Intervall $[x_i,x_{i+1}]$ folgt

$$S'(x)=-M_i\frac{(x_{i+1}-x)^2}{2h_{i+1}}+M_{i+1}\frac{(x-x_i)^2}{2h_{i+1}}+\frac{y_{i+1}-y_i}{h_{i+1}}-\frac{h_{i+1}}{6}(M_{i+1}-M_i)$$

und im Intervall $[x_{i-1},x_i]$:

$$S'(x)=-M_{i-1}\frac{(x_i-x)^2}{2h_i}+M_i\frac{(x-x_{i-1})^2}{2h_i}+\frac{y_i-y_{i-1}}{h_i}-\frac{h_i}{6}(M_i-M_{i-1}).$$

Erinnern wir uns an die Forderung, daß die Ableitungen der Spline-Funktionen stetig anschließen sollen. Dies zieht nach sich, daß im Punkte x_i beide Ableitungen übereinstimmen müssen. Wir haben also beide Formeln mit $x=x_i$ gleichzusetzen, woraus nach Umordnen folgt:

$$\frac{h_i}{6}M_{i-1}+\frac{h_i+h_{i+1}}{3}M_i+\frac{h_{i+1}}{6}M_{i+1}=\frac{y_{i+1}-y_i}{h_{i+1}}-\frac{y_i-y_{i-1}}{h_i},\ i=1,\ldots,n-1.$$

Zusammen mit $M_0 = 0, M_n = 0$ haben wir damit $n+1$ Gleichungen für $n+1$ unbekannte Momente gewonnen. Dem Gleichungssystem soll noch eine übersichtlichere Form gegeben werden. Wir setzen

$$\lambda_i = \frac{h_{i+1}}{h_i + h_{i+1}}, \quad \mu_i = 1 - \lambda_i = \frac{h_i}{h_i + h_{i+1}},$$

$$d_i = \frac{6}{h_i + h_{i+1}} \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right), \; i = 1, \ldots, n-1.$$

Dann lautet das System

$$\mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = d_i, \; i = 1, \ldots, n-1.$$

Setzen wir noch $\lambda_0 = d_0 = \mu_n = d_n$, so ergibt sich schließlich

$$\begin{aligned} 2M_0 + \lambda_0 M_1 &= d_0, \\ \mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} &= d_i, \; i = 1, \ldots, n-1, \\ \mu_n M_{n-1} + 2M_n &= d_n. \end{aligned}$$

Die Koeffizientenmatrix $\mathbf{A}$ dieses Systems ist tridiagonal:

$$\mathbf{A} = \begin{bmatrix} 2 & \lambda_0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \mu_1 & 2 & \lambda_1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & \mu_2 & 2 & \lambda_2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & \mu_{n-1} & 2 & \lambda_{n-1} \\ 0 & 0 & 0 & 0 & \ldots & 0 & \mu_n & 2 \end{bmatrix}$$

mit $\lambda_i \geqq 0, \mu_i \geqq 0, \lambda_i + \mu_i = 1$ $(i = 1, \ldots, n-1)$. Dies ist eine streng diagonal dominante Matrix, von der wir aus den Übungen in der linearen Algebra wissen, daß sie regulär ist. Damit sind die Momente als Lösungen eines linearen Gleichungssystems mit einer streng diagonal dominanten Tridiagonalmatrix eindeutig bestimmt. Wegen der Darstellung einer Spline-Funktion in Abhängigkeit von den Momenten ist hiermit der Satz bewiesen. □

Es sei noch erwähnt, daß unter den zweimal stetig differenzierbaren Interpolationsfunktionen φ die natürlichen kubischen Spline-Funktionen jene Interpolierenden sind, die den Wert des Integrals

$$\int_a^b (\varphi''(x))^2 dx$$

zum Minimum machen. Den Wert des genannten Integrals kann man als „Welligkeit“ der Funktion φ auffassen, so daß die Spline-Funktionen gerade jene sind, die unter den genannten die kleinste Welligkeit haben. Mit der Spline-Interpolation modellieren wir daher insbesondere das Zeichnen von möglichst „glatten“ Kurven mittels eines Kurvenlineals, wie wir es aus dem Schulunterricht kennen.

6.4. Numerische Integration

Wir wissen, daß es viele Funktionen gibt, die man nicht elementar integrieren kann. Es liegt daher nahe, für die Berechnung von

$$\int_a^b f(x)\, dx$$

einer auf dem Intervall $[a, b]$ stetigen Funktion f den Integranden durch eine geeignete Funktion zu ersetzen, um so Näherungswerte für das gesuchte bestimmte Integral zu erhalten. Bei den **Integrationsformeln von Newton-Cotes** wird der Integrand durch ein interpolierendes Polynom P ersetzt. Dazu brauchen wir ein System von Stützstellen $(x_i, y_i), i = 0, \ldots, n$. Es sei

$$h = \frac{b-a}{n}, \quad x_i = a + i \cdot h, \quad f(x_i) = y_i, \quad i = 0, \ldots, n$$

und P_n das Polynom vom Grade höchstens n mit $P_n(x_i) = y_i, i = 0, \ldots, n$.
Nach der Lagrange-Interpolationsformel gilt

$$P_n(x) = \sum_{i=0}^{n} y_i L_i(x), \quad L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}$$

und mit $x = a + s \cdot h$ erhält man

$$L_i(a + s \cdot h) = \varphi_i(s) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{s-j}{i-j}.$$

Damit folgt wegen $dx = h \cdot ds$:

$$\int_a^b P_n(x)dx = \sum_{i=0}^{n} y_i \int_a^b L_i(x)dx = h \cdot \sum_{i=0}^{n} y_i \int_0^n \varphi_i(s)ds = h \cdot \sum_{i=0}^{n} \alpha_i y_i$$

mit

$$\alpha_i = \int_0^n \varphi_i(s)ds, \quad i = 0, \ldots, n.$$

Die Gewichte α_i hängen nicht von der zu integrierenden Funktion f ab, sondern nur von der Anzahl n der Teilintervalle. Setzen wir speziell $f \equiv 1$, dann ist auch $P_n \equiv 1$ und somit

$$b - a = \int_a^b P_n(x)dx = h \cdot \sum_{i=0}^{n} \alpha_i y_i = \frac{b-a}{n} \sum_{i=0}^{n} \alpha_i,$$

also

$$\sum_{i=0}^{n} \alpha_i = n.$$

Mittels der Restgliedabschätzung für die Polynominterpolation können wir die Güte der erreichten Annäherung an das gesuchte Integral ermitteln. Dazu müssen wir wie oben annehmen, daß die Funktion f im Intervall $[a, b]$ $(n+1)$-mal stetig differenzierbar ist. Dann existiert eine Zahl $M > 0$ mit $|f^{(n+1)}(x)| < M$ für alle $x \in [a, b]$ und wir erhalten

$$\int_a^b (f(x) - P_n(x))\, dx = \int_a^b \frac{\omega(x) f^{(n+1)}(\xi(x))}{(n+1)!} dx$$

und mit $\omega(a + sh) = h^{n+1} s(s-1)(s-2) \cdot \ldots \cdot (s-n) = h^{n+1} \overline{\omega}(s)$:

$$\int_a^b (f(x) - P_n(x))dx = h^{n+2} \int_0^n \frac{\overline{\omega}(s) f^{(n+1)}(\xi(s))}{(n+1)!} ds = \mathcal{O}(h^{n+2}).$$

Es liegt somit ein Verfahren der Ordnung $n + 2$ vor.
In Abhängigkeit von n erhält man verschiedene Integrationsregeln; so im Falle $n = 1$ die **Trapezregel**, bei $n = 2$ die **Simpsonregel** und bei $n = 3$ die **3/8-Regel**. Für $n > 6$ treten negative Gewichte auf, so daß die Formeln numerisch unbrauchbar werden. Da die Fehlerabschätzung für die Näherung nur für kleine h wirksam ist, kann man die Formeln nicht auf das gesamte Intervall anwenden; man zerlegt es und addiert die Näherungen für die Teilintervalle. Am Beispiel der Trapezregel $(n = 1)$ soll der erreichbare Effekt untersucht werden. Für das Teilintervall $[x_i, x_{i+1}]$ einer Zerlegung $x_i = a + ih, i = 0, \ldots, N$ und $h = \frac{b-a}{N}$ erhält man den Näherungswert

$$I_i = \frac{h}{2} \left(f(x_i) + f(x_{i+1}) \right)$$

und für das gesamte Intervall damit

$$\begin{aligned} S(h) &= \sum_{i=0}^{N-1} \frac{h}{2} \left(f(x_i) + f(x_{i+1}) \right) \\ &= h \left(\frac{f(a)}{2} + f(a+h) + f(a+2h) + \ldots + f(b-h) + \frac{f(b)}{2} \right), \end{aligned}$$

die **Trapezsumme** zur Schrittweite h. Für jedes Teilintervall hat man einen Fehler von der Größe $\mathcal{O}(h^3)$, falls die Funktion f zweilmal stetig differenzierbar ist; daher ergibt sich als Gesamtfehler

$$|S(h) - \int_a^b f(x)dx| = \sum_{i=0}^{N-1} \mathcal{O}(h^3) = N \cdot \mathcal{O}(h^3) = \frac{b-a}{h} \mathcal{O}(h^3) = \mathcal{O}(h^2),$$

was uns sagt, daß ein Verfahren zweiter Ordnung vorliegt.
Falls die zu integrierende Funktion f im Intervall $[a,b]$ $(2m+2)$-mal stetig differenzierbar ist, kann man beweisen, daß die Trapezsumme eine asymptotische Entwicklung der folgenden Form hat:

$$S(h) = \sigma_0 + \sigma_1 h^2 + \sigma_2 h^4 + \ldots + \sigma_m h^{2m} + \mathcal{O}(h^{2m+2})$$

mit

$$\sigma_0 = \int_a^b f(x)dx.$$

Dabei sind die Faktoren σ_i von h unabhängig; ihre Berechnung ist nicht erforderlich, da wir ja über eine einfache Möglichkeit zur Berechnung der Trapezsumme verfügen. Vernachlässigt man das Restglied $\mathcal{O}(h^{2m+2})$, kann man die Trapezsumme also als ein Polynom in h^2 auffassen, das an der Stelle $h = 0$ den Wert des gesuchten Integrals hat. Das legt es nahe, den Wert $\sigma_0 = S(0)$ mittels Polynominterpolation näherungsweise zu bestimmen, d. h. auf die Schrittweite $h = 0$ zu extrapolieren. Dazu brauchen wir ein System von Stützstellen. Da hier ein Polynom in h^2 vorliegt, haben die Stützstellen die Form (h_i^2, S_i) mit $S_i = S(h_i)$. Zu einer gegebenen Schrittweitenfolge $h_0 > h_1 > \ldots > h_r > 0$ sei S_{rr} dasjenige Polynom in h^2, für das $S_{rr}(h_i) = S(h_i), i = 0, \ldots, r$ gilt. Der extrapolierte Wert $S_{rr}(0)$ wird dann ein guter Näherungswert für das gesuchte Integral sein. Die Extrapolation läßt sich nach dem Neville-Algorithmus ausführen. In den Formeln ist $x = 0$ und $x_i = h_i^2$ zu setzen. Die entsprechenden Formeln lauten damit:

$$S_{i0} = S(h_i), \quad i = 0, \ldots, r,$$
$$S_{ij} = S_{i,j-1} + \frac{S_{i,j-1} - S_{i-1,j-1}}{\left(\frac{h_{i-j}}{h_i}\right)^2 - 1}, \quad j = 1, \ldots, i.$$

Wählt man als Schrittweitenfolge

$$h_0 = b - a, \quad h_{i+1} = \frac{h_i}{2},$$

so erhält man die **Romberg-Integration**. Diese Folge strebt für unseren Zweck zu schnell gegen 0; daher wählt man besser die **Burlirsch-Folge**

$$h_0 = b - a, \quad h_1 = \frac{h_0}{2}, \quad h_2 = \frac{h_0}{3}, \quad h_i = \frac{h_{i-2}}{2} \ (i \geqq 3).$$

Für die Anwendung ist es noch wichtig zu wissen, bis in welche Tiefe das Schema berechnet werden sollte. Ein zu kleines r schöpft die Vorteile der Methode nicht aus; ein zu großes r verbietet sich einerseits wegen der sich aufschaukelnden Rechenfehler im Schema und andererseits wegen des schnell wachsenden Aufwandes bei der Berechnung von $S(h)$. In der Praxis wählt man bei doppelt genauer Rechnung $r = 6$ oder $r = 7$ und steuert die Schrittweite h entsprechend.
Eine genaue Fehleruntersuchung zeigt, daß

$$S_{rr} - \int_a^b f(x)dx = \mathcal{O}(h_{i-r}^2 \cdot h_{i-r+1}^2 \cdot \ldots \cdot h_i^2)$$

gilt, also ein Verfahren der Ordnung $2r + 2$ vorliegt.

6.5. Numerisches Differenzieren

Um eine Näherungsformel für die Ableitung $f'(a)$ einer im Punkte a ableitbaren Funktion f zu bestimmen, legt man eine Ersatzfunktion φ durch einige benachbarte Punkte und berechnet $\varphi'(a)$. Ist die Ersatzfunktion φ z. B. eine Parabel durch die Stützstellen $(a-h, f(a-h)), (a, f(a)), (a+h, f(a+h))$, dann läßt sie sich in der Form

$$\varphi(x) = f(a) + \frac{f(a+h) - f(a-h)}{2h}(x-a) + \frac{f(a+h) - 2f(a) + f(a-h)}{2h^2}(x-a)^2$$

darstellen und wir erhalten als Näherung für die erste Ableitung von f an der Stelle a den **zentralen Differenzenquotienten**

$$f'(a) \approx \varphi'(a) = \frac{f(a+h) - f(a-h)}{2h}.$$

Andere zentrale Differenzenformeln sind z. B.

$$f'(a) \approx \frac{1}{48h}(-f(a+3h) + 27f(a+h) - 27f(a-h) + f(a-3h)),$$

$$f'(a) \approx \frac{1}{12h}(-f(a+2h) + 8f(a+h) - 8f(a-h) + f(a-2h)).$$

Verwendet man äquidistante Stützwerte mit dem Abstand h und wählt eine Ersatzfunktion φ, die linear von den Funktionswerten der Funktion f abhängt, so erhält man Näherungsformeln in der Form

$$f'(a) \approx \frac{1}{h}\sum_{i=0}^{n} \alpha_i f(x_i),$$

wobei aus $f'(a) = 0$ bei einer konstanten Funktion f folgt, daß

$$\sum_{i=0}^{n} \alpha_i = 0$$

sein muß. Durch Taylor-Entwicklung der Funktion f stellt man fest, daß der Verfahrensfehler bei der ersten zentralen Differenzenformel von der Ordnung $\mathcal{O}(h^2)$, bei den anderen von der Ordnung $\mathcal{O}(h^4)$ und im letzteren Falle von der Ordnung $\mathcal{O}(h)$ ist. Bei der ersten zentralen Differenzenformel heben sich in der Taylor-Entwicklung des Verfahrensfehlers die Summanden mit einer ungeraden h-Potenz weg, so daß bei einer $(2m+2)$-mal stetig differenzierbaren Funktion f folgt:

$$\frac{f(a+h) - f(a-h)}{2h} = f'(a) + \frac{h^2}{3!}f^{(3)}(a) + \cdots + \frac{h^{2m}}{(2m+1)!}f^{(2m+1)}(a) + \mathcal{O}(h^{2m+2}).$$

Diese Tatsache legt es nun nahe, eine Extrapolation analog zur numerischen Integration auszuführen, wodurch man die Genauigkeit der Differenzenformel besser dem Verlauf der Funktion f anpassen kann.
Der Eingabefehler kann hier nur in ungenauen Funktionswerten auftreten. Werden anstelle der Eingabewerte $f(x_i)$ die Werte $\overline{f}(x_i)$ benutzt, so erhält man als Fehler

$$\begin{aligned}\left|\frac{1}{h}\sum_{i=0}^{n}\alpha_i\overline{f}(x_i) - \frac{1}{h}\sum_{i=0}^{n}\alpha_i f(x_i)\right| &\leqq \frac{1}{h}\sum_{i=0}^{n}|\alpha_i||\overline{f}(x_i) - f(x_i)| \\ &\leqq \frac{1}{h}\left(\sum_{i=0}^{n}|\alpha_i|\right)\max_i|\overline{f}(x_i) - f(x_i)|.\end{aligned}$$

Der Eingabefehler nimmt also bei abnehmender Schrittweite h umgekehrt proportional zu h zu. Der Gesamtfehler wird sich daher nur solange bei Verkleinerung der Schrittweite verringern, bis die Abnahme des Verfahrensfehlers durch die Zunahme des Eingabe- und des Rechenfehlers wettgemacht ist.
Die mathematische Aufgabe der Berechnung eines Ableitungswertes ist eine instabile Aufgabe. So kann sich die Ableitung von

$$\overline{f}(x) = f(x) + \varepsilon\sin(Mx)$$

von der Ableitung der Funktion f um εM unterscheiden, obwohl die Funktionswerte um höchstens ε voneinander abweichen. Die starke Fehlerfortpflanzung ist daher problemspezifisch und hängt nicht von dem gewählten Algorithmus ab.

6.6. Lineare Gleichungssysteme

Es sei $\mathbf{A} = (a_{ij})_{n,n}$ eine Matrix mit n Zeilen und n Spalten; ferner sei $\mathbf{b}$ ein Vektor mit n Komponenten. Wir wollen numerische und algorithmische Fragen im Zusammenhang mit dem Lösen des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$ studieren. Dabei setzen wir voraus, daß die Koeffizientenmatrix $\mathbf{A}$ regulär ist. Zunächst interessieren wir uns dafür, wie sich Eingabefehler in der rechten Seite $\mathbf{b}$ auf die Lösung des Systems auswirken. Es sei $\mathbf{x}^*$ die Lösung des Systems $\mathbf{Ax} = \mathbf{b}$ und $\overline{\mathbf{x}}$ die Lösung von $\mathbf{Ax} = \overline{\mathbf{b}}$. Unter Ausnutzung der in der linearen Algebra eingeführten Matrixnorm können wir abschätzen:

$$\|\overline{\mathbf{x}} - \mathbf{x}^*\| = \|\mathbf{A}^{-1}(\overline{\mathbf{b}} - \mathbf{b})\| \leqq \|\mathbf{A}^{-1}\| \cdot \|\overline{\mathbf{b}} - \mathbf{b}\|$$

und für die relative Änderung:

$$\frac{\|\overline{\mathbf{x}} - \mathbf{x}^*\|}{\|\mathbf{x}^*\|} \leqq \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|\frac{\|\overline{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{b}\|}.$$

Der Verstärkungsfaktor $\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ heißt **Kondition** der Matrix $\mathbf{A}$:

$$\mathrm{cond}(\mathbf{A}) = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|.$$

Die Kondition hängt von der gewählten Norm ab und ist nur mit höherem Aufwand als die Lösung der Aufgabe berechenbar.
Beispiel: Wir wählen

$$\mathbf{A} = \begin{bmatrix} 1.00 & 0.99 \\ 0.99 & 0.98 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1.99 \\ 1.97 \end{bmatrix}, \quad \overline{\mathbf{b}} = \begin{bmatrix} 1.989903 \\ 1.970106 \end{bmatrix}.$$

Die exakten Lösungen lauten

$$\mathbf{x}^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \overline{\mathbf{x}} = \begin{bmatrix} 3 \\ -1.0203 \end{bmatrix}.$$

Der absolute Fehler

$$\overline{\mathbf{b}} - \mathbf{b} = \begin{bmatrix} -0.000097 \\ 0.000106 \end{bmatrix}$$

in den Eingabedaten bewirkt eine absolute Lösungsänderung

$$\overline{\mathbf{x}} - \mathbf{x}^* = \begin{bmatrix} 2 \\ -2.0203 \end{bmatrix}.$$

In der Maximumnorm ist

$$\frac{\|\overline{\mathbf{x}} - \mathbf{x}^*\|}{\|\mathbf{x}^*\|} = 2.0203, \quad \frac{\|\overline{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{b}\|} = 0.000053266,$$

was als Quotient die untere Abschätzung 37928 für die Kondition liefert. Es ist

$$\mathbf{A}^{-1} = \begin{bmatrix} -9800 & 9900 \\ 9900 & -10000 \end{bmatrix},$$

also

$$\mathrm{cond}(\mathbf{A}) = 1.99 \cdot 19900 = 39601.$$

Dieses Beispiel zeigt uns insbesondere, daß die Fehlerabschätzung realistisch ist.
Für die Reduzierung der Rundungsfehler bei der numerischen Lösung eines linearen Gleichungssystems kann man die **Nachiteration** anwenden, die wie folgt arbeitet. Es sei $\overline{\mathbf{x}}$ das Maschinenergebnis bei der Lösung von $\mathbf{Ax} = \mathbf{b}$. Das **Residuum**

$$\mathbf{r} = \mathbf{A}\overline{\mathbf{x}} - \mathbf{b}$$

liefert beschränkt Auskunft über die Genauigkeit des Maschinenergebnisses; wegen $\mathbf{Ax}^* = \mathbf{b}$ folgt $\mathbf{A}(\overline{\mathbf{x}} - \mathbf{x}^*) = \mathbf{r}$ und damit $\overline{\mathbf{x}} - \mathbf{x}^* = \mathbf{A}^{-1}\mathbf{r}$; aber die inverse Matrix $\mathbf{A}^{-1}$ ist unbekannt und $\|\mathbf{A}^{-1}\|$ kann sehr groß sein. Also muß man das Residuum mit erhöhter Genauigkeit berechnen; mit dem so berechneten Vektor $\mathbf{r}$ kann man bei Vorliegen einer $\mathbf{LU}$-Zerlegung für die Matrix $\mathbf{A}$ das System $\mathbf{Ay} = \mathbf{r}$ lösen; es sei $\overline{\mathbf{y}}$ das Maschinenergebnis. Nun wird der Vektor $\overline{\mathbf{x}}^{(1)} = \overline{\mathbf{x}} - \overline{\mathbf{y}}$ als neue Näherung für die exakte Lösung $\mathbf{x}^*$ betrachtet. Der Prozeß läßt sich wiederholen:
Berechne $\mathbf{r}^{(1)} = \mathbf{A}\overline{\mathbf{x}}^{(1)} - \mathbf{b}$ mit erhöhter Genauigkeit, löse das System $\mathbf{Ay} = \mathbf{r}^{(1)}$ mit der vorliegenden $\mathbf{LU}$-Zerlegung und setze $\overline{\mathbf{x}}^{(2)} = \overline{\mathbf{x}}^{(1)} - \overline{\mathbf{y}}^{(1)}$.
Die mit der Nachiteration zu erreichende Genauigkeit muß vorsichtig beurteilt werden. Der Eingabefehler kann höhere Auswirkungen als der Rundungsfehler haben. Mit die Nachiteration nähert man sich höchstens der exakten Lösung jenes Systems, das sich im Rechner befindet.
Beispiel: Es sei

$$\mathbf{A} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Wir rechnen mit 3 Ziffern, zur Basis 10 und ohne Pivotisierung. Als **LU**-Zerlegung folgt

$$\overline{\mathbf{L}} = \begin{bmatrix} 1 & 0 & 0 \\ 0.55 & 1 & 0 \\ 0.333 & 1.01 & 1 \end{bmatrix}, \quad \overline{\mathbf{U}} = \begin{bmatrix} 1 & 0.500 & 0.333 \\ 0 & 0.830 & 0.0840 \\ 0 & 0 & 0.00520 \end{bmatrix}$$

mit der Maschinenlösung

$$\overline{\mathbf{x}} = \begin{bmatrix} 42.1 \\ -233 \\ 225 \end{bmatrix},$$

wobei aber

$$\mathbf{x}^* = \begin{bmatrix} 39 \\ -216 \\ 210 \end{bmatrix}$$

die exakte Lösung ist. Für das Residuum mit erhöhter Genauigkeit erhält man

$$\mathbf{A}\overline{\mathbf{x}} - \mathbf{b} = \begin{bmatrix} 0.475 \\ 0.298 \\ 0.230 \end{bmatrix}$$

und die Nachiteration liefert

$$\overline{\mathbf{x}}^{(1)} = \begin{bmatrix} 42.9 \\ -236 \\ 228 \end{bmatrix},$$

die offenbar keine Annäherung an die exakte Lösung darstellt. Man beachte jedoch, daß sich wegen des Eingabefehlers eine fehlerhafte Aufgabe im Rechner befindet; diese hat bei 6-stelliger Rechnung die exakte Lösung

$$\begin{bmatrix} 42.9542 \\ -236.459 \\ 229.055 \end{bmatrix}.$$

Für diese Aufgabe war die Nachiteration offensichtlich erfolgreich.
Für die folgenden Untersuchungen wählen wir als Vektornorm die euklidische. Eine damit verträgliche Matrixnorm ist

$$\|\mathbf{A}\| = \max_{\mathbf{x}\neq\mathbf{o}} \sqrt{\frac{\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x}}{\mathbf{x}^\mathsf{T}\mathbf{x}}}.$$

