Abschnitt 1

ἀεὶ ὁ ἄνθρωπος ἀριθμητίζει[1)]
(R. Dedekind)

Zur Geschichte des Zahlbegriffs

Die Geschichte des Zahlbegriffs ist ein faszinierendes Kapitel der Kulturgeschichte. Der Begriff wurde im Verlaufe der Zeit mehrfach erweitert. In der griechischen Mathematik hatte das Wort *Zahl* (ἀριθμός) seit Pythagoras (ca. 580 – 500 v.Chr.) eine feste Bedeutung, nämlich die einer positiven natürlichen Zahl als einer Menge von Einheiten, wie in den *Elementen* des Euklid (ca. 300 v.Chr.) definiert; dabei sei die Sonderrolle, welche die 1 als Repräsentant der Einheit hatte, außer acht gelassen. Die Bezeichnungen rationale, irrationale, negative, imaginäre und reelle Zahlen kamen erst im Mittelalter auf. Der Erweiterungsprozess verlief über mehrere Etappen, siehe auch [10, 12]. Seit der durch C. F. Gauss (1777 – 1855) vollzogenen Ausräumung der Bedenken gegenüber den von ihm so benannten komplexen Zahlen hat sich der Zahlbegriff in der algebraisch-analytischen Richtung kaum noch verändert. Natürlich hat man sich über den Zahlbegriff häufig geäußert; aber die Frage „Was sind und was sollen die Zahlen?“ vermochte erst gegen Ende des 19. Jahrhunderts, im Zusammenhang mit der Loslösung der Analysis von ihren geometrisch-physikalischen Wurzeln, die Gemüter durchgreifend zu bewegen. Vorher hatte man sich auf den Begriff der messbaren (stetigen) Größe in der geometrisch-physikalischen Welt berufen; eine reelle Zahl wurde als das definiert, womit man misst, wie bei Stevin, Newton und anderen nachzulesen ist.

G. Cantor verallgemeinerte den ursprünglichen Zahlbegriff in einer anderen Richtung, nämlich zu den transfiniten Kardinal- und Ordinalzahlen, siehe **11.5**. In gewissem Sinne stehen diese der altgriechischen Auffassung von Zahlen als einer Kollektion von Einheiten sogar näher. Jedenfalls sind die reellen Zahlen deswegen nicht reeller, weil sie so heißen. Sie sind auch nicht reeller als die komplexen Zahlen. Denn deren Distanz zu den reellen Zahlen ist aus heutiger Sicht um ein Vielfaches kleiner als die der reellen zu den natürlichen. Komplexe Zahlen lassen sich einfach als Paare reeller Zahlen verstehen, siehe **9.4**. Nicht so einfach ist gewiss der Übergang von der reellen zur komplexen Analysis. Hier gibt es aber höchstens sachliche, aber keine logischen Schwierigkeiten. Diese wohnen dem Begriff der Irrationalzahl bei genauerer Betrachtung grundsätzlich inne, auch wenn die Definition in Abschnitt **3** sehr einfach aussieht.

[1)] Der Mensch treibt ewig Arithmetik – Anspielung auf Platons Wort „Gott treibt ewig Geometrie“.

1.1 Die ältere Geschichte

Die Vorstellungen über Zahlen in den großen Kulturkreisen des Altertums, bei den Ägyptern, den Sumerern, Babyloniern und den Griechen, den Indern und Chinesen sowie bei den Mayas und Azteken in Amerika sind uns durch Dokumente aus Stein, Knochen, Lehm und Papyrus hinlänglich klar überliefert worden. Was sich außerhalb dieser Kulturkreise oder vor deren Entstehung abgespielt hat, darüber können wir nur Mutmaßungen anstellen. Eine dieser ziemlich verbreiteten und durch Observationen noch lebender so genannter primitiver Kulturen gestützten Mutmaßungen ist, dass der in unserem Sinne noch nicht kultivierte Mensch nur über Worte für „Eins“und „Zwei“ verfügte, alles andere war „mehr“ oder „viel mehr“, ähnlich wie ja auch einige Säugetiere und Vögel über einen anscheinend angeborenen Sinn für die Erfassung von einem oder zwei (bis zu fünf) Gegenständen verfügen. Ein Indiz dafür ist, dass für die Zahlen 1 und 2 in fast allen Sprachen besondere Namen existieren, unterschiedlich zugleich nach ihrer Funktion als Anzahl oder als Ordnungszahl, und dass sich erst bei größeren Zahlen eine zunehmende Regelmäßigkeit sprachlicher Konstruktion ausprägt.

Auch die Geschichte der symbolischen Zahlbezeichnungen in den verschiedenen Kulturkreisen ist aufschlussreich. Das Bezeichnungsprinzip ist in den antiken Kulturen fast überall dasselbe. Man kann es vielleicht am treffendsten mit dem Wort *Kollektionsprinzip* umschreiben. Kleine Zahlen (bis höchstens 10) wurden meistens durch Striche oder Punkte bezeichnet. Sodann wurde für eine bestimmte Anzahl, z.B. für 10, ein neues Symbol eingeführt. Mehrere dieser Anzahlen wurden wieder zusammengefasst und abermals mit einem neuen Symbol bezeichnet, usw. Typisch in dieser Hinsicht ist das über Jahrtausende verwendete Bezeichnungssystem im alten Ägypten, ebenso das der Griechen und schließlich auch der Römer. In seiner spätmittelalterlichen Form (siehe dazu [23]) wird es auch heute noch verwendet.

Vorteil der auf dem Kollektionsprinzip beruhenden Systeme ist eine unmittelbare Veranschaulichung der jeweils dargestellten Anzahl. Dem aber steht der entscheidende Nachteil gegenüber, dass das Rechnen in diesen Systemen recht kompliziert ist. Die Addition kleinerer Anzahlen ist einfach. Man muss nur die „Zahlhaufen“ zusammenfügen, unter eventueller Verwendung eines neuen Symbols für eine neu entstandene Gruppierungseinheit. Das Multiplizieren hingegen wird schon zu einem Problem, und das Dividieren schließlich ist nur noch eine den Eingeweihten vorbehaltene Kunst. Mit Sicherheit standen den Rechenexperten der damaligen Zeit gewisse technische Hilfsmittel zur Verfügung, z.B. Tafeln und einfache „digitale“ Rechengeräte, ähnlich dem heute an manchen Orten noch gebräuchlichen Abakus, und wahrscheinlich neigen wir nur wegen mangelnder Vertrautheit mit den Rechentechniken der damaligen Zeit dazu, die Schwierigkeiten zu überschätzen. Der präzise Kalender, aufgebaut auf astronomischen Beobachtungen und Berechnungen, die Steuereintreibung und die Landvermessung bei einem komplizierten Bewässerungssystem machen deutlich, dass umfangreiche Rechnungen ständig ausgeführt wurden. In der Regel geschah dies in besonderen Institutionen an den Höfen und Tempeln der Könige, Fürsten und Priester.

Man liest gelegentlich, dass einigen Zahlbezeichnungssystemen in der Vergangenheit eine andere Grundzahl zugrunde lag als 10. So wurde bei den Babyloniern für Bruchzahlen das *sexagesimale* System verwendet, in welchem die Zahl 60 die heutige Rolle der Zahl 10 spielte. In anderen Kulturen war 20 eine Art Grundzahl. Die Bezeichnungssysteme des Altertums hatten jedoch nicht die wesentlichen Merkmale eines g-adischen Positionssystem. Statt von Grundzahlen sollte man besser z.B. von *Kollektionseinheiten* sprechen. Meist wurden und werden in derselben Sprache unterschiedliche Kollektionseinheiten benutzt, je nach Art der Gegenstände, auf die sich die Zählung bezog. Wir beobachten das noch heute etwa an dem Wort *Dutzend*, das nur in bestimmten Zusammenhängen benutzt wird. Sprachreste in den germanischen Sprachen deuten nicht wirklich auf eine frühere Verwendung der 12 als Kollektionseinheit hin. So bedeuteten z.B. die gotischen Namen *àinlif* für 11 und *twalif* für 12 – verwandt mit dem deutschen *elf* und *zwölf* und dem englischen *eleven* und *twelve* – ursprünglich etwa „eins gelassen“ bzw. „zwei gelassen“ (nachdem alle Finger zum Zählen verbraucht sind).

Wir wissen recht wenig darüber, ob und in welcher Richtung in den vorgriechischen Kulturen schon eine Philosophie des Zahlbegriffs existierte, wie sie uns von den Pythagoreern (600 – 500 v.Chr.), von PLATON (ca. 429 – 348 v.Chr.) und anderen überliefert worden ist. Die Griechen unterschieden klar zwischen (natürlichen) Zahlen und *Größen*. Größenbeziehungen wurden durch vier-stellige Relationen, so genannte *Proportionen* beschrieben, aber diese in [8, Buch V] dargestellte, EUDOXOS (ca. 408 – 355 v.Chr.) zugeschriebene Proportionenlehre ist keine Lehre von den rationalen Zahlen. Größen waren anschauliche, in der Regel geometrische Objekte wie Strecken- Winkel- oder Flächengrößen (siehe **10.1**). Es war den Griechen bekannt, dass mit Zahlenproportionen (aus natürlichen Zahlen) die Geometrie nicht adäquat zu beschreiben war. Ausdruck dieser Unzulänglichkeit ist, dass die Länge der Diagonalen eines Quadrats sich zur Seitenlänge verhält wie $\sqrt{2}$ zu 1. EUKLID liefert einen korrekten Beweis dafür, dass Diagonale und Seite des Quadrats inkommensurabel sind, oder in heutiger Terminologie, dass $\sqrt{2}$ irrational ist. Wahrscheinlich wussten dies schon die Babylonier, die mit unterschiedlichen Näherungen für $\sqrt{2}$ rechneten. Die Alltagsnäherung als Sexagesimalbruch war $1 + \frac{25}{60}$. Diese ist doppelt so genau wie 1,41.

Vielleicht kannten die Babylonier auch folgende Überlegung zur Verbesserung einer rationalen Näherung r für eine Irrationalzahl der Gestalt $\sqrt{m}$: Für $r < \sqrt{m}$ ist $\sqrt{m} < \frac{m}{r}$ (also $r < \sqrt{m} < \frac{m}{r}$); falls aber $\sqrt{m} < r$ so ist $\frac{m}{r} < \sqrt{m} < r$, so dass $\sqrt{m}$ jedenfalls zwischen r und $\frac{m}{r}$ liegt. Daher ist $r' = \frac{1}{2}(r + \frac{m}{r})$ sicher eine bessere Näherung für $\sqrt{m}$.[2)] Für die erste sexagesimale Näherung $r = 1 + \frac{25}{60}$ von $\sqrt{2}$ erhält man nach Abrunden so die den Babyloniern ebenfalls bekannte Näherung $r' = 1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3}$. Dies ist die beste 3-stellige sexagesimale Näherung; sie ist auf 5 Dezimalstellen genau. Natürlich läßt sich aus der Kenntnis dieser Näherungen noch nicht schließen, die Babylonier seien im Besitz des angegebenen Iterationsverfahrens zur Berechnung von $\sqrt{m}$ gewesen.

[2)] Diese Überlegung liefert offensichtlich eine gegen $\sqrt{m}$ konvergierende Iterationsfolge $\langle r_n \rangle$ mit $r_{i+1} = \frac{1}{2}(r_i + \frac{m}{r_i})$, die in gewissem Sinne zufällig mit derjenigen des berühmten Newton-Verfahrens zur Berechnung der positiven Nullstelle von $x \mapsto x^2 - m$ übereinstimmt.

Das von den Größen handelnde Buch X der Elemente [8] beginnt mit der folgenden klaren Definition, die wir in **10.9** unverändert übernehmen werden:

> Kommensurabel heißen Größen, die von demselben Maß gemessen werden, und inkommensurabel solche, für die es kein gemeinsames Maß gibt.

Das Phänomen inkommensurabler (oder irrationaler) Streckenverhältnisse in der Geometrie war den Griechen also wohlvertraut, und die Vermutung, dass dies eine ernste Grundlagenkrise der griechischen Mathematik verursachte, ist unbegründet. Statt von inkommensurablen Zahlen hat man von inkommensurablen Größen gesprochen, was durchaus sinnvoll ist. Die These der Pythagoreer „Alles ist Zahl“ war natürlich inzwischen modifiziert worden. Mit den so genannten Proportionen konnten auch inkommensurable Größenverhältnisse näherungsweise bestimmt werden. Man darf daher die Größen- und Proportionenlehre mit Recht als die Urform der Analysis bezeichnen.

Zwar unterschied auch ARCHIMEDES (ca. 287 – 212 v.Chr.) deutlich zwischen Zahl im herkömmlichen Sinne und Messgröße als Verhältnis der gemessenen Größe zur Einheit, aber er hatte anscheinend eine Vorstellung hierüber, die etwa derjenigen über reelle Zahlen (als Maßzahlen für geometrisch-physikalische Größen) in der beginnenden Neuzeit entsprach. Er berechnete auf mathematische Weise nicht nur die Kreiszahl π $(= 3{,}141\cdots)$ approximativ zu $\frac{22}{7}$, sondern auch krummlinig begrenzte Flächenstücke der verschiedensten Art. Die griechische Mathematik war zu ihrer Zeit progressiv; in ihr waren auch Erkenntnisse früherer Kulturen des Mittelmeerraumes aufgehoben. Andererseits hat die griechische Zahlkonzeption, eingeordnet in ausgeklügelte philosophische Kategorien, sich später teilweise als Hemmschuh erwiesen. Angesichts des zögerlichen Fortschritts der Mathematik im mittelalterlichen Europa hat man den Eindruck, dass einige der griechischen Dogmen allzu ehrfürchtig beachtet wurden.

Allmählich gelangten neben gebrochenen Zahlen auch negative Zahlen in Gebrauch, die das Lösen von Gleichungen wesentlich vereinfachten. Ein großer Fortschritt, und dies nicht nur im numerischen Bereich, wurde etwa gleichzeitig bewirkt durch eine scheinbare Äußerlichkeit, nämlich die Erfindung des dezimalen Positionssystems.

1.2 Die Entwicklung des Dezimalsystems

Die Erklärung der reellen Arithmetik in den Abschnitten **4** und **5** beruht wesentlich auf dem dezimalen Positionssystem. Dieses System stammt in seiner Grundstruktur aus Indien, wurde später von den Arabern übernommen und gelangte um das Jahr 1000 herum über den damals arabisch beherrschten südlichen und westlichen Mittelmeerraum in die italienischen Handelszentren. Das Dezimalsystem wurde mit anderen Bezeichnungen auch in China verwendet und gelangte von dort z.B. nach Japan.

Dieser Prozess vollzog sich unter wechselvollen Umständen. Die Gestalten der Ziffern wurden mehrfach abgewandelt, so dass die im europäischen Raum etwa seit der Erfindung der Buchdruckerkunst standardisierten Ziffern kaum noch Ähnlichkeit haben

mit den heutigen im arabischen Raum noch verwendeten Ziffern. Diese äußerlichen Änderungen sind jedoch nicht wesentlich. Vielmehr ist dies die uns heute fast banal erscheinende Erkenntnis, dass natürliche Zahlen mit nur endlich vielen Ziffern bezeichnet werden können, wenn man die Zahlen durch eine *Anordnung* von Ziffern kodiert, unter Einschluss einer Ziffer für die Zahl 0. Dies bedeutet natürlich eine Abkehr von einer Bezeichnungsweise nach dem Kollektionsprinzip.

Trotz klarer Vorteile gegenüber dem in Europa fast überall benutzten römischen Zahlensystem war die Durchsetzung des indisch-arabischen Positionssystems selbst in den progressiven italienischen Handelszentren mit Widerständen verbunden. Dennoch begann sich im 13. Jahrhundert dieses System in Europa zu verbreiten, wobei das 1202 erschienene *Liber abbaci* [22] des LEONARDO von Pisa (auch FIBONACCI genannt, etwa 1170 – 1250) eine erhebliche Rolle gespielt hat. LEONARDO beginnt dieses Werk mit einer musterhaften Erklärung der Rechenalgorithmen im dezimalen Positionssystem.

Das Prinzip des Dezimalsystems, nämlich jede positive natürliche Zahl k als endliche Folge $z_0 z_1 \cdots z_n$ aus den Ziffern 0 bis 9 darzustellen, ist allgemein bekannt. Sei $g = 10$ und seien $z_0, \ldots, z_n$ diejenigen eindeutig bestimmten Ziffern mit $z_0 \neq 0$, so dass

$$k = z_0 \cdot g^n + z_1 \cdot g^{n-1} + \ldots + z_{n-1} \cdot g + z_n \,.$$

Nun hätte man anstatt $g = 10$ hier jede beliebige andere natürliche Zahl $g \geq 2$ als die sogenannte Grundzahl wählen können. Man spricht dann vom *g-adischen System*; siehe hierzu auch **11.6**. Für $g > 10$ benötigt man entsprechend mehr, für $g < 10$ weniger Ziffern [3)] Das so genannte *Binärsystem* ($g = 2$) hat G. W. LEIBNIZ (1646 – 1716) sehr geschätzt; es diente ihm als Beispiel für seine These, dass das innere Wesen der Dinge von göttlicher Harmonie geprägt ist. Das System hat nur die Ziffern $0, 1$ (die *Binärziffern*) und besitzt große praktische Bedeutung, denn die meisten digitalen Computer rechnen binär. Für $g = 8$ spricht man vom *Oktalsystem*. Hier entfallen einfach die Ziffern 8 und 9. In der folgenden Tabelle sind die natürlichen Zahlen von 1 beginnend in der oberen Reihe im Dezimalsystem, darunter im Binärsystem, und in der unteren Reihe im Oktalsystem geschrieben.

g	:=	10	1	2	3	4	5	6	7	8	9	10	11	12	$\cdots$
g	:=	2	1	10	11	100	101	110	111	1000	1001	1010	1011	1100	$\cdots$
g	:=	8	1	2	3	4	5	6	7	10	11	12	13	14	$\cdots$

Hierbei erhebt sich natürlich die von historischen Betrachtungen ganz unabhängige Frage, ob eine andere Grundzahl als 10 Vorteile im Alltagsleben hätte. Diese Frage muss klar bejaht werden, und zwar zugunsten des Oktalsystems. Wesentliche Vorteile des Oktalsystems sind folgende:

[3)] Die Länge $\ell_g k$ der Zifferfolge von k in g-adischer Darstellung nimmt entsprechend zu. Nach (17) in **7.4** ist $\ell_g k$ für größere k ungefähr gleich $\frac{\ell_{10} k}{\log g}$. Das ist für $g = 8$ nur das etwa 1,1-fache der dezimalen Länge. Für kleine natürliche Zahlen ist der Längenzuwachs besonders klein. Für die ersten 500 natürlichen Zahlen ergibt sich ein mittlerer Zuwachs der Ziffernlänge von weniger als 3%.

- Das Einmaleins würde sich um fast die Hälfte (über 40%) verkürzen.
- Interpolationen bei Messvorgängen oder bei der Benutzung von Tabellen wären erheblich schneller auszuführen.
- Die Herstellung von Tastaturen und Messeinrichtungen (Schreib- und Rechenmaschinen, Waagen, Gewichtssätze und Skalen aller Art) würde sich verbilligen und ihre Handhabung vereinfacht.
- Die Ziffern $0, \ldots, 7$ des Oktalsystems werden durch die acht Tripel $000, \ldots, 111$ aus Binärziffern umkehrbar eindeutig kodiert. Deshalb erfolgt die Umrechnung einer reellen Zahl vom Oktal- ins Binärsystem *ziffernweise* in Dreierblöcken, und umgekehrt. So ist $2{,}4375 = 2{,}34_8 = 10{,}011100_2$ (weil $2 = 010_2$, $3 = 011_2$ und $4 = 100_2$). Die von kleinen und großen Rechnern ständig vorgenommenen Umrechnungen vom Dezimal- ins Dualsystem und umgekehrt würden praktisch entfallen, und damit entfiele zugleich eine permanente Quelle von Rundungsfehlern.
- Die Kapazität der Binärspeicher für Zahlen in den elektronischen Rechnern wäre viel besser auslastbar, weil in einer sagen wir 16-Bit Zelle mit einem Bitplatz für das Vorzeichen (siehe **9.4**) jede der $2^{16} - 1 = 65\,535$ ganzen Zahlen $\pm n$ mit $0 \leq n \leq 2^{15} - 1 = 32\,767 = 77777_8$ Platz fände. Das ist für 4-stellige ganzzahlige Dezimalzahlen mit Vorzeichen zu viel, für 5-stellige zu wenig.

Nur am Rande sei ein im Zeitalter der Computer natürlich unerheblicher Vorteil erwähnt: Die 2-te Rundung der Kreiszahl π ist 3,14. Im Oktalsystem ist diese $3{,}11_8$. Sie ist nicht nur für Handrechnungen bequemer, sondern ihr relativer Fehler ist mit rund 0,031% um fast die Hälfte kleiner als derjenige der dezimalen Rundung (ca. 0,051%).

Nun, es lässt sich auch in einer Welt des Dezimalsystems leben. Hier wie dort lässt sich nämlich die Zifferndarstellung natürlicher Zahlen in der bekannten, anschaulichen Weise auch auf Zifferndarstellungen nichtganzer Zahlen erweitern, was Anlass gibt zu den Dezimalzahlen oder Kommazahlen, auch Dezimalbrüche genannt. Diese sind etwa im 15. Jahrhundert an verschiedenen Orten des damals schon riesigen Verbreitungsraumes der indisch-arabischen Ziffern aufgetaucht. Von erheblicher Bedeutung für die Popularisierung der Dezimalbrüche in Westeuropa war vermutlich das Rechenbuch *De Thiende* von S. Stevin (1548 – 1620), in welchem z.B. die Zahl 27,847 in der Weise 27(0)8(1)4(2)7(3) notiert wird. In Stevins Büchlein finden sich auch die bekannten, noch heute benutzten Algorithmen für Addition und Multiplikation abbrechender Dezimalzahlen, die in Abschnitt **4** des vorliegenden Buches zum Fundament der Dezimalzahlarithmetik gemacht werden. Die Schreibweise Stevins ist zur Erläuterung dieser Algorithmen anhand von Beispielen ausgezeichnet geeignet. Es ist nicht klar erkennbar, ob Stevin die Dezimalbrüche unabhängig von Vorläufern erfunden hat, die es nachweislich mindestens 100 Jahre früher gegeben hat. Jedenfalls erkennt und verdeutlicht er klar die Vorteile eines allgemeinen dezimalen Maßsystems. In der Überzeugung, dass sich dieses früher oder später durchsetzen werde, schreibt er in [34] einleitend:

> Ob nun hierdurch kostbare, nicht käufliche Zeit gewonnen werden wird... überlasse ich gern Ihrem Urteil. Nun könnte mir jemand sagen, dass viele Dinge sich auf den ersten Blick oft besonders gut anlassen, aber wenn man sie durchführen will, so kann man nichts damit ausrichten, ebenso wie es bei den Neuerungen der Revolutionäre oft zugeht, welche im kleinen gut sind, aber im großen nichts taugen. Denen antworten wir, dass solch Zweifel hier keinesfalls bestehen kann, weil es... nun täglich in der Praxis genug erprobt wird, nämlich durch verschiedene erfahrene Landmesser hier in Holland, denen wir es erklärt haben ...

Vermutlich sind die Dezimalbrüche im Prozess des Experimentierens mit Rechenalgorithmen oder in der Praxis des Messens eher zufällig entstanden, obwohl eine dezimale Maßskala noch zu Ende des 18. Jahrhunderts weitgehend unbekannt war. Sie wurde erst nach der Französischen Revolution eingeführt und den kontinentalen Europäern durch NAPOLEON sozusagen zwangsverordnet.

Die Dezimalbrüche wurden und werden noch heute unterschiedlich notiert. Bezeichnungen wie 2|47 oder $2\underline{47}$ oder 2(0)4(1)7(2) für 2,47 waren im Gebrauch. Heute ist neben dem Dezimalkomma der angelsächsische Dezimalpunkt die am weitesten verbreitete Notation. Sie hat einige Vorteile gegenüber der Kommaschreibweise, etwa in der Niederschrift der Menge $\{1, 1.4, 1.41, \dots\}$ der dezimalen Näherungen von $\sqrt{2}$.