Die Verträglichkeitsbedingung $\|\mathbf{A}\mathbf{x}\| \leqq \|\mathbf{A}\| \cdot \|\mathbf{x}\|$ sieht man sofort ein, wenn man sie in der Form

$$\frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \leqq \|\mathbf{A}\| \quad (\mathbf{x} \neq \mathbf{o})$$

schreibt und berücksichtigt, daß die rechte Seite dieser Ungleichung gerade das Maximum der linken ist. Wir zeigen als nächstes, daß diese Matrixnorm **submultiplikativ** ist, d. h. es gilt

$$\|\mathbf{A} \cdot \mathbf{B}\| \leqq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$$

für alle regulären (n, n)-Matrizen $\mathbf{A}, \mathbf{B}$. Die Ungleichung folgt durch Ausrechnen:

$$\begin{aligned} \|\mathbf{A} \cdot \mathbf{B}\| &= \max_{\mathbf{x}\neq\mathbf{o}} \sqrt{\frac{\mathbf{x}^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{B}\mathbf{x}}{\mathbf{x}^\mathsf{T}\mathbf{x}}} = \max_{\mathbf{x}\neq\mathbf{o}} \sqrt{\frac{(\mathbf{B}\mathbf{x})^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}(\mathbf{B}\mathbf{x})}{(\mathbf{B}\mathbf{x})^\mathsf{T}(\mathbf{B}\mathbf{x})} \cdot \frac{\mathbf{x}^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{B}\mathbf{x}}{\mathbf{x}^\mathsf{T}\mathbf{x}}} \\ &\leqq \max_{\mathbf{y}\neq\mathbf{o}} \sqrt{\frac{\mathbf{y}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{y}}{\mathbf{y}^\mathsf{T}\mathbf{y}}} \cdot \max_{\mathbf{x}\neq\mathbf{o}} \sqrt{\frac{\mathbf{x}^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{B}\mathbf{x}}{\mathbf{x}^\mathsf{T}\mathbf{x}}} = \|\mathbf{A}\| \cdot \|\mathbf{B}\|. \end{aligned}$$

Wegen

$$1 = \|\mathbf{E}\| = \|\mathbf{A}\mathbf{A}^{-1}\| \leqq \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| = \mathrm{cond}(\mathbf{A})$$

gilt für jede reguläre Matrix $\mathbf{A}$:

$$\mathrm{cond}(\mathbf{A}) \geqq 1.$$

Orthogonale Matrizen $\mathbf{P}$ sind bekanntlich durch die Bedingung $\mathbf{P}^\mathsf{T}\mathbf{P} = \mathbf{E}$ charakterisiert; daher folgt

$$1 = \|\mathbf{P}\| = \|\mathbf{P}^\mathsf{T}\| = \|\mathbf{P}^{-1}\|,$$

also $\mathrm{cond}(\mathbf{P}) = 1$ und damit $\|\mathbf{P}\cdot\mathbf{A}\| = \|\mathbf{A}\|$. Für jede orthogonale Matrix $\mathbf{P}$ ist also

$$\mathrm{cond}(\mathbf{PA}) = \mathrm{cond}(\mathbf{A}).$$

Es sei nun eine $\mathbf{LU}$-Zerlegung der Matrix $\mathbf{A}$ gegeben: $\mathbf{A} = \mathbf{LU}$. Dann haben wir als Abschätzung der Lösungsänderung des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$ bei Änderung der rechten Seite auf $\overline{\mathbf{b}}$:

$$\frac{\|\overline{\mathbf{x}} - \mathbf{x}^*\|}{\|\mathbf{x}^*\|} \leqq \mathrm{cond}(\mathbf{L})\cdot\mathrm{cond}(\mathbf{U})\cdot\frac{\|\overline{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{b}\|}$$

und wir erkennen, daß durch die $\mathbf{LU}$-Zerlegung der Eingabefehler in der numerischen Lösung verstärkt wird.

6.6.1. Householder-Transformation

Im folgenden werden wir sehen, daß man jede reguläre Matrix $\mathbf{A}$ in der Form $\mathbf{A} = \mathbf{QR}$ zerlegen kann, wobei $\mathbf{Q}$ eine orthogonale und $\mathbf{R}$ eine obere Dreiecksmatrix darstellen. Hat man eine solche Zerlegung konstruiert, so würden sich bei rundungsfehlerfreier Rechnung die Eingabefehler beim Lösen eines linearen Gleichungssystems mit der Koeffizientenmatrix $\mathbf{A}$ nicht verstärken. Wenn gar die orthogonale Matrix $\mathbf{Q}$ als Produkt von orthogonalen Matrizen konstruiert wird, ist gesichert, daß sich bei der schrittweisen Konstruktion der oberen Dreiecksmatrix $\mathbf{R}$ die Rundungsfehler eines Schrittes im nächsten nicht verstärken, da man die Rundungsfehler in einem Schritt als Eingabefehler für den nächsten interpretieren kann. Nach Householder kann man eine $\mathbf{QR}$-Zerlegung in folgender Weise erhalten.
Man wähle zu gegebenem Vektor $\mathbf{w}$ eine Matrix $\mathbf{P}$ in der Form

$$\mathbf{P} = \mathbf{E} - 2\mathbf{w}\mathbf{w}^\mathsf{T} \text{ mit } \mathbf{w}^\mathsf{T}\mathbf{w} = 1.$$

Hierin ist $\mathbf{w}\mathbf{w}^\mathsf{T}$ ein **dyadisches Produkt**:

$$\mathbf{w}\mathbf{w}^\mathsf{T} = \begin{bmatrix} w_1w_1 & w_1w_2\cdots w_1w_n \\ w_2w_1 & w_2w_2\cdots w_2w_n \\ w_3w_1 & w_3w_2\cdots w_3w_n \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ w_nw_1 & w_nw_2\cdots w_nw_n \end{bmatrix}.$$

Jede solche Matrix $\mathbf{P}$ ist orthogonal, denn wegen $\mathbf{P}^\mathsf{T} = \mathbf{P}$ folgt

$$\mathbf{P}^\mathsf{T}\mathbf{P} = (\mathbf{E} - 2\mathbf{w}\mathbf{w}^\mathsf{T})(\mathbf{E} - 2\mathbf{w}\mathbf{w}^\mathsf{T}) = \mathbf{E} - 4\mathbf{w}\mathbf{w}^\mathsf{T} + 4\mathbf{w}\mathbf{w}^\mathsf{T} = \mathbf{E}.$$

In der Matrix $\mathbf{P}$ können wir über den Vektor $\mathbf{w}$ frei verfügen. Wir versuchen daher, den Vektor $\mathbf{w}$ so zu bestimmen, daß ein gegebener Vektor in ein Vielfaches des ersten Einheitsvektors $\mathbf{e}_1$ transformiert wird:

$$\mathbf{Px} = \varrho\mathbf{e}_1.$$

Wir multiplizieren diese Gleichung skalar mit sich:

$$\varrho^2 = \mathbf{x}^\mathsf{T}\mathbf{x}, \text{ d. h. } |\varrho| = \|\mathbf{x}\|$$

und über das Vorzeichen von ϱ darf noch verfügt werden.
Wir setzen den Ansatz für $\mathbf{P}$ in die Gleichung $\mathbf{Px} = \varrho\mathbf{e}_1$ ein:

$$\mathbf{Px} = (\mathbf{E} - 2\mathbf{w}\mathbf{w}^\mathsf{T})\mathbf{x} = \mathbf{x} - 2(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w} = \varrho\mathbf{e}_1,$$

also

$$\mathbf{w} = \frac{\mathbf{x} - \varrho\mathbf{e}_1}{2\mathbf{w}^\mathsf{T}\mathbf{x}}.$$

Diese Gleichung multiplizieren wir skalar mit sich und berücksichtigen, daß $\mathbf{w}^\mathsf{T}\mathbf{w} = 1$ vorausgestzt ist:

$$1 = \mathbf{w}^\mathsf{T}\mathbf{w} = \frac{\|\mathbf{x} - \varrho\mathbf{e}_1\|^2}{(2\mathbf{w}^\mathsf{T}\mathbf{x})^2},$$

also $2\mathbf{w}^\mathsf{T}\mathbf{x} = \|\mathbf{x} - \varrho\mathbf{e}_1\|$ und damit

$$\mathbf{w} = \frac{\mathbf{x} - \varrho\mathbf{e}_1}{\|\mathbf{x} - \varrho\mathbf{e}_1\|}.$$

Es folgt weiter

$$\|\mathbf{x} - \varrho\mathbf{e}_1\| = \|\mathbf{x} \mp \|\mathbf{x}\|\mathbf{e}_1\| = \sqrt{(x_1 \mp \|\mathbf{x}\|)^2 + x_2^2 + \ldots + x_n^2}.$$

Damit keine Auslöschung eintritt, wählt man als Vorzeichen von ϱ das entgegengesetzte von x_1, falls $x_1 \neq 0$:

$$\varrho = -\operatorname{sign}(x_1) \cdot \|\mathbf{x}\|$$

bzw. $\varrho = \|\mathbf{x}\|$, falls $x_1 = 0$. Mit dieser Festsetzung folgt

$$(x_1 - \varrho)^2 = \|\mathbf{x}\|^2 + 2|x_1| \cdot \|\mathbf{x}\| + x_1^2$$

und

$$\|\mathbf{x} - \varrho\mathbf{e}_1\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{x}\||x_1|,$$

$$2\mathbf{w}\mathbf{w}^\mathsf{T} = 2\frac{(\mathbf{x} - \varrho\mathbf{e}_1)(\mathbf{x} - \varrho\mathbf{e}_1)^\mathsf{T}}{\|\mathbf{x} - \varrho\mathbf{e}_1\|^2} = \frac{\mathbf{u}\mathbf{u}^\mathsf{T}}{\|\mathbf{x}\|(\|\mathbf{x}\| + |x_1|)}$$

mit $\mathbf{u} = \mathbf{x} - \varrho\mathbf{e}_1$ und

$$\mathbf{P} = \mathbf{E} - \alpha\mathbf{u}\mathbf{u}^\mathsf{T}, \quad \alpha = \frac{1}{\|\mathbf{x}\|(\|\mathbf{x}\| + |x_1|)}.$$

Wir fassen das Ergebnis in einem Satz zusammen.

Satz 158 (Householder-Transformation). *Zu einem gegebenen Vektor* $\mathbf{x} \neq \mathbf{o}$ *sei*

$$\varrho = \begin{cases} -\operatorname{sign}(x_1)\|\mathbf{x}\|, & x_1 \neq 0 \\ \|\mathbf{x}\|, & x_1 = 0 \end{cases}$$

Dann wird der Vektor $\mathbf{x}$ *mittels der orthogonalen Matrix*

$$\mathbf{P} = \mathbf{E} - \frac{(\mathbf{x} - \varrho\mathbf{e}_1)(\mathbf{x} - \varrho\mathbf{e}_1)^\mathsf{T}}{\|\mathbf{x}\|(\|\mathbf{x}\| + |x_1|)}$$

in das ϱ*-fache des Einheitsvektors* $\mathbf{e}_1$ *transformiert:* $\mathbf{P}\mathbf{x} = \varrho\mathbf{e}_1$.

Die durch diesen Satz definierte Transformation bezeichnen wir bei Anwendung auf einen Vektor $\mathbf{x} \in \mathbb{R}^n$ mit $\mathbf{H}_n(\mathbf{x})$.
Die Householder-Transformation soll nun verwendet werden, um eine reguläre Matrix schrittweise auf eine obere Dreiecksmatrix zu transformieren. Dazu sei $\mathbf{A}^{(0)} = \mathbf{A}$ und $\mathbf{a}_1$ die erste Spalte der Matrix $\mathbf{A}^{(0)}$. Wir bilden die Householder-Transformation $\mathbf{P}_1 = \mathbf{H}_n(\mathbf{a}_1)$ und setzen

$$\mathbf{A}^{(1)} = \mathbf{P}_1\mathbf{A}^{(0)}.$$

Es sei bemerkt, daß unterhalb der Hauptdiagonalen in der 1. Spalte der Matrix $\mathbf{A}^{(1)}$ nur Null-Elemente stehen. Nach $r-1$ Schritten haben wir eine Matrix $\mathbf{A}^{(r-1)}$ der Form

$$\mathbf{A}^{(r-1)} = \begin{bmatrix} * & * \cdots * & * & \cdots * \\ 0 & * \cdots * & * & \cdots * \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 \cdots * & * & \cdots * \\ 0 & 0 \cdots 0 & a_{rr}^{(r-1)} & \cdots a_{rn}^{(r-1)} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 \cdots 0 & a_{nr}^{(r-1)} & \cdots a_{nn}^{(r-1)} \end{bmatrix} = \begin{bmatrix} \mathbf{D} & \mathbf{B} \\ \mathbf{0} & \overline{\mathbf{A}}^{(r-1)} \end{bmatrix}$$

mit einer oberen Dreiecksmatrix $\mathbf{D}$ der Ordnung $r-1$. Nun ermitteln wir eine Householder-Transformation für die erste Spalte $\overline{\mathbf{a}}_1^{(r-1)}$ der Matrix $\overline{\mathbf{A}}^{(r-1)}$:

$$\overline{\mathbf{P}}_r = \mathbf{H}_{n-r+1}(\overline{\mathbf{a}}_1^{(r-1)}).$$

Die orthogonale Matrix $\overline{\mathbf{P}}_r$ wird durch Einheitsvektoren zu einer orthogonalen Matrix $\mathbf{P}_r$ der Ordnung n ergänzt:

$$\mathbf{P}_r = \begin{bmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{P}}_r \end{bmatrix}.$$

Mit dieser Matrix bilden wir

$$\mathbf{A}^{(r)} = \mathbf{P}_r\mathbf{A}^{(r-1)} = \begin{bmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{P}}_r \end{bmatrix} \begin{bmatrix} \mathbf{D} & \mathbf{B} \\ \mathbf{0} & \overline{\mathbf{A}}^{(r-1)} \end{bmatrix} = \begin{bmatrix} \mathbf{D} & \mathbf{B} \\ \mathbf{0} & \overline{\mathbf{P}}_r\overline{\mathbf{A}}^{(r-1)} \end{bmatrix}.$$

Die Multiplikation der Matrix $\overline{\mathbf{A}}^{(r-1)}$ mit der Matrix $\overline{\mathbf{P}}_r = \mathbf{E} - \alpha_r\mathbf{u}^{(r)}\mathbf{u}^{(r)T}$ führt man so aus:

$$\mathbf{y} = \alpha_r\mathbf{u}^{(r)T}\overline{\mathbf{A}}^{(r-1)}, \quad \overline{\mathbf{P}}_r\overline{\mathbf{A}}^{(r-1)} = \overline{\mathbf{A}}^{(r-1)} - \mathbf{u}^{(r)}\mathbf{y}^{\mathsf{T}}.$$

Nach $n-1$ Householder-Transformationen erhält man auf diese Weise eine obere Dreiecksmatrix $\mathbf{R} = \overline{\mathbf{A}}^{(n-1)}$. Die $\mathbf{u}$-Vektoren aus den Transformationsmatrizen $\mathbf{P}_r$ werden auf die erzeugten Nullelemente in den Matrizen $\mathbf{A}^{(r)}$ gespeichert. Da der $\mathbf{u}$-Vektor im r-ten Schritt genau $n-r+1$ wesentliche Komponenten hat, wird zur Abspeicherung auch noch die Hauptdiagonale benötigt, so daß man die Diagonal-Elemente der Matrix $\mathbf{R}$ in einem besonderen Vektor ablegen muß. Zur Konstruktion der Matrix $\mathbf{Q}$ benötigt man außerdem die Faktoren α_r, die man zweckmäßigerweise in einem weiteren Vektor abgelegt. Falls während der Transformation festgestellt wird, daß die Matrix nicht regulär ist (was sich dadurch zeigt, daß die 1. Spalte der Matrix $\overline{A}^{(r)}$ eine Nullspalte ist), setzt man den entsprechenden Faktor α_r gleich Null und fährt mit der nächsten Spalte fort. Wegen

$$\mathbf{R} = \mathbf{A}^{(n-1)} = \mathbf{P}_{n-1}\cdots\mathbf{P}_2\mathbf{P}_1\mathbf{A} = \mathbf{Q}^{\mathsf{T}}\mathbf{A}$$

mit der orthogonalen Matrix $\mathbf{Q}^{\mathsf{T}} = \mathbf{P}_{n-1}\cdots\mathbf{P}_2\mathbf{P}_1$ folgt $\mathbf{A} = \mathbf{QR}$.
Die $\mathbf{QR}$-Zerlegung einer (n,n)-Matrix benötigt etwa $\frac{2}{3}n^3$ Operationen, also doppelt soviel wie die $\mathbf{LU}$-Zerlegung.
Abschließend noch eine wichtige Bermerkung.
Die Householder-Transformation kann auf die Spalten jeder (m,n)-Matrix $\mathbf{A}$ angewendet werden. Also existiert zu jeder Matrix eine orthogonale Matrix $\mathbf{Q}$ mit

$$\mathbf{Q}^{\mathsf{T}}\mathbf{A} = \begin{bmatrix} \mathbf{R} & \mathbf{S} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

wobei die obere Dreicksmatrix $\mathbf{R}$ die Ordnung $\mathrm{rg}(\mathbf{A})$ hat und daher der untere Nullteil sowie der rechte Teil auch fehlen können. Mit einer $\mathbf{QR}$-Zerlegung lassen sich auch jene Aussagen gewinnen, die man mittels einer $\mathbf{LU}$-Zerlegung erzielt.

6.6.2. Symmetrische Matrizen

Beim Lösen eines linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$ mit einer symmetrischen Matrix $\mathbf{A}$ kann man Rechenoperationen sparen, da beim Konstruieren einer $\mathbf{LU}$-Zerlegung unter Umständen die Symmetrie erhalten bleibt. Ist etwa die Restmatrix $\mathbf{A}^{(r)}$ symmetrisch, so folgt für die Elemente von $\mathbf{A}^{(r+1)}$:

$$a_{ij}^{(r+1)} = a_{ij}^{(r)} - \frac{a_{ir}^{(r)}}{a_{rr}^{(r)}}a_{rj}^{(r)}, \quad i,j = r+1,\ldots,n,$$

$$a_{ji}^{(r+1)} = a_{ji}^{(r)} - \frac{a_{jr}^{(r)}}{a_{rr}^{(r)}}a_{ri}^{(r)}, \quad i,j = r+1,\ldots,n,$$

also $a_{ij}^{(r+1)} = a_{ji}^{(r+1)}$, da $a_{ij}^{(r)} = a_{ji}^{(r)}$. Eventuelle Zeilen-Vertauschungen müssen mit entsprechenden Spalten-Vertauschungen kombiniert werden, um die Symmetrie zu erhalten; die Pivotsuche muß also entlang der Hauptdiagonalen geschehen, was aber nicht bei jeder symmetrischen Matrix zum Erfolg führen wird. Für eine praktisch wichtige Klasse symmetrischer Matrizen ist eine Pivotisierung entlang der Hauptdiagonalen möglich; dies sind die positiv definiten Matrizen. Eine symmetrische Matrix $\mathbf{A}$ heißt **positiv definit** , wenn $\mathbf{x}^{\mathsf{T}}\mathbf{Ax} > 0$ gilt für alle Vektoren $\mathbf{x} \neq \mathbf{o}$.
Es sei nun $\mathbf{A}$ eine reguläre, symmetrische Matrix der Ordnung n; ferner sei eine $\mathbf{LU}$-Zerlegung gegeben: $\mathbf{A} = \mathbf{LU}$. Wir schreiben die obere Dreiecksmatrix als Produkt einer Diagonalmatrix $\mathbf{D}$ und einer oberen Dreiecksmatrix $\mathbf{V}$, deren Hauptdiagonal-Elemente sämtlich gleich 1 sind:

$$\mathbf{D} = \mathrm{diag}(u_{ii})_{n,n} = \begin{bmatrix} u_{11} & 0 & \cdots & 0 \\ 0 & u_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} 1 & \frac{u_{12}}{u_{11}} & \cdots & \frac{u_{1n}}{u_{11}} \\ 0 & 1 & \cdots & \frac{u_{2n}}{u_{22}} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Mit diesen Bezeichnungen gilt $\mathbf{U} = \mathbf{DV}$ und damit $\mathbf{A} = \mathbf{LDV} = \mathbf{A}^\mathsf{T} = \mathbf{V}^\mathsf{T}\mathbf{D}\mathbf{L}^\mathsf{T}$, also $\mathbf{L}^\mathsf{T} = \mathbf{V}$. Damit haben wir den folgenden Satz bewiesen.

Satz 159. *Jede reguläre, symmetrische (n,n)-Matrix $\mathbf{A}$ hat eine Zerlegung der Form $\mathbf{A} = \mathbf{LDL}^\mathsf{T}$ mit einer Diagonalmatrix $\mathbf{D}$ und einer unteren Dreiecksmatrix $\mathbf{L}$, für welche $l_{ii} = 1, i = 1, \ldots, n$ gilt.*

Es sei nun die Matrix $\mathbf{A}$ außerdem noch positiv definit; dann erhalten wir für alle Vektoren $\mathbf{x} \neq \mathbf{o}$:

$$0 < \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{L}\mathbf{D}\mathbf{L}^\mathsf{T}\mathbf{x} = (\mathbf{L}^\mathsf{T}\mathbf{x})^\mathsf{T}\mathbf{D}(\mathbf{L}^\mathsf{T}\mathbf{x}).$$

Indem wir in dieser Ungleichung n-mal einen solchen Vektor $\mathbf{x}$ wählen, daß $\mathbf{L}^\mathsf{T}\mathbf{x} = \mathbf{e}_i$ gilt, erhalten wir $u_{ii} > 0, i = 1, \ldots, n$. Damit können wir

$$\sqrt{\mathbf{D}} = \operatorname{diag}\left(\sqrt{u_{ii}}\right)_{n,n}$$

setzen und schreiben $\mathbf{A} = \mathbf{L}\sqrt{\mathbf{D}}\sqrt{\mathbf{D}}\mathbf{L}^\mathsf{T} = \mathbf{G}^\mathsf{T}\mathbf{G}$ mit $\mathbf{G} = \sqrt{\mathbf{D}}\mathbf{L}^\mathsf{T}$, d. h.

$$\mathbf{G} = \begin{bmatrix} \sqrt{u_{11}} & \frac{u_{12}}{\sqrt{u_{11}}} & \cdots & \frac{u_{1n}}{\sqrt{u_{11}}} \\ 0 & \sqrt{u_{22}} & \cdots & \frac{u_{2n}}{\sqrt{u_{22}}} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sqrt{u_{nn}} \end{bmatrix}.$$

Diese Zerlegung heißt **Cholesky-Zerlegung** der Matrix $\mathbf{A}$. Es sei $\mathbf{g}_j$ die j-te Spalte der Matrix $\mathbf{G}$:

$$\mathbf{g}_j = \begin{bmatrix} g_{1j} \\ g_{2j} \\ \vdots \\ g_{jj} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Es ist dann $\mathbf{G}^\mathsf{T}\mathbf{G} = (\mathbf{g}_i^\mathsf{T}\mathbf{g}_j)_{n,n}$ und

$$a_{ij} = \mathbf{g}_i^\mathsf{T}\mathbf{g}_j = g_{ii}g_{ij} + \sum_{l=1}^{i-1} g_{li}g_{lj} \quad (i \leqq j),$$

was eine Berechnungsmöglichkeit der Zahlen g_{ij} liefert. Unser Ergebnis wird im folgenden Satz zusammengefaßt.

Satz 160 (Cholesky-Zerlegung). *Zu jeder symmetrischen, positiv definiten Matrix $\mathbf{A}$ der Ordnung n gibt es eine obere Dreiecksmatrix $\mathbf{G}$, so daß $\mathbf{A} = \mathbf{G}^\mathsf{T}\mathbf{G}$ gilt. Die Elemente g_{ij} der Matrix $\mathbf{G}$ kann man nach den Formeln*

$$g_{ii} = \sqrt{a_{ii} - \sum_{l=1}^{i-1} g_{li}^2}, \quad i = 1, \ldots, n$$

$$g_{ij} = \frac{1}{g_{ii}}\left(a_{ij} - \sum_{l=1}^{i-1} g_{li}g_{lj}\right), \quad j = i+1, \ldots, n$$

berechnen.

Neben den n Quadratwurzeln hat die Methode einen Aufwand von ca. $\frac{n^3}{6}$ Operationen. Nach dem vorangegangenen Satz kann man das Wurzelziehen vermeiden und eine **LDLT**-Zerlegung bestimmen. Für Anwendungen wichtig sind sog. Bandmatrizen; solche Matrizen haben lediglich entlang von einigen Nebendiagonalen von Null verschiedene Elemente. Die Bandstruktur bleibt bei der Faktorisierung erhalten, so daß sich ein Operationsaufwand $\mathcal{O}(n^2)$ ergibt.
Nach der Berechnung einer Faktorisierung hat man anstelle des Systems $\mathbf{Ax} = \mathbf{b}$ die beiden Systeme $\mathbf{Ly} = \mathbf{b}$ und $\mathbf{L}^\mathsf{T}\mathbf{x} = \mathbf{D}^{-1}\mathbf{y}$ zu lösen.
Abschließend beweisen wir noch ein hinreichendes Kriterium für positiv definite Matrizen.

Satz 161. *Jede symmetrische, streng diagonal-dominante Matrix* $\mathbf{A}$*, d. h.*

$$a_{ii} > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n$$

ist positiv definit.

Beweis. Es genügt offenbar zu zeigen, daß für die Matrix $\mathbf{A}$ eine $\mathbf{LU}$-Zerlegung ohne Pivotisierung möglich ist und beim Übergang von $\mathbf{A}^{(r)}$ zu $\mathbf{A}^{(r+1)}$ die Bedingung

$$a_{ii}^{(r)} > \sum_{\substack{j=r+1\\ j\neq i}}^{n} |a_{ij}^{(r)}|, \quad i = 1, \ldots, n$$

erhalten bleibt. Dazu brauchen wir nur den Fall $r = 0$, d. h. den Übergang von $\mathbf{A} = \mathbf{A}^{(0)}$ zu $\mathbf{A}^{(1)}$ zu betrachten. Es ist

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}}, \quad i, j = 2, \ldots, n.$$

Wir setzen die folgenden Hilfsgrößen

$$\varrho_i = \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|a_{ij}|}{a_{ii}}, \quad p_i = \frac{|a_{i1}|}{a_{ii}}, \quad i = 1, \ldots, n, \quad q = \frac{|a_{1i}|}{a_{11}}.$$

Offenbar gilt

$$0 \leqq \varrho_i < 1, \quad 0 \leqq p_i < 1, \quad i = 1, \ldots, n, \quad 0 \leqq q < 1$$

und

$$\sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}^{(1)}| \leqq \sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}| + |a_{i1}| \sum_{\substack{j=2\\ j\neq i}}^{n} \frac{|a_{1j}|}{a_{11}} = \sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}| + |a_{i1}|(\varrho_1 - q),$$

sowie

$$a_{ii}^{(1)} \geqq a_{ii} - \frac{|a_{i1}| \cdot |a_{1i}|}{a_{11}} = a_{ii}(1 - q \cdot p_i) > 0.$$

Zusammen folgt

$$\begin{aligned} \frac{\sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}^{(1)}|}{a_{ii}^{(1)}} &\leqq \frac{\sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}| + |a_{i1}|(\varrho_1 - q)}{a_{ii}(1 - q \cdot p_i)} \\ &= \frac{\varrho_i - p_i + p_i(\varrho_1 - q)}{1 - q \cdot p_i} \\ &= \varrho_i - p_i \frac{q(1 - \varrho_i) + (1 - \varrho_1)}{1 - q \cdot p_i} \leqq \varrho_i < 1, \end{aligned}$$

womit die Behauptung bewiesen ist. □

6.6.3. Große, schwach besetzte Matrizen

Bei der numerischen Behandlung angewandter Aufgaben treten oft sehr große Matrizen auf; so z. B. bei der Berechnung von Spannungen und Verformungen in Bauteilen, bei der Beschreibung grafischer Bilder auf dem Rechner oder bei der Optimierung von Produktionsplänen in großen Betrieben. Die dort auftretenden Matrizen haben nur eine sehr geringe Anzahl von Nichtnullelementen und diese sind oft noch innerhalb der Matrix in spezieller Form angeordnet, so z. B. in großer Nähe zur Hauptdiagonalen, so daß ein Band entsteht, d. h. die Nichtnullelemente der Matrix befinden sich in jeweils r oberen und unteren Nebendiagonalen um die Hauptdiagonale herum. Üblicherweise nennt man eine (m, n)-Matrix **schwach besetzt** (sparse-Matrix), wenn die Anzahl der Nichtnullelemente (NNE) von der Ordnung $\mathcal{O}(\max(m, n))$ ist. Wenn man bei der Berechnung einer $\mathbf{LU}$-Zerlegung die Pivotisierung geschickt wählt, kann man u. U. errreichen, daß sowohl die berechnete untere als auch die obere Dreiecksmatrix wenig Nichtnullelemente enthält. Gelingt es überdies noch, das Rechnen

mit Nullelementen möglichst zu vermeiden, kann man den Rechenaufwand bei der **LU**-Zerlegung von $\mathcal{O}(n^3)$ auf höchstens $\mathcal{O}(n^2)$ senken. Bei großer Matrixordnung n erlaubt eine solche Reduzierung des Rechenaufwandes überhaupt erst das numerische Lösen des linearen Gleichungssystems in einer akzeptablen Zeit.
Um beim Lösen von linearen Gleichungssysteme die schwache Besetztheit der Koeffizientenmatrix ausnutzen zu können, benötigt man Speicherungsformen für Matrizen, die das Abspeichern von Nichtnullelementen vermeiden. Eine solche Speicherungsform muß folgende Operationen ermöglichen, damit auf ihrer Grundlage eine Zerlegung der Matrix ermittelt werden kann:

- Wiederauffinden der NNE, d. h. man benötigt ein Programm, das bei vorgegebenen Indices i, j das Matrixelement a_{ij} ermittelt,
- Ändern von NNE, d. h. vorhandenen NNE werden neue Werte zugewiesen,
- Hinzufügen neuer NNE,
- Streichen von NNE (Nullsetzen),
- Zeilen- und spaltenweiser Elemente-Durchlauf.

Eine Matrix wird auf eine zweidimensionale Liste abgebildet, wobei die Null-Elemente nicht abgespeichert sind. Die Abbildung erfolgt in 2 Stufen: Die 1. Stufe enthält alle Funktionen, die nicht von der Tatsache Gebrauch machen, daß es sich um reelle Zahlen (Elemente eines Körpers) handelt. Hierbei entsteht eine Klassenvorlage `sp_list2`, in die auch die Speicherplatzverwaltung integriert ist.

```
#ifndef SP_LIST2
#define SP_LIST2
#include <string.h>
#include "ls_array2.h"
#include "sp_list1.h"
template<class T>
class sp_list2: public sp_list<T>
{ protected:
  list2<T>  *a; // a->i=m; a->j=n; a->r_next=a->c_next=b;
                // b+i : Anker der i-ten Zeile und i-ten Spalte
                // (b+i)->r_next : 1. Datenelement der i-ten Zeile;
                // (b+i)->c_next : 1. Datenelement der i-ten Spalte;
                // (b+i)->i : Element-Anzahl der i-ten Zeile;
                // (b+i)->j : Element-Anzahl der i-ten Spalte;
                // das letzte r_next bzw. c_next zeigt auf a;
  char ONAME[ls_len];    // Feld fuer den Objektnamen
  void set_data(list2<T> *aa){ a=aa;}
  public:
  char *name;
  sp_list2(ls_UINT =0, ls_UINT =0, char* ="sp_list2");
  sp_list2(sp_list2<T> &);
  sp_list2(T*, ls_UINT, ls_UINT, char* ="sp_list2");
  ~sp_list2();
  sp_list2<T>& swap(sp_list2<T> &);
  ls_UINT number_of_rows() const{ return a->i; }
  ls_UINT number_of_columns() const{ return a->j;}
  list2<T>* asList()const{ return a;}
  const sp_list2<T>& operator=(const sp_list2<T> &);
  sp_list2<T>& put(T, ls_UINT, ls_UINT);
  sp_list2<T>& put_row(const sp_list1<T> &, ls_UINT, ls_UINT =0);
  sp_list2<T>& put_row(const ls_array1<T> &, ls_UINT, ls_UINT =0);
  sp_list2<T>& put_row(T, ls_UINT, ls_UINT =0);
  sp_list2<T>& put_column(const sp_list1<T>&, ls_UINT, ls_UINT);
  sp_list2<T>& put_column(const ls_array1<T>&, ls_UINT, ls_UINT);
  sp_list2<T>& put_column(T, ls_UINT, ls_UINT);
  sp_list2<T>& put_diagonal(const sp_list1<T> &, ls_UINT =0, ls_UINT =0);
  sp_list2<T>& put_diagonal(const ls_array1<T> &, ls_UINT =0, ls_UINT =0);
  sp_list2<T>& put_diagonal(T, ls_UINT =0, ls_UINT =0);
  sp_list2<T>& put_array(const sp_list2<T> &, ls_UINT =0, ls_UINT =0);
  sp_list2<T>& put_array(const ls_array2<T> &, ls_UINT =0, ls_UINT =0);
  sp_list2<T>& put_array(T, ls_UINT =0, ls_UINT =0);
  T get(ls_UINT, ls_UINT) const;
  const sp_list2<T>& get_row(sp_list1<T> &,ls_UINT, ls_UINT =0) const;
  const sp_list2<T>& get_row(ls_array1<T> &,ls_UINT, ls_UINT =0) const;
  const sp_list2<T>& get_column(sp_list1<T> &,ls_UINT, ls_UINT) const;
  const sp_list2<T>& get_column(ls_array1<T> &,ls_UINT, ls_UINT) const;
```

```
  const sp_list2<T>& get_diagonal(sp_list1<T> &,ls_UINT =0, ls_UINT =0) const;
  const sp_list2<T>& get_diagonal(ls_array1<T> &,ls_UINT =0, ls_UINT =0) const;
  const sp_list2<T>& get_array(sp_list2<T>&, ls_UINT =0, ls_UINT =0) const;
  const sp_list2<T>& get_array(ls_array2<T>&, ls_UINT =0, ls_UINT =0) const;
  sp_list2<T>& append_row(ls_UINT =1);
  sp_list2<T>& append_column(ls_UINT =1);
  sp_list2<T>& append_array(ls_UINT =1, ls_UINT =1);
  sp_list2<T>& remove_row(ls_UINT);
  sp_list2<T>& remove_column(ls_UINT);
  sp_list2<T>& remove();
  sp_list2<T>& swap_row(ls_UINT, ls_UINT);
  sp_list2<T>& swap_column(ls_UINT, ls_UINT);
  const sp_list2<T>& write_row(ostream &) const;
  sp_list2<T>& read_row(istream &);
  const sp_list2<T>& write_column(ostream &) const;
  sp_list2<T>& read_column(istream &);
  const sp_list2<T>& operator>> (char *) const;
  sp_list2<T>& operator<< (char *);
  ls_REAL filling_density();
};
#ifndef SP_LIB
#include "sp_list2.cpp"
#endif
#endif
```

In einer zweidimensionalen Liste der Größe `(m,n)`, wo `m` die Zeilenanzahl und `n` die Spaltenanzahl darstellen, ist die Position eines Datenelementes durch zwei Indices `(i,j)` definiert: Der erste charakterisiert die Zeile, der zweite die Spalte, in der das Datenelement steht. Dabei beziehen sich diese Indices stets auf das entsprechende Urbild. Die Zeilen und Spalten sind jeweils vorwärts verkettet; in `a->r_next` und in `a->c_next` findet man den Zeiger auf das Ankerfeld `b`. Im Zeilenindex `(b+k)->i` des Ankers wird die Anzahl der verketteten Datenelemente in der `k`-ten Zeile, im Spaltenindex `(b+k)->j` die Anzahl der verketteten Datenelemente in der `k`-ten Spalte abgelegt. Der `next`-Zeiger des letzten Datenelementes einer Zeile oder Spalte zeigt auf `a`; außerdem gilt `a->i = m` und `a->j = n`. Typische Verarbeitungsprozesse einer Liste sind das zeilen- oder spaltenweise Suchen eines Elementes mit eventuellem Ein- oder Ausketten des betreffenden Elementes und das Durchlaufen einer Zeile oder Spalte. Mit der gewählten Organisation ist dies leicht möglich:
1. Gesucht sei das Datenelement aus der Zeile `i` mit dem Spaltenindex `j`:

```
list2<T> *ap=asList(), *aa; ap=ap->r_next+i;
while((aa=ap->r_next)->j < j) ap=aa;
if(aa->j == j) // vorhanden ...
```

2. Durchlaufen der vorhandenen Listenelemente der `j`-ten Spalte:

```
list2<T> *aa=asList(); aa=aa->c_next+j;
while((aa=aa->c_next) != a) ...
```

Es ist klar, daß auf diese Weise die obigen Oprationen ausführbar sind. Als 2. Abbildungsstufe erscheint die Klasse `sp_Matrix` als eine von `sp_list2` abgeleitete Klasse. Rechteck-Matrizen werden wie Datentypen deklariert: `...; sp_Matrix A(m,n); ....` Dabei sind `m` die Zeilen- und `n` die Spaltenanzahl der Matrix `A` mit `m` $\geqq$ `n`, die erst zur Laufzeit ihre Werte erhalten (dynamische Klassen-Komponenten). An den Operationen und Funktionen sind nur die NNE beteiligt. Zur Ein- und Ausgabe dienen die Funktionen `read_row` und `write_row`:

```
...; sp_Matrix A(3,2); ...
ofstream fo("bb"); A.write_row(fo); ...
ifstream fi("aa"); A.read_row(fi); ....
```

Hier wird die Matrix `A` mit ihrem Objektnamen `sp_Matrix` in die Datei `bb` geschrieben; sodann wird in der Datei `aa` das Wort `sp_Matrix` gesucht; danach muß zeilenweise die einzulesende Matrix als in geschweiften Klammern eingeschlossener Block folgen. Der entsprechende Block für die `(6,5)`-Matrix

$$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ -3 & 2 & 0 & 0 & 4 \\ 0 & -4 & 1 & -5 & 2 \\ -6 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \end{pmatrix}$$

hat den folgenden Aufbau:

```
{ number_of_rows: 6  number_of_columns: 5
```

```
  row: 0  number_of_elements: 2
       0  1  2 -1
  row: 1  number_of_elements: 2
       2 -2  3  1
  row: 2  number_of_elements: 3
       0 -3  1  2  4  4
  row: 3  number_of_elements: 4
       1 -4  2  1  3 -5  4  2
  row: 4  number_of_elements: 2
       0 -6  1  2
  row: 5  number_of_elements: 2
       3 -1  4 -1
}
```

Die Matrixelemente sind zeilenweise geschrieben mit Zeilennummer und NNE-Anzahl in der Zeile; danach folgen die NNE mit Spaltenindex und Wert. Zwischen den Daten sind die üblichen Trennzeichen erlaubt: Leerzeichen, Tabulatorzeichen, neue-Zeile-Zeichen. Natürlich gibt es auch die spaltenweise Version (`read_column`, `write_column`). Falls man keine Ansprüche an die Lese- oder Schreibeinstellung der Datei hat, darf man für Schreiben `v >> "bb"` und für Lesen `v << "aa"` verwenden; hier wird die Zeilen-Version aufgerufen. Alle Funktionen und Operationen, die nicht von der Tatsache Gebrauch machen, daß es sich bei den Daten um reelle Zahlen handelt, werden von der Klasse `sp_list2<ls_REAL>` geerbt; hierin steht `ls_REAL` für `float` oder `double`. Weitere Operationen und Funktionen sind die folgenden:

Multiplikation einer sparse-Matrix mit einem Vektor: `x = A*y`.
Multiplikation einer transponierten sparse-Matrix mit einem Vektor: `x = y*A`.

```
#ifndef SP_MATRIX
#define SP_MATRIX
#include "ls_Matrix.h"
#include "sp_Vector.h"
#include "sp_list2.h"
class sp_Matrix: public sp_list2<ls_REAL>
{ public:
  sp_Matrix(ls_UINT =0, ls_UINT =0, char* ="sp_Matrix");
  sp_Matrix(ls_REAL*, ls_UINT, ls_UINT, char* ="sp_Matrix");
  sp_Matrix(sp_Matrix &);
  ls_Vector row(ls_UINT i)const
  { ls_Vector u(a->j); get_row(u,i,0); return u;}
  ls_Vector column(ls_UINT j)const
  { ls_Vector u(a->i); get_column(u,0,j); return u;}
  ls_Vector upper_diagonal(ls_UINT j) const
  { ls_UINT n=(a->i<a->j)?a->i:a->j; ls_Vector u(n);
    get_diagonal(u,0,j); return u;}
  ls_Vector lower_diagonal(ls_UINT i)const
  { ls_UINT n=(a->i<a->j)?a->i:a->j; ls_Vector u(n);
    get_diagonal(u,i,0); return u;}
  const sp_Matrix& operator=(const sp_Matrix &);
  sp_Matrix operator+(const sp_Matrix &)const;   //A+B
  sp_Matrix operator-(const sp_Matrix &)const;   //A-B
  sp_Matrix& operator+=(const sp_Matrix &);      //A=A+B
  sp_Matrix& operator-=(const sp_Matrix &);      //A=A-B
  sp_Matrix operator*(const sp_Matrix &)const;   //A*B
  sp_Vector operator* (const sp_Vector &)const;  //A*x
  ls_Vector operator* (const ls_Vector &)const;  //A*x
  ls_UINT solve(ls_Vector&, ls_Vector&) const;      //cg-Verfahren
};
sp_Vector operator* (const sp_Vector &, const sp_Matrix &);
ls_Vector operator* (const ls_Vector &, const sp_Matrix &);
sp_Matrix operator*(ls_REAL, const sp_Matrix &);  // s*A
#ifndef SP_LIB
#include "sp_Matrix.cpp"
#endif
#endif
```

Natürlich wächst bei der Kompaktspeicherung der Organisationsaufwand. Man kann jedoch sagen, daß selbst bei einer Besetzung der Matrix mit ca. 20% NNE eine Kompaktspeicherung noch zu wesentlichen Zeiteinsparungen führt.
Bei symmetrischen Matrizen spart man dadurch weiteren Speicherplatz, daß man nur das untere bzw. obere Dreieck abspeichert. Es sei hier angemerkt, daß eine analoge Kompaktspeicherung auch bei Anwendung auf Datenbanken zu erheblichen Einsparungen an Rechenzeit und Speicherplatz führen kann.