Für Addition, Subtraktion und Multiplikation natürlicher Zahlen im Dezimalsystem wurden schon im frühen Mittelalter von verschiedenen Autoren im wesentlichen dieselben Verfahren empfohlen, die wir heute kennen. Hauptschwierigkeit war das Dividieren; hier gab es viele phantasievolle Methoden, deren Hauptprobleme insgesamt alle darauf hinaus liefen, mit den „Resten“ fertig zu werden; denn die Dezimalbrüche kamen ja erst viel später, nämlich im 17. Jahrhundert, allmählich in Gebrauch. LEONARDO von Pisa löst z.B. die Gleichung $x^3 + 2x^2 + 10x = 20$ mit dem sehr exakten Näherungswert

$$1.22^{\mathrm{I}}7^{\mathrm{II}}42^{\mathrm{III}}33^{\mathrm{IV}}4^{\mathrm{V}}40^{\mathrm{VI}},$$

einer bequemen, leicht durchschaubaren Schreibweise für $1+\frac{22}{60}+\frac{7}{60^2}+\frac{42}{60^3}+\frac{33}{60^4}+\frac{4}{60^5}+\frac{40}{60^6}$ (bei LEONARDO geschrieben als „unum et minuta .XXII. et secunda .VII. et tertia .XLII. et quarta .XXXIII. et quinta .IIII. et sexta .XL.“). Uns erscheint kurios, dass die natürliche Zahlen dezimal, Bruchzahlen aber zugleich sexagesimal dargestellt wurden, auch wenn die Zahl 60 besonders günstig ist hinsichtlich der relativen Anzahl ihrer Teiler. Dazu muss man aber bedenken, dass Bruchzahlen als wesentlich verschieden von natürlichen angesehen wurden, und nicht einmal Zahlen genannt wurden.

Es ist hier nicht der Platz darzustellen, welche ausgeklügelten Methoden erforderlich gewesen sind, um in diesem gemischten System Multiplikations- und Divisionsaufgaben zu lösen. Ein anderer Autor aus dem 16. Jahrhundert berechnet z.B. die Quadratwurzel aus 10 wie folgt: Er multipliziert zunächst 10 mit dem Wert 10^6 und findet, dass 3162^2 die am nächsten bei $10^7 = 10 \cdot 10^6$ liegende Quadratzahl ist, also $\sqrt{10^7} \approx 3162$ ($\approx$ meint *ungefähr gleich*). Er hätte nun schreiben können $\sqrt{10} \approx 3{,}162$, denn es ist

$$\sqrt{10} = \frac{\sqrt{10^7}}{\sqrt{10^6}} \approx \frac{3162}{10^3} = 3 + \frac{162}{1000}.$$

Statt dessen rechnet er $\frac{162}{1000}$ mit dem g-adischen Algorithmus aus **8.2** für $g = 60$ in einen Sexagesimalbruch um, $\frac{162}{1000} = 9^{\mathrm{I}}43^{\mathrm{II}}12^{\mathrm{III}}$, und erhält so $\sqrt{10} \approx 3.9^{\mathrm{I}}43^{\mathrm{II}}12^{\mathrm{III}}$. Dieses Ergebnis ist übrigens nicht die beste 3-stellige Sexagesimalbruchnäherung für $\sqrt{10}$. Vielmehr ist dies die Zahl $3.9^{\mathrm{I}}44^{\mathrm{II}}12^{\mathrm{III}} = 3{,}1622\,777\cdots$, die sich erst in der 7. Dezimalen von $\sqrt{10} = 3{,}1622\,776\cdots$ unterscheidet. Siehe hierzu auch **8.4**.

Reste des Sexagesimalsystems für Brüche haben sich bis in die heutige Zeit erhalten, insbesondere in der Winkelteilung und der Zeitmessung. Dezimale Winkelteilung schafft hier nur zusätzliche Verwirrung, denn Gewohnheiten sind nun einmal mächtig.

Die Geschichte der dezimalen Arithmetik ist ein Musterbeispiel dafür, wie wichtig scheinbare Äußerlichkeiten – im vorliegenden Falle die dezimale Darstellung reeller Zahlen – für den Fortschritt einer Wissenschaft sein können. Diese Darstellung hat mit Sicherheit geholfen, Schwierigkeiten mit den Irrationalzahlen als weniger gravierend zu empfinden. Schon die dezimale Darstellung natürlicher Zahlen war ein gewaltiger Fortschritt. Sie waren nun der Öffentlichkeit zugänglich und nicht mehr Statussymbol einer geistigen Elite. Die Griechen hatten fraglos das zahlentheoretische Wissen, um eine dezimale oder andere g-adische Darstellung natürlicher Zahlen einwandfrei zu begründen. Der EUKLIDische Algorithmus zur Bestimmung des größten gemeinsamen Teilers benutzt nämlich dasselbe Hilfsmittel, die so genannte Division mit Rest, siehe dazu **11.6**. Aber sie haben diesen Weg anscheinend nicht gesucht und auch nicht zufällig gefunden. Am Fehlen der Null lag es sicher nicht, denn durch eine Modifikation der g-adischen Darstellung kommt man auch ohne die Ziffer 0 ausgezeichnet zurecht, Übung 11.7. Im übrigen benutzten die Griechen und andere ein Symbol für die Null auch bei der Kennzeichnung fehlender Glieder in der sexagesimalen Bruchdarstellung.

1.3 Die neuere Geschichte

Eine Antwort auf die Frage nach dem *Wesen* des Zahlbegriffs, also dergestalt „Was sind die Zahlen“, hängt nicht unerheblich von dem abstrakten Begriffsgefüge ab, über das der Mensch einer jeweiligen Epoche unter Einbeziehung aller Erfahrungen der Vergangenheit verfügt. Die Maßstäbe wissenschaftlicher Strenge sind einem beständigen Wandel unterworfen, und sie werden fortfahren sich zu wandeln. Man darf sich z.B. nicht darüber wundern, dass GAUSS in mehreren der von ihm erstmals geführten Beweise des Fundamentalsatzes der Algebra bedenkenlos Zwischenwertargumente für stetige Funktionen benutzte, obwohl der Stetigkeitsbegriff erst später präzisiert wurde und noch später die Notwendigkeit des Beweises von Zwischenwertargumenten allgemeine Anerkennung fand. Mit einem Wort, seit den Zeiten von GAUSS haben sich unsere Ansichten über das, was einer Begründung bedarf, erheblich gewandelt.

NEWTON und auch LEIBNIZ verstanden unter einer reellen Zahl das Verhältnis zweier Größen gleicher Art, von denen eine als Einheit betrachtet wurde. Nach Newton ist eine reelle Zahl „das womit man misst“. Noch im 19. Jahrhundert wurden Irrationalzahlen als Größenverhältnisse von Strecken erklärt. Die Klärung des Begriffs einer reellen Zahl

nach modernen Maßstäben ist wesentlich das Werk R. DEDEKINDs (1831 – 1916), obwohl auch andere, wie etwa B. BOLZANO (1781 – 1848), A. CAUCHY (1789 – 1857), K. WEIERSTRASS (1815 – 1897), G. CANTOR (1845 – 1918) und D. HILBERT (1862 – 1943) in diesem Zusammenhang genannt werden müssen. Der Aufbau des Systems der reellen Zahlen nach modernen Maßstäben begrifflicher Strenge ist Gegenstand der Abschnitte **3** - **5** und muss daher an dieser Stelle nicht erläutert werden.

Jedem Schritt einer Erweiterung des Zahlbegriffs ging ein erster Schritt voraus, nämlich das formale Operieren mit Termen wie z.B. $7 - 12$, $2 + \sqrt{3}$, $5 + \sqrt{-15}$ [4)], verbunden mit ernsten Auseinandersetzungen über die Frage, ob solchen Rechengrößen nun eine reale Existenz zukommt oder ob sie nur fiktive oder imaginäre, in unserer Einbildung existierende Größen darstellen. Die begriffliche Erfassung irrationaler Größen und ihre Einordnung in das Zahlensystem war zweifellos der entscheidende und schwierigste Schritt. Lassen wir zum Problem der irrationalen Zahlen einen bekannten Mathematiker des ausgehenden Mittelalters zu Worte kommen, M. STIFEL (1486 – 1567). In seinem Hauptwerk [35] schreibt er:

> Mit Recht wird bei den irrationalen Zahlen darüber disputiert, ob sie wahre Zahlen sind oder nur fiktive. Denn bei Beweisen an geometrischen Figuren haben die irrationalen Zahlen noch Erfolg, wo uns die rationalen im Stich lassen, und sie beweisen genau das, was die rationalen Zahlen nicht beweisen konnten. Wir werden also veranlasst, ja gezwungen zuzugeben, dass sie in Wahrheit existieren, nämlich auf Grund ihrer Wirkungen, die wir als wirklich, gewiss, und feststehend empfinden.
>
> Aber andere Gründe veranlassen uns zu der entgegengesetzten Behauptung, dass wir nämlich bestreiten müssen, dass die irrationalen Zahlen Zahlen sind. Nämlich wenn wir versuchen, sie der Zählung zu unterwerfen und sie mit rationalen Zahlen in ein Verhältnis zu setzen, dann finden wir, dass sie uns fortwährend entweichen, so dass keine von ihnen sich genau erfassen lässt... Es kann aber nicht etwas eine wahre Zahl genannt werden, bei dem es keine Genauigkeit gibt und was zu wahren Zahlen kein bekanntes Verhältnis hat. So wie eine unendliche Zahl keine Zahl ist, so ist eine irrationale Zahl keine wahre Zahl, weil sie sozusagen unter einem Nebel der Unendlichkeit verborgen ist; doch ist das Verhältnis einer irrationalen Zahl zu einer rationalen nicht weniger unbestimmt als das einer unendlichen Zahl zu einer endlichen.

Die Vorstellung über Irrationalzahlen beschränkte sich bei STIFEL überdies nur auf Wurzelausdrücke wie z.B. $\sqrt[3]{2} + \sqrt{5}$. Eine Unterscheidung zwischen algebraischen und transzendenten Irrationalzahlen wie π existierte selbstredend noch nicht. Mit Beginn der Neuzeit, vor allem seit Erfindung der Logarithmen und der Differential- und Integralrechnung, hat sich ein allgemeinerer als nur durch Wurzelausdrücke oder geometrische Verhältnisse beschreibbarer Begriff der Irrationalzahl als unentbehrlich erwiesen. Diese speziellen Zahlen bilden nur ein Tröpfchen im riesigen Meer aller Irrationalzahlen.

[4)] zuerst aufgetreten in [4] und durch $5 \cdot \tilde{p} \cdot R \cdot \tilde{m} \cdot 15$ bezeichnet, $\tilde{p}$ = plus, $\tilde{m}$ = minus, R = Radix.

Wegen der Dichtheit der rationalen in der Menge $\mathbb{R}$ aller reellen Zahlen ist anschaulich klar, dass eine Irrationalzahl a durch das Paar (U_a, V_a) eindeutig bestimmt ist, wobei U_a die Menge aller rationalen Zahlen $\leq a$ und V_a diejenige aller rationalen Zahlen $> a$ bezeichne. Diesen natürlich schon lange bekannten Umstand hat DEDEKIND in [5] benutzt, um a als eben dieses Paar (von ihm *Schnitt* genannt) zu definieren [5)]. Auf der Menge dieser Schnitte lassen sich dann Rechenoperationen erklären, und das Resultat ist der Körper der reellen Zahlen (auch die Bezeichnung *Körper* stammt von DEDEKIND). Dies war aus historischer Sicht die erste, in allen Details begrifflich klare Konstruktion des Körpers der reellen Zahlen. Übrigens könnte man statt von Schnitten rationaler Zahlen z.B. auch von Schnitten abbrechender Dezimalzahlen ausgehen; man benötigt nicht alle rationalen Zahlen zur Ausführung der Konstruktion. Wesentlich an der Konstruktion ist nur die Wahl eines in der Menge der reellen Zahlen dicht liegenden überschaubaren Zahlenbereichs, in dem man wie gewohnt rechnen kann.

Im Anschluss an DEDEKIND wurden von anderen Autoren ähnliche mengentheoretische Konstruktionen vorgeschlagen, z.B. von CANTOR dasjenige mittels Fundamentalfolgen (CAUCHY-Folgen) und von P. BACHMANN das Verfahren der Intervallschachtelungen. Schließlich wurde – im wesentlichen von HILBERT in [17] – auch eine vollständige axiomatische Charakterisierung des Bereichs der reellen Zahlen angegeben. Diese wurde später in einzelnen Punkten verfeinert, siehe auch **10.5**. In [37] wurde erstmals auch auf die bis auf Isomorphie eindeutige Kennzeichnungsmöglichkeit der reellen Zahlen als Elemente einer stetig geordneten Gruppe hingewiesen, wodurch sich ein allseitig befriedigendes Verständnis des logarithmischen Rechnens ergibt. Dies wird in **10.6** im einzelnen dargelegt.

Die Analysis hat in der 2. Hälfte des 20. Jahrhunderts noch einmal eine wesentliche Bereicherung ihrer Grundlagen durch die Entwicklung der so genannten *Nichtstandard-Analysis* durch A. ROBINSON erfahren. Diese hat schon heute nicht allein nur theoretisches Interesse insofern, als durch sie die alte Idee von LEIBNIZ von den Differentialen als unendlich kleinen Zahlen in überraschender Weise Wirklichkeit geworden ist, sondern viele Phänomene aus dem Bereich der Molekularphysik und der Quantenfeldtheorie lassen sich mit ihrer Hilfe besser modellieren als mit Mitteln der traditionellen Analysis und Funktionalanalysis. Grundlage der Nichtstandard-Analysis ist die Existenz von Modellen einer jeden formalisierten oder formalisierbaren Theorie, welche ein ausgezeichnetes Modell der Theorie, meist das Standardmodell genannt, unter Erhalt aller dort gültigen Aussagen erweitern. Das betrifft insbesondere die Theorie der reellen Zahlen unter Einschluss eines ausreichenden Fragments der Mengenlehre. Darauf können wir im Rahmen dieser Darstellung aber nicht eingehen. Siehe [20] oder [32].

[5)] genauer, a wird durch diesen Schnitt *hervorgebracht*, wie DEDEKIND sich ausdrückte. DEDEKIND war der Meinung, Zahlen seien eine freie Schöpfung des menschlichen Geistes und es bedürfe keiner direkten Identifizierung von a mit dem Schnitt (U_a, V_a). Diese beträfe konsequenterweise dann auch rationale Schnitte, wodurch die rationalen Zahlen eine neue Qualität erhielten. Ob nun a mit (U_a, V_a) identifiziert wird oder nicht, ist im Prinzip unwesentlich; denn weder aus der einen noch aus der anderen Verfahrensweise lässt sich ein zusätzlicher Gewinn ziehen.

Abschnitt 2

Natürliche Zahlen und Rechenregeln für nichtnegative Zahlen

Mit den einfachsten Eigenschaften der natürlichen Zahlen $0, 1, 2, \ldots$ ist jedermann vertraut. Daher wollen wir uns an dieser Stelle nicht mit Erläuterungen darüber aufhalten, was diese Zahlen sind und woher die üblichen, in **2.1** genannten Rechenregeln kommen. Einzelheiten hierüber findet man im Anhang (Abschnitt **11**). Man lernt frühzeitig, wie natürliche Zahlen n, m im Dezimalsystem der Größe nach verglichen, addiert, multipliziert und – falls dies möglich ist – voneinander subtrahiert werden. Zu den Elementarkenntnissen im weiteren Sinne gehört auch eine gewisse Bekanntschaft mit Primzahlen, mit elementaren Teilbarkeitseigenschaften, der Potenz n^m, und dem überaus wichtigen Beweisverfahren durch vollständige Induktion.

$\mathbb{N}$ bezeichnet überall die Menge der natürlichen Zahlen zu denen wir auch die Zahl 0 rechnen. Die Buchstaben n, m, i, j, k werden ausschließlich als Variable für natürliche Zahlen verwendet. Daran sollte man sich stets erinnern, weil Zusätze wie $n \in \mathbb{N}$ in der Regel unterbleiben. Die auf n unmittelbar folgende natürliche Zahl $n + 1$ werde mit n^+ bezeichnet und heißt der *Nachfolger* von n. Eine von 0 verschiedene natürliche Zahl heiße *positiv*. $\mathbb{N}_+$ bezeichne die Menge der positiven natürlichen Zahlen.

In diesem Abschnitt befassen wir uns mit denjenigen arithmetischen Eigenschaften oder Rechenregeln des Zahlenbereichs $\mathbb{N}$, welche bei Erweiterung von $\mathbb{N}$ zu anderen Bereichen nichtnegativer Zahlen in jedem Falle ihre Gültigkeit bewahren sollen, insbesondere für die Bereiche der endlichen Dezimalzahlen und der nichtnegativen reellen Zahlen. Einige dieser Eigenschaften werden besonders hervorgehoben und in Gestalt eines Axiomensystems zusammengestellt, aus welchem alle übrigen interessierenden arithmetischen Eigenschaften hergeleitet werden. Aus dem Axiomensystem ergeben sich zwangsläufig auch gewisse Eigenschaften der nachträglich definierten Subtraktion und der Potenz. Von allen arithmetischen Eigenschaften sollen so viele wie möglich auch nach Einführung negativer Zahlen ihre Gültigkeit bewahren. Vorerst lassen wir die negativen Zahlen aus Gründen der Übersichtlichkeit jedoch aus dem Spiel.

2.1 Grundrechenregeln für nichtnegative Zahlen

Mit bloßen Fertigkeiten in der Handhabung des Zahlenrechnens kann man noch keine Mathematik betreiben oder anwenden. Man muss die Rechengesetze kennen, die dieser Handhabung zugrundeliegen. Die im folgendem Axiomensystem zusammengestellten Rechengesetze oder Rechenregeln gelten sämtlich im Bereich der natürlichen Zahlen. Sie sind im Ergebnis des Umgangs mit diesen Zahlen formuliert worden und insofern empirischer Herkunft. Nicht aufgeführt sind hierbei Gesetze von rein logischem (d.h. allgemeingültigem) Charakter, wie z.B. $a = b \Rightarrow a + c = b + c$.

Die angegebenen Axiome sind charakteristisch für alle Bereiche nichtnegativer Zahlen. Für Zahlbereiche unter Einschluss negativer Zahlen sind diese geringfügig zu modifizieren. Dadurch gelangt man zu den Rechengesetzen für geordnete Ringe, die in **9.1** vorgestellt werden. Deren Vorteil ist vor allem der Wegfall hinderlicher Einschränkungen hinsichtlich der Subtraktion.

Axiome des Rechnens mit nichtnegativen Zahlen

N^+ :	$a + 0 = a$	$\mathsf{N}^\times$: $a \cdot 1 = a$	(Neutralitätsgesetze),
K^+ :	$a + b = b + a$	$\mathsf{K}^\times$: $a \cdot b = b \cdot a$	(Kommutativgesetze),
A^+ :	$a + (b + c) = (a + b) + c$	$\mathsf{A}^\times$: $a \cdot (b \cdot c) = (a \cdot b) \cdot c$	(Assoziativgesetze),
D :	$(a + b) \cdot c = a \cdot c + b \cdot c$		(Distributivgesetz),
V :	*Entweder* $a = b$ *oder* $a < b$ *oder* $b < a$		(Vergleichbarkeit),
E :	*Es gibt ein* $d \neq 0$ *mit* $a + d = b$ *genau dann, wenn* $a < b$		
F :	$a \cdot b \neq 0$ *für* $a, b \neq 0$		(Nullteilerfreiheit),
G :	Es gibt ein $a \neq 0$.		

Diese Schreibweisen sind genau genommen Abkürzungen für gewisse *Aussagen*. So lautet etwa K^+ ausführlich *Für alle* a, b : $a + b = b + a$. Wir verwenden auch übliche Konventionen der Klammerersparnis, z.B. bei der Niederschrift des Terms $a \cdot c + b \cdot c$ in Axiom D die Verabredung, dass der Malpunkt stärker bindet als $+$. Den Gepflogenheiten folgend wird der Malpunkt künftig nur dann geschrieben, wenn dies der Deutlichkeit dienlich ist. Axiom E beschreibt die Existenzbedingung einer von 0 verschiedenen Lösung für Gleichungen der Gestalt $a + x = b$. E kann aber auch als explizite Definition von $<$ durch $+$ verstanden werden, so daß $<$ als Grundbegriff eigentlich entbehrlich ist.

Im Bereich der natürlichen Zahlen gelten weitere spezifische Gesetze, etwa dass zu jeder Zahl eine unmittelbar nachfolgende gibt, sowie insbesondere das schon erwähnte Beweisprinzip der vollständigen Induktion. Anders als diese Besonderheiten behalten jedoch die angegebenen Axiome und damit auch die daraus durch logische Folgerung herleitbaren Rechenregeln für alle Bereiche nichtnegativer Zahlen ihre Gültigkeit.

Unter einem *Bereich* verstehe man eine nichtleere Menge A – genannt die *Trägermenge* – mit gewissen auf A erklärten Operationen und Relationen, sowie gewissen ausgezeichneten Elementen aus A. Sind $+, \cdot$ solche Operationen, $<$ eine Relation, und $0, 1$

ausgezeichnete Elemente, und gelten dort alle oben aufgelisteten Axiome, so heiße dieser ein *Elementarbereich* oder kurz ein *$\mathcal{E}$-Bereich*. Diesen Namen wählen wir, weil obiges Axiomensystem und seine Folgerungen genau die Rechengesetze umfasst, die man im mathematischen Elementarunterricht, bewusst oder unbewusst noch vor Einführung der negativen Zahlen erlernt. In einer Fußnote in **9.1** wird erläutert, wie sich der Begriff des $\mathcal{E}$-Bereichs in die algebraische Standard-Terminologie einordnet.

Wir unterscheiden i.a. nicht zwischen der Bezeichnung eines $\mathcal{E}$-Bereichs und seiner Trägermenge A. Stets sei $A_+ = \{a \in A \mid a > 0\}$. Man beachte, daß hier $a > 0$ wie verabredet nur eine andere Schreibweise für $0 < a$ ist. 0 und 1 heißen in diesem Zusammenhang auch die *neutralen* Elemente, 0 das *Nullelement* und 1 das *Einselement*. Das prominenteste Beispiel eines $\mathcal{E}$-Bereichs ist $\mathbb{N}$. Weitere Beispiele, darunter die Bereiche der nichtnegativen rationalen und der reellen Zahlen, werden noch konstruiert.

2.2 Folgerungen aus den Grundrechenregeln

Es werden nun aus den angegebenen Axiomen weitere Rechengesetze als logische Folgerungen hergeleitet. Zunächst sei vermerkt, dass aufgrund der Assoziativgesetze A^+ und $\mathsf{A}^\times$ Klammern innerhalb von Termen, die entweder nur $+$ oder nur $\cdot$ enthalten, nicht geschrieben werden müssen. Auch kommt es wegen K^+ und $\mathsf{K}^\times$ auf die Reihenfolge der Summanden bzw. Faktoren nicht an. Zwei erste Folgerungen aus den Axiomen sind

H : $\quad a < a + b$, falls $b \neq 0$, $\qquad\qquad$ O : $\quad 0 \leq a$.

Es existiert nämlich ein $d \neq 0$ mit $a + d = a + b$, und zwar $d = b$, so dass $a < a + b$ nach Axiom E. Das ergibt mit a für b speziell $0 < 0 + a = a$, also $0 \leq a$. Ferner: $a + b = 0$ impliziert $a = b = 0$. Denn wäre etwa $b \neq 0$, liefert $a + b = 0$ offenbar $a < 0$ nach Axiom E. Dies aber ist unmöglich; denn nach O ist $0 \leq a$ und damit wird $a < 0$ durch V ausgeschlossen. Hiermit erhalten wir für $<$ nun leicht die Eigenschaft der *Transitivität*

T : $\quad$ *Wenn* $a < b$ *und* $b < c$, *so* $a < c$.

Denn sei $a < b$ und $b < c$, so dass $a + d = b$ und $b + e = c$ für gewisse $d, e \neq 0$ gemäß E. Dann erhält man $a + d + e = c$, und hieraus $a < c$ nach E, weil mit $e \neq 0$ immer auch $d + e \neq 0$. Wesentlich aus E folgen auch die so genannten *Monotoniegesetze*

M^+ : $\quad a < b \;\Rightarrow\; a + c < b + c$, $\qquad$ $\mathsf{M}^\times$: $\quad a < b \;\Rightarrow\; ac < bc \quad (c > 0)$ [1].

Denn sei $a < b$, also $a + d = b$ für ein $d \neq 0$. Dann folgt $a + c + d = b + c$ und mit E daher $a + c < b + c$. Das beweist M^+. Ferner ergibt $a + d = b$ offenbar $ac + dc = bc$ gemäß D. Ist nun auch $c \neq 0$, so gilt gemäß F auch $dc \neq 0$, also $ac < bc$, wiederum nach E. Das beweist $\mathsf{M}^\times$. Gelegentlich ist es nützlich, die folgenden offensichtlich beweisbaren Abschwächungen von M^+ und $\mathsf{M}^\times$ zur Verfügung zu haben:

$$a \leq b \;\Rightarrow\; a + c \leq b + c \quad ; \quad a \leq b \;\Rightarrow\; ac \leq bc, \quad \text{falls } c > 0.$$

[1] Im Interesse eines übersichtlichen Schriftbildes stehen Zusatzvoraussetzungen oft in Klammern, anstelle eines Textes wie „falls $c > 0$“ im vorliegenden Falle.