Durch die Anwendung einer Kompaktspeicherung treten bei der numerischen Berechnung einer Zerlegung der Matrix neue Probleme auf. Die Auswahl der Pivotelemente darf nicht nur dazu dienen, numerische Probleme zu reduzieren, sondern muß auch das Anwachsen der NNE-Anzahl in der berechneten Zerlegung gering halten. Eine in dieser Hinsicht schlechte Pivotisierung kann aus einer sparse-Matrix eine volle machen. So liefert die Cholesky-Zerlegung für eine Matrix der Form

$$\begin{bmatrix} * & * & * & * & * \\ * & * & 0 & 0 & 0 \\ * & 0 & * & 0 & 0 \\ * & 0 & 0 & * & 0 \\ * & 0 & 0 & 0 & * \end{bmatrix}$$

eine untere Dreiecksmatrix der Form

$$\begin{bmatrix} * & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 \\ * & * & * & 0 & 0 \\ * & * & * & * & 0 \\ * & * & * & * & * \end{bmatrix},$$

während nach Vertauschen von erster mit letzter Zeile und Spalte eine sparse-Dreiecksmatrix der Form

$$\begin{bmatrix} * & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & 0 \\ 0 & 0 & * & 0 & 0 \\ 0 & 0 & 0 & * & 0 \\ * & * & * & * & * \end{bmatrix}$$

entsteht. Effiziente Methoden zur Lösung eines linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$ mit einer positiv definiten Koeffizientenmatrix bestehen aus drei Schritten:

- Wahl einer geeigneten Permutation $\mathbf{P}$ der Matrix $\mathbf{A}$, so daß das sog. **Fill-in** (die Auffüllung mit NNE) für die Cholesky-Zerlegung möglichst gering ist,
- numerische Berechnung der unteren Dreiecksmatrix $\mathbf{L}$,
- Berechnung der Lösung $\mathbf{x}^*$ durch Lösen der gestaffelten Gleichungssysteme

$$\mathbf{Lz} = \mathbf{Pb}, \quad \mathbf{L}^\mathsf{T}\mathbf{u} = \mathbf{z}, \quad \mathbf{x} = \mathbf{P}^\mathsf{T}\mathbf{u}.$$

Für zahlreiche Anwendungen ist eine Bandmatrix typisch: Die NNE befinden sich in der Nähe der Hauptdiagonalen (in wenigen Nebendiagonalen), so daß außerhalb eines Bandes um die Hauptdiagonale alle Matrixelemente gleich 0 sind. Derartige Matrizen lassen sich insbesondere bei kleiner, von n unabhängiger Bandbreite d sehr schnell behandeln, da z. B. die Multiplikation einer Bandmatrix mit einem Vektor nur $\mathcal{O}(n)$ Operationen benötigt. Bei ihnen würde eine Pivotisierung die Bandstruktur zerstören, während die untere Dreiecksmatrix bei einer Cholesky-Zerlegung wieder eine Bandmatrix ist und mit $\mathcal{O}(n^2)$ Operationen berechnet werden kann. Bisher haben wir nur sog. direkte Methoden zum Lösen linearer Gleichungssysteme besprochen; das sind solche Methoden, die die Aufgabe in eine äquivalente überführen, deren Lösung in einem Schritt erhalten werden kann. Solche Methoden verwenden direkt die Koeffizientenmatrix, indem sie diese transformieren. Hier bieten sich auch iterative Methoden an, die anstelle der Koeffizientenmatrix ein Unterprogramm verwenden, das bei Eingabe eines Vektors als Ausgabe die Multiplikation des Vektors mit der Koeffizientenmatrix liefert. Dadurch braucht man im Algorithmus selbst die Koeffizientenmatrix nicht unmittelbar, und in dem genannten Unterprogramm kann man die spezielle Struktur der Matrix direkt ausnutzen, so daß ein Aufruf des Unterprogramms meist mit $\mathcal{O}(n)$ Operationen auskommt. Beispielhaft soll hier das **Gauß-Seidel**-Verfahren besprochen werden. Der Grundgedanke der Methode besteht darin, das Gleichungssystem in eine iterierfähige Form zu bringen, so daß man zeigen kann, daß die mit der Iteration erzeugte Vektorfolge gegen die Lösung der Aufgabe konvergiert. Wir setzen voraus, daß die Koeffizientenmatrix eine absolut streng diagonal-dominante Hauptdiagonale hat. Lösen wir die i-te Gleichung nach x_i auf:

$$x_i = -\frac{1}{a_{ii}} \left(\sum_{j=1}^{i-1} a_{ij} x_j + \sum_{j=i+1}^{n} a_{ij} x_j - b_i \right), \quad i = 1, \ldots, n.$$

Das legt die folgende Iteration nahe:

$$x_i^{r+1} = x_i^{(r)} - \frac{1}{a_{ii}} \left(\sum_{j=1}^{i-1} a_{ij} x_j^{(r+1)} + \sum_{j=i+1}^{n} a_{ij} x_j^{(r)} - b_i \right), \quad i = 1, \ldots, n.$$

Für Matrizen mit streng diagonal-dominanter Hauptdiagonale konvergiert diese Methode recht gut gegen die gesuchte Lösung. Muß man das System mit mehreren rechten Seiten lösen, so kann man natürlich eine bereits berechnete Lösung als Startvektor für den nächsten Durchlauf nehmen. Falls der Startvektor schon nahe an der wahren Lösung liegt, brauchen nur wenige Iterationsschritte ausgeführt zu werden.
Eine weitere, besonders effiziente Methode ist das **cg-Verfahren** (konjugiertes Gradientenverfahren) zur Lösung eines linearen Gleichungssystems

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

mit einer positiv definiten Koeffizientenmatrix $\mathbf{A}$. Die Methode liefert bei einem beliebigen Startvektor $\mathbf{x}_0$ eine Kette von Vektoren $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_l$, die nach spätestens n Schritten mit der gesuchten Lösung abbricht, falls man exakt rechnet. Der Operationsaufwand pro Schritt wird durch den Aufwand bei der Multiplikation der Matrix $\mathbf{A}$ mit einem Vektor bestimmt und ist daher attraktiv bei schwachbesetzten Matrizen. Wir wollen sogleich die Methode beschreiben:
Wähle $\mathbf{x}_0 \in \mathbb{R}^n$, setze $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ und berechne

$$\mathbf{cg}: \quad \begin{cases} \alpha_k &= \dfrac{\mathbf{r}_k^\mathsf{T}\mathbf{r}_k}{\mathbf{p}_k^\mathsf{T}\mathbf{A}\mathbf{p}_k}, \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k\mathbf{p}_k, \\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k\mathbf{A}\mathbf{p}_k, \\ \beta_k &= \dfrac{\mathbf{r}_{k+1}^\mathsf{T}\mathbf{r}_{k+1}}{\mathbf{r}_k^\mathsf{T}\mathbf{r}_k}, \\ \mathbf{p}_{k+1} &= \mathbf{r}_{k+1} + \beta_k\mathbf{p}_k. \end{cases}$$

bis $\mathbf{p}_k = \mathbf{o}$ gilt.
Eine formale Betrachtung der Methode zeigt, daß man 4 Vektoren speichern muß; den Operationsaufwand für Matrix mal Vektor und für 6 Skalarprodukte pro Schritt hat. Für dieses Verfahren gilt nun

Satz 162. *Es gibt eine kleinste natürliche Zahl $l, 0 \leqq l \leqq n$ mit $\mathbf{p}_l = \mathbf{o}$. Außerdem gilt*

1. $\mathbf{A}\mathbf{x}_l = \mathbf{b}$.
2. $\mathbf{r}_i^\mathsf{T}\mathbf{p}_k = 0 \quad (0 \leqq k < i \leqq l)$.
3. $\mathbf{r}_i^\mathsf{T}\mathbf{p}_i = \mathbf{r}_i^\mathsf{T}\mathbf{r}_i \quad (i \leqq l)$.
4. $\mathbf{p}_i^\mathsf{T}\mathbf{A}\mathbf{p}_k = 0 \quad (0 \leqq i < k \leqq l)$.
5. $\mathbf{p}_i^\mathsf{T}\mathbf{A}\mathbf{p}_i > 0 \quad (i < l)$.
6. $\mathbf{r}_i^\mathsf{T}\mathbf{r}_k = 0 \quad (0 \leqq i < k < l)$.
7. $\mathbf{r}_i^\mathsf{T}\mathbf{r}_i > 0 \quad (i < l)$.
8. $\mathbf{r}_i = \mathbf{b} - \mathbf{A}\mathbf{x}_i \quad (i \leqq l)$.

Nach diesem Satz sind die Vektoren $\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_l$ orthogonal; es können höchstens n Nicht-Nullvektoren zueinander orthogonal sein; daher muß die Methode nach spätestens n Schritten mit der gesuchten Lösung enden. Wegen der auftretenden Rundungsfehler wird dies numerisch nicht der Fall sein. Man setzt daher das Verfahren solange fort bis das Residuum $\mathbf{r}$ hinreichend klein geworden ist.
Das **cg**-Verfahren kann auch auf allgemeine Gleichungssysteme $\mathbf{A}\mathbf{x} = \mathbf{b}$ mit einer regulären Koeffizientmatrix angewendet werden. Da ein Vektor $\mathbf{x}^*$ das System $\mathbf{A}\mathbf{x} = \mathbf{b}$ genau dann löst, wenn er das System $\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} = \mathbf{A}^\mathsf{T}\mathbf{b}$ löst, kann man das **cg**-Verfahren auf letzteres System anwenden, zumal dieses eine positiv definite Koeffizientenmatrix besitzt. Die explizite Berechnung von $\mathbf{A}^\mathsf{T}\mathbf{A}$ kann dabei vermieden werden:
Es sei $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0,\ \mathbf{p}_0 = \mathbf{A}^\mathsf{T}\mathbf{r}_0$.

$$\mathbf{cg_u}: \quad \begin{cases} \alpha_k &= \dfrac{\mathbf{r}_k^\mathsf{T}\mathbf{r}_k}{\mathbf{p}_k^\mathsf{T}\mathbf{p}_k}, \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k\mathbf{p}_k, \\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k\mathbf{A}\mathbf{p}_k, \\ \beta_k &= \dfrac{\mathbf{r}_{k+1}^\mathsf{T}\mathbf{r}_{k+1}}{\mathbf{r}_k^\mathsf{T}\mathbf{r}_k}, \\ \mathbf{p}_{k+1} &= \mathbf{A}^\mathsf{T}\mathbf{r}_{k+1} + \beta_k\mathbf{p}_k. \end{cases}$$

Man kann zeigen, daß das **cg**-Verfahren umso schneller konvergiert, je kleiner die Kondition der Koeffizientenmatrix ist. Dieser Sachverhalt wird bei den sog. vorkonditionierten **cg**-Verfahren ausgenutzt. Man versucht, die positiv definite Koeffizientenmatrix $\mathbf{A}$ durch eine andere positiv definite Matrix $\mathbf{C}$ (Vorkonditionierungsmatrix) derart anzunähern, daß die Matrix $\mathbf{C}^{-1}\mathbf{A}$ näherungsweise die Einheitsmatrix ist. Dazu sei $\mathbf{C}$ eine unvollständige Cholesky-Zerlegung der Matrix $\mathbf{A}$: $\mathbf{C} = \mathbf{L}\mathbf{L}^\mathsf{T}$. Das System $\mathbf{A}\mathbf{x} = \mathbf{b}$ ist äquivalent zu $\overline{\mathbf{A}}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ mit $\overline{\mathbf{A}} = \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^{-1})^\mathsf{T}$, $\overline{\mathbf{x}} = \mathbf{L}^\mathsf{T}\mathbf{x}$, $\overline{\mathbf{b}} = \mathbf{L}^{-1}\mathbf{b}$. Unter Verwendung der Transformationsregeln erhält man sofort aus dem **cg**-Verfahren die neuen Regeln:
Es sei $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $\mathbf{p}_0 = (\mathbf{L}\mathbf{L}^\mathsf{T})^{-1}\mathbf{r}_0$, $\mathbf{q}_0 = \mathbf{p}_0$.

$$\textbf{u\_cg}: \quad \begin{cases} \alpha_k &= \dfrac{\mathbf{r}_k^\mathsf{T}\mathbf{q}_k}{\mathbf{p}_k^\mathsf{T}\mathbf{A}\mathbf{p}_k}, \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k\mathbf{p}_k, \\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k\mathbf{A}\mathbf{p}_k, \\ \mathbf{q}_{k+1} &= (\mathbf{L}\mathbf{L}^\mathsf{T})^{-1}\mathbf{r}_{k+1} \\ \beta_k &= \dfrac{\mathbf{r}_{k+1}^\mathsf{T}\mathbf{q}_{k+1}}{\mathbf{r}_k^\mathsf{T}\mathbf{q}_k}, \\ \mathbf{p}_{k+1} &= \mathbf{q}_{k+1} + \beta_k\mathbf{p}_k. \end{cases}$$

Wie man sieht, ist hier zusätzlich in jedem Schritt ein lineares Gleichungssystem $\mathbf{q} = (\mathbf{L}\mathbf{L}^\mathsf{T})^{-1}\mathbf{r}$ zu lösen. Für die Wahl der unvollständigen Cholesky-Zerlegung der Matrix $\mathbf{A}$ gibt es verschiedene Vorschläge. Der wohl bekannteste Vorschlag ist, nur die NNE der wahren Cholesky-Zerlegung für die NNE der Matrix $\mathbf{A}$ (oder einer Teilmenge davon) zu berechnen. Dieser Vorschlag läßt sich für diagonaldominante Matrizen $\mathbf{A}$ mit $a_{ii} > 0$ und $a_{ij} \leqq 0 (i \neq j)$ begründen. Alle diese Techniken sind dann effizient, wenn man $\mathbf{A}\mathbf{x}$ mit $\mathcal{O}(n)$ Operationen berechnen kann, wie z. B. bei schwachbesetzten Matrizen. Daher sind diese Techniken im System **SP** implementiert.

6.6.4. Ausgleichungsrechnung

Eine ziemlich typische angewandte Aufgabe ist die folgende:
Es sollen gewisse Werte $x_1, x_2, \ldots, x_n$ bestimmt werden; jedoch ist es praktisch nicht möglich, diese direkt zu messen. Vielmehr ist man gezwungen, sich mit der Messung einer anderen Größe y zu begnügen, wobei man annimmt, daß zwischen y und $x_1, x_2, \ldots, x_n$ sowie einstellbaren Versuchsbedingungen z ein funktionaler Zusammenhang besteht:

$$y = f(z, x_1, x_2, \ldots, x_n).$$

Unter $m, m \geqq n$ verschiedenen Versuchsbedingungen

$$z_1, z_2, \ldots, z_m$$

werden die entsprechenden Ergebnisse $y_1, y_2, \ldots, y_m$ gemessen. Dabei ist nicht zu erwarten, daß die Messungen dem wahren bzw. angenommenen funktionalen Zusammenhang entsprechen; dies kann verschiedene Ursachen haben, wie etwa Meßfehler, unscharfe Versuchsbedingungen oder eine ungenaue Schätzung des funktionalen Zusammenhangs. Man wird daher durch Rechnung die unbekannten Parameter $x_1, x_2, \ldots, x_n$ so bestimmen, daß der angesetzte funktionale Zusammenhang möglichst gut mit den Meßwerten übereinstimmt, wozu z. B. das Gütemaß

$$\sum_{i=1}^{n} \left(y_i - f_i(x_1, x_2, \ldots, x_n)\right)^2$$

verwendet werden kann mit

$$f_i(x_1, x_2, \ldots, x_n) = f(z_i, x_1, x_2, \ldots, x_n), \quad i = 1, \ldots, n.$$

Ein wichtiger Spezialfall liegt vor, wenn die Funktionen f_i linear von den Parametern abhängen, d. h. wenn es eine (m, n)-Matrix $\mathbf{A}$ gibt mit

$$\begin{bmatrix} f_1(x_1, x_2, \ldots, x_n) \\ f_2(x_1, x_2, \ldots, x_n) \\ \ldots \\ f_m(x_1, x_2, \ldots, x_n) \end{bmatrix} = \mathbf{A}\mathbf{x}.$$

Dieser Fall soll hier untersucht werden. Genauer liegt die folgende Aufgabestellung vor: Es sei $\|\cdot\|$ die euklidische Norm. Gegeben seien eine (m, n)-Matrix $\mathbf{A}$ mit $m \geqq n$, ein Vektor $\mathbf{y} \in \mathbb{R}^m$; dann ist die Funktion

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 = (\mathbf{y} - \mathbf{A}\mathbf{x})^\mathsf{T}(\mathbf{y} - \mathbf{A}\mathbf{x})$$

zu minimieren.
Im Zusammenhang mit dieser Aufgabe spielen die sog. **Normalgleichungen** eine wesentliche Rolle:

$$\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} = \mathbf{A}^\mathsf{T}\mathbf{y}.$$

Satz 163. *Das lineare Ausgleichungsproblem*

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|$$

hat stets eine Lösung $\mathbf{x}^*$ *und genau alle Lösungen genügen der Gleichung* $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}^*$. *Das Residuum* $\mathbf{r} = \mathbf{y} - \mathbf{A}\mathbf{x}$ *genügt der Gleichung* $\mathbf{A}^\mathsf{T}\mathbf{r} = \mathbf{o}$. *Ein Vektor* $\mathbf{x}^*$ *löst genau dann die Aufgabe, wenn er Lösung der Normalgleichungen ist.*

Beweis. Es sei $\mathcal{L} \subseteqq \mathbb{R}^m$ die lineare Hülle aus den Spaltenvektoren der Matrix $\mathbf{A}$:

$$\mathcal{L} = \{\ \mathbf{A}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\ \}$$

und $\mathcal{L}^\perp$ der zugehörige Orthogonalraum:

$$\mathcal{L}^\perp = \left\{\ \mathbf{r} \;\middle|\; \mathbf{r}^\mathsf{T}\mathbf{A} = \mathbf{o}\ \right\}.$$

Dann läßt sich jeder Vektor $\mathbf{y} \in \mathbb{R}^m$ eindeutig in der Form

$$\mathbf{y} = \mathbf{u} + \mathbf{r}, \quad \mathbf{u} \in \mathcal{L}, \mathbf{r} \in \mathcal{L}^\perp$$

darstellen. Wegen $\mathbf{u} \in \mathcal{L}$ existiert ein Vektor $\mathbf{x}^*$ mit $\mathbf{A}\mathbf{x}^* = \mathbf{u}$, woraus

$$\mathbf{A}^\mathsf{T} = \mathbf{A}^\mathsf{T}\mathbf{u} + \mathbf{A}^\mathsf{T}\mathbf{r} = \mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x}^*,$$

folgt, d. h. der vektor $\mathbf{x}^*$ erfüllt die Normalgleichungen. Umgekehrt entspricht jeder Lösung $\overline{\mathbf{x}}$ der Normalgleichungen eine Zerlegung

$$\mathbf{y} = \mathbf{u} + \mathbf{r}, \quad \mathbf{u} = \mathbf{A}\overline{\mathbf{x}}, \quad \mathbf{r} = \mathbf{y} - \mathbf{A}\overline{\mathbf{x}}, \quad \mathbf{u} \in \mathcal{L}, \mathbf{r} \in \mathcal{L}^\perp.$$

Da die Zerlegung eindeutig ist, haben wir damit gezeigt, daß für zwei Lösungen $\mathbf{x}^*, \overline{\mathbf{x}}$ der Normalgleichungen $\mathbf{A}\overline{\mathbf{x}} = \mathbf{A}\mathbf{x}^*$ gilt. Es sei nun $\mathbf{x}^*$ eine Lösung der Normalgleichungen und $\mathbf{x}$ beliebig. Wir setzen $\mathbf{z} = \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*$ und $\mathbf{r} = \mathbf{y} - \mathbf{A}\mathbf{x}^*$. Wegen $\mathbf{r}^\mathsf{T}\mathbf{z} = 0$ folgt:

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 = \|\mathbf{r} - \mathbf{z}\|^2 = \|\mathbf{r}\|^2 + \|\mathbf{z}\|^2 \geqq \|\mathbf{y} - \mathbf{A}\mathbf{x}^*\|^2,$$

d. h. der Vektor $\mathbf{x}^*$ minimiert die Funktion $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2$. □
Es seien nun die Spalten der Matrix $\mathbf{A}$ linear unabhängig; dann gilt $\mathbf{A}\mathbf{x} \neq \mathbf{o}$ für alle Vektoren $\mathbf{x} \neq \mathbf{o}$ und die Matrix $\mathbf{A}^\mathsf{T}\mathbf{A}$ ist regulär, sogar positiv definit, da in diesem Falle

$$\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|^2 > 0, \quad \forall \mathbf{x} \neq \mathbf{o}$$

gilt. Daher sind dann die Normalgleichungen eindeutig lösbar:

$$\mathbf{x}^* = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{y}$$

und die Lösung kann man über eine Cholesky-Zerlegung der Matrix $\mathbf{A}^\mathsf{T}\mathbf{A}$ bestimmen. Diese Vorgehensweise ist aber numerisch nicht gutartig, da sich der Eingabefehler aus der Matrix $\mathbf{A}$ durch die Matrizenmultiplikation wesentlich verstärken kann. Ein anderer Weg ist vorzuziehen: Das lineare Ausgleichsproblem kann mittels Householder-Transformation gelöst werden. Dazu transformiert man die gegebene Matrix $\mathbf{A}^{(0)} = \mathbf{A}$ und den Vektor $\mathbf{y}^{(0)} = \mathbf{y}$ durch eine Folge von Householder-Transformationen $\mathbf{P}_r$

$$\mathbf{A}^{(r)} = \mathbf{P}_r\mathbf{A}^{(r-1)}, \quad y^{(r)} = \mathbf{P}_r y^{(r-1)}$$

in eine Matrix

$$\mathbf{A}^{(n)} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$$

mit einer oberen (n,n)-Dreiecksmatrix $\mathbf{R}$ und einen Vektor $\mathbf{h} = \mathbf{y}^{(n)}$; letzterer wird entsprechend zu $\mathbf{A}^{(n)}$ aufgespalten:

$$\mathbf{h} = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix}, \quad \mathbf{h}_1 \in \mathbb{R}^n, \quad \mathbf{h}_2 \in \mathbb{R}^{m-n}.$$

Die Matrix $\mathbf{P} = \mathbf{P}_n \cdots \mathbf{P}_1$ ist wieder orthogonal und es gilt $\mathbf{A}^{(n)} = \mathbf{PA}, \mathbf{h} = \mathbf{Py}$. Eine orthogonale Matrix läßt die Längen von Vektoren unverändert; also gilt

$$\|\mathbf{y} - \mathbf{Ax}\| = \|\mathbf{P}(\mathbf{y} - \mathbf{Ax})\| = \|\mathbf{y}^{(n)} - \mathbf{A}^{(n)}\mathbf{x}\| = \left\| \begin{array}{c} \mathbf{h}_1 - \mathbf{Rx} \\ \mathbf{h}_2 \end{array} \right\|.$$

Folglich wird die Länge genau dann minimal, wenn der Vektor $\mathbf{x}$ so gewählt wird, daß $\mathbf{h}_1 = \mathbf{Rx}$ ausfällt. Die Matrix $\mathbf{R}$ ist genau dann regulär, wenn die Spalten der Matrix $\mathbf{A}$ linear unabhängig sind. In diesem Falle erhält man aus dem System $\mathbf{h}_1 = \mathbf{Rx}$ genau eine Lösung, die das lineare Ausgleichsproblem löst. Sind die Spalten der Matrix $\mathbf{A}$ linear abhängig, so hat System $\mathbf{h}_1 = \mathbf{Rx}$ unendlich viele Lösungen, die alle das Ausgleichsproblem lösen. Für den Fehler erhält man

$$\|\mathbf{y} - \mathbf{Ax}\| = \|\mathbf{h}_2\|.$$

6.6.5. Implementierung linearer Systeme

Mittels der Programmiersprache `C++` werden Klassenvorlagen für ein- und zweidimensionale sowie obere Dreiecksfelder definiert, die der Verwaltung beliebiger Datentypen dienen. Davon abgeleitet sind die Klassen für Vektoren, allgemeine Matrizen, quadratische Matrizen und symmetrische Matrizen (inklusive symmetrische Bandmatrizen). Die Methoden und Operatoren dieser Klassen gestatten die Darstellung von Operationen analog zur Matrizenrechnung. Als Gleichungslöser gibt es in Abhängigkeit von der Klasse das Lösen mittels `LU`-Zerlegung, mittels `LDLT`-Zerlegung, mittels `QR`-Zerlegung (auch Lösen linearer Ausgleichsprobleme, Regularisierung) und mittels konjugiertem Gradientenverfahren. Die Klassen arbeiten mit dynamischen Komponenten; durch Streichen oder Hinzufügen von Funktionen lassen sie sich leicht spezifischen Aufgaben anpassen.

Vektoren auf Rechnern

Vektoren lassen sich i. a. nicht auf Rechnern darstellen. Dies erkennt man bereits daran, daß man z. B. keine stetigen Funktionen auf einem Rechner darstellen kann. Man ist daher gut beraten, sich auf solche Vektoren zu beschränken, die `n`-Tupel mit Komponenten gleichen Typs sind. Außerdem soll der Speicherplatz für ein `n`-Tupel dynamisch angefordert sein. Schließlich sollten die `n`-Tupel im Interesse einer schnelleren Fehlerfindung logisch unterscheidbar sein. Damit erfolgt die Klassen-Darstellung von Vektoren auf dem Rechner in 2 Stufen: Zunächst braucht man eine Klasse für `n`-Tupel von Daten und zu ihnen gehörende Methoden, die für jeden Datentyp ausführbar sind. Falls mit den Daten solche Operationen ausführbar sind, wie sie für `n`-dimensionale Vektoren gelten, ergeben sich aus den `n`-Tupeln Vektoren. Zur Darstellung von Vektoren auf Rechnern wird hier die Programmiersprache `C++` verwendet. Selbst wenn dem Leser die Konstrukte dieser objektorientierten Sprache fremd sein sollten, wird er diesen Text mit Gewinn studieren können, sofern Interesse und Verständnis für eine Programmiersprache vorhanden sind.
Die folgenden Darlegungen gelten in analoger Weise für weitere Klassen und ihre Objekte.
Die Aneinanderreihung von Daten gleichen Typs ist ein `n`-Tupel. Da der Typ der in einem `n`-Tupel abgelegten Daten zunächst unbekannt ist, wird eine Klassenvorlage `ls_array1<T>` für dynamische Felder des Typs `T` definiert.

```
#ifndef LS_ARRAY1
#define LS_ARRAY1
#include "ls_error.h"
template <class T>
class ls_array1 // Eindimensionale Felder
{ protected:
  T *a;
  char ONAME[ls_len];                          // Feld für den Objektnamen
  ls_UINT dim;                                 // Felddimension
  void set_data(T *aa=NULL,ls_UINT d=0){ a=aa, dim=d;}
  public:
  char* name;                                  // Zeiger auf Objektnamen
  ls_array1(ls_UINT =0,char* ="ls_array1"); // Standardkonstruktor
  ls_array1(ls_array1<T> &);                   // move-Konstruktor
  ls_array1(T*,ls_UINT,char* ="ls_array1"); // Feld-Übernahme
  ~ls_array1();
  ls_array1<T>& swap(ls_array1<T> &);          // Datenfelder-Austausch
  ls_UINT dimension() const{ return dim;}
  T* asArray()const{ return a;}
  T& operator[](ls_UINT) const;                // indizierter Zugriff
  const ls_array1<T>& operator=(const ls_array1<T> &);
  ls_array1<T>& put_array(const ls_array1<T>&,ls_UINT =0);
                                               // Feldeingabe ab Position
```

```
  ls_array1<T>& put_array(T, ls_UINT =0);   // Elementeingabe ab Position
  ls_array1<T>& put(T, ls_UINT);            // Elementeingabe auf Position
  T get(ls_UINT);                           // Elementausgabe von Position
  const ls_array1<T>& get_array(ls_array1<T> &, ls_UINT =0) const;
                                           // Feldausgabe ab Position
  ls_array1<T>& append(ls_UINT =1);         // Nullelemente anhängen
  ls_array1<T>& remove(ls_UINT);            // Element entfernen
  ls_array1<T>& remove();                   // Feld entfernen
  const ls_array1<T>& write(ostream&)const; // in Datei schreiben
  ls_array1<T>& read(istream &);            // aus Datei lesen
  const ls_array1<T>& operator>>(char*)const;// in namentl. Datei schreiben
  ls_array1<T>& operator<<(char*);          // aus namentl. Datei lesen
};
#ifndef LS_LIB
#include "ls_array1.cpp"
#endif
#endif
```

Wie hat man sich die Funktionsweise dieser Klassenvorlage vorzustellen? Aus der Vorlage wird eine Klasse, indem der Parameter `T` einen Wert erhält; dies geschieht dadurch, daß in einem Programm ein Konstrukt der Form `ls_array1<int> kind` auftritt. Wir wählen als Datentyp `int`. Zunächst sei eine Instanz deklariert:

```
ls_array1<int> v(n);.
```

Diese Deklaration wird als Aufruf des Unterprogramms `ls_array1<int>` mit dem Parameter `n` und Referenzen, die auf die Daten des Objektes zeigen, übersetzt; Danach ist `v` logisch ein `n`-Tupel des Typs `int` und der Anfangsbelegung `0`. Jedes Objekt erhält einen Namen; standardmäßig wird der Klassenname als Objektname vergeben; er darf mittels `strcpy` geändert werden: `strcpy(v.name,"v-Feld")`. Das `n`-Tupel ist dynamisch angelegt; die Komponente `a` (Zeiger) enthält die Referenz auf das angeforderte Feld; die Komponente `dim` enthält die aktuelle Dimension des Datenfeldes in Einheiten des Datentyps. Eine Instanz darf auch mit einem Datenfeld der Form `T *A` und seiner Feldlänge instanziiert werden. Die Dimension der dynamischen Komponente wird durch `v.dimension()` erfragt.
Die Tatsache, daß dynamische Komponenten verwendet werden, zwingt dazu, der Klasse eigene Konstruktoren, einen eigenen Zuweisungsoperator und einen eigenen Destruktor zu geben. In einem Konstruktor wird u. a. die dynamische Komponente angelegt und im Destruktor der Speicherplatz für die dynamische Komponente freigegeben. Üblicherweise ist auch ein sog. `copy`-Konstruktor vonnöten, der aufgerufen wird, falls man ein Objekt mittels eines anderen instanziiert.
Sehr wichtig ist es zu bemerken, daß hier in Klassen mit dynamischen Komponenten der `copy`-Konstruktor als `move`-Konstruktor arbeitet: Die Datenfelder werden in das neue Objekt übernommen; der Objektname wird kopiert; danach ist das Quellobjekt ohne Daten. Diese Form wird vorteilhaft beim Instanziieren des Rückkehrwertes verwendet. Angenommen, die in einem Unterprogramm erzeugte Instanz `A` mit dynamischen Komponenten soll Rückkehrwert sein. Bei der Anweisung `return A;` passiert folgendes: Durch den Aufruf des `copy`-Konstruktors wird auf dem Stack eine Kopie von `A` erzeugt; danach wird mittels Destruktor das Original zerstört und in das aufrufende Programm zurückgekehrt. Durch den realisierten `copy`-Konstruktor wird die Stack-Kopie von `A` die dynamischen Komponenten übernehmen ohne neuen Speicherplatz anzufordern und der für `A` aufgerufene Destruktor kann die dynamischen Komponenten nicht freigeben. Dadurch vermeidet man zeitweilige Doppelungen von Datenfeldern und Laufzeitfehler wegen fehlendem Speicherplatz. Soll der so implementierte `copy`-Konstruktor nicht aufgerufen werden, sollte man die Objekte per `call_by_reference` an ein Programm übergeben oder den Inhalt des Objektes vor dem Aufruf retten. Soll das Quellobjekt erhalten bleiben, so hat man anstelle von `ls_array1<int> A(B)` die Anweisungen `ls_array1<int> A; A=B;` zu codieren. Diese Anweisungen dürfen nicht zu `ls_array<int> A=B` verkürzt werden, da bei letzterer der `copy`-Konstruktor aufgerufen wird.
Eine weitere Möglichkeit, dynamische Datenfelder an andere Objekte zu übergeben besteht darin, die Methode `swap` anzuwenden: Durch den Aufruf `A.swap(B)` werden alle Datenfelder aus der Instanz `B`, einschließlich Objektname, mit dem Objekt `A` getauscht. Durch Anwenden der `swap`-Methode innerhalb eines Unterprogramms auf zwei Objekte, wobei das eine innerhalb und das andere außerhalb des Unterprogramms instanziiert wurde, werden Daten an das aufrufende Programm übermittelt, ohne (zeitweilig) zusätzlich Speicherplatz anzufordern. Der Vorteil einer solchen Vorgehensweise ist offensichtlich: Wenn ein Objekt `A` mehrere Objekte anderer Klassen enthält, können diese zunächst außerhalb von `A` erzeugt werden, um sie dann beim Instanziieren von `A` ohne Anfordern von zusätzlichem Speicherplatz zu übernehmen. Hätte man diesen Mechanismus nicht, müßte man im Interesse von Speicherplatzeinsparung Methoden zur Konstruktion der einzelnen Unterobjekte in die Klassendefinition von `A` aufnehmen.
Datenaustausch gibt es mit anderen Instanzen und mit Dateien. Zunächst soll der Datenaustausch mit anderen Instanzen kommentiert werden. Besonders wichtig ist der elementweise Zugriff:

```
i = v[3]; v[3] = k; .
```

Zur Ein- oder Ausgabe mehrerer Datenelemente dienen die Funktionen `put_array` und `get_array`. Im Aufruf beider sind das Datenfeld und die Position des ersten Datenelementes in der Quelle bzw. im Ziel anzugeben. Beim Aufruf `u.put_array(v,3)` werden Daten ab `v[0]` nach `u` ab `u[3]` kopiert; die Elementanzahl richtet sich nach dem Minimum aus der Dimension von `v` und der Dimension von `u` minus Anfangsposition. Der Aufruf `u.get_array(v,3)` schreibt die ab `u[3]` stehenden Daten in das Feld `v` ab `v[0]`; die geschriebene Elementanzahl berechnet sich in analoger Weise. Das Belegen aller Datenelemente von `u` mit dem Wert `s` wird durch `u.put_array(s)` oder `u.put_array(s,0)` erreicht.
Die Funktionen `write`, `read`, `<<` und `>>` dienen dem Datenaustausch mit Dateien. Will man z. B. die Instanz `v` in eine Datei namens `hallo` schreiben, so hat man einfach `v >> "hallo"` zu codieren; die Anweisung `v >> ""` führt zur Ausgabe auf das Standard-Ausgabemedium. Sollte man besondere Wünsche an die Voreinstellung der Datei haben, wie z. B. an die formatierte Ausgabe von Gleitpunktzahlen, sind die Funktionen `v.write(fout)` bzw. `v.read(fin)` zu verwenden; hierin bedeuten `fout` und `fin` Filedeskriptoren. In einer Datei haben Objektdaten eine standardisierte Darstellung. Sie beginnen mit dem Objektnamen, gefolgt von einem in geschweiften Klammern eingeschlossenen Block; der Block beginnt mit einer Zeile in der die Werte der Konstruktionsparameter mit vorangestelltem Schlüsselwort stehen, gefolgt von den Datenelementen. So hat das 4-Tupel

```
(1.1, 2.22, 3.333, 4.4444)
```

als Objekt mit dem Namen `huhu` die externe Darstellung

```
huhu { dimension: 4 1.1 2.22 3.333 4.4444 }.
```

Die einzelnen Daten sind durch übliche Trennzeichen getrennt. Sollte ein Objekt Instanzen enthalten, stehen im Block die Daten der entsprechenden Instanzen in analoger Form.
Ein Objekt, das Daten einlesen möchte, muß leer sein. Hat das Zielobjekt keinen Namen, wird das erste eingelesene Wort als Objektname verwendet; andernfalls wird zunächst nach dem Objektnamen gesucht. Durch diese Technik braucht das Programm vor dem Einlesen nicht den Objektnamen zu kennen:

```
ls_array1<int> k; strcpy(k.name,""); k<<"meine_datei";.
```

Vor Anwendung einer Klassenvorlage auf einen konkreten Datentyp müssen folgende Bedingungen erfüllt sein: Für den Datentyp muß ein Zuweisungsoperator definiert sein. Die `>>`- und `<<`-Operatoren für Dateien müssen definiert sein:

```
fin << v; fout >> v.
```

Für die standardmäßig vorhandenen Basistypen sind diese Bedingungen erfüllt.
Die Klassenvorlage `ls_array1<T>` vereinigt in sich Daten und Methoden, die unabhängig vom konkreten Datentyp sind. So kann die Klasse `ls_array1<unsigned short>` Basisklasse für eindimensionale Felder sein, deren Daten natürliche Zahlen sind. Eine konkrete Klasse mit diesen Datenelementen wird jedoch noch weitere Methoden beinhalten, wie z. B. die komponentenweise Addition von Feldern.
Wir werden erkennen, daß diese Klassenvorlage Basis für die Darstellung weiterer Objekte der linearen Algebra ist.
Basisklasse für einen Vektor ist die Klasse `ls_array1<ls_REAL>`, wobei `ls_REAL` für `float` bzw. `double` steht:

```
#ifndef LS_VECTOR
#define LS_VECTOR
#include "ls_array1.h"
class ls_Matrix; class ls_sMatrix;
class ls_Vector: public ls_array1<ls_REAL>
{ public:
  ls_REAL eps;
  ls_Vector(ls_UINT =0, char* ="ls_Vector");
  ls_Vector(ls_Vector &);
  ls_Vector(ls_REAL*, ls_UINT, char* ="ls_Vector");
  ~ls_Vector(){}
  const ls_Vector& operator=(const ls_Vector &);
  ls_Vector operator-() const;                 //-x
  ls_Vector operator*(ls_REAL) const;          //x*s
  ls_Vector& operator*=(ls_REAL);              //x=x*s
  ls_Vector operator+(const ls_Vector &)const; //x+y
  ls_Vector operator-(const ls_Vector &)const; //x-y
  ls_Vector& operator+=(const ls_Vector &);    //x=x+y
  ls_Vector& operator-=(const ls_Vector &);    //x=x-y
  ls_REAL operator*(const ls_Vector &)const;   //x*y
  ls_Matrix dyad(const ls_Vector &)const;      //dyad.
```

```
  ls_sMatrix dyad()const;                       //dyad.
};
ls_Vector operator*(ls_REAL, const ls_Vector &); //s*x
#ifndef LS_LIB
#include "ls_Vector.cpp"
#endif
#endif
```

Die Klasse `ls_Vector` erbt zunächst alle Daten und Methoden ihrer Basisklasse. Zusätzliche Methoden - hier meist arithmetische - führen von der Basisklasse zur Vektor-Klasse. Sind nun `x`, `y` Instanzen gleicher Dimension der Klasse `ls_Vector` und `s` eine `ls_REAL`-Zahl, so sind auch `x + y`, `x - y`, `s*x`, `x*s`, `u += s*x`, `u -= y`, `u *= s` Instanzen der gleichen Klasse. Daraus folgt insbesondere, daß sich Linearkombinationen von Vektoren analog zu mathematischen Formeln darstellen lassen:

```
x = y + s*u + t*v.
```

Die `dyad`-Funktionen erzeugen als dyadisches Produkt eine Matrix.

Matrizen auf Rechnern

Um eine Matrix auf einem Rechner darstellen zu können, benötigt man zunächst ein zweidimensionales Datenfeld mit Daten beliebigen Typs. Auf dem Rechner gibt es aber nur die aufeinander folgende Anordnung von Datenelementen, wie sie durch die Klassenvorlage `ls_array1<T>` erfaßt ist. Folglich muß ihr eine neue Struktur aufgeprägt werden: Wir wollen von Zeilen und Spalten sprechen dürfen und brauchen Methoden, die dieser Struktur angepaßt sind. Daraus ergibt sich eine Klassenvorlage `ls_array2<T>`, die aus `ls_array1<T>` abgeleitet ist:

```
#ifndef LS_ARRAY2
#define LS_ARRAY2
#include "ls_array1.h"
template <class T>
class ls_array2: public ls_array1<T>        // Zweidimensionale Felder
{ protected:
  ls_UINT m, n;
  void set_data(ls_UINT mm=0, ls_UINT nn=0){ m=mm; n=nn;}
  public:
  ls_array2(ls_UINT =0, ls_UINT =0, char* ="ls_array2");
  ls_array2(ls_array2<T> &);                // move-Konstruktor
  ls_array2<T>(T*, ls_UINT, ls_UINT, char* ="ls_array2");// Feld-Übernahme
  ~ls_array2(){}
  ls_array2<T>& swap(ls_array2<T> &);
  T* operator[](ls_UINT i) const;           // indizierter Zugriff
  ls_UINT number_of_rows() const{ return m;}
  ls_UINT number_of_columns() const{ return n;}
  const ls_array2<T>& operator=(const ls_array2<T> &);
  ls_array2<T>& put(T, ls_UINT, ls_UINT);   // Elementeingabe
  ls_array2<T>& put_row(const ls_array1<T>&, ls_UINT, ls_UINT =0);
  ls_array2<T>& put_row(T, ls_UINT, ls_UINT =0);
  ls_array2<T>& put_column(const ls_array1<T>&, ls_UINT, ls_UINT);
  ls_array2<T>& put_column(T, ls_UINT, ls_UINT);
  ls_array2<T>& put_diagonal(const ls_array1<T> &, ls_UINT =0, ls_UINT =0);
  ls_array2<T>& put_diagonal(T, ls_UINT =0, ls_UINT =0);
  ls_array2<T>& put_array(const ls_array2<T> &, ls_UINT =0, ls_UINT =0);
  ls_array2<T>& put_array(T, ls_UINT =0, ls_UINT =0);
  T get(ls_UINT, ls_UINT) const;            // Element-Ausgabe
  const ls_array2<T>& get_row(ls_array1<T>&, ls_UINT, ls_UINT =0) const;
  const ls_array2<T>& get_column(ls_array1<T>&, ls_UINT, ls_UINT) const;
  const ls_array2<T>& get_diagonal(ls_array1<T>&,ls_UINT =0,ls_UINT =0) const;
  const ls_array2<T>& get_array(ls_array2<T>&, ls_UINT =0, ls_UINT =0) const;
  ls_array2<T>& append_row(ls_UINT =1);       // Null-Zeilen anhaengen
  ls_array2<T>& append_column(ls_UINT =1);    // Null-Spalten anhaengen
  ls_array2<T>& append_array(ls_UINT =1, ls_UINT =1);// Null-Feld anhaengen
  ls_array2<T>& remove_row(ls_UINT i);        // Zeile streichen
  ls_array2<T>& remove_column(ls_UINT i);     // Spalte streichen
  ls_array2<T>& remove();                     // Datenfeld streichen
  ls_array2<T>& swap_row(ls_UINT, ls_UINT);   // Zeilentausch
  ls_array2<T>& swap_column(ls_UINT, ls_UINT); // Spaltentausch
  const ls_array2<T>& write_row(ostream &) const;
  ls_array2<T>& read_row(istream &);
  const ls_array2<T>& write_column(ostream &) const;
```

```
  ls_array2<T>& read_column(istream &);
  const ls_array2<T>& operator>>(char *) const;// zeilenweise Ausgabe in Datei
  ls_array2<T>& operator<<(char *);            // zeilenweise Eingabe aus Datei
};
#ifndef LS_LIB
#include "ls_array2.cpp"
#endif
#endif
```

Wählen wir als Datentyp `int`, so lautet die Deklaration einer Instanz:

```
ls_array2<int> A(mm, nn).
```

Hierin geben `mm` die Zeilenanzahl und `nn` die Spaltenanzahl des Datenfeldes an; diese Daten werden auf `m` und `n` abgelegt, so daß sich als Gesamtlänge des angeforderten Feldes `dim = m*n` ergibt. Ein Vergleich mit der Klassenvorlage `ls_array1<T>` zeigt die Analogien. Der Direktzugriff erfolgt über Doppelindices: `A[i][j] = s; s = A[i][j];`. Die Zeilen- bzw. Spaltenanzahl wird mit `A.number_of_rows()` bzw. `A.number_of_columns()` abgefragt. Methoden der Basisklasse werden entweder übernommen oder sinnvoll durch andere überlagert. Da dieses Datenfeld in Zeilen und Spalten strukturiert ist, gibt es auch Diagonalen. Dem wird dadurch entsprochen, daß es Methoden gibt, die mit Zeilen, Spalten oder Diagonalen arbeiten:

- Eingabe von Werten und Feldern als Zeile, Spalte oder Diagonale:

  `put_row`, `put_column`, `put_diagonal`,

- Anhängen von Null-Zeilen bzw. Null-Spalten:

  `append_row`, `append_column`,

- Eingabe eines zweidimensionalen Feldes als Unterfeld:

  `put_array`,

- Ausgabe von Feldern, die als Zeilen, Spalten, Diagonalen, Unterfeldern in der Instanz vorkommen:

  `get_row`, `get_column`, `get_diagonal`, `get_array`,

- Vertauschen von Zeilen oder Spalten:

  `swap_row`, `swap_column`,

- Streichen von Zeilen oder Spalten:

  `remove_row`, `remove_column`.

Der externe Datenaustausch (Instanz mit Datei) erfolgt nun zeilen- oder spaltenweise, jedoch in der gleichen äußeren Form wie in der Basisklasse.
Beispiel: Das zweidimensionale Datenfeld mit 4 Zeilen und 5 Spalten

$$\begin{pmatrix} 1 & 1 & 2 & 2 & 0 \\ 1 & 2 & 3 & 2 & 1 \\ 2 & 1 & 0 & 3 & 1 \\ 0 & 0 & 1 & 1 & 3 \end{pmatrix}$$

und dem Namen `hello` hat die externe Darstellung

```
hello
{ number_of_rows: 4 number_of_columns: 5
  1 1 2 2 0
  1 2 3 2 1
  2 1 0 3 1
  0 0 1 1 3
}.
```

Bis hier enthält diese Klassenvorlage weitgehend datentyp-unabhängige Methoden. Aus ihr wird eine **Matrix** (Rechteckmatrix), wenn die Datenelemente aus einem algebraischen Körper genommen werden. Diese Datentyp-Spezifikation erlaubt es, arithmetische Operationen auszuführen und man erhält die abgeleitete Klasse `ls_Matrix`.

```
#ifndef LS_MATRIX
#define LS_MATRIX
#include "ls_Vector.h"
#include "ls_array2.h"
```

```
class ls_Matrix: public ls_array2<ls_REAL>
{ public:
  ls_Matrix(ls_UINT =0, ls_UINT =0, char* ="ls_Matrix");
  ls_Matrix(ls_Matrix &);
  const ls_Matrix& operator=(const ls_Matrix &);
  ls_Vector row(ls_UINT i)const
  { ls_Vector u(n); get_row(u,i); return u;}
  ls_Vector column(ls_UINT j)const
  { ls_Vector u(n); get_column(u,0,j); return u;}
  ls_Vector upper_diagonal(ls_UINT j)const
  { ls_Vector u(n); get_diagonal(u,0,j); return u;}
  ls_Vector lower_diagonal(ls_UINT i)const
  { ls_Vector u(n); get_diagonal(u,i,0); return u;}
  ls_Matrix& operator+=(const ls_Matrix &);    //A=A+B
  ls_Matrix operator-(const ls_Matrix &)const; //A-B
  ls_Matrix& operator-=(const ls_Matrix &);    //A=A-B
  ls_Matrix operator+(const ls_Matrix &) const;//A+B
  ls_Matrix operator*(const ls_Matrix &) const;//A*B
  ls_Vector operator*(const ls_Vector &) const;//A*x
  ls_Matrix operator*(ls_REAL) const;          //A*s
  ls_Matrix& operator*=(ls_REAL);              //A=A*s
  ls_UINT solve(ls_Vector &x, ls_Vector &b) const;//cg-Verfahren
};
ls_Matrix operator*(ls_REAL, const ls_Matrix &);  // s*A
ls_Vector operator*(const ls_Vector &, const ls_Matrix &);// x*A
#ifndef LS_LIB
#include "ls_Matrix.cpp"
#endif
#endif
```

Man sieht, daß die Klasse `ls_array2<ls_REAL>` lediglich um arithmetische Operationen mit Vektoren, Matrizen und Zahlen ergänzt ist. Sind nun `A`, `B`, `C` Rechteckmatrizen (Instanzen der Klasse `ls_Matrix`), `x`, `y` `z` Vektoren passender Dimension und `s, t` Zahlen, so wird durch `A*x` ein Vektor erzeugt (Matrix mal Vektor); ebenso durch `y*A` (transponierte Matrix mal Vektor); bei `C = A*B` wird der Matrix `C` ein Matrizenprodukt Matrix mal Matrix zugewiesen. Aber auch zusammengesetzte Operationen lassen sich problemlos in einem Programm notieren: `z = s*(A*x) - t*(y*B)`.