Ist A ein beliebiger $\mathcal{E}$-Bereich, und hat $f : A \to A$ die Eigenschaft $a < b \;\Rightarrow\; fa < fb$ für alle $a, b \in A$, so folgt allein aufgrund von V schon die Umkehrung hiervon; also $fa < fb \;\Rightarrow\; a < b$. Denn $a \not< b \Rightarrow b \leq a \Rightarrow fb \leq fa \Rightarrow fa \not< fb$. Also gilt die Umkehrung von M^+: $a + c < b + c \;\Rightarrow\; a < b$, und analog von $\mathsf{M}^\times$. Ähnlich folgt die

Streichungsregel: $\quad c + a = c + b \;\Rightarrow\; a = b.$

Denn für $a \neq b$ gilt $a < b$ oder $b < a$ gemäß V, was mit M^+ in beiden Fällen zu einem Widerspruch zu $c + a = c + b$ führt. Mit einem analogen Schluss folgt aus $\mathsf{M}^\times$ die

Kürzungsregel: $\quad ca = cb \;\Rightarrow\; a = b \qquad (c \neq 0).$

Als nächstes zeigen wir $a0 = 0a = 0$ für alle a. N^+, $\mathsf{N}^\times$, sowie $\mathsf{K}^\times$ und D ergeben nämlich $a + 0 = a = 1a = (1 + 0)a = 1a + 0a = a + 0a$. Also $0a = 0$ nach der Streichungsregel. Erst hieraus ergibt sich übrigens $0 \neq 1$, also auch $0 < 1$ nach O. Denn $1 = 0$ impliziert $a = a1 = a0 = 0$ für alle a was Axiom G widerspricht. Es sei vermerkt, dass Axiom G weder ein logisches ist noch folgt es aus den übrigen Axiomen. Ohne dieses Axiom wäre die wichtige Ungleichung $0 < 1$ unbeweisbar. Wiederholte Anwendung von M^+ auf $0 < 1$ ergibt insbesondere leicht $0 < 1 < 1 + 1 < 1 + 1 + 1 < \ldots$

Bemerkung. Es ist sinnvoll, in einem $\mathcal{E}$-Bereich A mit n zugleich auch das n-Vielfache von $1 \in A$, also $1 + \ldots + 1$ mit genau n Summanden zu bezeichnen. Man kann nämlich ohne Einschränkung der Allgemeinheit davon ausgehen, dass diese Vielfachen, einschließlich 0 und 1 selbst, mit den entsprechenden natürlichen Zahlen identisch sind. Leser, die mit dem Isomorphiebegriff noch nicht vertraut sind, sollten nur $\mathcal{E}$-Bereiche betrachten, die diese Bedingung erfüllen und das folgende Argument dafür, dass dies keine Einschränkung der Allgemeinheit ist, einfach überlesen: Man zeigt unschwer, dass die Gesamtheit der 1-Vielfachen eines beliebigen $\mathcal{E}$-Bereichs A ein isomorphes Bild des $\mathcal{E}$-Bereichs $\mathbb{N}$ ist. Auf diese Weise verschwindet die Zweideutigkeit in der Betrachtung von n als natürliche Zahl einerseits und als n-Vielfaches von 1 in einem $\mathcal{E}$-Bereich andererseits. Kurzum, die natürliche Zahl n darf mit dem aus A stammenden n-Vielfachen von 1 identifiziert werden, ähnlich wie in Abschnitt 9 die komplexen Zahlen ohne Imaginärteil mit den reellen Zahlen identifiziert werden.

Seien a, b Elemente eines $\mathcal{E}$-Bereichs, $b \leq a$. Dann gibt es ein d mit $b + d = a$ (für $a = b$ wähle man $d = 0$). Aus $b + d = b + e = a$ folgt $d = e$ mit der Streichungsregel, also gibt es genau ein d mit $b + d = a$. Dieses d wird mit $a - b$ bezeichnet und heißt die *Differenz* von a und b. Definitionsgemäß gilt also $(a - b) + b = b + (a - b) = a$, sowie

(1) $\quad a - b = d \;\Leftrightarrow\; a = d + b \qquad (a \geq b).$

Mit dieser auch *Subtraktion* genannten, nur partiell erklärten Operation lässt sich wie gewohnt rechnen – nur muss die Definiertheit der vorkommenden Terme gesichert sein. Zum Beispiel ist $a(b - c) = ab - ac$ falls $b \geq c$; denn dann ist auch $ab \geq ac$ und die Behauptung folgt mit (1) und D aus $ac + a(b - c) = a(c + (b - c)) = ab$. Ähnlich folgt

(2) $\quad (a - c) + (c - b) = a - b \qquad (a \geq c \geq b).$

Dies ergibt sich mit (1) aus $(a - c) + (c - b) + b = (a - c) + c = a$. Man beachte, (1) und (2) sind vorerst nur unter den in Klammern gesetzten Zusatzbedingungen sinnvoll.

2.3 Die Potenz

Sei A ein beliebiger $\mathcal{E}$-Bereich und $a \in A$. Dann wird a^n wie folgt rekursiv erklärt:

$$a^0 = 1 \quad ; \quad a^{n+1} = a^n \cdot a,$$

so dass $a^1 = a^0 \cdot a = a,\ a^2 = a^1 \cdot a = a \cdot a$, usw. Insbesondere ist $0^0 = 1$. Dies ist eine bequeme Konvention, weil dann z.B. $1 + c + c^2 + \ldots + c^n = \sum_{i \le n} c^i$ ohne zusätzliche Verabredungen auch für $n = c = 0$ gilt. Für $n \neq 0$ hingegen ist stets $0^n = 0$.

Es gelten die folgenden Regeln der Potenzrechnung in A mit Exponenten aus $\mathbb{N}$:

$$\begin{array}{lll} \mathsf{P}^+ :\ a^{m+n} = a^m a^n, & \mathsf{P}^\times :\ a^{mn} = (a^m)^n, & \mathsf{P}_\times :\ (ab)^n = a^n b^n, \\ \mathsf{P}^< :\ n < m \ \Rightarrow\ a^n < a^m \quad (a > 1), & \mathsf{P}_< :\ a < b \ \Rightarrow\ a^n < b^n \quad (n \neq 0), & \\ \mathsf{P}^= :\ a^n = a^m \ \Rightarrow\ n = m \quad (a \neq 0, 1), & \mathsf{P}_= :\ a^n = b^n \ \Rightarrow\ a = b \quad (n \neq 0). & \end{array}$$

$\mathsf{P}^+, \mathsf{P}^\times, \mathsf{P}_\times$ beweist man meist durch vollständige Induktion. Es geht aber auch mit folgendem einfachen Satz, der die Induktion vorverlegt und oft deutlicher erkennen lässt, warum gewisse Identitäten gelten. A kann hierin eine beliebige Menge sein.

Satz 2.1. *Seien $F : A \times \mathbb{N} \to A$, $f : \mathbb{N} \to A$ und $g : \mathbb{N} \to A$ Funktionen derart, dass*

$$(*) \quad f0 = g0 \ \ ; \ \ f(n+1) = F(fn, n) \ \ ; \ \ g(n+1) = F(gn, n), \quad \textit{für alle } n \in \mathbb{N}.$$

Dann sind f und g identisch.

Beweis. Wir zeigen $fn = gn$ durch Induktion über n. Es ist $f0 = g0$. Sei $fn = gn$ nach Induktionsannahme. $(*)$ liefert dann $f(n+1) = F(fn, n) = F(gn, n) = g(n+1)$; also ist auch $f(n+1) = g(n+1)$. Folglich gilt $fn = gn$ für alle n, d.h. $f = g$. ∎

Analoges gilt auch, wenn im Satz $\mathbb{N}$ durch $\mathbb{N}_+$ und $f0 = g0$ durch $f1 = g1$ ersetzt wird. Kurzum, es ist $f = g$, falls f und g dieselben Anfangswerte haben und derselben Rekursionsgleichung genügen. Um die Beweiskraft dieses Satzes an einem simplen Beispiel zu verdeutlichen, bezeichne fn die Summe der ersten n ungeraden Zahlen, so dass $f(n+1) = fn + 2n + 1 = F(fn, n)$ mit $F(x, n) = x + 2n + 1$. Ferner sei $gn = n^2$. Dann ist auch $g1 = 1$ und wegen $g(n+1) = (n+1)^2 = gn + 2n + 1$ genügen f und g derselben Rekursionsgleichung. Also $f = g$, d.h. $1 = 1^2,\ 1 + 3 = 2^2,\ 1 + 3 + 5 = 3^2$ usw.

Zum Nachweis von P^+ betrachte man $f : \mathbb{N} \to A$ mit $fn = a^{m+n}$ und $g : \mathbb{N} \to A$ mit $gn = a^m \cdot a^n$, wobei A ein $\mathcal{E}$-Bereich ist und $a \in A$ sowie $m \in \mathbb{N}$ beliebig aber fest gewählt seien. Dann gilt gewiss $f0 = a^m = g0$, und nach den Definitionen ist

$$f(n+1) = a^{m+n} \cdot a = f(n) \cdot a \quad ; \quad g(n+1) = a^m a^{n+1} = a^m a^n a = g(n) \cdot a.$$

Daher genügen f und g der Voraussetzung $(*)$ von Satz 2.1, mit $F(x, n) = x \cdot a$. Also $f = g$ und P^+ ist bewiesen. Genauso zeigt man $\mathsf{P}^\times$ und $\mathsf{P}_\times$ durch den Nachweis, dass $f : n \mapsto (a^m)^n$ und $g : n \mapsto a^{mn}$ bzw. $f : n \mapsto (ab)^n$ und $g : n \mapsto a^n b^n$ die Bedingung $(*)$ für geeignetes F erfüllen. Mit derselben Methode ergibt sich auch die nützliche Formel

$$(3) \quad 1 - c^{n+1} = (1 - c)(1 + c + c^2 + \ldots + c^n) \qquad (c \le 1).$$

Denn für $f : n \mapsto 1 - c^{n+1}$ und $g : n \mapsto (1 - c)(1 + c + \ldots + c^n)$ ist sicher $f0 = g0$, und

$$\begin{array}{rcl} f(n+1) & = & 1-c^{n+2} = (1-c^{n+1}) + (c^{n+1}-c^{n+2}) \;=\; fn + (1-c)c^{n+1}, \\ g(n+1) & = & (1-c)(1+c+c^2+\ldots+c^n+c^{n+1}) \;=\; gn + (1-c)c^{n+1}. \end{array}$$

Damit gilt $(*)$ in Satz 2.1 mit $F(x,n) = x + (1-c)c^{n+1}$, und (3) ist bewiesen.

Wichtig und in allen $\mathcal{E}$-Bereichen gültig ist ferner die so genannte *binomische Formel*

$$(4) \qquad (a+b)^n = \textstyle\sum_{k\le n} \binom{n}{k} a^{n-k}b^k = a^n + \binom{n}{1}a^{n-1}b + \ldots + \binom{n}{n-1}ab^{n-1} + b^n.$$

Diese verallgemeinert die Gleichung $(a+b)^2 = a^2+2ab+b^2$, welche für $n=2$ durch (4) mit erfasst wird. Hier sind die so genannten *Binomialkoeffizienten* $\binom{n}{k}$ für jedes Paar (n,k) mit $n \ge k$ definierte natürliche Zahlen, die wie folgt rekursiv erklärt sind:

$$\binom{n}{0} = \binom{n}{n} = 1 \quad ; \quad \binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1} \quad (k<n).$$

$$\begin{array}{ccccccccccc} & & & & 1 & & & & & & \\ & & & 1 & & 1 & & & & & \\ & & 1 & & 2 & & 1 & & & & \\ & 1 & & 3 & & 3 & & 1 & & & \\ 1 & & 4 & & 6 & & 4 & & 1 & & \\ & & & & \vdots & & & & & & \end{array}$$

PASCAL*sches Dreieck*

Deren Entstehung lässt sich mit dem berühmten PASCALschen Dreieck leicht veranschaulichen. Die 1 an der Spitze gilt als 0-te Zeile und es sei $\binom{0}{0} = 1$. In der n-ten Zeile des Dreiecks für $n>0$ mit den Randzahlen $\binom{n}{0} = 1$ und $\binom{n}{n} = 1$ findet man die $\binom{n}{k}$ für $k = 1,\ldots,n-1$ durch Addition der beiden jeweils darüber stehenden Zahlen. So ist $\binom{2}{1} = \binom{1}{0} + \binom{1}{1} = 1+1 = 2$ und $\binom{4}{2} = 3+3 = 6$. Induktion zeigt $\binom{n}{1} = n = \binom{n}{n-1}$ für $n \ge 1$, sowie die nützliche Rekursionsformel $\binom{n+1}{2} = n + \binom{n}{2}$ für $n \ge 2$.

Es ist sinnvoll, $\binom{n}{k} = 0$ zu setzen für $n < k$. Dann ist $\binom{n}{k}$ für alle $n, k \in \mathbb{N}$ definiert.

2.4 Übungen

1. In den klammerfreien Termen dieser Übung sei Linksklammerung gemeint (im Zweifelsfalle auch sonstwo). Man folgere aus den Axiomen für $\mathcal{E}$-Bereiche

   (5) $a+(b-c) = a+b-c \quad (b \ge c)$, (6) $a \ge b+c \;\Leftrightarrow\; a-b \ge c \quad (a \ge b)$,

   (7) $a+c \ge b \;\Leftrightarrow\; a \ge b-c \quad (b \ge c)$, (8) $a-(b+c) = a-b-c \quad (a \ge b+c)$,

   (9) $a-(b-c) = a+c-b \quad (a+c \ge b \ge c)$.

   Man beachte, (6),(7) garantieren die Wohldefiniertheit aller Terme in (8),(9).

2. Man beweise die in **2.3** formulierten Eigenschaften $\mathsf{P}^<, \mathsf{P}_<, \mathsf{P}^=$ und $\mathsf{P}_=$.

3. Man bestätige in $\mathcal{E}$-Bereichen $(a-b)^2 = a^2+b^2-2ab$ $(a \ge b)$, und $a^2+b^2 \ge 2ab$.

4. Es bezeichne f jeweils eine der beiden auf ganz $\mathbb{N}$ erklärten Funktionen

   $$n \mapsto 0+1+2+\ldots+n, \qquad n \mapsto \tbinom{n+1}{2}.$$

   Beide genügen der Bedingung $f0 = 0$. Man zeige, beide Funktionen erfüllen die Rekursionsgleichung $f(n+1) = fn+n+1$ und sind nach Satz 2.1 daher identisch. Kurzum, $\sum_{i\le n} i = \binom{n+1}{2}$ für alle n.

5. Man beweise durch Induktion über n oder mit Satz 2.1 die Formel (4).

Abschnitt 3

Reelle Zahlen und ihre Anordnung

Endliche oder abbrechende Dezimalzahlen, auch Kommazahlen genannt, sind uns von Kindheit an vertraut. Gleichungen wie zum Beispiel $1{,}23\,m = 123\,cm$ haben einen unmittelbar anschaulichen Sinn. Auch lernt man frühzeitig, wie man solche Zahlen addiert, multipliziert und der Größe nach vergleicht.

Nun ist bekannt, dass sich z.B. bei den so genannten nicht aufgehenden Divisionen und anderen Rechenverfahren mit endlichen Dezimalzahlen nichtabbrechende Dezimalfolgen ergeben, die sich nach dem Komma unbegrenzt fortsetzen. Dabei stellt sich heraus, dass gewisse dieser Folgen als Ergebnis solcher Operationen nicht erscheinen, nämlich diejenigen mit einer „Neunerperiode“. Diese werden wir bei unserem Vorgehen als unzulässig ausschließen, und zwar nicht nur weil man sie im Prinzip nicht benötigt, sondern weil sie unser intuitives Bild über die dichte Anordung reeller Zahlen stören würden, siehe **3.2**. Alle übrigen Dezimalfolgen nennen wir kurzerhand reelle Zahlen.

Mit einer bloßen Benennung ist es natürlich nicht getan. Worauf es nämlich ankommt ist der Nachweis, dass man mit diesen Zahlen wie gewohnt rechnen kann und dass sie eben alle jene Eigenschaften besitzen, welche die reellen Zahlen auszeichnen. Diesen Nachweis werden wir erbringen, und zwar ohne uns auf unendliche Reihen zu berufen, ja nicht einmal auf die Bruchrechnung. Diese wird sozusagen übersprungen. In diesem Abschnitt befassen wir uns zunächst mit der Anordnung dieser Zahlen. Gerechnet wird erst in den Abschnitten **4** und **5**. Wegen des anschaulichen Charakters der Thematik werden hier von Anfang an gleich die nichtabbrechenden Dezimalfolgen in die Betrachtung mit einbezogen. Dadurch lässt sich auch die Sonderrolle der Dezimalfolgen mit Neunerperiode sehr gut veranschaulichen.

Anders als in der geordneten Menge der endlichen Dezimalzahlen gilt in der geordneten Menge aller Dezimalzahlen der Satz von der oberen Grenze, der sehr einfach beweisbar ist. Dieser garantiert nicht nur die unbeschränkte Ausführbarkeit der Division – von der Division durch 0 abgesehen – sondern auch die Existenz von Funktionen wie z.B. der Exponentialfunktion, und macht dadurch Analysis überhaupt erst möglich.

3.1 Reelle Zahlen als Dezimalfolgen

Eine *Folge* von Elementen einer Menge M sei eine Abbildung, welche jedem $n \in \mathbb{N}$ ein Element $a_n \in M$ zuordnet. Eine solche Folge werde durch $\langle a_n \rangle$, oder z.B. auch $\langle a_i \rangle$, bezeichnet. Gelegentlich betrachten wir auch durch $\langle a_n \rangle_{n \geq k}$ bezeichnete Folgen mit dem Definitionsbereich $\{n \in \mathbb{N} \mid n \geq k\}$, sowie auch *endliche Folgen* $\langle a_0, \ldots, a_n \rangle$ oder $\langle a_i \rangle_{i \leq n}$ der Länge $n+1$. Es sind dies Abbildungen von $\{0, \ldots, n\}$ in eine Menge.

Unter einer *Dezimalfolge* sei nun eine Folge $\langle z_n \rangle$ verstanden, so dass z_i für jedes $i > 0$ eine Dezimalziffer ist, die eine natürliche Zahl zwischen 0 und 9 einschließlich bezeichnet (welche man sich mit der entsprechenden Ziffer identifiziert denken kann); hingegen darf z_0 eine beliebige natürliche Zahl in Dezimaldarstellung sein, also eine nicht mit 0 beginnende endliche Ziffernfolge, oder aber $z_0 = 0$. Die Folge $\langle z_n \rangle$ werde wie gewohnt als $z_0{,}z_1 z_2 \cdots$ notiert und heiße dann eine *Dezimalzahl*, auch *Kommazahl* genannt. Ist $a = z_0{,}z_1 z_2 \cdots$, darf z_n der Deutlichkeit halber auch mit z_n^a bezeichnet werden. z_n $(= z_n^a)$ wird für $n > 0$ die *n-te Dezimale* (oder *Kommastelle*) von a genannt. n heiße der *Index* (genauer, *Stellenindex*) von z_n. Die natürliche Zahl z_0, der *ganzzahlige Teil von a*, werde mit $\operatorname{Int} a$ bezeichnet (von *Integer*). Für $a = 13{,}746\cdots$ z.B. ist $\operatorname{Int} a = z_0 = 13$, $z_1 = 7$ die erste Dezimale, und so fort.

Wenn im folgenden von einer Dezimalzahl $a = z_0{,}z_1 \cdots z_n zzz \cdots$ die Rede ist, so ist stets gemeint, dass $z_i = z$ für *alle* $i > n$. Sei a von der Gestalt $z_0{,}z_1 \cdots z_n 000 \cdots$, also $z_i = 0$ für alle $i > n$. Dann heißt a eine *endliche* oder *abbrechende Dezimalzahl der Stellenzahl n*. Diese denken wir uns mit der im anschaulichen Sinne endlichen Dezimalfolge $\langle z_i \rangle_{i \leq n}$ identifiziert, welche wir in gewohnter Weise als $z_0{,}z_1 \cdots z_n$ schreiben. Doch soll die Gleichung $a = z_0{,}z_1 \cdots z_n$ nicht etwa zum Ausdruck bringen, n sei die kleinste Stellenzahl von a, also $z_n \neq 0$. Die kleinste Stellenzahl von a ist vielmehr das kleinste m mit $z_{m+i} = 0$ für alle $i > 0$. Demnach sind z.B. sowohl 3,14 als auch 3,140 Bezeichnungen für die Dezimalzahl $3{,}14000\cdots$, und diese Zahl hat die Stellenzahlen 2, 3, ... Aufgrund dieser Verabredung haben je zwei endliche Dezimalzahlen sicher eine gemeinsame Stellenzahl, und natürlich mehr als nur eine.

Es bezeichne $\mathbb{E}$ die Menge aller endlichen Dezimalzahlen. Die Zahl $z_0{,}000\cdots$ der Stellenzahl 0 identifizieren wir mit der natürlichen Zahl z_0. Man kann dies auch so ausdrücken, dass $z_0{,}000\cdots$ als neue Bezeichnung für $z_0 \in \mathbb{N}$ verwendet werden darf. Insbesondere ist $0{,}000\cdots = 0$. Die natürlichen Zahlen gehören daher zu $\mathbb{E}$. Sie sind in diesem Kontext als die endlichen Dezimalzahlen der Stellenzahl 0 gekennzeichnet.

Bemerkung. In diesem Stadium der Theorie hat die Schreibweise $z_0{,}z_1 \cdots z_n$ noch nichts mit einer Darstellung $z_0{,}z_1 \cdots z_n = z_0 + z_1 \cdot \frac{1}{10^1} + \ldots + z_n \cdot \frac{1}{10^n}$ zu tun. Erst nachdem die Terme $\frac{1}{10^n}$ und die Rechenoperationen auf $\mathbb{E}$ erklärt wurden, ist die Frage nach der eben erwähnten Identität sinnvoll. Vorläufig wissen wir nicht oder geben vor dies nicht zu wissen, was $\frac{1}{10}$ ist, noch wissen wir was z.B. die Summe $3 + \frac{1}{10}$ bedeutet. Wir vermeiden deshalb auch die Bezeichnung Dezimalbruch, denn Dezimalzahlen sind gemäß obiger Auffassung gewisse Folgen natürlicher Zahlen und keine Brüche im herkömmlichen Sinne.

Dezimalzahlen der Form $z_0{,}z_1 \cdots z_n 999 \cdots$ mit einer Neunerperiode heißen aus in **3.2** erläuterten Gründen *unzulässig* (in dem Sinne, dass sie vorerst von gewissen Betrachtungen ausgeschlossen werden); alle übrigen Dezimalzahlen heißen *zulässig*. Diese Unterscheidung kann nach Einführung der unendlichen Reihen in **7.1** wieder aufgehoben werden. Es bezeichne $\mathbb{D}$ die Menge aller zulässigen Dezimalzahlen, die wir fortan einfach einfach nur (nichtnegative) *reelle Zahlen* nennen. Dazu gehören insbesondere die endlichen Dezimalzahlen, kurz $\mathbb{E} \subseteq \mathbb{D}$. Eine von 0 verschiedene Zahl aus $\mathbb{D}$ heiße *positiv*. $\mathbb{E}_+$ und $\mathbb{D}_+$ bezeichnen die Mengen der positiven Elemente von $\mathbb{E}$, bzw. $\mathbb{D}$. Wir haben damit den Objekten unserer Betrachtung im voraus die Bezeichnung *reelle Zahlen* zugestanden. Dies wird nach entsprechender Vorarbeit in Abschnitt **5** gerechtfertigt.

Ist $a = z_0{,}z_1 z_2 \cdots$, so heiße $z_0{,}z_1 \cdots z_n$ die *n-te kanonische Näherung* (weil sie für alle $a \in \mathbb{D}$ in gleicher Weise definiert ist). Diese werde fortan mit $a_{.n}$ bezeichnet. Offenbar ist $a \in \mathbb{E}$ genau dann, wenn es ein n gibt mit $a = a_{.n}$. Denn $a = a_{.n} = z_0{,}z_1 \cdots z_n$ heißt ja $z_{n+i} = 0$ für alle $i > 0$. Speziell ist $a \in \mathbb{N}$ genau dann, wenn $a = a_{.0}$ $(= \operatorname{Int} a)$.

Beispiel. Für $a = 3{,}14000 \cdots$ ist $a_{.0} = 3$, $a_{.1} = 3{,}1$ und $a_{.2} = a_{.3} = \ldots = a = 3{,}14$.

3.2 Die Anordnung der Dezimalzahlen

Allgemein versteht man unter einer *Anordnung* oder *Ordnung* einer Menge M eine Relation $<$ auf M mit den Eigenschaften T und V der Transitivität und der Vergleichbarkeit aus **2.1**. V lässt sich hierbei auch ersetzen durch die beiden Forderungen I: $a \not< a$ (*Irreflexivität*), sowie V': $a < b$ *oder* $a = b$ *oder* $b > a$, einer Abschwächung von V. Wegen I schließen sich die Fälle $a < b$, $b < a$, $a = b$ nämlich gegenseitig aus. Wäre etwa $a < b$ und $b < a$, folgt $a < a$ nach T, ein Widerspruch zu I. Dass anderseits V auch I impliziert, ist klar: wegen $a = a$ entfällt die Möglichkeit $a < a$.