Symmetrische Matrizen auf Rechnern

Symmetrische Matrizen zeichnen sich gegenüber quadratischen Matrizen dadurch aus, daß die Daten an der Hauptdiagonalen gespiegelt sind; daher braucht auch nur das obere Dreieck der Matrix abgespeichert zu werden. Dem obigen Vorgehen folgend ist also zunächst aus der Klassenvorlage `ls_array1<T>` eine Klassenvorlage für ein zweidimensionales, oberes Dreiecksfeld abzuleiten.

```
// oberes Dreiecksfeld
#ifndef LS_ARRAYU
#define LS_ARRAYU
#include "ls_array2.h"
template <class T>
class ls_arrayU: public ls_array1<T>
{ protected:
  ls_UINT n;
  void set_data(ls_UINT nn=0){ n=nn;}
  public:
  ls_arrayU(ls_UINT =0, char* ="ls_arrayU");
  ls_arrayU(ls_arrayU<T>&);
  ~ls_arrayU(){}
  ls_arrayU<T>& swap(ls_arrayU<T>&);
  T* operator[](ls_UINT k) const;
  ls_UINT dimension() const{ return n;}
  const ls_arrayU<T>& operator=(const ls_arrayU<T> &);
  ls_arrayU<T>& put(T, ls_UINT, ls_UINT);
  ls_arrayU<T>& put_row(const ls_array1<T>&, ls_UINT, ls_UINT);
  ls_arrayU<T>& put_row(T, ls_UINT, ls_UINT);
  ls_arrayU<T>& put_column(const ls_array1<T>&, ls_UINT, ls_UINT);
  ls_arrayU<T>& put_column(T, ls_UINT, ls_UINT);
  ls_arrayU<T>& put_diagonal(const ls_array1<T>&, ls_UINT, ls_UINT);
  ls_arrayU<T>& put_diagonal(T, ls_UINT, ls_UINT);
  T get(ls_UINT, ls_UINT) const;
```

```
  const ls_arrayU<T>& get_row(ls_array1<T>&, ls_UINT, ls_UINT) const;
  const ls_arrayU<T>& get_column(ls_array1<T>&, ls_UINT, ls_UINT) const;
  const ls_arrayU<T>& get_diagonal(ls_array1<T>&, ls_UINT, ls_UINT) const;
  ls_arrayU<T>& append_column(ls_UINT =1);
  ls_arrayU<T>& swap(ls_UINT, ls_UINT);
  ls_arrayU<T>& remove(ls_UINT);
  ls_arrayU<T>& remove();
  const ls_arrayU<T>& write_row(ostream &) const;
  ls_arrayU<T>& read_row(istream &);
  const ls_arrayU<T>& write_column(ostream &) const;
  ls_arrayU<T>& read_column(istream &);
  const ls_arrayU<T>& operator>> (char *) const;
  ls_arrayU<T>& operator<< (char *);
};
#ifndef LS_LIB
#include "ls_arrayU.cpp"
#endif
#endif
```

Man erkennt die große Ähnlichkeit mit der Vorlage `ls_array2<T>`. Es gibt aber wichtige Änderungen: Die Indizierung der Datenelemente erfolgt so, als ob die Daten aus dem unteren Dreieck vorhanden wären: Bei jeder Positionsangabe mittels Zeilenindex `i` und Spaltenindex `j` muß stets `i <= j` gelten. Dem Nutzer ist es überlassen, ob es sich dabei um ein symmetrisches oder unsymmetrisches oberes Dreiecksfeld handelt. Aus dieser Klasse wird eine symmetrische Matrix abgeleitet.

```
// symmetrische Matrix (nur oberes Dreieck)
#ifndef LS_SMATRIX
#define LS_SMATRIX
#include "ls_Vector.h"
#include "ls_arrayU.h"
class ls_Matrix;
class ls_sMatrix: public ls_arrayU<ls_REAL>
{ public:
  ls_sMatrix(ls_UINT =0, char* ="ls_sMatrix");
  ls_sMatrix(ls_sMatrix &);
  ls_Vector row(ls_UINT i)const
  { ls_Vector u(n); get_row(u,i,i+1), get_column(u,0,i); return u;}
  ls_Vector diagonal(ls_UINT j)const
  { ls_Vector u(n); get_diagonal(u,0,j); return u;}
  const ls_sMatrix& operator=(const ls_sMatrix &); //A=B
  ls_sMatrix& operator+=(const ls_sMatrix&);       //A=A+B
  ls_sMatrix operator-(const ls_sMatrix&) const;   //A-B
  ls_sMatrix& operator-=(const ls_sMatrix&);       //A=A-B
  ls_sMatrix operator+(const ls_sMatrix&) const;   //A+B
  ls_sMatrix& operator*=(ls_REAL);                 //A=A*s
  ls_Vector operator*(const ls_Vector&) const;     //A*x
  ls_sMatrix operator*(ls_REAL) const;             //A*s
  ls_Matrix operator*(const ls_sMatrix &) const;   //A*B
  ls_UINT solve(ls_Vector &x, ls_Vector &b);       //cg-Verfahren
};
ls_sMatrix operator*(ls_REAL, const ls_sMatrix&);  //s*A
#include "ls_Matrix.h"
#ifndef LS_LIB
#include "ls_sMatrix.cpp"
#endif
#endif
```

Hier wird das obere Dreiecksfeld symmetrisch interpretiert. Dies wirkt sich insbesondere auf das Lösen eines entsprechenden linearen Gleichungssystems aus. Desweiteren wird aus der Klassenvorlage `ls_arrayU<T>` eine oberer Dreiecksmatrix abgeleitet.

```
// obere Dreiecksmatrix
#ifndef LS_UMATRIX
#define LS_UMATRIX
#include "ls_Vector.h"
#include "ls_arrayU.h"
class ls_Matrix;
class ls_uMatrix: public ls_arrayU<ls_REAL>
{ public:
  ls_uMatrix(ls_UINT =0, char* ="ls_uMatrix");
  ls_uMatrix(ls_uMatrix &);
```

```
  ls_Vector row(ls_UINT i)const
  { ls_Vector u(n); get_row(u,i,i); return u;}
  ls_Vector column(ls_UINT j)const
  { ls_Vector u(n); get_column(u,0,j); return u;}
  ls_Vector diagonal(ls_UINT j)const
  { ls_Vector u(n); get_diagonal(u,0,j); return u;}
  const ls_uMatrix& operator=(const ls_uMatrix &); //A=B
  ls_uMatrix& operator+=(const ls_uMatrix&);       //A=A+B
  ls_uMatrix operator-(const ls_uMatrix&) const;   //A-B
  ls_uMatrix& operator-=(const ls_uMatrix&);       //A=A-B
  ls_uMatrix operator+(const ls_uMatrix&) const;   //A+B
  ls_uMatrix& operator*=(ls_REAL);                 //A=A*s
  ls_Vector operator*(const ls_Vector&) const;     //A*x
  ls_uMatrix operator*(ls_REAL) const;             //A*s
  ls_Matrix operator*(const ls_uMatrix &) const;   //A*B
  ls_UINT solve(ls_Vector &, ls_Vector &);         //cg-Verfahren
  void backward(ls_Vector &, ls_Vector &);         // Rueckwaertseinsetzen
};
ls_uMatrix operator*(ls_REAL, const ls_uMatrix&);  //s*A
ls_Vector operator*(const ls_Vector&, const ls_uMatrix&); //x*A
#include "ls_Matrix.h"
#ifndef LS_LIB
#include "ls_uMatrix.cpp"
#endif
#endif
```

Symmetrischen Bandmatrizen mit Diagonalspeicherung sind für Diskretisierungsmethoden wichtig; daher ist ihnen eine besondere Klasse gewidmet. Zunächst sei eine Klassenvorlage namens `ls_array1M<T>` definiert, in der eine Sammlung von eindimensionalen Datenfeldern verwaltet wird:

```
#ifndef LS_ARRAY1M
#define LS_ARRAY1M
#include "ls_array1.h"
template <class T>
struct array1M{ ls_UINT l; T *f;};
template <class T>
class ls_array1M          // Reihung Eindimensionaler Felder
{ protected:
  array1M<T> *a;          // Datenfeld
  ls_UINT   n;            // Dimension von a
  char ONAME[ls_len];
  void set_data(array1M<T> *aa=NULL,ls_UINT nn=0){ a=aa; n=nn;}
  public:
  char *name;             // Objektname
  ls_array1M(ls_UINT =0, char* ="ls_array1M");
  ls_array1M(ls_array1M<T> &);
  ~ls_array1M();
  ls_array1M<T>& swap(ls_array1M<T> &);
  ls_UINT number_of_arrays() const{ return n;}
  ls_UINT length(ls_UINT i) const{ return (a+i)->l;}
                                  // Länge des i-ten Feldes
  array1M<T>* asArray()const{ return a;}
  T* operator[](ls_UINT i){ return (a+i)->f;}
                                  // indizierter Zugriff auf das i-te Feld
  const ls_array1M<T>& operator=(const ls_array1M<T>&);
  ls_array1M<T>& put_array(const ls_array1<T>&v, ls_UINT i);
                                  // Eingabe des i-ten Feldes
  ls_array1M<T>& put(const T*, ls_UINT l, ls_UINT i);
                                  // Eingabe eines Feldes als i-tes Feld
  ls_array1M<T>& put(T, ls_UINT); // Eingabe eines Wertes als i-tes Feld
  const ls_array1M<T>& get_array(ls_array1<T>&v, ls_UINT i) const;
                                  // Ausgabe des i-ten Feldes
  ls_array1M<T>& remove_array(ls_UINT i);
                                  // i-tes Feld entfernen
  ls_array1M<T>& remove();        // alle Felder entfernen
  const ls_array1M<T>& write(ostream &) const;
  ls_array1M<T>& read(istream &);
  const ls_array1M<T>& operator>> (char *) const;
  ls_array1M<T>& operator<< (char *);
};
```

```
#ifndef LS_LIB
#include "ls_array1M.cpp"
#endif
#endif
```

Ein Objekt dieser Klasse enthält `n` eindimensionale Datenfelder; jedes Einzelfeld hat eine eigene Dimension. Aus dieser Klasse wird eine Klasse `ls_bMatrix` abgeleitet, die symmetrische Bandmatrizen repräsentiert, wobei in den eindimensionalen Feldern die oberen Diagonalen abgespeichert sind. Beim Initialisieren ist die Anzahl der Diagonalen anzugeben. Als Operation ist hier nur die Operation Matrix mal Vektor hinzugefügt.

```
#ifndef LS_BMATRIX
#define LS_BMATRIX
#include "ls_Vector.h"
#include "ls_array1M.h"
class ls_bMatrix: public ls_array1M<ls_REAL>
{ public:
  ls_bMatrix(ls_UINT =0,char* ="ls_bMatrix");
  ls_bMatrix(ls_bMatrix &);
  operator ls_array1M<ls_REAL>&(){ return *this;}
  const ls_bMatrix& operator=(const ls_bMatrix &);//A=B
  ls_UINT number_of_diagonals() const{ return n;}
  ls_Vector operator* (ls_Vector &);              //A*x
  ls_UINT solve(ls_Vector &x, ls_Vector &b);      //cg-Verfahren
};
#ifndef LS_LIB
#include "ls_bMatrix.cpp"
#endif
#endif
```

Das Speicherabbild einer symmetrischen Bandmatrix mit Diagonalspeicherung hat eine Besonderheit: Falls in einer Diagonalen nur ein Element abgelegt ist, werden alle Elemente der Diagonalen als gleich diesem angesehen. Extremal kann so jede Diagonale durch jeweils ein Element repräsentiert sein. Damit belegen solche Matrizen minimalen Speicherplatz. In der Operation Matrix mal Vektor ist diese Möglichkeit entsprechend berücksichtigt.

Gleichungslöser

Die Spezifik einer Matrix findet in den Methoden zur Lösung eines linearen Gleichungssystems ihre Fortsetzung. Zunächst gehört zu jeder Matrix-Klasse als Methode das konjugierte Gradientenverfahren (`solve`). Beim Aufruf dieser Methode ist als 1. Parameter der Startvektor und als 2. Parameter die rechte Seite anzugeben; auf dem Startvektor findet man als Ausgabe die gefundene Lösung, auf der rechten Seite das Residuum. Im Falle einer allgemeinen Koeffizientenmatrix wird gegebenenfalls die Quadratmittellösung bestimmt.
Jede andere Methode verändert die Koeffizientenmatrix; daher entpricht ihr eine Klasse; alle Löser-Klassen haben einen einen einheitlichen Aufbau; sie enthalten insbesondere die Koeffizientenmatrix in unveränderter Form und die zugehörige Faktorisierung, gegebenenfalls mit Hilfsfeldern, damit eine Lösungsberechnung möglich ist. Betrachten wir z. B. die `QR`-Faktorisierung nach Householder:

```
#ifndef LS_QR
#define LS_QR
#include "ls_Matrix.h"
class ls_QR
{ protected:
  ls_Matrix A, F;                 // Matrix und QR-Faktorisierung
  ls_array1<ls_REAL> gamma, rho;  // Hilfsfelder
  char ONAME[ls_len];
  public:
  char *name, *A_name, *F_name, *gamma_name, *rho_name;
  ls_REAL eps;
  int rc;                           // Rueckkehrwert nach Faktorisierung
  ls_QR(char* ="ls_QR");
  ls_QR(ls_Matrix &, ls_REAL =0., char* ="ls_QR");
                                  // Matrix-Übernahme (move-Konstruktor)
                                  // Faktorisierung (mit Regularisierung)
  ls_QR(ls_QR &);
  ls_QR& swap(ls_QR &);
  unsigned char good()const{ return !rc;} // Erfolgssignal
  const ls_QR& operator=(const ls_QR &);
  const ls_QR& solve(ls_Vector &x, const ls_Vector &b) const;
                                  // Gleichungslöser
  ls_UINT post_iteration(ls_Vector &x, ls_Vector &b) const;
```

```
                                  // Nachiteration
  ls_Vector residuum(const ls_Vector &x, const ls_Vector &b) const;
  ls_Vector Qx(const ls_Vector &x) const; // Q*x
  ls_Vector xQ(const ls_Vector &x) const; // x*Q
  ls_Vector Rx(const ls_Vector &x) const; // R*x
  ls_Vector xR(const ls_Vector &x) const; // x*R
  const ls_QR& write_row(ostream &) const;
  ls_QR& read_row(istream &);
  const ls_QR& write_column(ostream &) const;
  ls_QR& read_column(istream &);
  const ls_QR& operator>> (char *) const;
  ls_QR& operator<< (char *);
};
#ifndef LS_LIB
#include "ls_QR.cpp"
#endif
#endif
```

Um einen Zugriff auf die Objektnamen der eingebetteten Objekte zu ermöglichen, gibt es die entsprechenden Zeiger: `A_name` enthält eine Zeiger auf den Objektnamen des eingebetteten Objektes `A` usw. . Bei der Anweisung `ls_QR C(B);` mit einer Matrix `B` wird der `move`-Konstruktor aufgerufen und die Faktorisierung ausgeführt. Nach erfolgreicher Deklaration, was mittels der `good`-Funktion überprüft werden kann, ruft man die Funktion `solve` mit zwei Vektoren auf, wobei auf dem ersten die gefundene Lösung abgelegt wird und auf dem zweiten die rechte Seite für das entsprechende Gleichungssystem zu übergeben ist. Im Falle einer quadratische Koeffizientenmatrix wird die Lösung des Gleichungssystems, bei einem überbestimmten Gleichungssystem die Quadratmittel-Lösung bestimmt.
Bei der Anweisung `ls_QR C(B,s);` mit einer kleinen, positiven, reellen Zahl `s` wird eine Regularisierung angewendet. Diese Vorgehensweise empfiehlt sich sehr bei Gleichungssystemen mit schlechtkonditionierter Koeffizientenmatrix.
Gegebenenfalls darf eine Nachiteration durchgeführt werden:

`post_iteration(ls_Vector &x, ls_Vector &b)`.

Dabei ist auf `x` die aktuelle Lösung und auf `b` die rechte Seite zu übergeben. Als Ergebnis erhält man die nachiterierte Lösung und das Residuum; der Rückkehrwert liefert die Anzahl der ausgeführten Iterationen. Bei der Nachiteration ist zu berücksichtigen, daß versucht wird, die gefundene Rechnerlösung der im Rechner befindlichen Aufgabe anzupassen. Hat die Koeffizientenmatrix eine große Kondition, wird eine Nachiteration wenig erfolgreich sein. Meist wird eine schlechtkonditionierte Aufgabe bereits bei der Zerlegungsberechnung dadurch erkannt, daß die Berechnung abbricht, da die Matrix numerisch singulär ist. Es ist dringend empfohlen, nach der Zerlegungsberechnung der Rückkehrwert zu testen.
Für weitere Anwendungen mit quadratischen Faktorisierungen benötigt man oft bei gegebener `QR`-Faktorisierung die Operationen `Q*x`, `x*Q`, `R*x` und `x*R`; daher sind die entsprechenden Funktionen hinzugefügt worden.
Über die Lese-Schreib-Funktionen kann man die Faktorisierung (und die Matrix) retten, um zu einem späteren Zeitpunkt weitere Lösungen zu berechnen. Alle anderen Klassen für Gleichungslöser arbeiten nach dem gleichen Muster und enthalten analoge Funktionen.
Die `LU`-Faktorisierung für eine quadratische Matrix gibt es in 5 Varianten:

- ohne Pivotisierung (`ls_LU`),
- Spalten-Pivotisierung (`ls_LU_column`),
- Zeilen-Pivotisierung (`ls_LU_row`),
- Diagonal-Pivotisierung (`ls_LU_diagonal`).
- Total-Pivotisierung (`ls_LU_total`).

Die verschiedenen Pivotisierungen verwenden dabei eine fiktive Skalierung.
Beispielhaft sei die Klasse `ls_LU_column` notiert.

```
#ifndef LS_LU_COLUMN
#define LS_LU_COLUMN
#include "ls_Matrix.h"
class ls_LU_column
{ protected:
  ls_Matrix A, F;
  ls_array1<ls_UINT> ind;
  char ONAME[ls_len];
  public:
```

```
  char *name, *A_name, *F_name, *ind_name;
  ls_REAL eps;                    // Genauigkeitsschranke
  int rc;                           // Rueckkehrwert nach Faktorisierung
  ls_LU_column(char* ="ls_LU_column");
  ls_LU_column(ls_Matrix &, ls_REAL =0., char* ="ls_LU_column");
                                  // Matrix-Übernahme (move-Konstruktor)
                                  // LU-Faktor. mit Spalten-Pivotisierung
  unsigned char good()const{ return !rc;}// Erfolgssignal
  ls_LU_column(ls_LU_column &);    // move-Konstruktor
  ls_LU_column& swap(ls_LU_column &);
  const ls_LU_column& operator=(const ls_LU_column &);
  const ls_LU_column& solve(ls_Vector &x, const ls_Vector &b) const;
                                  // Gleichungslöser
  ls_UINT post_iteration(ls_Vector &, ls_Vector &) const;
                                  // Nachiteration
  ls_Vector residuum(const ls_Vector &x, const ls_Vector &b) const;
  ls_Vector Lx(const ls_Vector &x) const; // L*x
  ls_Vector xL(const ls_Vector &x) const; // x*L
  ls_Vector Ux(const ls_Vector &x) const; // U*x
  ls_Vector xU(const ls_Vector &x) const; // x*U
  const ls_LU_column& write_row(ostream &) const;
  ls_LU_column& read_row(istream &);
  const ls_LU_column& write_column(ostream &) const;
  ls_LU_column& read_column(istream &);
  const ls_LU_column& operator>> (char *) const;
  ls_LU_column& operator<< (char *);
};
#ifndef LS_LIB
#include "ls_LU_column.cpp"
#endif
#endif
```

Man erkennt die Analogie zur Klasse `ls_QR`.
Daneben gibt es mit den gleichen Pivotisierungsvarianten die Invertierung einer Matrix (`ls_INV`).
Ein kleines Beispielprogramm möge die Anwendung illustrieren.

```
//: $CC -o beispiel1 beispiel1.cpp -lm
//#define ls_REAL float
#include "ls_QR.h"
int main()
{ ls_UINT n=8; ls_REAL s=0;
try
  {
    ls_Vector x(n,"Loesung"), b(n,"rechte_Seite");
    ls_Matrix A(n,n,"Koeffizienten-Matrix");
    for(ls_UINT j, i=0; i < n; i++)
      for(j=i; j < n; j++) A[i][j] = A[j][i]=1./(i+j+1.); // Hilbert-Matrix
    x.put_array(1.);          // alle Komponenten von x gleich 1
    b = A*x;                  // rechte Seite ist die Zeilensumme
    b >> "";                  // Ausgabe auf Standard-Ausgabe-Medium
    x.put_array(0.);          // alle Komponenten von x sind gleich 0
    ls_QR AA(A,s);            // QR-Faktorisierung mit Regularisierung
    if(AA.good())             // normal weiter, falls erfolgreich
    { AA.solve(x,b);          // Gleichungssystem loesen; alle gleich 1.
      x >> "";                // Loesung anschauen
      AA.residuum(x,b) >> "";// Residuum anschauen
    }
    else cerr << "method indicates singular." << endl;
  }
  catch(...){};               // erkannte Fehler auffangen.
  cin >> n;
  return 0;
}
```

Es sei erwähnt, daß selbst beim Lösen eines Systems mit einer `(500,500)`-Hilbertmatrix mittels Regularisierung eine Lösung berechnet wird, die `5` erste genaue Ziffern hat.
Zum Gleichungslösen von Systemen mit symmetrischer Koeffizienten-Matrix dienen die Klassen

- `ls_LDLT`: LDLT-Faktorisierung ohne Pivotisierung,
- `ls_LDLT_diagonal`: LDLT-Faktorisierung mit Pivotisierung entlang der Hauptdiagonalen.

```
#ifndef LS_LDLT_DIAGONAL
#define LS_LDLT_DIAGONAL
#include "ls_sMatrix.h"
class ls_LDLT_diagonal
{ protected:
  ls_sMatrix A, F;                // Matrix und Faktorisierung
  ls_array1<ls_UINT> ind;         // wahre Diagonal-Indices
  char ONAME[ls_len];
  public:
  char *name, *A_name, *F_name, *ind_name;
  ls_REAL eps;
  int rc;                           // Rueckkehrwert nach Faktorisierung
  ls_LDLT_diagonal(char* ="ls_LDLT_diagonal");
  ls_LDLT_diagonal(ls_sMatrix &, ls_REAL =0., char* ="ls_LDLT_diagonal");
                                  // Matrix-Übernahme (move-Konstruktor)
                                  // LDLT-Faktor. mit Diagonal-Pivot.
  unsigned char good()const{ return !rc;} // Erfolgssignal
  ls_LDLT_diagonal(ls_LDLT_diagonal &);
  ls_LDLT_diagonal& swap(ls_LDLT_diagonal &);
  const ls_LDLT_diagonal& operator=(const ls_LDLT_diagonal&);
  const ls_LDLT_diagonal& solve(ls_Vector &x, const ls_Vector &b) const;
                                  // Gleichungslöser
  ls_UINT post_iteration(ls_Vector &, ls_Vector &) const;
                                  // Nachiteration
  ls_Vector residuum(const ls_Vector &x, const ls_Vector &b) const;
  ls_Vector Lx(const ls_Vector &x) const; // L*x (L: Cholesky-Faktor)
  ls_Vector xL(const ls_Vector &x) const; // x*L (L: Cholesky-Faktor)
  const ls_LDLT_diagonal& write_row(ostream &) const;
  ls_LDLT_diagonal& read_row(istream &);
  const ls_LDLT_diagonal& write_column(ostream &) const;
  ls_LDLT_diagonal& read_column(istream &);
  const ls_LDLT_diagonal& operator>> (char *) const;
  ls_LDLT_diagonal& operator<< (char *);
};
#ifndef LS_LIB
#include "ls_LDLT_diagonal.cpp"
#endif
#endif
```

Das obige Beispiel läßt sich sehr leicht zu einem Beispiel für die LDLT-Faktorisierung machen.

```
//: $CC -o beispiel2 beispiel2.cpp -lm
//#define ls_REAL float
#include "ls_LDLT_diagonal.h"
int main()
{ ls_UINT n=6, i, j; ls_REAL s=1.e-10;
try
  {
    ls_Vector x(n,"Loesung"), b(n,"rechte_Seite");
    ls_sMatrix A(n,"Koeffizienten-Matrix");
    for(i=0; i < n; i++)
      for(j=i; j < n; j++) A[i][j] = 1./(i+j+1.); // Hilbert-Matrix
/*
    ls_Matrix AA(n,n);
    for(i=0; i < n; i++)
      for(j=i; j < n; j++) AA[i][j] = AA[j][i] = A[i][j];
    AA>>"";
    A.swap(1,2);
    for(i=0; i < n; i++)
      for(j=i; j < n; j++) AA[i][j] = AA[j][i] = A[i][j];
    AA>>"";
*/
    x.put_array(1.);         // alle Komponenten von x gleich 1
    b = A*x;                 // rechte Seite ist die Zeilensumme
    b >> "";                 // Ausgabe auf Standard-Ausgabe-Medium
    x.put_array(0.);         // alle Komponenten von x sind gleich 0
    ls_LDLT_diagonal AA(A,s);// LDLT-Faktorisierung mit Diagonalpivotisierung
    if(AA.good())            // normal weiter, falls erfolgreich
    { AA.solve(x,b);         // Gleichungssystem loesen; alle gleich 1.
      x >> "";               // Loesung anschauen
      AA.residuum(x,b) >> "";// Residuum anschauen
```

```
    }
    else cerr << "method indicates singular." << endl;
  }
  catch(...){};              // erkannte Fehler auffangen.
  cin >> n;
  return 0;
}
```

Die folgende Tabelle belegt, in welcher Weise bei den obigen Gleichungslösern die Regularisierung wirkt, sofern man mit reellen Zahlen vom Typ `double` rechnet. Dazu wurde für wachsende Ordnung `n` der Koeffizienten-Hilbertmatrix (`n=6, 10, 50, 100, 500`) das entsprechende Gleichungssystem mit den Regularisierungsparametern

`s=0., 1.e-7, 1.e-8, 1.e-9, 1.e-10, 1.e-11, 1.e-12`

in der Methoden-Reihenfolge

`LU (1)`, `LU_column (2)`, `LU_row (3)`, `LU_diagonal (4)`, `LU_total (5)`, `QR (6)`, `LDLT (7)`, `LDLT_diagomal (8)`

gelöst. Gemessen und ausgegeben wurde der relative Fehler.

	0.00e+00	1.00e-07	1.00e-08	1.00e-09	1.00e-10	1.00e-11	1.00e-12	1.00e-13
n=6								
(1)	7.45e-11	1.87e-04	3.25e-05	3.52e-06	3.54e-07	3.56e-08	3.55e-09	1.97e-10
(2)	4.64e-10	1.87e-04	3.25e-05	3.52e-06	3.54e-07	3.57e-08	3.16e-09	4.19e-10
(3)	2.42e-10	1.87e-04	3.25e-05	3.52e-06	3.55e-07	3.56e-08	2.98e-09	5.07e-10
(4)	3.91e-10	1.87e-04	3.25e-05	3.52e-06	3.55e-07	3.53e-08	3.51e-09	9.18e-11
(5)	1.90e-10	1.87e-04	3.25e-05	3.52e-06	3.55e-07	3.51e-08	3.19e-09	4.63e-10
(6)	8.07e-10	1.87e-04	3.25e-05	3.52e-06	3.55e-07	3.60e-08	2.42e-09	3.93e-10
(7)	4.25e-10	1.87e-04	3.25e-05	3.52e-06	3.55e-07	3.47e-08	9.51e-10	1.53e-09
(8)	3.80e-10	1.87e-04	3.25e-05	3.52e-06	3.55e-07	3.52e-08	2.75e-09	4.36e-10
(9)	1.10e-09	1.87e-04	3.25e-05	3.52e-06	3.55e-07	3.42e-08	2.02e-09	3.81e-10
(10)	1.10e-09	1.87e-04	3.25e-05	3.52e-06	3.55e-07	3.42e-08	2.02e-09	3.81e-10
(11)	4.11e-10	1.87e-04	3.25e-05	3.52e-06	3.55e-07	3.53e-08	3.19e-09	1.54e-10
(12)	2.63e-10	1.87e-04	3.25e-05	3.52e-06	3.54e-07	3.53e-08	3.46e-09	1.18e-10
(13)	2.60e-10	1.87e-04	3.25e-05	3.52e-06	3.55e-07	3.53e-08	3.45e-09	1.58e-10
n=10								
(1)	6.30e-04	1.82e-04	5.13e-05	1.74e-05	4.68e-06	4.06e-06	5.73e-05	1.76e-04
(2)	4.12e-04	1.82e-04	5.13e-05	1.74e-05	5.50e-06	4.51e-06	4.28e-05	1.51e-04
(3)	4.41e-04	1.82e-04	5.13e-05	1.74e-05	5.16e-06	3.32e-06	4.97e-05	2.44e-05
(4)	8.01e-06	1.82e-04	5.13e-05	1.74e-05	5.15e-06	5.02e-06	6.11e-05	1.70e-04
(5)	2.00e-04	1.82e-04	5.13e-05	1.74e-05	4.74e-06	4.92e-06	2.17e-05	3.83e-04
(6)	6.10e-03	1.82e-04	5.13e-05	1.77e-05	4.76e-06	4.48e-05	7.56e-05	3.80e-03
(7)	9.78e-03	1.82e-04	5.14e-05	1.74e-05	6.02e-06	2.66e-05	8.62e-04	2.71e-03
(8)	4.44e-04	1.82e-04	5.13e-05	1.74e-05	5.22e-06	1.71e-05	6.69e-05	4.78e-04
(9)	2.22e-04	1.82e-04	5.13e-05	1.74e-05	5.43e-06	2.78e-06	6.77e-05	2.35e-04
(10)	2.22e-04	1.82e-04	5.13e-05	1.74e-05	5.43e-06	2.78e-06	6.77e-05	2.35e-04
(11)	3.38e-04	1.82e-04	5.13e-05	1.74e-05	5.22e-06	2.79e-06	2.98e-05	1.45e-04
(12)	2.08e-04	1.82e-04	5.13e-05	1.74e-05	5.18e-06	4.16e-06	2.65e-05	2.38e-04
(13)	1.22e-04	1.82e-04	5.13e-05	1.74e-05	5.15e-06	4.96e-06	2.64e-05	2.18e-04
n=50								
(1)	5.24e+02	1.79e-04	5.67e-05	1.80e-05	6.63e-06	2.76e-05	2.61e-04	2.09e-03
(2)	1.01e+02	1.79e-04	5.67e-05	1.80e-05	6.05e-06	1.87e-05	2.23e-04	1.65e-03
(3)	9.72e+02	1.79e-04	5.67e-05	1.80e-05	5.96e-06	2.11e-05	2.35e-04	2.00e-03
(4)	1.05e+03	1.79e-04	5.67e-05	1.79e-05	6.38e-06	2.61e-05	2.17e-04	2.20e-03
(5)	1.56e+02	1.79e-04	5.67e-05	1.80e-05	6.21e-06	2.38e-05	1.50e-04	2.38e-03
(6)	2.10e+09	1.79e-04	5.68e-05	8.69e-05	1.80e-03	3.85e-02	9.59e-01	1.63e+01
(7)	5.77e+09	1.79e-04	9.10e-05	1.83e-03	7.20e-02	2.30e+00	5.81e+01	1.42e+03
(8)	3.99e+03	1.79e-04	5.67e-05	1.80e-05	9.99e-06	8.48e-05	1.08e-03	7.15e-03
(9)	1.50e+03	1.79e-04	5.67e-05	1.80e-05	9.23e-06	6.07e-05	5.33e-04	6.19e-03
(10)	9.59e+16	1.79e-04	5.67e-05	1.80e-05	9.23e-06	6.07e-05	5.33e-04	6.19e-03
(11)	1.23e+02	1.79e-04	5.67e-05	1.80e-05	6.02e-06	1.44e-05	1.44e-04	1.36e-03
(12)	8.25e+01	1.79e-04	5.67e-05	1.80e-05	5.87e-06	1.49e-05	1.53e-04	1.22e-03
(13)	singular	1.79e-04	5.67e-05	1.80e-05	5.96e-06	1.42e-05	1.49e-04	1.24e-03
n=100								
(1)	singular	1.79e-04	5.64e-05	1.79e-05	6.98e-06	3.12e-05	3.48e-04	3.47e-03
(2)	1.59e+02	1.79e-04	5.64e-05	1.79e-05	6.49e-06	2.35e-05	2.35e-04	2.30e-03
(3)	7.06e+02	1.79e-04	5.64e-05	1.79e-05	6.49e-06	3.14e-05	3.49e-04	2.49e-03
(4)	2.43e+02	1.79e-04	5.64e-05	1.79e-05	6.91e-06	3.65e-05	3.20e-04	3.45e-03
(5)	9.26e+02	1.79e-04	5.64e-05	1.79e-05	7.53e-06	3.44e-05	3.45e-04	4.25e-03
(6)	singular	1.79e-04	5.66e-05	1.58e-04	3.94e-03	1.03e-01	2.09e+00	8.09e+01
(7)	5.30e+11	1.78e-04	3.40e-04	8.98e-03	2.24e-01	1.46e+01	1.34e+03	3.27e+04
(8)	2.25e+03	1.79e-04	5.64e-05	1.77e-05	1.63e-05	1.35e-04	1.29e-03	1.60e-02
(9)	singular	1.79e-04	5.64e-05	1.79e-05	1.05e-05	8.38e-05	9.38e-04	8.20e-03
(10)	singular	1.79e-04	5.64e-05	1.79e-05	1.05e-05	8.38e-05	9.38e-04	8.20e-03
(11)	1.27e+02	1.79e-04	5.64e-05	1.79e-05	6.09e-06	1.95e-05	1.71e-04	1.61e-03
(12)	8.94e+03	1.79e-04	5.64e-05	1.79e-05	6.15e-06	1.92e-05	1.85e-04	1.68e-03
(13)	singular	1.79e-04	5.64e-05	1.79e-05	6.19e-06	1.97e-05	1.93e-04	1.68e-03

n=500								
(1)	singular	1.79e-04	5.65e-05	1.79e-05	1.04e-05	8.22e-05	8.14e-04	8.00e-03
(2)	singular	1.79e-04	5.65e-05	1.79e-05	9.02e-06	5.84e-05	6.54e-04	6.28e-03
(3)	singular	1.79e-04	5.65e-05	1.79e-05	9.29e-06	6.43e-05	6.09e-04	6.35e-03
(4)	singular	1.79e-04	5.65e-05	1.79e-05	9.93e-06	7.56e-05	7.39e-04	8.52e-03
(5)	singular	1.79e-04	5.65e-05	1.79e-05	1.19e-05	9.78e-05	9.83e-04	9.88e-03
(6)	singular	1.79e-04	5.70e-05	2.31e-04	6.94e-03	1.92e-01	5.73e+00	1.67e+02
(7)	singular	1.79e-04	1.08e-03	4.17e-02	2.08e+00	1.69e+02	1.32e+04	6.86e+05
(8)	6.91e+04	1.79e-04	5.65e-05	2.21e-05	8.72e-05	6.91e-04	9.62e-03	5.84e-02
(9)	singular	1.79e-04	5.65e-05	1.80e-05	2.57e-05	2.54e-04	2.00e-03	2.52e-02
(10)	singular	1.79e-04	5.65e-05	1.80e-05	2.57e-05	2.54e-04	2.00e-03	2.52e-02
(11)	singular	1.79e-04	5.65e-05	1.79e-05	8.15e-06	4.49e-05	3.81e-04	3.98e-03
(12)	singular	1.79e-04	5.65e-05	1.79e-05	7.61e-06	4.04e-05	3.89e-04	3.68e-03
(13)	singular	1.79e-04	5.65e-05	1.79e-05	7.43e-06	3.92e-05	3.74e-04	3.69e-03

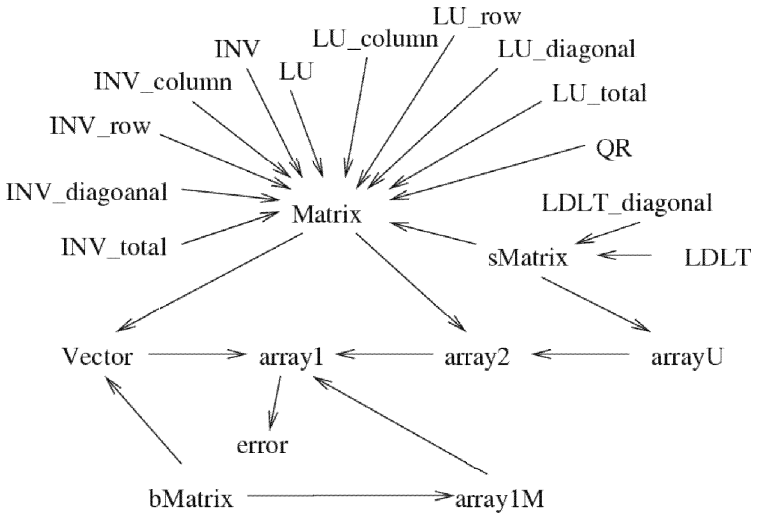
Für dieses Beispiel belegt die Tabelle, daß die Regularisierung wesentlich stärker die Genauigkeit erhöht als die Pivotisierung. Selbst beim Rechnen mit Zahlen vom Typ `float` liefert die Regularisierung noch brauchbare Ergbenisse, wie die folgende Tabelle zeigt.

	0.00e+00	1.00e-03	1.00e-04	1.00e-05	1.00e-06	1.00e-07	1.00e-08	1.00e-09
n=6								
(1)	2.39e-02	1.82e-02	5.16e-03	3.54e-03	1.68e-02	1.27e-01	1.64e-02	2.39e-02
(2)	1.36e-01	1.82e-02	5.19e-03	2.91e-03	2.30e-02	4.03e-02	2.36e-01	1.36e-01
(3)	1.15e-01	1.82e-02	5.09e-03	2.93e-03	7.60e-03	5.28e-02	2.40e-01	1.15e-01
(4)	2.25e-02	1.82e-02	5.17e-03	3.35e-03	2.30e-03	4.28e-02	1.07e-01	2.25e-02
(5)	9.20e-03	1.82e-02	5.17e-03	2.57e-03	5.49e-03	3.50e-02	3.18e-01	9.20e-03
(6)	1.10e-01	1.82e-02	5.01e-03	4.13e-03	6.58e-02	5.87e-02	9.64e-01	1.10e-01
(7)	2.04e-01	1.82e-02	5.34e-03	9.47e-03	2.93e-02	2.98e-01	2.17e-01	2.04e-01
(8)	2.89e-01	1.82e-02	5.49e-03	3.21e-03	6.15e-02	2.17e-01	0.00e+00	2.89e-01
(9)	2.93e-01	1.82e-02	5.34e-03	1.83e-03	1.28e-02	1.02e-01	2.99e-01	2.93e-01
(10)	2.93e-01	1.82e-02	5.34e-03	1.83e-03	1.28e-02	1.02e-01	2.99e-01	2.93e-01
(11)	3.60e-02	1.82e-02	5.07e-03	1.94e-03	1.07e-02	2.71e-02	2.16e-02	3.60e-02
(12)	2.44e-02	1.82e-02	5.12e-03	2.55e-03	4.88e-03	1.85e-02	3.82e-02	2.44e-02
(13)	6.17e-02	1.82e-02	5.10e-03	2.60e-03	4.78e-03	4.38e-02	6.67e-02	6.17e-02
n=10								
(1)	1.80e+01	1.77e-02	5.84e-03	4.05e-03	3.12e-02	1.77e-01	1.72e+00	1.80e+01
(2)	2.71e+01	1.77e-02	5.79e-03	7.03e-03	5.97e-02	4.25e-01	9.06e+00	2.71e+01
(3)	1.48e+01	1.77e-02	5.97e-03	5.36e-03	4.32e-02	2.21e-01	4.28e+00	1.48e+01
(4)	8.43e+00	1.77e-02	5.93e-03	4.48e-03	2.91e-02	6.48e-01	3.59e+00	8.43e+00
(5)	2.44e+01	1.77e-02	5.84e-03	4.46e-03	2.86e-02	2.10e-01	8.08e-01	2.44e+01
(6)	1.05e+02	1.77e-02	5.52e-03	1.36e-02	1.11e-01	2.11e+00	2.02e+01	1.05e+02
(7)	2.05e+02	1.78e-02	6.07e-03	8.40e-03	3.02e-01	1.38e+00	2.49e+01	2.05e+02
(8)	1.72e+02	1.77e-02	5.82e-03	5.28e-03	8.21e-02	6.18e-01	8.63e+00	1.72e+02
(9)	1.03e+02	1.77e-02	5.94e-03	7.57e-03	5.04e-02	9.91e-01	6.86e+00	1.03e+02
(10)	1.03e+02	1.77e-02	5.94e-03	7.57e-03	5.04e-02	9.91e-01	6.86e+00	1.03e+02
(11)	8.98e+01	1.77e-02	5.78e-03	2.90e-03	2.14e-02	1.15e-01	1.70e+00	8.98e+01
(12)	5.19e+01	1.77e-02	5.81e-03	2.79e-03	2.44e-02	6.83e-02	3.02e+00	5.19e+01
(13)	8.92e+00	1.77e-02	5.89e-03	3.33e-03	1.35e-02	1.44e-01	1.89e+00	8.92e+00
n=50								
(1)	8.41e+01	1.80e-02	5.72e-03	1.23e-02	1.05e-01	1.21e+00	4.64e+01	8.05e+01
(2)	8.00e+01	1.80e-02	5.85e-03	9.32e-03	1.16e-01	8.51e-01	4.04e+02	5.15e+02
(3)	2.41e+02	1.80e-02	5.84e-03	1.11e-02	7.71e-02	7.60e-01	2.26e+02	9.20e+02
(4)	2.54e+02	1.80e-02	5.89e-03	1.10e-02	1.34e-01	1.24e+00	2.62e+01	1.15e+02
(5)	1.90e+02	1.80e-02	5.79e-03	1.77e-02	1.10e-01	1.31e+00	5.06e+01	2.36e+02
(6)	4.90e+05	1.80e-02	6.76e-03	9.28e-02	2.23e+00	5.40e+01	1.03e+04	2.99e+05
(7)	1.88e+06	1.80e-02	9.60e-03	2.79e-01	1.60e+01	3.78e+02	3.65e+04	3.50e+07
(8)	6.60e+02	1.80e-02	7.61e-03	3.81e-02	4.03e-01	5.52e+00	2.08e+02	2.21e+03
(9)	8.22e+02	1.80e-02	6.01e-03	3.14e-02	2.32e-01	3.44e+00	1.44e+02	1.06e+03
(10)	2.20e+08	1.80e-02	6.01e-03	3.14e-02	2.32e-01	3.44e+00	3.06e+07	5.28e+08
(11)	3.60e+01	1.80e-02	5.73e-03	4.13e-03	5.39e-02	4.06e-01	6.65e+00	1.47e+02
(12)	2.22e+02	1.80e-02	5.62e-03	4.08e-03	4.62e-02	4.72e-01	4.83e+01	6.15e+01
(13)	6.38e+01	1.80e-02	5.66e-03	3.76e-03	4.40e-02	5.69e-01	1.54e+01	8.47e+01
n=100								
(1)	7.97e+02	1.80e-02	5.97e-03	1.65e-02	1.88e-01	1.76e+00	1.11e+02	2.16e+02
(2)	2.71e+02	1.80e-02	5.80e-03	8.11e-03	8.73e-02	1.04e+00	4.11e+01	2.61e+02
(3)	2.60e+02	1.80e-02	5.77e-03	9.05e-03	1.06e-01	9.58e-01	1.21e+02	2.44e+02
(4)	2.49e+02	1.80e-02	6.11e-03	1.58e-02	1.87e-01	1.97e+00	9.61e+01	1.23e+03
(5)	2.44e+02	1.80e-02	6.03e-03	2.00e-02	2.91e-01	2.23e+00	1.61e+02	3.90e+02
(6)	1.68e+07	1.80e-02	7.51e-03	1.10e-01	3.00e+00	8.72e+01	7.35e+03	4.01e+06
(7)	9.14e+06	1.80e-02	1.21e-02	4.15e-01	4.45e+01	1.24e+03	6.32e+06	1.33e+07
(8)	2.22e+03	1.80e-02	7.93e-03	8.55e-02	7.88e-01	7.32e+00	5.85e+03	2.89e+03
(9)	5.91e+03	1.80e-02	6.99e-03	4.28e-02	4.67e-01	4.46e+00	5.05e+03	5.31e+03
(10)	5.20e+08	1.80e-02	6.99e-03	4.28e-02	4.67e-01	4.46e+00	7.05e+08	1.32e+09
(11)	3.36e+02	1.80e-02	5.65e-03	1.07e-02	4.79e-02	5.58e-01	5.73e+00	8.18e+01
(12)	6.87e+01	1.80e-02	5.66e-03	5.75e-03	6.51e-02	5.94e-01	2.38e+01	9.97e+01

(13)	2.04e+02	1.80e-02	5.65e-03	6.30e-03	5.46e-02	5.76e-01	2.99e+01	1.60e+02
n=500								
(1)	6.04e+03	1.81e-02	6.69e-03	3.75e-02	3.87e-01	3.92e+00	2.50e+02	4.24e+03
(2)	2.46e+04	1.80e-02	6.84e-03	2.16e-02	3.03e-01	3.24e+00	5.32e+01	4.09e+03
(3)	2.16e+04	1.80e-02	6.97e-03	2.06e-02	2.77e-01	2.85e+00	3.58e+02	2.83e+03
(4)	4.45e+03	1.80e-02	7.09e-03	6.39e-02	9.04e-01	6.16e+00	1.79e+02	8.19e+03
(5)	9.30e+03	1.80e-02	9.19e-03	6.06e-02	2.23e+00	1.34e+01	2.59e+02	3.81e+03
(6)	1.45e+08	1.80e-02	1.42e-02	1.76e-01	4.35e+00	1.26e+02	6.15e+04	2.20e+07
(7)	1.06e+08	1.81e-02	2.06e-02	5.84e-01	6.58e+01	4.24e+03	8.68e+05	7.08e+08
(8)	5.17e+04	1.80e-02	1.77e-02	1.78e-01	2.74e+00	3.69e+01	1.66e+03	3.80e+05
(9)	2.37e+05	1.81e-02	1.55e-02	1.27e-01	1.26e+00	1.29e+01	1.42e+03	9.63e+04
(10)	4.19e+09	1.81e-02	1.55e-02	1.27e-01	1.26e+00	1.29e+01	2.98e+07	3.15e+09
(11)	9.97e+02	1.81e-02	6.83e-03	2.35e-02	2.65e-01	2.14e+00	8.13e+01	2.15e+03
(12)	1.10e+03	1.80e-02	5.79e-03	1.03e-02	9.43e-02	9.13e-01	4.23e+01	8.23e+02
(13)	7.29e+02	1.80e-02	5.73e-03	6.78e-03	6.59e-02	6.70e-01	2.88e+01	1.00e+03

Hinweise

Eine Klasse ist hier durch eine `h`-Datei und eine `cpp`-Datei repräsentiert. Die `h`-Datei enthält die Klassendefinition; in der `cpp`-Datei findet man die Implementation der Methoden. In der Anwendung der Klassen ist es nur erforderlich, die `h`-Datei der jeweils höchsten Klasse in ein Programm einzubinden, da die sonst noch benötigten Klassen nachgeladen werden. Die folgende Darstellung zeigt den Nachladegraphen.