Wir gehen aus von der üblichen (lexikographischen) Anordnung der Dezimalfolgen: Sind $a = z_0{,}z_1 z_2 \cdots$ und $a' = z'_0{,}z'_1 z'_2 \cdots$ voneinander verschieden und ist $z_0 \neq z'_0$, so soll $a < a'$ bzw. $a' > a$ sein, je nachdem ob $z_0 < z'_0$ oder $z_0 > z'_0$. Wenn aber $z_0 = z'_0$, so richten wir uns nach den Werten von z_1 und z'_1 und allgemein nach dem kleinsten Index n, so dass $z_n \neq z'_n$. Wir fassen dies zusammen in der folgenden

Definition. Für $a = z_0{,}z_1 z_2 \cdots$ und $a' = z'_0{,}z'_1 z'_2 \cdots$ sei $a < a'$ genau dann, wenn $a \neq a'$ und wenn $z_n < z'_n$ für den kleinsten Index n mit $z_n \neq z'_n$.

Offenbar ist $a_{.n} \leq a$ für alle n. Ist $a < b$, so gibt es gewiss ein n mit $a < b_{.n}$, und umgekehrt folgt hieraus auch $a < b$. So ist $2{,}6398 \cdots < 2{,}640197 \cdots$, weil bereits $2{,}6398 \cdots < 2{,}64$. Es ist klar, dass $<$ die Eigenschaften V und T hat, also eine Ordnung ist, insbesondere auf $\mathbb{D}$. Ferner ist klar, dass sich für natürliche Zahlen genau die dort bereits vorhandene Anordnung ergibt. Anders als in $\mathbb{N}$ gibt es jedoch kein kleinstes positives Element in $\mathbb{D}$. Dies erkennt man leicht durch Betrachtung der folgenden Zahlenreihe, die im weiteren eine wichtige Rolle spielen wird:

$$\varepsilon_0 = 1; \quad \varepsilon_1 = 0{,}1; \quad \varepsilon_2 = 0{,}01; \quad \varepsilon_3 = 0{,}001 \text{ usw.}$$

Es ist unmittelbar klar, dass $\varepsilon_0 > \varepsilon_1 > \ldots > 0$ und dass zu jedem $a \in \mathbb{D}_+$ ein n existiert mit $\varepsilon_n < a$. Offenbar ist ε_n unter allen von 0 verschiedenen endlichen Dezimalzahlen der Stellenzahl n gerade die kleinste.

Nunmehr kann erklärt werden, warum die Dezimalfolgen $a = z_0{,}z_1 \cdots z_n 999 \cdots$ eine Sonderrolle in den Betrachtungen spielen und als unzulässig bezeichnet wurden. Es gibt nämlich keine Dezimalfolge, die im Sinne der soeben erklärten Anordnung zwischen $a := z_0{,}999 \cdots$ und $a' := z_0 + 1$ liegt. Man verdeutliche sich dies an Beispielen, etwa $a = 2{,}4999 \cdots$ und $a' = 3$ $(= 3{,}000 \cdots)$. Ebenso gibt es für $a := z_0{,}z_1 \cdots z_n 999 \cdots$ im Falle $0 < n$ und $z_n < 9$ keine Dezimalfolge, die zwischen a und $a' := z_0{,}z_1 \cdots z_n^{+}$ [1)] liegt. Aus $a \leq x \leq a'$ folgt nämlich $x = a$ oder $x = a'$ für alle $x \in \mathbb{D}$. Um der anschaulich klaren Forderung gerecht zu werden, dass zwischen zwei reellen Zahlen mindestens eine weitere liegt, ist man genötigt, eine der beiden folgenden Alternativen zu wählen: Entweder man schließt die Dezimalfolge a von den Betrachtungen aus (was man z.B. durch ihre Bezeichnung als unzulässige Dezimalzahlen erreicht), oder aber man identifiziert a und a', d.h. man sieht sie als Bezeichnungen für dasselbe Objekt an.

Ein Nachteil der zweiten Alternative ist, dass uns etwa für die Identifizierung von 2,5 mit 2,4999... gegenwärtig noch kein überzeugendes Argument zur Verfügung steht. Das gewinnen wir erst durch den sehr einfachen Nachweis in **7.1**, dass Dezimalzahlen als unendliche Reihen verstanden werden können.

Obige anschauliche Betrachtung macht deutlich, dass $\mathbb{D}$ hinsichtlich der Relation $<$ dicht geordnet ist, d.h. zwischen je zwei derartigen Zahlen liegt noch wenigstens eine, und damit unendlich viele weitere. Ganz allgemein heißt eine geordnete Menge *dicht geordnet*, wenn sie mindestens zwei Elemente enthält und zwischen je zwei Elementen noch ein weiteres Element liegt. Die Dichtheit der Anordnung von $\mathbb{D}$ lässt sich nun sogar zu folgender Aussage verschärfen, welche zugleich die Dichtheit von $\mathbb{E}$ offenbart:

Satz 3.1. *Zwischen je zwei Elementen $a, b \in \mathbb{D}$ mit $a < b$ liegt wenigstens ein $c \in \mathbb{E}$, d.h. es ist $a < c < b$. Mit anderen Worten, $\mathbb{E}$ liegt dicht in $\mathbb{D}$.*

Beweis. Sei $a = z_0{,}z_1 z_2 \cdots < b$ und k kleinster Index mit $a < b_{.k}$. Falls $b_{.k} < b$, wähle man $c = b_{.k}$ womit die Behauptung offenbar erfüllt ist. Falls aber $b_{.k} = b$, gibt es wegen der Zulässigkeit von a sicher ein $n > k$ mit $z_n \neq 9$. Dann gilt für $c = z_0{,}z_1 \cdots z_k \cdots z_n^{+}$ sicher $a < c$, aber auch $c < b_{.k} = b$. ∎

Es sei dem Anfänger geraten, diesen Beweis anhand von Beispielen zu verfolgen. Sei etwa $a = 2{,}49996 \cdots$ und $b = 2{,}5001 \cdots$, so dass $a < b_{.1}$. Eine gemäß Beweisvorschrift konstruierte Zahl ist $c = 2{,}5$. Ist hingegen b selbst gleich 2,5 wähle man $c = 2{,}49997$.

[1)] Die Nachfolgeroperation $^{+}$ soll sich hier nur auf die letzte Ziffer beziehen. Da wir nicht strikt zwischen einer Ziffer als Symbol und der durch sie bezeichneten natürlichen Zahl unterscheiden, ist der Nachfolger z^{+} von $z < 9$ wieder eine Ziffer, so dass $z_0{,}z_1 \cdots z_n^{+}$ im Falle $n > 0$ nur für $z_n \neq 9$ wohlerklärt ist. Zum Beispiel ist $0{,}98^{+} = 0{,}99$ und $0{,}99^{+}$ ist nicht erklärt. Weil $z_0{,}z_1 \cdots z_n = z_0$ im Falle $n = 0$, sei naturgemäß $z_0{,}z_1 \cdots z_n^{+} = z_0 + 1$ in diesem Falle.

3.3 Der Satz von der oberen Grenze

Dieser Satz ist grundlegend für die gesamte Theorie der reellen Zahlen und Funktionen. Zu seiner Formulierung in Satz 3.2 werden einige Begriffe benötigt, die wir gleich für eine beliebige geordnete Menge M formulieren, obwohl sie zunächst nur für den Spezialfall der geordneten Mengen $\mathbb{E}$ und $\mathbb{D}$ wichtig sind.

Definition. Eine Teilmenge $X \subseteq M$ einer geordneten Menge M heißt *nach oben beschränkt*, wenn ein $b \in M$ existiert mit $x \leq b$ für alle $x \in X$. Jedes derartige b heißt eine *obere Schranke* für X. Ein Element $s \in M$ heißt *obere Grenze* oder *Supremum* für $X \subseteq M$, symbolisch $s = \sup X$, wenn s eine kleinste obere Schranke von X ist, d.h. es gelten (a): s ist obere Schranke von X, und (b): $s \leq b$ für jede obere Schranke b von X. **Achtung:** (b) ist wegen V gleichwertig zu (b'): Zu jedem $u < s$ gibt es ein $x \in X$ mit $u < x$, d.h. man kommt s von links mit Elementen aus X beliebig nahe.

Beispiel 1. Sei $a \in \mathbb{D}$. Ferner sei $X = \{a_{.0}, a_{.1}, a_{.2}, \ldots\}$ die durch a beschränkte Menge der kanonischen Näherungen von a. Wir behaupten $\sup X = a$. Ist nämlich $u < a$ beliebig gewählt, so ist $u < a_{.n}$ für hinreichend großes n gemäß Definition von $<$, so dass Bedingung (b') erfüllt ist. Insbesondere ist $\sup\{0{,}3\,;\, 0{,}33\,;\ldots\} = 0{,}333\cdots$.

Beispiel 2. Sei $X = \{z_0{,}9\,;\, z_0{,}99\,;\, z_0{,}999\,;\ldots\}$. Wir behaupten $\sup X = z_0 + 1$. Gewiß ist $z_0 + 1$ obere Schranke von X. Sei nun $u < z_0 + 1$. Falls $\operatorname{Int} u < z_0$ ist sicher $u < z_0$. Aber auch im Falle $\operatorname{Int} u = z_0$ gibt es – weil u keine Neunerperiode hat – offenbar ein n mit $u < z_0{,}\underbrace{9\cdots9}_{n}$. Allgemeiner sei $X \subseteq \mathbb{D}$, $n \geq 0$ und $s = z_0{,}z_1\cdots z_n^+$ obere Schranke von X. Ferner existiere zu jedem m ein $x \in X$ mit $z_0{,}z_1\cdots z_n\underbrace{9\cdots9}_{m} < x$. Dann ist $\sup X = s$. Denn es gibt ein hinreichend großes k mit $u < z_0{,}z_1\cdots z_n\underbrace{9\cdots9}_{k}$ und damit gilt auch $u < x$ für ein gewisses $x \in X$.

Völlig analog definiert man die Begriffe nach *unten beschränkt, untere Schranke* und *untere Grenze (Infimum)*. Eine eventuell existierende untere Grenze für X wird mit $\inf X$ bezeichnet. Eine *beschränkte* Teilmenge von M ist eine solche, die zugleich nach oben und nach unten beschränkt ist. Eine nach oben beschränkte Teilmenge X von M muss eine obere Grenze besitzen. Wenn sie aber eine solche hat, so ist diese eindeutig bestimmt. Denn sind a, b obere Grenzen für X, so ist $a \leq b$, weil a kleinste obere Schranke ist. Aus analogem Grunde ist auch $b \leq a$, also $a = b$. Obere und untere Grenze von X können, aber müssen nicht zu X gehören.

Jede Teilmenge von $\mathbb{D}$ ist nach unten beschränkt, z.B. durch 0. Daher bedeutet für $\mathbb{D}$ beschränkt dasselbe wie nach oben beschränkt. Die leere Teilmenge von $\mathbb{D}$ hat außerdem gerade das Supremum 0. Deswegen ist die Voraussetzung $X \neq \emptyset$ in Satz 3.2 unten eigentlich überflüssig! Hingegen hat $\emptyset$ in $\mathbb{D}$ kein Infimum. Das liegt natürlich daran, dass $\mathbb{D}$ ein kleinstes aber kein größtes Element hat. In einer geordneten Menge ohne kleinstes und größtes Element hat $\emptyset$ weder ein Supremum noch ein Infimum.

Es lässt sich sehr einfach zeigen, dass nicht jede beschränkte Teilmenge von $\mathbb{E}$ ein Supremum in $\mathbb{E}$ besitzt. Dieser Nachweis ist viel einfacher als in $\mathbb{Q}$ (siehe **6**). Hat nämlich $X \subseteq \mathbb{E}$ ein Supremum s bereits in $\mathbb{E}$, so ist s auch Supremum für X in $\mathbb{D}$, wie aus der Dichtheit von $\mathbb{E}$ in $\mathbb{D}$ sofort folgt. Hätte also etwa $X = \{0{,}3\,;\, 0{,}33\,;\, \ldots\}$ ein Supremum $s \in \mathbb{E}$, so wäre s auch Supremum für X in $\mathbb{D}$. Das aber kann nicht sein, denn $\sup X = 0{,}333\cdots$ liegt nicht in $\mathbb{E}$. Folglich hat X kein Supremum in $\mathbb{E}$. Anders sind die Verhältnisse in $\mathbb{D}$. Denn hier gilt

Satz 3.2. *Jede nichtleere beschränkte Teilmenge X von $\mathbb{D}$ besitzt ein Supremum.*

Beweis. Entweder hat X ein Supremum in $\mathbb{E}$ oder nicht. Im ersten Fall ist nichts zu beweisen. Es liege also der zweite Fall vor. Wir konstruieren jetzt schrittweise die Ziffern einer nichtabbrechenden Dezimalzahl $s = z_0{,}z_1z_2\cdots$ und beweisen $s = \sup X$.

Da X nichtleer und beschränkt ist, gibt es gewiss eine größte natürliche Zahl z_0, so dass gerade noch $z_0 \leq x$ für wenigstens ein $x \in X$. Seien nun $z_0, \ldots, z_n$ schon definiert, und zwar so, dass $z_0{,}z_1\cdots z_n \leq x$ für ein $x \in X$. Dann sei z_{n+1} die größte Ziffer, so dass noch $z_0{,}z_1\cdots z_{n+1} \leq x$ für ein $x \in X$. Weil nach Voraussetzung $z_0{,}z_1\cdots z_n0 \leq x$ für ein $x \in X$, ist z_{n+1} wohlbestimmt, und falls $z_{n+1} < 9$, ist $z_0{,}z_1\cdots z_n^+$ sogar obere Schranke von X. Damit ist $s = z_0{,}z_1z_2\cdots$ wohlerklärt. Als erstes zeigen wir, dass s zulässig ist. Wäre etwa $s = z_0{,}999\cdots$, so wäre nach Beispiel 2 $z_0 + 1$ Supremum von X, im Widerspruch dazu, dass wir uns im zweiten Fall befinden. Analog schließt man $s = z_0{,}z_1\cdots z_n999\cdots$ für $n > 0$ aus. Als nächstes zeigen wir, s ist obere Schranke für X. Sei $a \in X$. Falls $a = s$, gilt $a \leq s$ trivial. Falls $a \neq s$, sei k der kleinste Index mit $z_k \neq z_k^a$ ($= k$-te Ziffer von a). Aufgrund der Bestimmung von z_k gilt $z_k^a \leq z_k$, also $z_k^a < z_k$, daher $a < s$. Schließlich ist noch zu bestätigen, dass s wirklich das Supremum von X ist. Sei $u < s$ und etwa $u < z_0{,}z_1\cdots z_n$. Dann gibt es gemäß Konstruktion ein $x \in X$ mit $z_0{,}z_1\cdots z_n \leq x$, also $u < x$. Damit $s = \sup X$ und der Satz bewiesen. ∎

Man kann die Fallunterscheidung in diesem Beweis auch vermeiden, indem man die Konstruktion eines unzulässigen s in Kauf nimmt und s sodann durch die entsprechende zulässige Dezimalzahl ersetzt. Jedenfalls liefert dieser Beweis ein Verfahren zur schrittweisen Berechnung von $\sup X$ für konkret gegebenes X. So folgt nach Definition der Multiplikation in **5** unschwer, dass $s^2 = a$, wobei $s = \sup\{x \in \mathbb{E} \mid x^2 \leq a\}$. Deshalb kann man $s = \sqrt{a}$ für explizit gegebenes $a \in \mathbb{D}$ auch ohne ein spezielles Verfahren mit Satz 3.2 näherungsweise und mit beliebiger Genauigkeit notfalls von Hand ausrechnen. Der Satz ist aber ein mächtiges Existenzprinzip eben auch dann, wenn man bestenfalls nur weiß, dass X beschränkt ist. Dazu das subtile und lehrreiche Beispiel 3 unten. Dort wie auch anderswo benötigen wir folgende Definitionen.

Sind M und N geordnete Mengen, so heißt $f : M \to N$ *monoton wachsend* oder einfach *wachsend*, wenn $x < y \;\Rightarrow\; fx \leq fy$, für alle $x, y \in M$. Gilt $x < y \;\Rightarrow\; fx < fy$, so heißt f *streng monoton wachsend*, oder auch *ordnungstreu*. Teilmengen von M der Gestalt $[u, v] := \{x \in M \mid u \leq x \leq v\}$ mit $u < v$ heißen *Intervalle* oder deutlicher *abgeschlossene Intervalle*. Ist c ein Wert mit $fc = c$, so heißt c ein *Fixpunkt* von f.

Beispiel 3. Sei $I = [u, v]$ ein Intervall in $\mathbb{D}$ und $f : I \to I$ wachsend. Wir behaupten, f hat einen Fixpunkt in I. Es ist $s = \sup X$ ein solcher, mit $X := \{x \in I \mid x \leq fx\}$. Weil $u \leq fu$, ist $u \leq s$. Da v obere Schranke für X ist, gilt auch $s \leq v$, also $s \in I$. Daher ist fs wohlerklärt. $x \in X$ impliziert $fx \in X$, denn mit $x \leq fx$ ist auch auch $fx \leq f(fx)$. Damit gilt $x \leq fx \leq s$ für alle $x \in X$ und so $fx \leq fs$. Also ist fs obere Schranke für X und deshalb $s \leq fs$. Daher ist auch $s \in X$, also $fs \in X$ und somit $fs \leq s$. Insgesamt ergibt sich $s = fs$, d.h. s ist in der Tat ein Fixpunkt von f.

Von vielen Anwendungen des Satzes 3.2 nennen wir die Erklärungsmöglichkeit von Potenzen b^x mit beliebiger Basis $b > 0$ und irrationalen Exponenten x in **7.3**. Die Konstruktion von b^x ist im Grunde viel harmloser als die in Beispiel 3, das ziemlich unanschaulich ist weil die Funktion f dort einen total unstetigen Verlauf haben kann. b^x wird in **8.4** konstruktiv definiert. Hingegen ist ein Fixpunkt von $f : I \to I$ i.a. nur berechenbar, wenn f über bloße Monotonie hinaus weitere Zusatzeigenschaften hat.

3.4 Lückenlosigkeit der reellen Zahlen

Es ist nützlich, sich den Satz von der oberen Grenze auf unterschiedliche Weise zu veranschaulichen, z.B. durch die Lückenlosigkeit der Ordnung von $\mathbb{D}$. Dazu die folgenden ganz elementaren, auf DEDEKIND [5] zurückgehenden Definitionen.

Unter einem *Schnitt* einer geordneten Menge M versteht man ein Paar nichtleerer Teilmengen U, V von M derart, dass

$$(1)\ \ U \cup V = M, \quad (2)\ \ U \cap V = \emptyset, \quad (3)\ a \in U, b \in V \ \Rightarrow\ a < b.$$

U heißt die *Unterklasse* und V die *Oberklasse* des Schnitts (U, V). Offensichtlich gibt es nur die folgenden drei Typen von Schnitten:

I. U hat ein größtes und V ein kleinstes Element,
II. U hat kein größtes und V kein kleinstes Element,
III. U hat ein größtes und V kein kleinstes Element, oder umgekehrt.

Schnitte vom Typ I heißen *Sprünge*, solche vom Typ II *Lücken*, und Schnitte vom Typ III seien DEDEKIND*sche Schnitte* genannt. Figur 1 veranschaulicht die drei Typen.

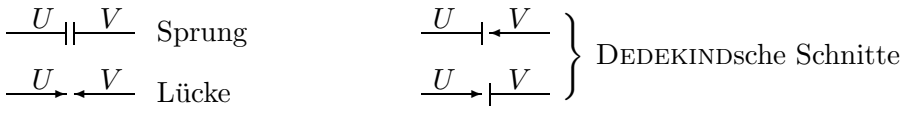


Fig. 1 Schnitt-Typen geordneter Mengen

Jedem DEDEKINDschen Schnitt (U, V) ist eindeutig ein so genanntes *Schnittelement* s zugeordnet, womit das größte Element von U oder das kleinste Element von V gemeint ist. Man sieht mit bloßem Auge, dass s sowohl Supremum von U als auch Infimum von V ist, unabhängig davon ob s zu U oder zu V gehört.

Sei M nun dicht geordnet, oder gleichwertig, M habe keine Sprünge und enthält mehr als ein Element. Hat in einem solchen Falle M auch keine Lücken, d.h. ist M *lückenlos* geordnet, so verbleiben offenbar nur DEDEKINDsche Schnitte. Umgekehrt hat eine geordnete Menge M mit nur DEDEKINDschen Schnitten natürlich keine Lücken. M erfüllt in diesem Falle auch den Satz von der oberen Grenze (d.h. Satz 3.2 mit M für $\mathbb{D}$). Denn sei $X \subseteq M$ nichtleer und beschränkt, V die Menge ihrer oberen Schranken, und $U := M \backslash V$. Dann ist das Schnittelement s des Schnitts (U, V) gerade das Supremum von M, wie leicht zu erkennen ist. Gilt andererseits in M der Satz von der oberen Grenze, so ist M auch lückenlos; denn das Supremum der Unterklasse eines Schnittes ist offenbar dessen Schnittelement. Damit haben wir folgenden Satz bewiesen:

Satz 3.3. *Für eine dicht geordnete Menge M sind gleichwertig*

(i) *M ist lückenlos geordnet,*
(ii) *Jeder Schnitt von M ist ein* DEDEKIND*scher,*
(iii) *In M gilt der Satz von der oberen Grenze.*

3.5 Übungen

1. Man bestätige durch Induktion über $n \geq 2$ folgende Behauptungen:
   (4) $(1+a)^n > 1 + na \quad (a > 0,$ BERNOULLI*sche Ungleichung*$)$.
   (5) $1 - (1-a)^n < na \quad (0 < a \leq 1)$,
   (6) $(1-a)^n + na \leq 1 + \binom{n}{2} a^2 \quad (0 < a \leq 1)$.

2. Der *Abstand* $|a-b|$ zweier Elemente eines $\mathcal{E}$-Bereichs sei wie folgt erklärt: Es sei $|a-b| = a-b$ für $a \leq b$ und $|a-b| = b-a$ sonst, d.h. $a < b$. Offenbar gelten dann $|a-b| = |b-a|$, sowie $|a-b| = 0 \Leftrightarrow a = b$. Man beweise die so genannte *Dreiecksungleichung* $|a-b| \leq |a-c| + |b-c|$.

3. Ein $\mathcal{E}$-Bereich A heißt *archimedisch* (geordnet), wenn es zu allen $a, b \in A_+$ ein n gibt mit $na > b$. Sicher ist $\mathbb{N}$ archimedisch. Man beweise, dass in einem archimedischen $\mathcal{E}$-Bereich A zu allen $a, b \in A_+$ mit $a > 1$ ein n mit $a^n > b$ existiert. Letzteres folgt noch nicht aus den in **2.3** aufgelisteten Potenzgesetzen.

4. Man beweise, ein lückenlos geordneter $\mathcal{E}$-Bereich A ist archimedisch (die Umkehrung gilt i.a. nicht wie das Beispiel des $\mathcal{E}$-Bereichs $\mathbb{E}$ zeigen wird).

5. Sei $f : \mathbb{E} \to \mathbb{D}$ monoton wachsend. Die Funktion $\bar{f} : \mathbb{D} \to \mathbb{D}$ sei erklärt durch
   $$\bar{f}a = \sup\{f(a_{.n}) \mid n \in \mathbb{N}\}$$
   und heiße die *natürliche Fortsetzung* von f auf $\mathbb{D}$. Man beweise
   (a) $\bar{f}$ ist Fortsetzung von f, d.h. $\bar{f}a = fa$ für alle $a \in \mathbb{E}$,
   (b) $\bar{f}$ ist wachsend, und mit f ist auch $\bar{f}$ strikt wachsend.

Abschnitt 4

Arithmetik der abbrechenden Dezimalzahlen

Vor einer Erklärung der Rechenoperationen für beliebige Dezimalzahlen definieren wir diese zuerst im Bereich $\mathbb{E}$ der endlichen oder abbrechenden Dezimalzahlen. Dies ist mehr als nur eine technische Vorbereitung auf die reelle Arithmetik. Das Rechnen mit endlichen Dezimalzahlen verdient schon deswegen spezielle Aufmerksamkeit, weil wir damit im Alltag ständig zu tun haben. Auch rechnen sowohl Taschenrechner als auch Hochleistungscomputer mit endlichen Dezimalzahlen [1)].