Jeder Quell-Datei ist ein Identifikator zugeordnet: Der Identifikator für die Datei `ls_Vector.h` ist `LS_VECTOR`; für die Datei `ls_Vector.cpp` ist es `LS_VECTORC` usw.. Ein solcher Identifikator ist genau dann definiert, wenn die betreffende Datei geladen ist. Durch Abfragen des Identifikators vermeidet man, daß eine Datei doppelt geladen wird. Zu jeder `h`-Datei gehört eine `cpp`-Datei. Falls eine `h`-Datei geladen ist, wird an ihrem Ende die zugeordnete `cpp`-Datei geladen. Auf diese Weise wird einerseits erreicht, daß alle und nur die Dateien geladen sind, die zum Übersetzen benötigt werden; andererseits kann jede `cpp`-Datei separat übersetzt werden. Außerdem braucht das System beim Übersetzen eines Programms keine separat übersetzte `cpp`-Datei, da die benötigten Dateien während der Übersetzung eingebunden werden. Dies kann man als Vor- oder als Nachteil ansehen. Ein Nachteil ist, daß mit einer `h`-Datei stets auch die `cpp`-Datei geladen wird, was inhaltlich bedeutet, daß beide zu einer Datei vereinigt und die in einer Objektmodulbibliothek abgelegten Übersetzungen unbrauchbar sind. Der Identifikator `LS_LIB` behebt diesen Nachteil: Wenn er gesetzt ist, wird keine `cpp`-Datei geladen. Es ist sehr zu empfehlen, in einem solchen Falle für die `LS`-Klassen eine separate Objektmodulbibliothek zu erstellen.

Es sei noch erwähnt, daß ein sehr analoges System namens `SP` für schwachbesetzte Matrizen unter der gleichen Bezugsquelle verfügbar ist.

6.7. Nullstellen nichtlinearer Gleichungen

Es wird die Aufgabe betrachtet, zu einer auf einem Intervall $[a, b]$ definierten Funktion f einen Punkt $x^* \in [a, b]$ zu finden, der Nullstelle der Funktion f ist: $f(x^*) = 0$.
Zunächst soll untersucht werden, wie die Lösung der Aufgabe von den Eingabedaten abhängt. Die Eingabedaten bestehen hier aus der Funktion f. Es sei also $f + \overline{f}$ eine Funktion mit der Nullstelle $x^* + \overline{x}$; dann folgt in erster Näherung, falls die Funktion f ableitbar ist:

$$0 = f(x^* + \overline{x}) + \overline{f}(x^* + \overline{x}) = f(x^*) + f'(x^*)\overline{x} + \overline{f}(x^*),$$

also

$$\overline{x} \doteq -\frac{1}{f'(x^*)}\overline{f}(x^*).$$

Die Aufgabe ist also schlecht konditioniert, wenn $|f'(x^*)|$ sehr klein ausfällt. Im Falle $f'(x) = 0$ in der Nähe des Punktes x^* ist keine Abschätzung der Form $|\overline{x}| \leqq K|\overline{f}(x^*)|$ möglich, was eine extrem schlechte Kondition der Aufgabe bedeutet. Bei solchen Aufgaben werden daher alle Methoden mehr oder weniger schnell versagen.
Viele Verfahren zur Lösung der Nullstellen-Aufgabe laufen nach dem folgenden Schema ab: Es sei $x^{(0)}$ als Näherung von x^* bekannt; durch

$$x^{(r+1)} = \Phi\left(x^{(r)}\right), \quad r = 0, 1, \ldots$$

wird eine Folge $\left(x^{(r)}\right)$ erzeugt, die gegen x^* konvergiert. Abgesehen von der Bestimmung einer geeigneten Iterationsfunktion Φ muß der Algorithmus noch einen geeigneten Startpunkt $x^{(0)}$ ermitteln und die Iteration mit einer brauchbaren Näherung abbrechen.
In einer punktierten Umgebung $U(x^*)$ einer gesuchten, einfachen Nullstelle x^* gelte $f(x) \neq 0$. Wählt man als Iterationsfunktion

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

so erhält man das **Newton-Verfahren**:

$$x^{(r+1)} = x^{(r)} - \frac{f\left(x^{(r)}\right)}{f'\left(x^{(r)}\right)},$$

wofür der folgende Konvergenzsatz gilt.

Satz 164. *Die zweimal stetig differenzierbare Funktion f habe im Intervall $[a, b]$ eine Nullstelle x^*; es mögen Konstanten m, M mit*

$$|f'(x)| \geqq m > 0, \quad |f''(x)| \leqq M \quad \forall x \in [a, b]$$

geben; ferner gelte für $x^{(r)} \in [a, b]$ auch $x^{(r+1)} \in [a, b]$. Dann gilt für die durch das Newton-Verfahren erzeugte Punktfolge:

$$|x^{(r+1)} - x^*| \leqq \frac{M}{2m}|x^{(r)} - x^*|^2.$$

Beweis. Mit dem Mittelwertsatz der Differentialrechnung folgt für ein gewisses $\xi^{(r)} \in [x^{(r)}, x^*]$:

$$0 = f(x^*) = f(x^{(r)}) + f'(x^{(r)})(x^* - x^{(r)}) + \frac{1}{2}f''(\xi^{(r)})(x^* - x^{(r)})^2.$$

Wegen

$$x^{(r)} = x^{(r+1)} + \frac{f\left(x^{(r)}\right)}{f'\left(x^{(r)}\right)}$$

folgt weiter

$$0 = f(x^{(r)}) + f'(x^{(r)})\left(x^* - x^{(r+1)} - \frac{f\left(x^{(r)}\right)}{f'\left(x^{(r)}\right)}\right) + \frac{1}{2}f''(\xi^{(r)})(x^* - x^{(r)})^2$$

und nach Umstellen:

$$x^{(r+1)} - x^* = \frac{1}{2}\frac{f''(\xi^{(r)})}{f'\left(x^{(r)}\right)}(x^* - x^{(r)})^2.$$

Die Voraussetzungen an die Funktion f sind so gewählt, daß man mit ihnen sofort die behauptete Abschätzung erhält. □
Nach diesem Satz konvergiert das Newton-Verfahren quadratisch, falls es überhaupt konvergieren sollte. Die Konvergenz des Newton-Verfahrens ist stets nur eine lokale, da gefordert wird,daß die erzeugten Punkte das gegebene Intervall nicht verlassen dürfen; ein Intervall mit dieser Eigenschaft ist aber oft sehr klein.
Beispiel. Für die Quadratwurzelgleichung $x^2 - a = 0 \quad (a > 0)$ lautet das Newton-Verfahren

$$x^{(r+1)} = \frac{1}{2}\left(x^{(r)} + \frac{a}{x^{(r)}}\right).$$

Als Startpunkt sollte man

$$x^{(0)} = \frac{1+a}{2} \geqq \sqrt{a}$$

wählen. Dann folgt für alle r: $x^{(r)} \geqq \sqrt{a}$ und der Satz liefert die Abschätzung

$$x^{(r+1)} - \sqrt{a} \leqq \frac{1}{2\sqrt{a}} \left(x^{(r)} - \sqrt{a}\right)^2$$

oder

$$\frac{x^{(r+1)} - \sqrt{a}}{\sqrt{a}} \leqq \frac{1}{2} \left(\frac{x^{(r)} - \sqrt{a}}{\sqrt{a}}\right)^2,$$

d. h. die Anzahl der richtigen Stellen verdoppelt sich mit jedem Schritt.
Sollte die Auswertung der Funktion f' zu aufwendig sein, darf man eine Konstante $m \approx f'(x^*)$ verwenden und erhält das vereinfachte Newton-Verfahren:

$$x^{(r+1)} = x^{(r)} - \frac{f'(x^{(r)})}{m},$$

das noch linear konvergiert, d. h.

$$|x^{(r+1)} - x^*| \leqq q|x^{(r)} - x^*|$$

mit

$$q = \max_x \left|1 - \frac{f'(x)}{m}\right|.$$

Damit Konvergenz gesichert ist muß $q < 1$ sein, was dann garantiert ist, wenn man die Zahl m so wählt, daß

$$\max_x \left|1 - \frac{f'(x)}{m}\right| < 1$$

ausfällt.
Anstelle der Ableitung f' darf man auch einen Differenzenquotienten benutzen:

$$f'(x^{(r)}) \approx \frac{f(x^{(r)}) - f(x^{(r-1)})}{x^{(r)} - x^{(r-1)}},$$

wodurch das Newton-Verfahren in die **Regula falsi** übergeht:

$$x^{(r+2)} = x^{(r+1)} - \frac{f(x^{(r+1)})}{\dfrac{f(x^{(r+1)}) - f(x^{(r)})}{x^{(r+1)} - x^{(r)}}}.$$

Unter geeigneten Voraussetzungen konvergiert dieses Verfahren mit der Ordnung

$$q = \frac{1+\sqrt{5}}{2} \approx 1.618$$

und

$$|x^{(r+1)} - x^*| \leqq \left(\frac{M}{2m}\right)^{q-1} |x^{(r)} - x^*|^q.$$

Trotz der Tatsache, daß die Regula falsi etwas schlechter konvergiert als das Newton-Verfahren, kann es diesem überlegen sein, insbesondere dann, wenn die Berechnung von Ableitungen erheblich aufwendiger ist als die Berechnung von Funktionswerten.
Weitere Iterationsfunktionen Φ kann man nach folgendem Prinzip gewinnen: Man setze

$$\Phi(x) = x - q(x) \cdot f(x)$$

mit einem **Relaxationsparameter** $q(x)$. Die Funktion q ist hier so zu wählen, daß Φ eine kontrahierende Abbildung des Intervalls $[a, b]$ in sich wird, wobei der Kontraktionsparameter möglichst klein sein sollte. Beim Newton-Verfahren lautet der Relaxationsparamter

$$q(x) = \frac{1}{f'(x)}$$

und beim vereinfachten Newton-Verfahren

$$q(x) = \frac{1}{m}.$$

Ein vollständiger Algorithmus zur Lösung einer Nullstellenaufgabe muß noch ein geeignetes Anfangsintervall ermitteln, in dem die Konvergenz der verwendeten Methode eintritt. Dazu eignet sich etwa das im Teil Analysis besprochene Bisektionsverfahren.
Kann die Funktion f nur mit einem absoluten Fehler $\varepsilon > 0$ berechnet werden, so erhält man mit jedem Verfahren nur eine Näherung $\overline{x}^*$ mit

$$|\overline{x}^* - x^*| \leqq \frac{\varepsilon}{|f'(x^*)|},$$

wodurch die erzielbare Genauigkeit unmittelbar begrenzt wird. Das Rechnen mit erhöhter Genauigkeit im Algorithmus ist nur dann sinnvoll, wenn auch die Funktionswerte mit entsprechender Genauigkeit vorliegen.

6.8. Übungen

1. Welche Grundgesetze der Arithmetik sind bei Rechneroperationen i.a. nicht mehr gültig (Begründung) ?

2. Man ermittle die Fehlerfortpflanzungsformel für die Grundoperationen $(+, -, \cdot, /)$.Die Werte c_1 und c_2 sind derart zu bestimmen, daß
   $\varepsilon^z \approx c_1\varepsilon_x + c_2\varepsilon_y$ für $x \approx \tilde{x}$ und $y \approx \tilde{y}$ gilt, wobei

   $$z = x \cdot y,\ \tilde{z} = \tilde{x} \cdot \tilde{y},\ \varepsilon_z = \frac{\tilde{z} - z}{z},\ \varepsilon_y = \frac{\tilde{y} - y}{y},\ \varepsilon_x = \frac{\tilde{x} - x}{x}$$

   sind.

3. Man forme die folgenden Ausdrücke so um, daß ihre Auswertung möglichst ohne Auslöschung vorgenommen werden kann:

   (a)

   $$\frac{1}{2x+1} - \frac{1-x}{1+x},$$

   (b)

   $$\frac{1-\cos x}{x},$$

   (c)

   $$\sqrt{x + \frac{1}{x}} - \sqrt{x - \frac{1}{x}}.$$

4. Es werden die Folgen

   $$e_n(x) = \sum_{i=0}^{n} \frac{x^i}{i!}, \quad f_n(x) = \left(1 + \frac{x}{n}\right)^n,$$

   $$g_n(x) = \left(1 + \frac{x}{n}\right)^{n+1}, \quad h_n(x) = \frac{1}{2}\left(f_n(x) + g_n(x)\right)$$

   betrachtet, die gleichmäßig gegen e^x für $x \in \mathbb{R}$ konvergieren. Jede dieser Folgen soll als Grundlage für die Berechnung des Funktionswertes der Exponentialfunktion an einer gegebenen Stelle x gewählt werden. Bei der Suche nach möglichst guten Algorithmen lassen wir uns von folgender Überlegung leiten. Auf einem Rechner ist der Funktionswert wegen Exponentenunterlauf bzw. Exponentenüberlauf nur für x-Werte aus einem beschränkten Intervall $(x_{\min}, x_{\max})$ berechenbar. Wenn man eine Genauigkeit *ianz*, gemessen in der Anzahl der richtigen Mantissenstellen, und die Anzahl r der Iterationen vorgibt, so existiert hierzu bei jedem Algorithmus ein Arbeitsintervall $(\underline{x}, \overline{x})$ mit der Eigenschaft, daß der Algorithmus mit r Iterationen *ianz* richtige Mantissenstellen liefert, falls der Wert x aus dem Arbeitsintervall vorgegeben wird. Natürlich sollte das Arbeitsintervall maximal berechnet sein. Für die Durchführung der Iterationen benötigt ein Algorithmus *iop* Operationen, falls der Wert x im Arbeitsintervall liegt. Die Daten $(\underline{x}, \overline{x})$, *iop* sind algorithmenspezifisch und können unabhängig von einer Anwendung bestimmt werden. Wenn man

mit dem Arbeitsintervall startet, kann man somit für jeden Algorithmus eine Zerlegung des Ausgangsintervalls $(x_{\min}, x_{\max})$ finden; pro Algorithmus entstehe dabei eine endliche Folge $x_0, x_1, \ldots, x_p$. Wenn nun $x \in (x_i, x_{i+1})$ gilt, so folgt

$$e^x = e^{x_i} e^y$$

und y liegt im Arbeitsintervall. Hat man daher die Werte $e^{x_0}, e^{x_1}, \ldots, e^{x_p}$ bereits a priori berechnet (und abgespeichert), so transformiert man mit der obigen Formel den x-Wert in sein zugeordnetes Arbeitsintervall, berechnet mit r Iterationen einen Funktionswert mit *ianz* richtigen Mantissenstellen und erhält mit dem bereits vorhandenen Funktionswert den gesuchten. Zusammenfassend werden daher bei gegebener Genauigkeit die Speichereffizienz eines Algorithmus durch die Länge des Arbeitsintervalls und die Operationseffizienz durch die Anzahl der Operationen pro Iteration beschrieben. Die Aufgabe lautet nun: Man finde aus den obigen Algorithmen den besten.

5. Die Funktion $f(x) = \tan \pi x$ soll an den Stützstellen $x_0 = 0$, $x_1 = 1/6$ und $x_2 = 1/4$ gegeben sein. Man löse das Interpolationsproblem entsprechend der Definition für folgende Ansätze:

   (a)
   $$P(x) = a_o + a_1 x + a_2 x^2 \,,$$

   (b)
   $$Q(x) = b_0 + b_1 x + b_2 \frac{1}{x - 1/2} \,.$$

   Welche Näherungen ergeben sich hieraus für $\tan 20^0$?

6. Man berechne $P(\bar{x})$ aus der vorherigen Aufgabe für $\bar{x} = 0,2$

   (a) nach der Methode von Lagrange,

   (b) nach der Methode von Newton,

   (c) nach dem Neville-Algorithmus.

7. Man schätze den Fehler von $P(\bar{x})$ aus der vorherigen Aufgabe ab.

8. Die Funktion $\ln x$ werde quadratisch interpoliert. Stützstellen seien $x_0 = 10$, $x_1 = 11$ und $x_2 = 12$.

   (a) Man schätze den Interpolationsfehler für $x = 11,1$ ab.

   (b) Wie hängt das Vorzeichen des Interpolationsfehlers von x ab?

9. Aus den Werten von $f(x) = \sqrt{x}$ an den Stellen $x_0 = 0$, $x_1 = 1$ und $x_2 = 4$ berechne man den Näherungswert für $\sqrt{3}$ und $\sqrt{1/3}$ durch

   (a) Polynominterpolation,

   (b) Berechnung, Auswertung der kubischen Spline-Interpolierenden; wobei $S''(0) = S''(4) = 0$ gelten möge.

   Man diskutiere das Ergebnis.

10. Für die Zerlegung des Intervalls $I = (0,1)$ durch $x_k = k \cdot h$ mit $k = 0,1,2,3,4$ und $h = 1/4$ ist die Splinefunktion S mit $S''(0) = S''(1) = 0$ und

    $$S(x) = \begin{cases} 1 & x = x_0 \\ 0 & x = x_k \,,\ k = 1,2,3,4 \end{cases}$$

    auf I zu berechnen (in Form von Formeln für die Teilintervalle). Man berechne insbesondere $S(1/8)$ und $S(3/8)$.

11. (a) Man berechne die Koeffizienten in den Newton-Cotes-Formeln für $n = 2$ (Simpson-Regel) und $n = 4$ (Milne-Formel).

    (b) Welche Näherungswerte ergeben sich nach den Newton-Cotes-Formeln für $n = 1,2,3,4$ bei der Berechnung von $\int_0^1 \sin \pi x \, dx$?

12. Man leite analog zum Vorgehen bei der Trapezsumme die zusammengesetzte Simpson-Regel

$$S(h) = \frac{h}{3}\{f(a) + f(b) + 2[f(a+2h) + \cdots + f(b-2h)] + 4[f(a+h) + \cdots + f(b-h)]\}$$

mit

$$h = \frac{b-a}{2n}$$

her.

13. Man zeige, daß das Romberg-Verfahren mit den Schrittweiten $h_0 = b - a$ und $h_1 = (b-a)/2$ gerade die Simpson-Regel liefert.

$$T_{i,k} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left[\frac{h_{i-k}}{h_i}\right]^2 - 1}\,.$$

14. Für die lineare Abbildung $f(x) = a + bx$ mit $a \neq 0$, $b \neq 0$ soll die erste Ableitung $f'(0) = b$ nach der Differentiationsformel

$$D_h f(0) = \frac{f(h) - f(-h)}{2h}$$

in dualer Gleitpunktarithmetik berechnet werden. Dabei seien a und b gegebene duale Gleitpunktzahlen. h sei eine Potenz von 2, so daß Multiplikation mit h und Divison durch $2h$ exakt ausgeführt werden. Man gebe eine Schranke für den relativen Fehler von$D_h f(0)$ an. Wie verhält sich diese Schranke für $h \to 0$?

15. Mit dem Householder-Verfahren löse man das Gleichungssystem

$$\begin{bmatrix} 1/3 & -1 & 5/6 \\ 2/3 & 0 & 1/6 \\ 2/3 & 1/5 & 1/6 \end{bmatrix} x = \begin{bmatrix} 1/6 \\ 5/6 \\ 31/30 \end{bmatrix}.$$

16. Gegeben seien $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ mit $\mathbf{a} \neq b$ und $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2$ ($\|\cdot\|_2$ - euklidische Norm). Man konstruiere eine Housholder-Transformation $\mathbf{H}$, für die $\mathbf{Ha} = \mathbf{b}$ gilt.

17. Es sei $\mathbf{H} = \mathbf{E} - \alpha\mathbf{u}\mathbf{u}^\mathsf{T}$ eine Housholder-Transformation. Das Produkt $\mathbf{Hy}$ kann nach folgenden Algorithmen berechnet werden:

    (a) Berechne $\mathbf{H} = \mathbf{E} - \alpha\mathbf{u}\mathbf{u}^\mathsf{T}$ und berechne $\mathbf{z} = \mathbf{Hy}$,
    (b) Berechne $\beta = \alpha\mathbf{u}^\mathsf{T}\mathbf{y}$ und $\mathbf{z} = \mathbf{y} - \beta\mathbf{u}$.

    Man zeige, daß beide Algorithmen äquivalent sind und vergleiche die Algorithmen hinsichtlich der Anzahl der Operationen und des benötigten Speicherplatzes.

18. Man berechne die Cholesky-Zerlegung der Matrix

$$\mathbf{A} = \begin{bmatrix} 16 & 4 & 4 \\ 4 & 5 & 3 \\ 4 & 3 & 11 \end{bmatrix}.$$

19. Für das Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ mit

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 0,99 \end{bmatrix} \quad \text{und} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

ist

$$\mathbf{A}^{-1} = \begin{bmatrix} -99 & 100 \\ 100 & -100 \end{bmatrix} \quad \text{und} \quad x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Für die Störungen

$$\delta\mathbf{A} = 10^{-3}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad \text{und} \quad \delta\mathbf{b} = 10^{-3}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

löse man das Gleichungssystem $(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$ und berechne $\delta\mathbf{x}$ und $\|\delta\mathbf{x}\|_\infty$. Man vergleiche die berechnete Störung $\|\delta\mathbf{x}\|_\infty$ mit den Schranken aus der Abschätzung

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq \frac{\operatorname{cond}\mathbf{A}}{1 - \|\delta\mathbf{A}\|\|\mathbf{A}^{-1}\|}\left(\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} + \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}\right).$$

20. Es liege das folgende mathematische Gesetz vor: $y = x_1 z + x_2$ mit zwei unbekannten Parametern x_1 und x_2. Ferner sei ein Satz von Meßdaten gegeben: (y_l, z_l), $l = 1 \ldots m$ mit $z_l = l$.
Man versuche, mittels linearer Ausgleichsrechnung die Parameter x_1, x_2 aus den Meßdaten zu bestimmen.

    (a) Wie lauten die Normalgleichungen für das lineare Ausgleichsproblem?

    (b) Man führe die Cholesky-Zerlegung der Matrix der Normalgleichung $\mathbf{B} = \mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{G}^\mathsf{T}\mathbf{G}$ durch.

21. Man zeige: Hat die $m \times n$-Matrix $\mathbf{A}$ den Rang n, so ist $\mathbf{A}^\mathsf{T}\mathbf{A}$ positiv definit.

22. Man zeige: $\lim\limits_{i \to \infty} x_i = 2$ für $x_0 = 0$ und $x_{i+1} = \sqrt{2 + x_i}$, $i = 0, 1 \ldots$

23. Man zeige, daß die Iteration $x_{k+1} = \cos(x_k)$ für alle $x_0 \in \mathbb{R}$ gegen den einzigen Fixpunkt $\xi\,(\xi = \cos\xi)$ konvergiert.

24. Man bestimme die drei Lösungen der Gleichung $F(x) = 3x^2 - e^x = 0$ mit dem Newton-Verfahren, wobei die Startwerte

    (a) $x_0 = 0,25\,,$

    (b) $x_0 = 0,3\,,$

    (c) $x_0 = 0,35$

    zu benutzen sind (max. 10 Iterationen bzw. bis zur Rechnergenauigkeit).

25. Man prüfe, ob $\Phi(x) = \ln 3 + 2\ln|x|$ eine geeignete Iterationsfunktion zur Bestimmung der Lösung von $F(x) = 3x^2 - e^x = 0$ ist.

Index