Bei der Summierung der Dezimalzahlen $z_0{,}z_1 \cdots z_n$ und $z'_0{,}z'_1 \cdots z'_m$ verfährt man bekanntlich wie folgt: Zuerst wird durch „Anhängen von Nullen“ die Stellenzahl angeglichen; kurz, es darf hierbei $m = n$ angenommen werden. Sodann summiert man die natürlichen Zahlen $z_0 z_1 \cdots z_n$ und $z'_0 z'_1 \cdots z'_n$ – d.h. man denkt sich die Kommata vorübergehend entfernt – und trennt in dieser Summe die letzten n Stellen durch das Komma ab. Ist hingegen das Produkt von $z_0{,}z_1 \cdots z_n$ und $z'_0{,}z'_1 \cdots z'_m$ zu bilden, wird $z_0 z_1 \cdots z_n$ mit $z'_0 z'_1 \cdots z'_m$ multipliziert, und in diesem Resultat werden sodann $n + m$ Stellen durch das Komma abgetrennt.

Diese einfachen Kommasetzungsregeln lassen sich leicht formal fassen. Auf diese Weise wird nicht nur die Ausführung der Rechenoperationen mit Dezimalzahlen auf diejenige mit natürlichen Zahlen zurückgeführt, sondern es lassen sich auch die Rechengesetze auf eine verblüffend einfache Weise beweisen. Rationale Zahlen oder auch nur eine spezielle Form der Bruchrechung treten hierbei nicht in Erscheinung. Es zeigt sich, dass entgegen einer landläufigen Meinung die Dezimalzahlarithmetik keineswegs der Bruchrechnung als eines theoretischen Fundaments bedarf. Was man hier benötigt ist lediglich die Gewissheit, dass die mehr als 1000 Jahre alten Rechenverfahren für Summe und Produkt natürlicher Zahlen im Dezimalsystem tatsächlich das gewünschte Ergebnis liefern, und dies setzen wir selbstverständlich voraus.

[1)] genau genommen, falls deren Chips speziell für kaufmännisches Rechnen ausgelegt sind; sonst wird intern mit Binärzahlen gerechnet, was aber keinen äußerlich erkennbaren Unterschied ausmacht.

4.1 Rechnen mit endlichen Dezimalzahlen

Zuerst führen wir, und zwar gleich für beliebige Dezimalzahlen, die Operationen $\overset{n}{\rightarrow}$ und $\overset{n}{\leftarrow}$ der n-fachen *Kommaverschiebung nach rechts* beziehungsweise *nach links* ein. Sei $a \in \mathbb{D}$. Dann entstehe $a\overset{n}{\rightarrow}$ aus a dadurch, dass in a das Komma um n Stellen nach rechts verschoben wird. Analog sei $a\overset{n}{\leftarrow}$ das Resultat der Kommaverschiebung in a um n Stellen nach links. Abbrechende Dezimalzahlen werden durch die Operationen $\overset{n}{\rightarrow}, \overset{n}{\leftarrow}$ wieder in abbrechende, nichtabbrechende in nichtabbrechende überführt. Eine n-stellige Dezimalzahl wird durch n-fache Kommaverschiebung nach rechts offenbar in eine natürliche Zahl überführt. So erhält man unter Beachtung von $\varepsilon_n = 0,\underbrace{0\cdots01}_{n}$

$$25{,}3\overset{2}{\rightarrow} = 2530 \;\;;\;\; 25{,}3\overset{3}{\leftarrow} = 0{,}0253 \;\;;\;\; 1\overset{n}{\rightarrow} = 10^n \;\;;\;\; 1\overset{n}{\leftarrow} = \varepsilon_n \;\;;\;\; \varepsilon_n\overset{n}{\rightarrow} = 1.$$

Es ist offensichtlich und sei nur nebenbei erwähnt, dass jedes $a \in \mathbb{E}_+$ sich in der Weise $a = z_0{,}z_1\cdots z_n\overset{k}{\rightarrow}$ oder $a = z_0{,}z_1\cdots z_n\overset{k}{\leftarrow}$ für geeignetes k so schreiben lässt, dass z_0 eine Dezimalziffer $\neq 0$ ist, die so genannte *Gleitkommadarstellung*. Z.B. ist $27 = 2{,}7\overset{1}{\rightarrow}$. Auf dem Monitor oder gegebenenfalls auch im Taschenrechner-Display erscheint a im ersten Falle als $z_0{,}z_1\cdots z_n\mathrm{E}k$, im zweiten Falle als $z_0{,}z_1\cdots z_n\mathrm{E}{-}k$. Rechnerintern wird das Exponentensymbol E lediglich als Kennzeichnung einer Kommaverschiebung (*shifting*) interpretiert. Diesbezüglich gelten offenbar die folgenden Regeln:

S_0 : $a\overset{0}{\rightarrow} = a\overset{0}{\leftarrow} = a$,

S_1 : $a\overset{n}{\rightarrow}\overset{m}{\rightarrow} = a\overset{n+m}{\longrightarrow}$; $a\overset{n}{\leftarrow}\overset{m}{\leftarrow} = a\overset{n+m}{\longleftarrow}$,

S_2 : $a\overset{m}{\leftarrow}\overset{n}{\rightarrow} = a\overset{n}{\rightarrow}\overset{m}{\leftarrow} = \begin{cases} a\overset{n-m}{\longrightarrow} & \text{falls } n \geq m, \\ a\overset{m-n}{\longleftarrow} & \text{falls } m \geq n, \end{cases}$

S_3 : $a\overset{n}{\rightarrow} < b\overset{n}{\rightarrow} \;\Leftrightarrow\; a < b \;\Leftrightarrow\; a\overset{m}{\leftarrow} < b\overset{m}{\leftarrow}$; $a\overset{n}{\rightarrow} = b\overset{n}{\rightarrow} \;\Leftrightarrow\; a = b \;\Leftrightarrow\; a\overset{m}{\leftarrow} = b\overset{m}{\leftarrow}$,

S_4 : $a\overset{n}{\rightarrow} = a \cdot 10^n$ für natürliche Zahlen a.

Wer $\mathsf{S}_1 - \mathsf{S}_4$ nicht für offensichtlich hält, kann auch streng induktiv vorgehen und muss nur beachten, dass $\overset{1}{\rightarrow}$ eine Bijektion von $\mathbb{E}$ ist mit der Umkehroperation $\overset{1}{\leftarrow}$, und dass $\overset{n}{\rightarrow}$ die n-fache Wiederholung von $\overset{1}{\rightarrow}$ darstellt. S_1 liefert $a\overset{n}{\rightarrow}\overset{m}{\rightarrow} = a\overset{n+m}{\longrightarrow} = a\overset{m+n}{\longrightarrow} = a\overset{m}{\rightarrow}\overset{n}{\rightarrow}$, und analog ist $a\overset{n}{\leftarrow}\overset{m}{\leftarrow} = a\overset{m}{\leftarrow}\overset{n}{\leftarrow}$. S_4 ergibt sich aus der dezimalen Darstellung natürlicher Zahlen und liefert die nützliche, später wesentlich verallgemeinerte Gleichung

S_5 : $(a+b)\overset{n}{\rightarrow} = a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow}$ für natürliche Zahlen a, b.

Es werde nun die Addition in $\mathbb{E}$ wie folgt erklärt:

Definition. Die Summe $a + b$ von $a, b \in \mathbb{E}$ sei erklärt durch $(a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow})\overset{n}{\leftarrow}$, wobei n eine beliebige gewählte gemeinsame Stellenzahl von a und b ist.

Danach ist z.B. $3{,}14 + 8{,}7 = (314 + 870)\overset{2}{\leftarrow} = 1184\overset{2}{\leftarrow} = 11{,}84$. Diese Definition ist nichts anderes als eine präzise Formulierung des üblichen Additionsverfahrens, das im Prinzip auch in den Chips elektronischer Rechner „verdrahtet“ ist. Der Dezimalpunkt bei Eingabe der Operanden bewirkt intern nur eine entsprechende Shift-Operation.

Die angegebene Definition hängt nur scheinbar von n ab. Wählt man einmal n, ein andermal n' in der geforderten Weise und ist etwa $n' = n + k$, so ergibt sich

$$\begin{aligned}(a\overset{n'}{\rightarrow} + b\overset{n'}{\rightarrow})\overset{n'}{\leftarrow} &= (a\overset{n}{\rightarrow}\overset{k}{\rightarrow} + b\overset{n}{\rightarrow}\overset{k}{\rightarrow})\overset{k}{\leftarrow}\overset{n}{\leftarrow} && \text{(gemäß } \mathsf{S}_1\text{)}\\ &= (a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow})\overset{k}{\rightarrow}\overset{k}{\leftarrow}\overset{n}{\leftarrow} && \text{(gemäß } \mathsf{S}_5\text{, weil } a\overset{n}{\rightarrow}, b\overset{n}{\rightarrow} \in \mathbb{N})\\ &= (a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow})\overset{n}{\leftarrow} && \text{(weil } c\overset{k}{\rightarrow}\overset{k}{\leftarrow} = c \text{ gemäß } \mathsf{S}_2\text{).}\end{aligned}$$

Definition. Das Produkt $a \cdot b$ von $a, b \in \mathbb{E}$ sei erklärt durch $(a\overset{n}{\rightarrow} \cdot b\overset{m}{\rightarrow})\overset{n+m}{\longleftarrow}$, wobei n und m Stellenzahlen von a bzw. b seien.

Beispiele. $2{,}31 \cdot 1{,}2 = (2{,}31\overset{2}{\rightarrow} \cdot 1{,}2\overset{1}{\rightarrow})\overset{3}{\leftarrow} = (231 \cdot 12)\overset{3}{\leftarrow} = 2772\overset{3}{\leftarrow} = 2{,}772$. Ferner ist $a \cdot 10^n = (a\overset{m}{\rightarrow} \cdot 10^n\overset{0}{\rightarrow})\overset{m}{\leftarrow} = (a\overset{m}{\rightarrow}\overset{n}{\rightarrow})\overset{m}{\leftarrow} = a\overset{n}{\rightarrow}$ für m-stelliges a, speziell $10\varepsilon_{n+1} = \varepsilon_n$. Analog ist $a \cdot \varepsilon_n = (a\overset{m}{\rightarrow} \cdot \varepsilon_n\overset{n}{\rightarrow})\overset{m+n}{\longleftarrow} = (a\overset{m}{\rightarrow} \cdot 1)\overset{m+n}{\longleftarrow} = a\overset{m}{\rightarrow}\overset{m+n}{\longleftarrow} = a\overset{n}{\leftarrow}$. Daher lässt sich zu jedem $a \in \mathbb{E}$ und jedem $b \in \mathbb{D}_+$ offenbar ein n so wählen, dass $a\varepsilon_n \leq b$. Ferner ist $(ab)\overset{n}{\rightarrow} = a \cdot b \cdot 10^n = a(b\overset{n}{\rightarrow})$; es ist also gleichgültig, wie $ab\overset{n}{\rightarrow}$ geklammert wird.

Auch das Produkt hängt von der Wahl der Stellenzahlen von a und b nicht ab. Wichtig vor allem ist, dass die Einschränkungen der so erklärten Operationen auf den Bereich $\mathbb{N}$ mit den dort bereits gegebenen Operationen übereinstimmen. Dies ist aber klar, denn man braucht in obigen Definitionen für den Fall $a, b \in \mathbb{N}$ ja nur $n = m = 0$ zu wählen. Die Operationen $+, \cdot$ auf $\mathbb{E}$ sind also Fortsetzungen derjenigen von $\mathbb{N}$. Deswegen dürfen sie auch mit den gleichen Symbolen bezeichnet werden.

Bemerkung. Es sei mit Nachdruck darauf hingewiesen, dass in den angegebenen Definitionen keine Willkür liegt. Für die Summe z.B. ergibt sich die Definition bis auf äquivalente Formulierung zwangsläufig aus ihrer Entsprechung zur Addition in Größenbereichen. Sie ist durch Dezimalteilung von Einheiten in Größenbereichen leicht zu veranschaulichen. Es genügt vorerst, sich dies anhand von anschaulichen Beispielen zu verdeutlichen, wie etwa in $3{,}14\,\mathtt{m} + 2{,}7\,\mathtt{m} = 314\,\mathtt{cm} + 270\,\mathtt{cm} = 584\,\mathtt{cm} = 5{,}84\,\mathtt{m}$. In **10.3** wird rigoros bewiesen, dass obige Definition der Summe sozusagen eine zwingende ist.

Hat man sich einmal überzeugt, dass die Summe so und nicht anders als angegeben definiert werden kann, so lässt sich allein aufgrund der Rechengesetze schon beweisen, dass auch das Produkt bis auf äquivalente Formulierung so und nicht anders definiert werden kann als dies geschehen ist, Übung 4. Ein anderes und ausschließlich didaktisches Problem ist es, die Multiplikation sachgerecht zu *veranschaulichen*. Möglichkeiten hierzu ergeben sich nicht nur bei der Flächenberechung und in der elementaren Kombinatorik, sondern ebenso bei der Deutung der Dezimalzahlen als Maßzahlen (Gleichung (1°) in **10.3**), oder als Operatoren, bei welchen die Multiplikation als Abbildungsprodukt in Erscheinung tritt. Auf diese Interpretation gehen wir in **10.7** ein.

Wir werden nun den einfachen Nachweis erbringen, dass $\mathbb{E}$ mit den oben erklärten Operationen einen $\mathcal{E}$-Bereich bildet, also sämtliche in **2.1** angegebenen Axiome überprüfen. Danach darf man mit den endlichen Dezimalzahlen wie gewohnt rechnen. Natürlich ist damit nicht nur bloßes Ausrechnen gemeint, sondern das bewusste und zielorientierte Anwenden der bekannten Rechengesetze, zum Beispiel bei der Termumformung.

4.2 Nachweis der Rechengesetze

Die Eigenschaften V, F und G sind offensichtlich. Es seien nun a, b, c beliebige Elemente aus $\mathbb{E}$. Ferner sei n gemeinsame Stellenzahl von a, b, c, so dass $a\overset{n}{\rightarrow}$, $b\overset{n}{\rightarrow}$, $c\overset{n}{\rightarrow}$ natürliche Zahlen sind. Dann ergeben sich die weiteren Rechengesetze in folgender Weise, wobei ausschließlich deren Gültigkeit in $\mathbb{N}$ verwendet wird.

N^+: $\quad a + 0 = (a\overset{n}{\rightarrow} + 0\overset{n}{\rightarrow})\overset{n}{\leftarrow} = a\overset{n}{\rightarrow}\overset{n}{\leftarrow} = a.$

$\mathsf{N}^\times$: $\quad a \cdot 1 = (a\overset{n}{\rightarrow} \cdot 1\overset{0}{\rightarrow})\overset{n}{\leftarrow} = a\overset{n}{\rightarrow}\overset{n}{\leftarrow} = a.$

K^+: $\quad a + b = (a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow})\overset{n}{\leftarrow} = (b\overset{n}{\rightarrow} + a\overset{n}{\rightarrow})\overset{n}{\leftarrow} = b + a.$

Analog verläuft der Nachweis von Axiom $\mathsf{K}^\times$.

$$\begin{aligned}\mathsf{A}^+: \quad a + (b + c) &= a + (b\overset{n}{\rightarrow} + c\overset{n}{\rightarrow})\overset{n}{\leftarrow} \\ &= (a\overset{n}{\rightarrow} + (b\overset{n}{\rightarrow} + c\overset{n}{\rightarrow})\overset{n}{\leftarrow}\overset{n}{\rightarrow})\overset{n}{\leftarrow} = (a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow} + c\overset{n}{\rightarrow})\overset{n}{\leftarrow}.\end{aligned}$$

Derselbe Wert ergibt sich für $(a + b) + c$, womit A^+ bestätigt wurde.

$\mathsf{A}^\times$: $\quad a \cdot (b \cdot c) = (a\overset{n}{\rightarrow} \cdot (b\overset{n}{\rightarrow} \cdot c\overset{n}{\rightarrow})\overset{2n}{\leftarrow}\overset{2n}{\rightarrow})\overset{3n}{\leftarrow} = (a\overset{n}{\rightarrow} \cdot b\overset{n}{\rightarrow} \cdot c\overset{n}{\rightarrow})\overset{3n}{\leftarrow}.$

Derselbe Wert ergibt sich für $(a \cdot b) \cdot c$.

$$\begin{aligned}\mathsf{D}: \quad (a + b) \cdot c &= (a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow})\overset{n}{\leftarrow} \cdot c \\ &= \big((a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow})\overset{n}{\leftarrow}\overset{n}{\rightarrow} \cdot c\overset{n}{\rightarrow}\big)\overset{2n}{\leftarrow} = \big((a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow}) \cdot c\overset{n}{\rightarrow}\big)\overset{2n}{\leftarrow} \\ &= (a\overset{n}{\rightarrow} \cdot c\overset{n}{\rightarrow} + b\overset{n}{\rightarrow} \cdot c\overset{n}{\rightarrow})\overset{2n}{\leftarrow} \qquad \text{(weil } \mathsf{D} \text{ in } \mathbb{N} \text{ gilt)} \\ &= (a\overset{n}{\rightarrow} \cdot c\overset{n}{\rightarrow})\overset{2n}{\leftarrow} + (b\overset{n}{\rightarrow} \cdot c\overset{n}{\rightarrow})\overset{2n}{\leftarrow} \qquad \text{(Definition der Summe)} \\ &= a \cdot c + b \cdot c \qquad \text{(Definition des Produkts).}\end{aligned}$$

Es verbleibt der Nachweis von Axiom E. Für $c \in \mathbb{E}$ ist $c\overset{n}{\rightarrow} = c \cdot 10^n$. Also folgt mit D

(1) $\quad (a + b)\overset{n}{\rightarrow} = a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow} \qquad (a, b \in \mathbb{E}).$

Seien nun a, b n-stellig und zuerst angenommen, $a + d = b$ mit $d \neq 0$. (1) ergibt $a\overset{n}{\rightarrow} + d\overset{n}{\rightarrow} = b\overset{n}{\rightarrow}$. Weil $a\overset{n}{\rightarrow}, b\overset{n}{\rightarrow} \in \mathbb{N}$, ist $d\overset{n}{\rightarrow} \in \mathbb{N}$, und weil mit $d \neq 0$ auch $d\overset{n}{\rightarrow} \neq 0$, ist $a\overset{n}{\rightarrow} < b\overset{n}{\rightarrow}$. Folglich ist $a < b$ gemäß S_3. Sei nun das Letztere vorausgesetzt und $a\overset{n}{\rightarrow}, b\overset{n}{\rightarrow} \in \mathbb{N}$. Nach S_3 ist $a\overset{n}{\rightarrow} < b\overset{n}{\rightarrow}$. Daher gibt es ein $c \in \mathbb{N}_+$ mit $a\overset{n}{\rightarrow} + c = b\overset{n}{\rightarrow}$. Für $d := c\overset{n}{\leftarrow}$ ergibt sich dann $a + d = (a\overset{n}{\rightarrow} + d\overset{n}{\rightarrow})\overset{n}{\leftarrow} = (a\overset{n}{\rightarrow} + c)\overset{n}{\leftarrow} = b\overset{n}{\rightarrow}\overset{n}{\leftarrow} = b$.

Damit sind sämtliche Axiome nachgewiesen und $\mathbb{E}$ hat sich als ein Elementarbereich erwiesen, der im Gegensatz zu $\mathbb{N}$ dicht geordnet ist. Es gelten für $\mathbb{E}$ dann automatisch auch alle Folgerungen der Axiome, wie etwa die Monotoniegesetze. Eine einfache, kaum bemerkbare Anwendung von M^+ liefert z.B. für jedes n die folgende, im weiteren oft verwendete Ungleichung, deren linke Hälfte natürlich klar ist:

(2) $\quad a_{.n} \leq a < a_{.n} + \varepsilon_n \qquad (a \in \mathbb{D}).$

Denn sei $a = z_0{,}z_1z_2\cdots$ und $a_{.n} + \varepsilon_n = z_0'{,}z_1'\cdots z_n'$. Weil $a_{.n} < a_{.n} + \varepsilon_n$, gibt es ein $k \leq n$ mit $z_i = z_i'$ für $i < k$, aber $z_k < z_k'$. Also gilt $a < a_{.n} + \varepsilon_n$ nach Definition von $<$.

4.3 Division endlicher Dezimalzahlen

Die Division ist im Zahlbereich $\mathbb{E}$ i.a. nicht ausführbar wie sich an einfachen Beispielen verdeutlichen lässt. So ist zum Beispiel dort die Gleichung $3 \cdot x = 1$ unlösbar. Denn angenommen es gibt ein $z_0{,}z_1 \cdots z_n$ mit $3 \cdot z_0{,}z_1 \cdots z_n = (3 \cdot z_0 \cdots z_n)\overset{n}{\leftarrow} = 1$, oder gleichwertig $3 \cdot z_0 \cdots z_n = 10^n$, wobei $z_n \neq 0$ angenommen werden kann. Für keine von 0 verschiedene Ziffer z_n endet aber $3 \cdot z_n$ auf 0 was notwendig der Fall sein müsste, wäre $3 \cdot z_0 \cdots z_n = 10^n$. Alternativ lässt sich hier auch zahlentheoretisch argumentieren.

Falls die Gleichung $b \cdot x = a$ für $a, b \in \mathbb{E}$ mit $b \neq 0$ lösbar ist, so nur auf eine Weise; denn $bc_1 = bc_2 \;\Rightarrow\; c_1 = c_2$ nach der Kürzungsregel. Die Lösung wird im Falle ihrer Existenz mit $\frac{a}{b}$ bezeichnet und heißt der *Quotient von a durch b*. In **6** wird bewiesen, dass für $a, b \in \mathbb{E}$ die Gleichung $b \cdot x = a$ mit $b \neq 0$ *in* $\mathbb{D}$ stets lösbar ist. Nur ist der Quotient $\frac{a}{b}$ dann in der Regel nicht abbrechend. Anders als in $\mathbb{N}$ ist in $\mathbb{E}$ für $a \in \mathbb{E}$ und jedes n immerhin die Gleichung $10^n x = a$ lösbar, nämlich durch $x = a\overset{n}{\leftarrow}$; denn $10^n a\overset{n}{\leftarrow} = a\overset{n}{\rightarrow}\overset{n}{\leftarrow} = a$. Man kann also in $\mathbb{E}$ stets durch 10^n dividieren und das Resultat entsteht durch n-fache Kommaverschiebung nach links; kurzum, $\frac{a}{10^n} = a\overset{n}{\leftarrow} = a\varepsilon_n$. Ist a speziell eine Dezimalziffer z, liefert dies $\frac{z}{10^n} = z\overset{n}{\leftarrow} = 0{,}\underbrace{0\cdots0z}_{n}$. Demnach gilt

$$z_0 + \tfrac{z_1}{10} + \ldots + \tfrac{z_n}{10^n} = z_0 + 0{,}z_1 + 0{,}0z_2 + \ldots + 0{,}0\cdots0z_n.$$

Die Ausrechnung des rechten Terms mit dem vertrauten Additionsverfahren für mehrere Summanden, das ja nur eine zweckmäßige Schreibweise unseres auf n Summanden erweiterten Additionsverfahrens für endliche Dezimalzahlen darstellt, ergibt

$$\begin{array}{l} \quad z_0 \\ +\ 0{,}z_1 \\ +\ 0{,}0z_2 \\ \quad\vdots \\ +\ 0{,}0\cdots0z_n \\ \hline = z_0{,}z_1 \cdots z_n. \end{array}$$

Damit haben wir die wichtige und wohlbekannte, in **3.1** schon erwähnte Gleichung

$$(3) \qquad z_0{,}z_1 \cdots z_n = z_0 + \tfrac{z_1}{10} + \ldots + \tfrac{z_n}{10^n}$$

bewiesen, die uns in Abschnitt **7** zu den unendlichen Reihen führen wird. Man möge sich vergegenwärtigen, dass eine Bruchrechnung für die Herleitung dieser Gleichung nicht erforderlich war, denn die Terme $\frac{z_n}{10^n}$ $(= z_n\overset{n}{\leftarrow})$ sind wohlbestimmte Dezimalzahlen [2)].

In $\mathbb{E}$ kann nicht nur durch 10, sondern auch durch 2, durch 5, und allgemein durch jede natürliche Zahl der Form $2^i \cdot 5^k$ dividiert werden, und i.a. nur durch diese, Übung 1. So gilt $\frac{a}{2} = 5a\overset{1}{\leftarrow}$, weil $a = 2 \cdot 5 \cdot a\overset{1}{\leftarrow}$, und analog ist $\frac{a}{5} = 2a\overset{1}{\leftarrow}$. Die letzte Gleichung ist nützlich für das praktische Rechnen: um eine Zahl durch 5 zu teilen, muss diese nur verdoppelt und im Ergebnis das Komma um eine Stelle verschoben werden.

[2)] Wer besonders pingelig sein will, kann (3) auch induktiv über n beweisen. Dafür benötigt man nur die leicht direkt beweisbare Gleichung $z_0{,}z_1 \cdots z_n = z_0{,}z_1 \cdots z_{n-1} + z_n\overset{n}{\leftarrow}$ $(n > 0)$.

Ganz wie in $\mathbb{E}$ hat die Gleichung $b \cdot x = a$ für $b \neq 0$ wegen der Streichungsregel in jedem $\mathcal{E}$-Bereich nur höchstens eine Lösung, welche stets mit $\frac{a}{b}$ bezeichnet wird, sofern sie existiert. Weil zum Beispiel $(1-c)\cdot(1+c+\ldots+c^n) = 1-c^{n+1}$ für $0<c<1$ nach (3) in **2**, gilt in jedem $\mathcal{E}$-Bereich die wichtige Formel

(4) $\quad \frac{1-c^{n+1}}{1-c} = 1+c+c^2+\ldots+c^n \quad (0<c<1).$

Auch $2x = n(n-1)$, und allgemeiner $k!\cdot x = n(n-1)\ldots(n-k+1)$, ist für $k \leq n$ bereits in $\mathbb{N}$ lösbar, nämlich durch $x = \binom{n}{k}$ (Übung 5). Dabei heißt die durch durch $0! = 1$ und $(k+1)! = k!\cdot(k+1)$ definierte Funktion von $\mathbb{N}$ nach $\mathbb{N}$ die *Fakultät-Funktion*, kurz $k! = 1\cdot 2\cdot\ldots\cdot k$. Die soeben behauptete Identität $k!\cdot\binom{n}{k} = n(n-1)\ldots(n-k+1)$ ist nach Definition des Quotienten gleichwertig mit der leichter merkbaren Formel

(5) $\quad \binom{n}{k} = \frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{1\cdot 2\cdot\ \ldots\ \cdot k}.$

Der Zähler rechts in (5) ist genau genommen erklärt als $\prod_{i<k}(n-i)$. Damit erkennt man sofort, dass dieser auch für $k=0$ wohldefiniert ist und den Wert 1 hat. Die Anfangsbedingungen der beiden folgenden, von einer gegebenen Termfolge $t_0, t_1, \ldots$ ausgehenden rekursiven Definitionen sind nämlich in der Mathematik überall die gleichen:

$$\textstyle\sum_{i<0} t_i = 0\ ,\ \sum_{i<k+1} t_i = \sum_{i<k} t_i + t_k \quad ; \quad \prod_{i<0} t_i = 1\ ,\ \prod_{i<k+1} t_i = \prod_{i<k} t_i \cdot t_k.$$

Im Sinne der Konvention $\prod_{i<0} t_i = 1$ ist übrigens dann auch $k! = \prod_{i=1}^{k} i = 1$ für $k=0$.

4.4 Übungen

1. Man zeige, die Gleichung $n\cdot x = m$ ist für teilerfremde positive m, n in $\mathbb{E}$ lösbar genau dann, wenn n außer 2 und 5 keine weiteren Primteiler hat.

2. Seien $a, b \in \mathbb{E}$ n-stellig. Man beweise $a < b \ \Leftrightarrow\ a + \varepsilon_n \leq b$.

3. Sei $b \in \mathbb{D}$ und $c \in \mathbb{E}$ n-stellig. Man zeige: $c = b_{.n}$ genau dann, wenn $c \leq b < c+\varepsilon_n$. Für $n=0$ bedeutet dies speziell $k = \operatorname{Int} b$ genau dann, wenn $k \leq b < k+1$.

4. Sei $a+b$ für $a, b \in \mathbb{E}$ wie im Text definiert. Man zeige unter alleiniger Benutzung der Rechengesetze für $\mathcal{E}$-Bereiche, dass notwendigerweise $a\cdot b = (a\xrightarrow{n}\cdot\, b\xrightarrow{m})\xleftarrow{n+m}$ für n-stelliges a und m-stelliges b.

5. Man beweise (5). Demnach gilt speziell $1+2+\ldots+n = \binom{n+1}{2} = \frac{n(n+1)}{2} \quad (n>0)$.

6. Die n-te *Rundung* $a_{(n)}$ einer Dezimalzahl $a = z_0{,}z_1z_2\cdots \in \mathbb{D}$ ist wie folgt erklärt:

$$a_{(n)} = \begin{cases} a_{.n} & \text{falls } z_{n+1} \leq 4, \\ a_{.n} + \varepsilon_n & \text{falls } z_{n+1} \geq 5. \end{cases}$$

Man beweise $a_{(n)} = (a_{.n+1} + 5\varepsilon_{n+1})_{.n}$ $\quad\big(= (\operatorname{Int} a\xrightarrow{n+1} + 5)\xleftarrow{n+1}{}_{.n}\big)$.

Für die wohlbekannte Kreiszahl $\boldsymbol{\pi} = 3{,}141592653589793\cdots$ ergibt sich demnach

$$\boldsymbol{\pi}_{(2)} = (3141+5)\xleftarrow{3}{}_{.2} = 3146\xleftarrow{3}{}_{.2} = 3{,}14 \quad ; \quad \boldsymbol{\pi}_{(3)} = (31415+5)\xleftarrow{4}{}_{.3} = 3{,}142.$$

Abschnitt 5

Arithmetik der nichtnegativen reellen Zahlen

Angesichts der im letzten Abschnitt deutlich gewordenen Mängel des Zahlenbereichs $\mathbb{E}$ hinsichtlich Division werden nunmehr die arithmetischen Operationen auf $\mathbb{D}$ erweitert. Damit erreicht man mehr, als man bei oberflächlicher Betrachtung zu erhoffen wagte. $\mathbb{D}$ wird nämlich zu einem $\mathcal{E}$-Bereich, in welchem nicht nur unbeschränkt dividiert werden kann – Division durch 0 natürlich ausgenommen – sondern darüber hinaus hat $\mathbb{D}$ wegen der Gültigkeit des Satzes 3.2 von der oberen Grenze alle jene Eigenschaften, welche die reelle Analysis überhaupt erst ermöglichen. In diesem Sinne ist die im vorliegenden Abschnitt begründete reelle Arithmetik im Verbund mit Satz 3.2 das Fundament, auf dem sich das Riesengebäude der Analysis erhebt. Erst dieser Abschnitt rechtfertigt, die Elemente aus $\mathbb{D}$ als reelle Zahlen zu bezeichnen. Zwar fehlen zunächst noch die negativen Zahlen, aber dies ist unwesentlich; mehr noch, es macht die mit dem Grenzwert zusammenhängenden Betrachtungen übersichtlicher. Was für die Definitionen der arithmetischen Operationen benötigt wird, ist übrigens sehr bescheiden. Nur die Grenzwerte von leicht beherrschbaren monotonen und beschränkten Folgen sind zu betrachten.

Anschaulich ist klar, dass die Summe $a + b$ reeller Dezimalzahlen a, b durch die Summe $a_{.n} + b_{.n}$ um so besser approximiert wird, je größer n ist. In genau dieser Weise – und entsprechende Bemerkungen beziehen sich auf das Produkt – wurde die Summe reeller Zahlen erstmals in der Praxis definiert, zum Beispiel bei der Erstellung der Logarithmentafeln zu Beginn des 17. Jahrhunderts. Das geschah in einem aus heutiger Sicht blindem Glauben an die Rechtmäßigkeit dieses Verfahrens. Tatsächlich sind ja diese Verhältnisse anschaulich so klar, dass die Frage berechtigt ist, ob es hier überhaupt einer Rechtfertigung bedarf. Diese Frage muss allerdings klar bejaht werden. Denn es geht hier nicht um die Ausführbarkeit irgendwelcher Operationen, sondern um den Nachweis gewisser, meistens stillschweigend benutzter Rechengesetze. Diese Gesetze wurden von den Rechenexperten des 17. Jahrhunderts bei der Berechnung von Logarithmen- und anderen Tafeln gewissermaßen nur experimentell überprüft.

5.1 Schlichte Folgen

Wir betrachten zuerst Folgen $\langle a_n \rangle$ aus einer beliebigen geordneten Menge M. Eine solche Folge M heißt (monoton) *wachsend*, falls stets $a_n \leq a_{n+1}$, (monoton) *fallend*, falls stets $a_n \geq a_{n+1}$, und *beschränkt*, wenn $\{a_n \mid n \in \mathbb{N}\}$ in M beschränkt ist.

Definition. Eine wachsende beschränkte Folge $\langle a_n \rangle$ heiße eine *schlichte* Folge. Falls $a = \sup\{a_n \mid n \in \mathbb{N}\}$ existiert, heiße a der *Grenzwert* oder *Limes* von $\langle a_n \rangle$, symbolisch $a = \lim_{n\to\infty} a_n$, oder $a = \lim\langle a_n \rangle$. Fällt $\langle a_n \rangle$ monoton und existiert $b = \inf\{a_n \mid n \in \mathbb{N}\}$, so heiße b der Grenzwert von $\langle a_n \rangle$. In beiden Fällen sagt man, $\langle a_n \rangle$ ist *konvergent* oder *konvergiert* oder *strebt* gegen ihren Grenzwert.

In diesem Abschnitt betrachten wir nur schlichte Folgen aus $\mathbb{D}$. Derartige Folgen sind nach Satz 3.2 stets konvergent, und nach Übung 3.4 ebenso beliebige fallende Folgen aus $\mathbb{D}$. Zum Beispiel ist für $a \in \mathbb{D}$ die Näherungsfolge $\langle a_{.n} \rangle$ schlicht, denn sie wächst und ist durch a beschränkt. Daher ist definitionsgemäß $\lim\langle a_{.n} \rangle = a$. Falls $\langle a_n \rangle$ ab einer gewissen Stelle k konstant ist, d.h. falls $a_n = a_k$ für alle $n \geq k$, so ist $\lim\langle a_n \rangle$ natürlich mit a_k identisch. Dieser Fall liegt für die Folge $\langle a_{.n} \rangle$ genau dann vor, wenn $a \in \mathbb{E}$.

Konvergenz im Sinne der obigen Definition wird beweisbar, wenn man von der folgenden, etwas allgemeineren Definition ausgeht, welche keine Ordnung, dafür aber das Vorhandensein einer geeigneten Abstandsfunktion $|\ |: M \times M \to \mathbb{D}$ voraussetzt. Diese haben wir zwar in jedem $\mathcal{E}$-Bereich (Übung 3.2), aber wir wollen ja erst zeigen, dass $\mathbb{D}$ ein solcher ist! Deswegen benutzen wir vorerst nur die anfängliche Definition. Die allgemeinere, statt auf einer Ordnung auf einer Abstandsfunktion beruhende Definition lautet: $\langle a_n \rangle$ *konvergiert* und es sei $\lim\langle a_n \rangle = a$, falls es zu jedem $\varepsilon \in \mathbb{D}_+$ ein k gibt mit $|a - a_n| < \varepsilon$ für alle $n \geq k$. Dass für schlichte Folgen aus $\mathbb{D}$ ihr auf die eine oder andere Weise definierter Grenzwert mit dem Supremum übereinstimmt, ist der Grund, warum diese Folgen in der Regel so einfach zu beherrschen sind.

Alle Begriffe übertragen sich sinngemäß auf Folgen $\langle a_n \rangle_{n\geq k}$. Mit $\langle a_n \rangle$ ist offenbar auch $\langle a_n \rangle_{n\geq k}$ schlicht, und weil $\sup\{a_n \mid n \in \mathbb{N}\} = \sup\{a_n \mid n \geq k\}$, gilt $\lim\langle a_n \rangle = \lim\langle a_n \rangle_{n\geq k}$ für jedes k. Kurzum, man darf von einer schlichten Folge ein beliebiges „Anfangsstück“ weglassen und erhält eine schlichte Folge mit demselben Grenzwert.

5.2 Erweiterung der Rechenoperationen

Im folgenden werden wir zunächst den Fall schlichter Folgen vor uns haben, deren Glieder sämtlich aus $\mathbb{E}$ stammen. Sind $\langle a_n \rangle$ und $\langle b_n \rangle$ zwei derartige Folgen, so können auch die Folgen $\langle a_n + b_n \rangle$ und $\langle a_n \cdot b_n \rangle$ betrachtet werden. Es ist wegen der in $\mathbb{E}$ gültigen Monotoniegesetze offensichtlich, dass es sich hierbei wieder um schlichte Folgen handelt, die in $\mathbb{D}$ demnach wohlbestimmte Grenzwerte haben.

Definition. Die Summe $a + b$ reeller Zahlen $a, b \in \mathbb{D}$ sei erklärt als $\lim\langle a_{.n} + b_{.n} \rangle$, das Produkt $a \cdot b$ als $\lim\langle a_{.n} \cdot b_{.n} \rangle$.

So ist $3 \cdot 0{,}33\cdots = \lim\langle 3 \cdot 0, \underbrace{3\cdots 3}_{n}\rangle = \lim\langle 0, \underbrace{9\cdots 9}_{n}\rangle = 1$. Also löst $x = 0{,}333\cdots$ die Gleichung $3 \cdot x = 1$. Nach der in **4.3** eingeführten Bezeichnung ist also $\frac{1}{3} = 0{,}333\cdots$. Aus der Definition ergeben sich ferner sehr einfach die Formeln $a\overset{n}{\rightarrow} = a \cdot 10^n$ und $a\overset{n}{\leftarrow} = \frac{a}{10^n}$ (Übung 1). Genau wie in $\mathbb{E}$ bedeutet also auch in $\mathbb{D}$ Kommaverschiebung dasselbe wie Multiplikation bzw. Division mit einer entsprechenden Potenz von 10.

Vor weiteren Betrachtungen haben wir uns zunächst erst einmal zu überzeugen, dass Summe und Produkt – eingeschränkt auf $\mathbb{E}$ – mit den dort bereits gegebenen Operationen übereinstimmen. Dies aber ist leicht erkennbar. Sei k so groß, dass $a = a_{.n}$ und $b = b_{.n}$ für alle $n \geq k$. Dann ist die Folge $\langle a_{.n} + b_{.n}\rangle$ von der Stelle k ab konstant; also $\lim\langle a_{.n} + b_{.n}\rangle = a_{.k} + b_{.k} = a + b$, wo die rechts stehende Summe diejenige im alten Sinne ist. Das zeigt die Übereinstimmung. Dieselbe Überlegung gilt auch für das Produkt.

Für den Nachweis einiger Rechenregeln verwenden wir vorübergehend folgendes nützliche Kriterium. Darin ist ε_n die in **3.2** definierte n-stellige Dezimalzahl $0{,}0\cdots 01$.

Kriterium. *Sei $c \in \mathbb{E}_+$ und seien $\langle a_n\rangle$, $\langle b_n\rangle$ schlichte Folgen endlicher Dezimalzahlen mit $\lim\langle a_n\rangle = a$ und $\lim\langle b_n\rangle = b$. Dann sind äquivalent*

(i) $a \leq b$; (ii) *zu jedem n gibt es ein m mit $a_n \leq b_m + c\varepsilon_n$.*

Beweis. Sei (i) erfüllt und n vorgegeben. Entweder ist $a_n \leq c\varepsilon_n$ oder aber $c\varepsilon_n < a_n$. Im ersten Falle gilt natürlich $a_n \leq b_m + c\varepsilon_n$, sogar für beliebiges m. Im zweiten Falle ist $a_n - c\varepsilon_n < a_n \leq a \leq b$. Daher ist $a_n - c\varepsilon_n \leq b_m$ für ein gewisses m, also $a_n \leq b_m + c\varepsilon_n$ wie verlangt. Umgekehrt sei (ii) erfüllt und angenommen, dass $b < a$, also $b < a_k$ für ein gewisses k. Sei $d \in \mathbb{E}$ so gewählt, dass $b < d < a_k$. Für ein hinreichend großes $n \geq k$ ist dann $c\varepsilon_n \leq a_k - d$, also $d + c\varepsilon_n \leq a_k \leq a_n$, d.h. $d + c\varepsilon_n \leq a_n$. Da $b_m < d$, ergibt sich $b_m + c\varepsilon_n < a_n$ für alle m, was (ii) widerspricht. ∎

Satz 5.1. *Sind $\langle a_n\rangle, \langle b_n\rangle$ schlichte Folgen endlicher Dezimalzahlen, so gelten*

(1) $\lim\langle a_n + b_n\rangle = \lim\langle a_n\rangle + \lim\langle b_n\rangle$; (2) $\lim\langle a_n \cdot b_n\rangle = \lim\langle a_n\rangle \cdot \lim\langle b_n\rangle$.

Beweis. Sei $a := \lim\langle a_n\rangle$, $b := \lim\langle b_n\rangle$, also $\lim\langle a_n\rangle = \lim\langle a_{.n}\rangle$, $\lim\langle b_n\rangle = \lim\langle b_{.n}\rangle$. Es ist $\lim\langle a_n + b_n\rangle = \lim\langle a_{.n} + b_{.n}\rangle$ zu beweisen. Nach dem Kriterium (angewandt mit $c = 1$) gibt es zu jedem n ein m mit $a_n \leq a_{.m} + \varepsilon_n$ und $b_n \leq b_{.m} + \varepsilon_n$ (o.B.d.A. kann hier dasselbe m für beide Folgen genommen werden). Daher ist $a_n + b_n \leq a_{.m} + b_{.m} + 2\varepsilon_n$ für alle n, und abermalige Anwendung des Kriteriums ergibt $\lim\langle a_n + b_n\rangle \leq \lim\langle a_{.n} + b_{.n}\rangle$. Ganz analog beweist man $\lim\langle a_{.n} + b_{.n}\rangle \leq \lim\langle a_n + b_n\rangle$ und (1) ist bewiesen. Zum Beweis von (2) beachte man, $a_n \leq a_{.m} + \varepsilon_n$ und $b_n \leq b_{.m} + \varepsilon_n$ ergeben

$$\begin{aligned} a_n b_n &\leq (a_{.m} + \varepsilon_n)(b_{.m} + \varepsilon_n) \\ &= a_{.m}b_{.m} + (a_{.m} + b_{.m} + \varepsilon_n)\varepsilon_n \leq a_{.m}b_{.m} + (r + s + 1)\varepsilon_n, \end{aligned}$$

wobei $r, s \in \mathbb{E}$ mit $a \leq r$, $b \leq s$ beliebig gewählt seien. Nach dem Kriterium – man setze dort $c = r + s + 1$ – ist mithin $\lim\langle a_n \cdot b_n\rangle \leq \lim\langle a_{.n} \cdot b_{.n}\rangle$. Völlig analog zeigt man $\lim\langle a_{.n} \cdot b_{.n}\rangle \leq \lim\langle a_n \cdot b_n\rangle$, womit auch (2) bewiesen ist. ∎

5.3 Nachweis der Rechengesetze

Nunmehr werden wir die recht einfache Aufgabe erledigen, auch für $\mathbb{D}$ alle Axiome für $\mathcal{E}$-Bereiche aus **2.1** nachzuweisen. V und G sind trivial. Axiom F folgt daraus, dass für $a, b \neq 0$ sicher $a_{.n}, b_{.n} > 0$ für ein n, also $0 < a_{.n} \cdot b_{.n} \leq \lim\langle a_{.n} \cdot b_{.n}\rangle = a \cdot b$.

N^+: Weil $0_{.n} = 0$ für alle n, ist $a + 0 = \lim\langle a_{.n} + 0\rangle = \lim\langle a_{.n}\rangle = a$. Völlig analog beweist man $\mathsf{N}^\times$. Auch sind die Kommutativgesetze unmittelbar klar.

A^+: $a + (b + c) = \lim\langle a_{.n}\rangle + \lim\langle b_{.n} + c_{.n}\rangle = \lim\langle a_{.n} + b_{.n} + c_{.n}\rangle$ nach Satz 5.1. Derselbe Wert ergibt sich für $(a + b) + c$. Ganz analog beweist man auch $\mathsf{A}^\times$.

D: $(a + b) \cdot c = \lim\langle a_{.n} + b_{.n}\rangle \cdot \lim\langle c_{.n}\rangle = \lim\langle (a_{.n} + b_{.n}) \cdot c_{.n}\rangle = \lim\langle a_{.n} \cdot c_{.n} + b_{.n} \cdot c_{.n}\rangle$. Derselbe Wert ergibt sich erwartungsgemäß für $a \cdot c + b \cdot c$.

E: Sei $a + d = b$ mit $d \neq 0$ und n so gewählt, dass $d_{.n} \neq 0$, also $\varepsilon_n \leq d_{.n}$. Dann ist

$$a < a_{.n} + \varepsilon_n \leq a_{.n} + d_{.n} \leq a + d = b.$$

Daher $a < b$. Etwas subtiler ist die andere Richtung von E, also der Existenzbeweis von $b - a$ für $a < b$. Zwar konvergiert $\langle b_{.n} - a_{.n}\rangle$ tatsächlich gegen $b - a$, diese Folge ist aber i.a. nicht monoton. Hier hilft nun die Einbeziehung einer anderen Näherung für a, nämlich $a^{\cdot n} := a_{.n} + \varepsilon_n$. Sicherlich ist $a_{.n+1} = a_{.n} + z^a_{n+1}\varepsilon_{n+1}$. Damit erhalten wir $a_{.n+1} + \varepsilon_{n+1} \leq a_{.n} + (z^a_{n+1} + 1)\varepsilon_{n+1} \leq a_{.n} + 10\varepsilon_{n+1} = a_{.n} + \varepsilon_n$. Kurzum, $\langle a^{\cdot n}\rangle$ fällt monoton. Sei nun $a < b$, also $a_{.n} \leq b_{.n}$ für alle n. Wir betrachten die Folge $\langle d_n\rangle$ mit

$$d_n = \begin{cases} 0 & \text{solange} \quad a_{.n} = b_{.n}, \\ b_{.n} - a^{\cdot n} & \text{falls} \quad\ a_{.n} < b_{.n}. \end{cases}$$

d_n ist wohldefiniert, weil $a_{.n} < b_{.n} \;\Leftrightarrow\; a_{.n} + \varepsilon_n \leq b_{.n} \;\Leftrightarrow\; a^{\cdot n} \leq b_{.n}$, Übung 4.2. Wir zeigen nun, $\langle d_n\rangle$ ist beschränkt, wächst monoton, und $a + d = b$ mit $d := \lim\langle d_n\rangle$.

Ersteres ist klar, weil $d_n \leq b_{.n} \leq b$. Auch ist $d_n \leq d_{n+1}$, denn wegen $a^{\cdot n} \geq a^{\cdot n+1}$ ist $b_{.n} - a^{\cdot n} \leq b_{.n+1} - a^{\cdot n} \leq b_{.n+1} - a^{\cdot n+1}$, solange $a^{\cdot n} \leq b_{.n}$ (d.h. $a_{.n} < b_{.n}$). Sei nun k nun minimal mit $a_{.k} < b_{.k}$. Nach Satz 5.1 ist $a + d = \lim\langle a_{.n}\rangle + \lim\langle d_n\rangle = \lim\langle a_{.n} + d_n\rangle$. Daher genügt zu zeigen $\lim\langle a_{.n} + d_n\rangle = b$. Für $n \geq k$ ist offenbar

$$a_{.n} + d_n = a_{.n} + b_{.n} - a^{\cdot n} = b_{.n} - \varepsilon_n \leq b_{.n},$$

aber $a_{.n} + d_n \leq b_{.n}$ gilt auch für $n < k$, weil dann $a_{.n} = b_{.n}$ und $d_n = 0$. Folglich ist $a + d = \lim\langle a_{.n} + d_n\rangle \leq \lim\langle b_{.n}\rangle = b$. Daher ist nur noch $\lim\langle b_{.n}\rangle \leq \lim\langle a_{.n} + d_n\rangle$ zu zeigen. Für $n \geq k$ ist $b_{.n} = a^{\cdot n} + d_n = a_{.n} + d_n + \varepsilon_n$, erst recht also $b_{.n} \leq (a_{.n} + d_n) + \varepsilon_n$, und dies gilt natürlich auch für $n < k$. Daher ist Bedingung (ii) des Kriteriums mit $m = n$ und $c = 1$ erfüllt. Folglich ist $b \leq a + d$, und alles ist bewiesen.

Unser wesentliches Ziel ist hiermit erreicht. $\mathbb{D}$ hat sich mit der oben erklärten Addition und Multiplikation als ein $\mathcal{E}$-Bereich erwiesen, welcher den Satz von der oberen Grenze erfüllt. Anders formuliert: $\mathbb{D}$ ist ein lückenloser (Satz 3.3) und damit auch archimedischer $\mathcal{E}$-Bereich (Übung 4 in **3.5**). In **10** wird zudem bewiesen, dass es bis auf Isomorphie nur einen solchen Bereich gibt. Damit haben wir volles Recht, die Elemente von $\mathbb{D}$ als reelle

Zahlen zu bezeichnen. Mit diesen darf man von nun an wie gewohnt rechnen, unter Beachtung der vorläufig noch gültigen Einschränkungen hinsichtlich Subtraktion. Zum Beispiel folgt $a - \operatorname{Int} a < 1$ aus $a < \operatorname{Int} a + 1$ nur deshalb, weil der Term $a - \operatorname{Int}$ wegen $a \geq \operatorname{Int} a$ wohldefiniert ist. Auf die Definiertheit solcher Terme wird künftig nur dann explizit hingewiesen, wenn dies (wie in Abschnitt **9**) wesentlich ist.

5.4 Rechnen mit Näherungen

Bekanntlich operieren große und kleine Rechner mit endlichen Dezimalzahlen (eigentlich Dualzahlen, aber das ist hier weniger wichtig). Die hieraus resultierenden Probleme der Resultatsverfälschung bei umfangreichen numerischen Rechnungen sind theoretisch nur schwer zu beherrschen und erfordern eine sorgfältige Planung und Fehleranalyse. Es ist aus diesem und anderen Gründen unerlässlich, eine genauere Vorstellung von den Beziehungen der arithmetischen Operationen im Bereich reeller Zahlen und ihrer näherungsweisen Realisierung durch entsprechende Operationen in $\mathbb{E}$ zu besitzen.

In der Regel wird mit Rundungen einer durch die Hardware des Rechners festgelegten Stellenzahl n gerechnet. Wir erörtern dies nicht in Einzelheiten, sondern interessieren uns hier nur für die „sicheren“ Kommastellen des Ergebnisses bei der Ausführung der elementaren Rechenoperationen. Gewiss ist $a_{.n} \leq a < a_{.n} + \varepsilon_n$ nach Formel (2) in **4.2**. Also ist $a_{.n} + b_{.n} \leq a + b < a_{.n} + b_{.n} + 2\varepsilon_n$. Es gilt gemäß Übung 5 aber eine etwas bessere Ungleichung für die n-te Näherung $(a+b)_{.n}$ der Summe, nämlich

$$(3) \quad a_{.n} + b_{.n} \leq (a+b)_{.n} \leq a_{.n} + b_{.n} + \varepsilon_n,$$

was für $n = 0$ die Ungleichung $\operatorname{Int} a + \operatorname{Int} b \leq \operatorname{Int}(a+b) < \operatorname{Int} a + \operatorname{Int} b + 1$ umfasst. $(a+b)_{.n}$ wird also durch die Summe der n-ten Näherungen von a, b mit einem minimalen Fehler in der letzten Dezimalen angegeben. Für $a = 1{,}213\cdots$ und $b = 2{,}381\cdots$ z.B. hat $(a+b)_{.3}$ gemäß (3) den Wert 3,594 oder 3,595. Die ersten zwei Dezimalen sind hier absolut gesichert, und die letzte hat eine minimale Unsicherheit. Ungeachtet dessen kann es vorkommen, dass für $a = z_0{,}z_1z_2\cdots$ und $b = z_0'{,}z_1'z_2'\cdots$ nicht einmal $\operatorname{Int}(a+b)$ exakt bestimmt werden kann, obwohl a und b ziffernweise berechenbar sind. Denn nach Übung 4 müssen zur Bestimmung von $\operatorname{Int}(a+b)$ die z_i und z_i' im Falle $z_i + z_i' = 9$ für $i = 1, 2, \ldots$ solange berechnet werden, bis erstmals entweder $z_n + z_n' < 9$, oder aber $z_n + z_n' > 9$, oder es ist beständig $z_i + z_i' = 9$. Im ersten Fall ist $\operatorname{Int}(a+b) = \operatorname{Int} a + \operatorname{Int} b$, in den beiden letzten Fällen ist $\operatorname{Int}(a+b) = \operatorname{Int} a + \operatorname{Int} b + 1$. Offenbar ist die Entscheidung, welcher Fall vorliegt, unter Umständen nicht zu treffen.

Ähnlich lässt sich leicht bestätigen, dass $a - b$ für $a > b$ und $a_{.n} > b_{.n}$ im Intervall mit den Grenzen $a_{.n} - b_{.n} - \varepsilon_n$ und $a_{.n} - b_{.n} + \varepsilon_n$ liegt. Bezüglich des Produkts ergibt sich

$$a_{.n}b_{.n} \leq ab \leq a_{.n}b_{.n} + \varepsilon_n(a_{.n} + b_{.n} + \varepsilon_n) < a_{.n}b_{.n} + \varepsilon_n(a+b) + \varepsilon_n^2.$$

Hieraus ersieht man, dass der Fehler in der Approximation von ab durch $a_{.n}b_{.n}$ sehr groß werden kann. Will man ab auf m Kommastellen genau, so muss $n \geq m$ so gewählt

werden, dass $\varepsilon_n(a+b) < \varepsilon_m$, also $a+b < 10^{n-m}$. Dies ergibt $\lg(a+b) < n-m$, oder $n > m + \lg(a+b)$; dabei wurde das quadratische Glied ε_n^2 vernachlässigt, weil es sich in der Regel erst in der $2n$-ten Dezimalen auswirkt. Die hier ausnahmsweise im voraus benutzte logarithmische Funktion lg wird in **7.4** behandelt.

Beispiel. Sei $a = 33333{,}3333\,3333\cdots$, $b = 3{,}6996\,3369\cdots$, und es soll $a \cdot b$ auf 2 Kommastellen genau ausgerechnet werden. Dazu benötigt man sogar die 7-ten Näherungen. Es ist nämlich $2 + \lg 33337{,}03\cdots = 6{,}5\cdots$. Rechnet man etwa nur mit $a_{.2}$ und $b_{.2}$, so erhält man $a_{.2} \cdot b_{.2} = 122999{,}9877$, während tatsächlich $a \cdot b = 123321{,}12\cdots$. Der (absolute) Fehler $ab - a_{.2}b_{.2}$ ist also größer als 321.

Diese hier nur angedeuteten Probleme des näherungsweisen Rechnens werden auch nicht aus der Welt geschafft, wenn mit den n-ten Rundungen statt mit den n-ten kanonischen Näherungen gerechnet wird. Eine genaue Kontrolle gewinnt man erst durch die so genannte Intervallarithmetik, auf die wir hier aber nicht eingehen können.

Meistens betrachtet man nicht den *absoluten Fehler* $|a' - a|$ einer Näherung a' für a, sondern den *relativen Fehler* $r = |\frac{a'-a}{a}|$, der häufig in Prozenten angegeben wird. In obigem Beispiel ist dieser mit etwa 0,26% zwar noch recht klein, aber bei Tausenden von Multiplikationen in der Ausführung eines Rechenprogramms können sich auch diese Fehler erheblich summieren. Siehe etwa [38] für diesen Themenkreis.

5.5 Übungen

1. Man beweise $a \cdot 10^n = a\overset{n}{\rightarrow}$ und $a \cdot \varepsilon_n = a\overset{n}{\leftarrow}$ für *alle* $a \in \mathbb{D}$ und schließe daraus $(a+b)\overset{n}{\rightarrow} = a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow}$. Demnach gilt $a+b = (a\overset{n}{\rightarrow} + b\overset{n}{\rightarrow})\overset{n}{\leftarrow}$ für beliebige $a, b \in \mathbb{D}$.

2. Eine Folge $\langle x_n \rangle$ aus $\mathbb{D}$ heiße eine *Nullfolge*, genauer eine *fallende Nullfolge*, wenn $x_0 \geq x_1 \geq \ldots$ und wenn es zu jedem k ein m gibt mit $x_m < \varepsilon_k$ (womit wegen der fallenden Monotonie dann $x_n < \varepsilon_k$ für alle $n \geq m$). Seien nun $a_n, a \in \mathbb{D}$ und sei die Folge $\langle a_n \rangle$ monoton wachsend. Man beweise die Äquivalenz von
   (i) $\langle a_n \rangle$ konvergiert und $\lim\langle a_n \rangle = a$,
   (ii) $a_n \leq a$ für alle n und $\langle a - a_n \rangle$ ist eine Nullfolge.

3. Man beweise: Satz 5.1 gilt für beliebige schlichte Folgen $\langle a_n \rangle, \langle b_n \rangle$ aus $\mathbb{D}$.

4. Sei $a = z_0{,}z_1z_2\cdots$ und $b = z_0'{,}z_1'z_2'\cdots$, also $\operatorname{Int} a = z_0$, $\operatorname{Int} b = z_0'$. Man zeige
$$\operatorname{Int}(a+b) = \begin{cases} \operatorname{Int} a + \operatorname{Int} b, & \text{falls es ein } k > 0 \text{ gibt mit } z_k + z_k' < 9 \text{ und} \\ & z_i + z_i' = 9 \text{ für alle } i \text{ mit } 0 < i < k, \\ \operatorname{Int} a + \operatorname{Int} b + 1 & \text{sonst.} \end{cases}$$
   Demnach gilt im allgemeinen Falle nur $\operatorname{Int}(a+b) \geq \operatorname{Int} a + \operatorname{Int} b$.

5. Man beweise die Ungleichung $a_{.n} + b_{.n} \leq (a+b)_{.n} \leq a_{.n} + b_{.n} + \varepsilon_n$.

Abschnitt 6

Division und rationale Zahlen

Wir werden als erstes zeigen, dass die Division in $\mathbb{D}$ stets ausführbar ist, wenn nur der Divisor nicht Null ist. Genauer, die Gleichung $b \cdot x = a$ ist stets lösbar für $b \neq 0$, und wegen der Kürzungsregel dann auch eindeutig lösbar. Die Division ist, ähnlich wie die Subtraktion, eine nachträglich eingeführte partiell definierte Operation. Die für $b \neq 0$ eindeutig bestimmte Lösung von $b \cdot x = a$ heißt der *Quotient* von a und b und wird mit $\frac{a}{b}$ bezeichnet. Die Namen „Zähler“ und „Nenner“ für a bzw. b haben hier mit zählen und nennen gar nichts zu tun und sind insofern nur Relikte der Vergangenheit.

Aus der Definition des Quotienten gewinnt man leicht die Regeln der so genannten Bruchrechnung, die das Dividieren in den bisherigen Rechenkalkül einbeziehen. Es wäre verschwendete Mühe, diese für Quotienten natürlicher Zahlen extra zu formulieren, denn sie gelten für beliebige reelle Zähler und Nenner, und allgemeiner, für Elemente eines beliebigen $\mathcal{E}$-Bereichs mit ausführbarer Division, und sind für Quotienten aus natürlichen Zahlen auch nicht etwa einfacher zu beweisen. Den Algorithmus zur expliziten Berechnung von $\frac{a}{b}$ für reelle a, b behandeln wir in **8.1**.

Die Quotienten $\frac{m}{n}$ für $m, n \in \mathbb{N}$ heißen *rationale* Zahlen. Deren Gesamtheit werde mit $\mathbb{Q}$ bezeichnet. In **9** ändern wir die Bezeichnung. Dann wird $\mathbb{Q}$ auch alle negativen rationalen Zahlen umfassen. Stets ist $\mathbb{Q}_+$ die Menge der positiven Elemente von $\mathbb{Q}$. Sowohl rationale als auch *irrationale* Zahlen (Elemente von $\mathbb{D}\backslash\mathbb{Q}$) sind im Sinne unserer Auffassung wohlbestimmte Dezimalzahlen. Auch ein elektronischer Rechner ignoriert die Einteilung in rationale und irrationale Zahlen, es sei denn, er werde speziell programmiert. Sonst erscheint in der Anzeige oder auf dem Bildschirm im Falle einer Division von m durch n immer nur das, was eine Bruchzahl gemäß unserer Auffassung ist, nämlich eine Dezimalzahl; genauer, es erscheint eine gewisse Rundung dieser Zahl.

Jede endliche Dezimalzahl a ist rational; denn hat a die Stellenzahl n, ist $10^n a = a\overset{n}{\longrightarrow}$ eine natürliche Zahl, also $a = \frac{a\overset{n}{\longrightarrow}}{10^n} \in \mathbb{Q}$. Aber es gibt sicher nichtabbrechende rationale Zahlen, z.B. $\frac{1}{3} = 0{,}333\cdots$, siehe **5.2**. Anderseits sind gewiss nicht alle Dezimalzahlen rational. Denn nach **6.2** ist jede rationale Dezimalzahl periodisch. Wir werfen schließlich einen kurzen Blick auf so genannte Stammbruchdarstellungen rationaler Zahlen, einer von vielen interessanten Anwendungen des Bruchrechnens in diesem Abschnitt.

6.1 Division und Bruchrechnung

Durch Anwendung des Satzes 3.2 von der oberen Grenze erhalten wir

Satz 6.1. *Für alle $a, b \in \mathbb{D}$ mit $b \neq 0$ gibt es genau ein $c \in \mathbb{D}$ mit $b \cdot c = a$.*

Beweis. Den Existenzbeweis kann man auf den Fall $a = 1$ beschränken. Denn ist ein $s \in \mathbb{D}$ mit $b \cdot s = 1$ konstruiert, so gilt für $c := sa$ dann $bc = bsa = 1a = a$. Sei also $b \neq 0$. Dann ist $X := \{x \in \mathbb{D} \mid bx \leq 1\}$ beschränkt. Denn ist $\varepsilon_m < b$ und $x \in X$, so folgt $\varepsilon_m x \leq bx \leq 1 = \varepsilon_m \cdot 10^m$, also $x \leq 10^m$. Sei $s := \sup X$. Wir beweisen $bs = 1$. Angenommen $bs < 1$. Wählt man n derart, dass $b\varepsilon_n = b\overleftarrow{{}^n} \leq 1 - bs$, ist $b(s + \varepsilon_n) \leq 1$. Also wäre $s + \varepsilon_n \in X$, was unmöglich ist. Aber auch $bs > 1$ ist ausgeschlossen. Denn ist $b\varepsilon_n \leq bs - 1$ und $\varepsilon_n < s$, so folgt $bx \leq 1 \leq b(s - \varepsilon_n)$ für $x \in X$, also $x \leq s - \varepsilon_n$. Damit wäre bereits $s - \varepsilon_n$ obere Schranke für X, was nicht angeht. Also ist $bs = 1$. Die Eindeutigkeitsaussage folgt mit der Kürzungsregel aus $bc_1 = bc_2 \;\Rightarrow\; c_1 = c_2$. ∎

Diese Beweismethode ist sehr verallgemeinerungsfähig. So lässt sich ganz analog zeigen, dass die Gleichung $x^2 = a$ für gegebenes $a \in \mathbb{D}$ genau eine, mit $\sqrt{a}$ bezeichnete Lösung hat, **7.2**. Schon in der Antike wusste man im Prinzip, dass nicht nur $\sqrt{2} = 1{,}41\cdots$, sondern auch $\sqrt{n}$ – repräsentiert durch eine geeignete Streckenlänge – immer dann irrational ist, wenn n einen Primfaktor in ungerader Potenz enthält.

Der nach Satz 6.1 für $b \neq 0$ existierende, in gewissen Zusammenhängen auch mit b^{-1} bezeichnete Quotient $\frac{1}{b}$ heißt der *Kehrwert* von b, und man sagt, b und $\frac{1}{b}$ seien *reziprok*. Es ist $\frac{1}{b} < 1$ für $b > 1$, und $\frac{1}{b} > 1$ für $0 < b < 1$ gemäß $\mathsf{M}^\times$. Besonders einfach lassen sich die Quotienten $\frac{1}{n}$ veranschaulichen, für $n > 1$ auch *Stammbrüche* genannt. Da

$$\underbrace{\tfrac{1}{n} + \ldots + \tfrac{1}{n}}_{n} = n \cdot \tfrac{1}{n} = 1,$$

ist $\frac{1}{n}$ nichts anderes als die n-Teilung der Zahl 1. So ist z.B. $3 \cdot \frac{1}{3} = 1$ und bereits in **5.2** wurde beispielhaft gezeigt, dass $\frac{1}{3} = 0{,}333\cdots$. Summiert man den Stammbruch $\frac{1}{n}$ m mal, erhält man die rationale Zahl $\frac{m}{n}$. So sind rationale Zahlen historisch gesehen entstanden. Ihre praktische Alltagsbedeutung hat sich angesichts der allgegenwärtigen automatischen Rechner jedoch erheblich vermindert und auch für die Begründung der reellen Arithmetik sind sie entbehrlich wie wir gesehen haben. Selbstverständlich ist die *rationale Arithmetik*, d.h. das Rechnen mit Bruchtermen – so heißen Schreibfiguren $\frac{s}{t}$ mit beliebigen Termen s, t – nach wie vor von eminenter Bedeutung.

Für $a, b \in \mathbb{D}$ und $b \neq 0$ bezeichnet $\frac{a}{b}$ die wohlbestimmte reelle Zahl c mit $bc = a$. Unterschiedliche Bruchterme können, ebenso wie etwa die Differenzterme $3 - 2$ und $4 - 3$, dieselbe Zahl bezeichnen. Genaueres wird unten in Q_1 formuliert.

Mitunter ist es von Vorteil, es bei der Darstellung eines Quotienten $c = \frac{a}{b}$ als Bruchterm zu belassen, vor allem dann, wenn a und b natürliche Zahlen sind. Oft ist die Kenntnis der dezimalen Darstellung von c gar nicht erforderlich, z.B. weil sich Zähler oder Nenner im Verlaufe arithmetischer Umformungen gegen andere Faktoren wegheben. Im Sinne

unserer Auffassung ist aber z.B. $\frac{1}{3}$ nur eine andere Schreibweise für $0{,}333\cdots$, die in besonders prägnanter Weise den Sachverhalt $3 \cdot 0{,}333\cdots = 1$ zum Ausdruck bringt.

Wir formulieren und beweisen jetzt die bereits erwähnten Regeln der Bruchrechnung. Hierin können a, b, c, d beliebige natürliche Zahlen, aber ebenso Dezimalzahlen, ja Elemente eines völlig beliebigen $\mathcal{E}$-Bereichs bedeuten, in welchem die Division ausführbar ist. Denn weder die Formulierungen noch die Beweise dieser Regeln hängen von irgendeiner Zahldarstellung ab. Deshalb sollte man diese Regeln besser die *Grundregeln der rationalen Arithmetik* nennen. Einzige Einschränkung in den Regeln ist $b, d \neq 0$.

Q_0 : $x = \frac{a}{b} \Leftrightarrow bx = a$, speziell $b\frac{a}{b} = a$,

Q_1 : $\frac{a}{b} = \frac{c}{d} \Leftrightarrow ad = bc$, speziell $\frac{ae}{be} = \frac{a}{b}$ für $e \neq 0$ (Kürzungsregel),

Q_2 : $\frac{a}{b} < \frac{c}{d} \Leftrightarrow ad < bc$,

Q_3 : $\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$, $\quad \frac{a}{b} - \frac{c}{d} = \frac{ad-bc}{bd}$ (falls $\frac{a}{b} \geq \frac{c}{d}$ und damit $ad - bc \geq 0$),

Q_4 : $\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$, $\quad \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$ $\quad (b, c, d \neq 0)$,

Q_5 : $(\frac{a}{b})^n = \frac{a^n}{b^n}$ $\quad (b \neq 0)$.

Es folgen die einfachen Beweise von Q_1 - Q_5 (Q_0 gibt lediglich die Definition wieder).

Q_1: $\frac{a}{b} = \frac{c}{d} \Leftrightarrow bd\frac{a}{b} = bd\frac{c}{d}$ (wegen der Kürzungregel in **2.2**)
$\Leftrightarrow ad = bc$ (weil $bd\frac{a}{b} = b\frac{a}{b}d = ad$ und $bd\frac{c}{d} = bc$ gemäß Q_0).

Q_2: $\frac{a}{b} < \frac{c}{d} \Leftrightarrow bd\frac{a}{b} < bd\frac{c}{d} \Leftrightarrow ad < bc$ (wegen $\mathsf{M}^\times$ und Q_0).

Q_3: $bd(\frac{a}{b} + \frac{c}{d}) = db\frac{a}{b} + bd\frac{c}{d} = ad + bc$. Daraus folgt die erste Gleichung mit Q_0. Die zweite ergibt sich ebenso, weil $bd(\frac{a}{b} - \frac{c}{d}) = bd\frac{a}{b} - bd\frac{c}{d} = ad - bc$.

Q_4: $bd \cdot \frac{a}{b}\frac{c}{d} = b\frac{a}{b} \cdot d\frac{c}{d} = ac$. Daraus folgt die erste Gleichung mit Q_0. Entsprechend folgt die zweite, denn nach dem Bewiesenen und der Kürzungsregel ist $\frac{ad}{bc} \cdot \frac{c}{d} = \frac{adc}{bcd} = \frac{a}{b}$.

Q_5: Nach $\mathsf{P}_\times$ aus **2.3** ist $b^n(\frac{a}{b})^n = (b\frac{a}{b})^n = a^n$, und die Behauptung folgt dann aus Q_0.

Diese Regeln lassen sich auf elegante Art auch so beweisen: Man schreibe $b^{-1}a$ für $\frac{a}{b}$ (also $\frac{1}{b} = b^{-1}$) und bestätigte zuerst $b^{-1}d^{-1} = (bd)^{-1}$ und $(c^{-1})^{-1} = c$. Dann erhält man z.B. Q_3 aus $b^{-1}a + d^{-1}c = b^{-1}d^{-1}da + b^{-1}bd^{-1}c = (bd)^{-1}(ad + cb)$.

Wegen Q_1 hat jedes $r \in \mathbb{Q}_+$ eine *gekürzte* Darstellung $r = \frac{m}{n}$, womit gemeint ist, m und n seien teilerfremd. Aber was noch wichtiger ist: Q_3 und Q_4 zeigen, dass $\mathbb{Q}$ gegenüber Addition und Multiplikation abgeschlossen ist. Ebenso ist die Abgeschlossenheit gegenüber Subtraktion gewährleistet, falls diese (in $\mathbb{D}$) ausführbar ist. Dies genügt um festzustellen, dass $\mathbb{Q}$ als Teilbereich von $\mathbb{D}$ selbst einen $\mathcal{E}$-Bereich bildet; denn die Axiome in **2.1** sind einfach nachzuprüfen. So gilt K^+ in $\mathbb{D}$, also speziell auch in $\mathbb{Q}$.

Von nun an verwenden wir die Rechenregeln in $\mathcal{E}$-Bereichen und die rationale Arithmetik ohne besonderen Hinweis. Eine etwas anspruchsvollere Anwendung der letzteren ist der unten geführte Beweis, dass $\langle e_n \rangle$ eine schlichte Folge ist, wobei $e_n := (1 + \frac{1}{n})^n$ für

$n > 0$ und $e_0 := 1$. Deren Grenzwert wird allgemein mit **e** bezeichnet und heißt die EULER*sche Zahl*. Also $\mathbf{e} = \lim\langle e_n\rangle$. In **7.1** wird ein bequemes Berechnungsverfahren für **e** angegeben, und in **8.3** wird nachgewiesen, dass **e** irrational ist. Hier zeigen wir nur $2 < \mathbf{e} < 3$. Nicht nur **e**, sondern auch die e_n, und allgemeiner die Zahlen $(1 + \frac{x}{n})^n$ für reelle x treten in vielen Anwendungen der Mathematik explizit auf.

Die Ungleichung (12) in **2.4** liefert offenbar $(1 - \frac{1}{n^2})^n > 1 - \frac{1}{n} = \frac{n-1}{n}$ für $n > 1$. Auch beachte man $1 + \frac{1}{n-1} = \frac{n}{n-1}$, sowie $\frac{1+\frac{1}{n}}{1+\frac{1}{n-1}} = \frac{(n+1)(n-1)}{n^2} = 1 - \frac{1}{n^2}$. Damit ergibt sich das (strikte) Wachstum von $\langle e_n\rangle$ aus der folgenden Ungleichung. Darin sei $n > 1$, weil $e_0 < e_1$ $(= (1 + \frac{1}{1})^1 = 2)$ ohnehin klar ist und damit offenbar auch $2 < \mathbf{e}$.

$$\frac{e_n}{e_{n-1}} = \frac{(1+\frac{1}{n})^n \cdot (1+\frac{1}{n-1})}{(1+\frac{1}{n-1})^n} = \left(\frac{1+\frac{1}{n}}{1+\frac{1}{n-1}}\right)^n \cdot (1 + \frac{1}{n-1}) = (1 - \frac{1}{n^2})^n \cdot \frac{n}{n-1} > \frac{n-1}{n} \cdot \frac{n}{n-1} = 1.$$

Die Beschränktheit von $\langle e_n\rangle$ folgt wegen $\binom{n}{k} = \frac{\prod_{i<k}(n-i)}{k!}$ (Formel (5) in **4.3**) aus

$$\begin{aligned}
(1 + \tfrac{1}{n})^n &= \textstyle\sum_{k\le n} \binom{n}{k} \frac{1}{n^k} && \text{(Binomische Formel)}\\
&\le \textstyle\sum_{k\le n} \frac{n^k}{k!} \frac{1}{n^k} = \sum_{k\le n} \frac{1}{k!} \le 1 + \sum_{k<n} (\frac{1}{2})^k && (n^k \le \textstyle\prod_{i<k}(n-i))\\
&= 1 + \tfrac{1-(\frac{1}{2})^n}{1-\frac{1}{2}} < 1 + \tfrac{1}{1-\frac{1}{2}} = 3 && \text{(Formel (4) in } \mathbf{4.3}).
\end{aligned}$$

6.2 Periodische Dezimalzahlen

Die reelle Zahl $1{,}2074\,074\,074\cdots$ ist Beispiel einer so genannten *periodischen* Dezimalzahl. Allgemein heiße $a = z_0{,}z_1z_2\cdots$ *periodisch*, wenn es Indizes $p \ge 1$ und $q \ge 0$ gibt mit $z_{p+q+i} = z_{q+i}$ für $i = 1, 2, \ldots$ Sind p, q minimal mit dieser Eigenschaft, so heiße p die *Periodenlänge*, und q auch der *Periodenindex* von a. Es ist demnach $a = z_0{,}z_1\cdots z_q z_{q+1}\cdots z_{q+p}z_{q+1}\cdots z_{q+p}z_{q+1}\cdots$ die allgemeine Form einer periodischen Dezimalzahl, und man schreibt meistens etwas kürzer $a = z_0{,}z_1\cdots z_q\overline{z_{q+1}\cdots z_{q+p}}$.

Die Folge $\langle z_{q+1}, \ldots, z_{q+p}\rangle$ heiße die *Periode* und die natürliche Zahl $P = z_{q+1}\cdots z_{q+p}$ der *Periodenwert*, wobei die Nullen, mit denen $z_{q+1}\cdots z_{q+p}$ eventuell beginnt, zu unterdrücken sind. Ist $q = 0$, heißt a *reinperiodisch*.

Beispiele. Für $a = 1{,}2\overline{074}$ ist $q = 1$ und $p = 3$. Der Periodenwert P ist 074, also 74. Auch $a = 3{,}14 = 3{,}14000\cdots$ ist periodisch, mit $q = 2$, $p = 1$ und Periodenwert 0.

Satz 6.2. *Eine periodische Dezimalzahl* $a = z_0{,}z_1\cdots z_q\overline{z_{q+1}\cdots z_{q+p}}$ *ist rational. Genauer, ist* $P = z_{q+1}\cdots z_{q+p}$ *der Periodenwert, so gilt*

$$(1)\quad a = a_{.q} + \frac{P}{(10^p-1)\cdot 10^q} = a_{.q} + \frac{P}{\underbrace{9\cdots9}_{p}\,\underbrace{0\cdots0}_{q}}.$$

Beweis. Sei (a) $R := 0{,}z_{q+1}\cdots z_{q+p}z_{q+1}\cdots = (a - a_{.q})\xrightarrow{q} = 10^q(a - a_{.q})$. Offenbar ist $10^pR = R\xrightarrow{p} = z_{q+1}\cdots z_{q+p}{,}z_{q+1}\cdots = P + R$, also (b) $R = \frac{P}{10^p-1}$. (a) und (b) ergeben $10^q(a - a_{.q}) = \frac{P}{10^p-1}$, also $a - a_{.q} = \frac{P}{(10^p-1)10^q}$, und damit (1). ∎

Nach (1) ist z.B. $1{,}2\overline{074} = 1{,}2 + \frac{74}{9990} = \frac{163}{135}$. Irrationalzahlen wie $\sqrt{2}$, $\sqrt{10}$, $\mathbf{e}$ usw. sind also nicht periodisch. Obwohl das schon hier unschwer beweisbar wäre, wird sich in **8.1** ganz nebenbei ergeben, dass umgekehrt jede rationale Zahl auch periodisch ist. Die rationalen Dezimalzahlen sind demnach als die periodischen vollständig gezeichnet. Für eine reinperiodische Zahl a ist $q = 0$, also $10^q = 1$; daher vereinfacht sich (1) zu

(2) $a = z_0 + \frac{P}{\underbrace{99\cdots9}_{p}}$ $\qquad (a \in \mathbb{D}$ reinperiodisch).

Nach dieser Formel ist z.B. $3{,}333\cdots = 3 + \frac{3}{9} = \frac{10}{3}$, aber $3{,}0303\cdots = 3 + \frac{3}{99} = \frac{100}{33}$. Ferner ist nach (2) z.B. $\frac{z}{9} = 0{,}zzz\cdots$, d.h. $\frac{1}{9} = 0{,}111\cdots$, $\frac{2}{9} = 0{,}222\cdots$ usw. Analog ist $\frac{z_1z_2}{99} = 0{,}z_1z_2z_1z_2\cdots$ und allgemein $\frac{z_1\cdots z_n}{10^n-1} = 0{,}\overline{z_1\cdots z_n}$. Auch ist $\frac{n}{11} = \frac{9n}{99} = 0{,}z_1z_2z_1z_2\cdots$ für $n \leq 10$, mit $z_1z_2 = 9 \cdot n$. Demnach ist $\frac{1}{11} = 0{,}\overline{09}$ $(= 0{,}0\overline{90})$, $\frac{2}{11} = 0{,}\overline{18}$ $(= 0{,}1\overline{81})$ usw. Taschenrechner liefern diese und weitere Einsichten sozusagen experimentell. Zum Beispiel sind die Topzahlen allen Aberglaubens, die Primzahlen 7 und 13, dadurch ausgezeichnet, dass gekürzte Bruchzahlen $\frac{m}{n} < 1$ mit Primnenner n nur für $n = 7$ und $n = 13$ die Periodenlänge 6 haben. Nach Übung 8.2 ist nämlich p die kleinste natürliche Zahl, so dass $10^p - 1$ durch n geteilt wird. Für $n = 7$ ist dies die Zahl 6, denn 7 teilt $10^6 - 1 = 99999$, nicht aber $10^i - 1$ für $0 < i < 6$. Dasselbe gilt für $n = 13$. Weitere Möglichkeiten gibt es nicht. Denn es gilt $10^6 - 1 = 3^3 \cdot 7 \cdot 11 \cdot 13 \cdot 37$, und $\frac{1}{3} = 0{,}\overline{3}$, $\frac{1}{11} = 0{,}\overline{09}$, $\frac{1}{37} = 0{,}\overline{027}$ haben die Periodenlängen $p = 1$, $p = 2$, bzw. $p = 3$.

6.3 Stammbruchapproximation

Brüche wurden im alten Ägypten durch Stammbruchsummen dargestellt. Für genauere Rechnungen gab es Rechentabellen wie z.B. der Papyrus Rhind belegt. So wie wir heute reelle Zahlen gern durch 2-stellige Kommazahlen approximieren (etwa $\boldsymbol{\pi}$ durch 3,14), verwandte man damals oft 2-gliedrige Stammbruchsummen. Diese Methode ist für viele Zwecke ausreichend genau. Weil keine 2-gliedrige Stammbruchsumme echt zwischen $\frac{1}{2} + \frac{1}{3} = \frac{5}{6}$ und 1 fällt, liegt der Approximationsfehler für Werte r mit $0 < r < 1$ durchgehend allerdings nur unterhalb $1 - \frac{5}{6} = \frac{1}{6}$. Vermutlich deshalb war $\frac{2}{3}$ der einzige häufiger anzutreffende Nicht-Stammbruch. Denn wegen $\frac{2}{3} + \frac{1}{4} = \frac{11}{12}$ ist dann der Fehler bei 2-gliedriger Approximation nur noch kleiner als $\frac{1}{12}$, wie man sich leicht überlegt.

Leonardo gibt in [22] u.a. folgendes Rezept an zur Verwandlung einer rationalen Zahl $0 < r < 1$ in eine Summe aus paarweise verschiedenen Stammbrüchen:

Man suche den größten Stammbruch $\frac{1}{g_0+1}$ mit $\frac{1}{g_0+1} \leq r$, so dass g_0 eindeutig gekennzeichnet ist durch $\frac{1}{g_0+1} \leq r < \frac{1}{g_0}$. Ist nicht schon $r = \frac{1}{g_0+1}$, suche man den größten Stammbruch $\frac{1}{g_1+1}$ mit $\frac{1}{g_1+1} \leq r - \frac{1}{g_0+1}$ $(< \frac{1}{g_1})$, und so fort. Stets ist $g_i \geq 1$. Nach endlich vielen (sagen wir $k+1$) Schritten steht das Gleichheitszeichen und es ergibt sich

(3) $r = \frac{1}{g_0+1} + \ldots + \frac{1}{g_k+1}$ $\quad (0 < r < 1$ rational).

So erhält man z.B. $\frac{5}{6} = \frac{1}{2} + \frac{1}{3}$. Wir zeigen, dass dieses Verfahren tatsächlich abbricht, und

zwar induktiv über den Zähler m von $\frac{m}{n}$ $(0 < m < n)$. Für $m = 1$ ist nichts zu zeigen. Sei angenommen, alle $r < 1$ $(r \in \mathbb{Q}_+)$ mit einem Zähler $< m$ hätten eine Darstellung (3). Sei $\frac{1}{g+1}$ größter Stammbruch mit $\frac{1}{g+1} \leq \frac{m}{n}$. Dann hat $r := \frac{m}{n} - \frac{1}{g+1} = \frac{m(g+1)-n}{n(g+1)}$ einen Zähler $< m$; denn $\frac{m}{n} < \frac{1}{g}$, also $mg < n$ und so $m(g+1) - n < m$. Daher hat das soeben gewählte r eine Darstellung (3) gemäß Induktionsvoraussetzung, und wir erhalten somit $\frac{m}{n} = r + \frac{1}{g+1} = \frac{1}{g_0+1} + \ldots + \frac{1}{g_k+1} + \frac{1}{g+1}$. Damit wurde bestätigt, dass LEONARDOs Prozedur stets abbricht. Auch sind die g_i paarweise verschieden. Denn nach Definition ist $\frac{1}{g_{i+1}+1} \leq (r - \frac{1}{g_0+1} - \ldots - \frac{1}{g_{i-1}+1}) - \frac{1}{g_i+1} < \frac{1}{g_i} - \frac{1}{g_i+1} = \frac{1}{g_i(g_i+1)}$. Also ist $g_{i+1} + 1 > g_i(g_i + 1)$. Daher genügen die Zahlen g_i der Bedingung

(4) $g_{i+1} \geq g_i(g_i + 1)$ für alle $i < k$.

Weil $g_i + 1 \geq 2$, gilt nach (4) sogar $g_{i+1} \geq 2g_i$. Stammbruchdarstellungen sind in der Regel nicht eindeutig. Zum Beispiel ist $\frac{7}{12} = \frac{1}{3} + \frac{1}{4} = \frac{1}{2} + \frac{1}{12}$. Siehe auch Übung 4. Eine Darstellung (3) mit der Eigenschaft (4) ist nach Übung 5 jedoch eindeutig.

Startet man LEONARDOs Verfahren mit irrationalem $r > 0$ (genauer, mit $r - \operatorname{Int} r$), so bricht das Verfahren nicht ab und man erhält eine Darstellung von r als unendliche Stammbruchreihe, siehe **7.1**. Es ist z.B. $\boldsymbol{\pi} = 3 + \frac{1}{8} + \frac{1}{61} + \ldots$ Übrigens ist der relative Fehler der 2-ten Stammbruchapproximation $3 + \frac{1}{8} + \frac{1}{61}$ für $\boldsymbol{\pi}$ mit etwa 0,006% um fast eine Größenordnung besser als die dezimale Approximation 3,14 mit etwa 0,05%.

6.4 Übungen

1. Man zeige: $\langle a_n \rangle$ mit $a_n \neq 0$ für alle n ist (fallende) Nullfolge genau dann, wenn $\langle \frac{1}{a_n} \rangle$ monoton und unbeschränkt wächst. Damit ist $\langle c^n \rangle$ für $c < 1$ eine Nullfolge, weil nach Übung 3.3 die Folge $\langle \frac{1}{c^n} \rangle = \langle (\frac{1}{c})^n \rangle$ unbeschränkt wächst.

2. Sei $e'_n = (1 - \frac{1}{n})^n$ mit $e'_0 := 0$. Man zeige: $\langle e'_n \rangle$ ist schlicht und $\lim \langle e'_n \rangle = \frac{1}{\mathbf{e}}$.

3. Sei $x > 0$ und $x_n = (1 + \frac{x}{n})^n$. Man zeige: $\langle x_n \rangle$ wächst strikt monoton und es ist $x_n \leq \sum_{k \leq n} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!}$. Ferner beweise man, $\langle \sum_{k \leq n} \frac{x^k}{k!} \rangle$ und damit auch $\langle x_n \rangle$ sind beschränkte und folglich schlichte Folgen.

4. Sei $\langle h_i \rangle$ eine Folge von Zahlen aus $\mathbb{N}_+$ mit $h_{i+1} = h_i(h_i + 1)$. Man zeige, für jedes n gilt $\frac{1}{h_0} = \sum_{i<n} \frac{1}{h_i+1} + \frac{1}{h_n}$. Nach dieser Formel ist z.B. für $h_0 = 1$ und $n = 0, 1, 2, \ldots$

$$1 = 0 + 1 = \tfrac{1}{2} + \tfrac{1}{2} = \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{6} = \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{7} + \tfrac{1}{42} = \ldots$$

   Dem entnimmt man leicht verschiedene Stammbruchdarstellungen von $\frac{1}{2}$, die aber Bedingung (4) verletzen. Zum Beispiel ist $41 \geq 6 \cdot 7$ nicht erfüllt.

5. Man zeige, eine Darstellung (3) mit der Eigenschaft (4) *ist* die LEONARDOsche und folglich eindeutig bestimmt.

Abschnitt 7

Beginnende Analysis – Unendliche Reihen, Potenzen, Logarithmen

Obwohl die negativen Zahlen noch nicht zur Verfügung stehen, lassen sich gewisse Elemente der analytischen, d.h. mit dem Grenzwert zusammenhängenden Betrachtungsweise bereits jetzt bequem entwickeln. Mehr noch, das Nichtvorhandensein negativer Zahlen macht den Kern mancher Konstruktionen durch den vorläufigen Wegfall von Vorzeichenbetrachtungen deutlicher sichtbar. Die Erweiterung der Begriffe unter Einschluss negativer reeller Zahlen ist später nur ein kleiner Schritt. Man benötigt lediglich Grenzwerte oder Suprema schlichter Folgen, um unendliche Reihen mit positiven Gliedern, Potenzen und Logarithmen einfach und gründlich behandeln zu können. Speziell ist der Nachweis von $z_0{,}z_1z_2\cdots=\sum_{i=0}^{\infty}\frac{z_i}{10^i}$ zu Beginn von **7.1** geradezu banal.

Ein Hauptziel dieses Abschnitts ist die Definition der Exponentialfunktion durch den lückenlosen Nachweis aller Potenzgesetze, auf denen z.B. die logarithmischen Formeln beruhen. Die Einführung der n-ten Wurzeln, beliebiger Potenzen und Logarithmen beruht meistens auf dem so genannten Zwischenwertsatz für stetige Funktionen. Nun haben die elementaren Funktionen eine die Gültigkeit dieses Satzes garantierende Eigenschaft, welche etwas einfacher handhabbar ist als die Stetigkeit und welche vor allem für die numerische Analysis von Bedeutung ist. Sie sind Lipschitz-stetig in allen Intervallen ihres Definitionsbereichs. Auch genügt es vorübergehend, sich auf monotone Funktionen dieser Art zu beschränken. Das führt uns zu dem Begriff der schlichten Funktion, einer sinngemäßen Verallgemeinerung schlichter Folgen, die z.B. auch nützlich ist für den Konvergenzssatz 8.5, einer Variante des BANACHschen Fixpunktsatzes für Funktionen, die nicht notwendig Kontraktionen sind.

Natürlich wird bei dieser Gelegenheit auch der für die höhere Analysis wesentliche Begriff der stetigen Funktion eingeführt. Schlichte Funktionen sind stetig und die Sätze dieses Abschnitts bleiben richtig, wenn schlichte Funktion überall durch stetige Funktion ersetzt wird - nur sind die Beweise dann weniger einfach. Auch lassen sich die Exponential- und Logarithmenfunktion auf gänzlich andere Weise ohne explizite Zwischenwertargumente einführen, nämlich als spezielle Isomorphismen, siehe hierzu **10.5**.

7.1 Unendliche Reihen

Nach Formel (3) in **4.3** haben wir die Darstellung $a_{.n} = z_0 + \frac{z_1}{10^1} + \ldots + \frac{z_n}{10^n}$ für die n-te kanonische Näherung der Dezimalzahl $a = z_0{,}z_1z_2\cdots$. Diese Summe liegt anschaulich um so näher bei a, je größer n ist. Daher stellt sich die Frage, ob man der Gleichung

(0) $\quad a = z_0 + \frac{z_1}{10} + \frac{z_2}{100} + \ldots$

einen vernünftigen Sinn zu geben imstande ist. Dies ist in der Tat möglich. Bei dem Term auf der rechten Seite von (0) handelt es sich nämlich um eine spezielle unendliche Reihe. Diese haben eine lange Tradition. Sie waren in der Mathematik schon vor Erfindung des Differentialkalküls aufgetreten und spielen auch heute eine wichtige Rolle, etwa bei der Reihenentwicklung von Funktionen. Hier einige bekannte Beispiele:

(1) $\quad 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$

(2) $\quad 1 + \frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \ldots$

(3) $\quad 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots$

(4) $\quad \frac{1}{2} + \frac{1}{3} + \frac{1}{7} + \ldots + \frac{1}{k+1} + \frac{1}{k(k+1)+1} + \ldots$

Mit diesen Reihen, die man meistens in der Form $a_0 + a_1 + \ldots$ oder $\sum_i a_i$ schreibt [1)], wurde lange Zeit recht unbefangen operiert. Aber erst der Satz von der oberen Grenze liefert das geeignete Werkzeug zur Beherrschung dieser Reihen, und zwar gemäß folgender Idee: Statt einer vermeintlich unendlichen Summe $a_0 + a_1 + \ldots$ betrachtet man die wohldefinierten Summen $s_n = a_0 + \ldots + a_n$, welche die *Partialsummen* der Reihe heißen. Die Folge $\langle s_n \rangle$ ist offenbar monoton. Wenn sie nun überdies beschränkt ist, konvergiert sie gegen einen wohlbestimmten Grenzwert s. In diesem Falle heißt auch die Reihe *konvergent* (andernfalls *divergent*), und man nennt s die *Summe* dieser Reihe, symbolisch $s = a_0 + a_1 + a_2 + \ldots$, oder kurz $s = \sum_i a_i$.

Für die Reihe rechts in (0) ist $s_n = z_0 + \frac{z_1}{10^1} + \ldots + \frac{z_n}{10^n} = z_0{,}z_1 \cdots z_n = a_{.n}$. Die Partialsummenfolge $\langle s_n \rangle$ ist also mit der Näherungsfolge $\langle a_{.n} \rangle$ identisch und konvergiert damit gegen a. Der Term rechts in (0) erhält so nicht nur einen wohldefinierten Sinn, sondern darüber hinaus trifft (0) auch zu. Damit entpuppt sich in unserem Aufbau eine nichtabbrechende Dezimalzahl gewissermaßen nachträglich als diejenige unendliche Reihe, durch welche sie traditionsgemäß definiert wird.

Durch die angegebene Präzisierung erübrigt sich auch die spekulative Frage nach einer Summation unendlich vieler Reihenglieder, so dass eine vermeintliche Beantwortung in diesem oder jenem Sinne das Geschehen in keiner Weise beeinflußt. An ursprüngliche Vorstellungen erinnert nur noch die traditionelle Notation $a_0 + a_1 + a_2 + \ldots$ Allerdings ist hierbei zu beachten, dass dieser Term, ebenso wie $\sum_i a_i$, in doppelter Bedeutung auftritt. Er bezeichnet nicht nur die Reihe selbst, genauer deren Partialsummenfolge,

[1)] auch $\sum_{i\geq 0} a_i$ oder $\sum_{i=0}^{\infty} a_i$. Dabei heißt a_n das *n-te Glied* oder der *n-te Summand* der Reihe. Falls i einen anderen Definitionsbereich hat, z.B. $\{i \in \mathbb{N} \mid i \geq k\}$, ist dies klar zu kennzeichnen, z.B. durch $\sum_{i\geq n} a_i$. Die Laufvariable sei stets die unmittelbar neben dem $\sum$-Zeichen stehende.

sondern im Falle der Konvergenz auch deren Grenzwert. Eine solche Doppeldeutigkeit hat auch ihre guten Seiten; beabsichtigt oder nicht, zwingt sie zu erhöhter Aufmerksamkeit. Von zahlreichen Konvergenzkriterien erwähnen wir die folgenden:

1. $\sum_i a_i$ konvergiert genau dann, wenn $\sum_{i>n} a_i = a_{n+1} + a_{n+2} + \ldots$ für beliebiges n konvergiert, und im Falle der Konvergenz ist $\sum_i a_i = \sum_{i\leq n} a_i + \sum_{i>n} a_i$. Der Beweis ist so einfach, dass er gänzlich dem Leser überlassen werden kann.

2. Mit $\sum_i a_i$ ist auch $\sum_i ca_i$ für beliebiges $c \in \mathbb{D}$ konvergent und es gilt $\sum_i ca_i = c \cdot \sum_i a_i$. Dies ergibt sich unmittelbar aus $\sum_{i\leq n} ca_i = c \cdot \sum_{i\leq n} a_i$ für alle n.

3. (*Majorantenkriterium*). Sind (a) $\sum_i a_i$ und (b) $\sum_i b_i$ unendliche Reihen mit $a_i \leq b_i$ für alle i – es heißt (b) dann eine *Majorante* für (a) – ist mit (b) auch (a) konvergent. Denn sind s_n, t_n die Partialsummen von (a) bzw. (b), so gilt $s_n \leq t_n$, so dass mit $\langle t_n \rangle$ auch $\langle s_n \rangle$ beschränkt ist. Danach sind z.B. mit (2) auch (3) und (4) konvergent. Bei (3) betrachte man die Reihe $\frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots$ für die (2) eine Majorante ist, so dass mit (2) nach Kriterium **1** auch (3) konvergiert. (2) ist tatsächlich konvergent; sie hat die Partialsumme $s_n = 2 - \frac{1}{2^n}$ wie man der Formel (4) in **6.2** leicht entnimmt. Daher konvergiert $\langle s_n \rangle$ gegen 2, mit anderen Worten $\sum_i \frac{1}{2^i} = 2$.

4. (*Nullfolgenkriterium*). Die Reihe $\sum_i a_i$ konvergiert genau dann mit der Summe s, wenn $s_n = a_0 + \ldots + a_n \leq s$ für alle n und wenn die Folge $\langle r_n \rangle$ der so genannten Restglieder $r_n := s - s_n$ eine Nullfolge ist. Dies folgt unmittelbar aus Übung 5.2.

Von den angegebenen Reihen ist nur die erste, die so genannte *harmonische* Reihe, divergent. Ihre Partialsummenfolge wächst unbeschränkt; ist nämlich $a > 0$ vorgegeben, so ist wegen $2^i - 2^{i-1} = 2^{i-1}$ für jedes $n > 2a$

$$\begin{array}{rcccccccccc} s_{2^n} = & 1 & + & \frac{1}{2} & + & \frac{1}{2+1} + \frac{1}{2^2} & + \ldots + & \frac{1}{2^{n-1}+1} + \ldots + \frac{1}{2^n} & & \\ > & \frac{1}{2} & + & \frac{1}{2} & + & \frac{2}{2^2} & + \ldots + & \frac{2^{n-1}}{2^n} & = & \frac{1}{2} \cdot n > a. \end{array}$$

Die harmonische Reihe entgeht in gewissem Sinne sehr knapp der Konvergenz und divergiert außerordentlich langsam, wovon man sich mit Hilfe eines programmierbaren Taschenrechners leicht überzeugt. Zum Beispiel ist immer noch $s_{10\,000} < 10$. Die harmonische Reihe zeigt uns klar, dass es für die Konvergenz von $\sum_i a_i$ in der Regel nicht hinreicht, dass die Glieder a_i selbst eine Nullfolge bilden.

In Lichte der unendlichen Reihen wollen wir uns eine aus der Antike überlieferte Paradoxie näher ansehen, nämlich den Wettlauf des Achilles mit der Schildkröte. Diese erhält einen Vorsprung von 100 m. Achilles hat die Aufgabe, die Schildkröte bei gleichzeitigem Start einzuholen. Die konstante Laufgeschwindigkeit des Achilles sei 100 m in 10 sec. Die Schildkröte hingegen bewege sich mit einer konstanten Geschwindigkeit von 10 m in 10 sec, also 1 m in einer Sekunde. Jedermann wird sagen, dass Achilles die Schildkröte nach einem Durchlauf von wenig mehr als 110 m eingeholt haben wird.

Nun wird aber gerade diese Banalität durch folgendes Gedankenexperiment problematisiert. Nach Durchlauf von 100 m ist die Schildkröte offenbar 10 m weit gekommen; nachdem Achilles diesen Standort erreicht hat, ist die Schildkröte 1m vorangekommen.

Sobald Achilles diesen Ort erreicht hat, hat die Schildkröte abermals ein Stückchen Boden gutgemacht, und so ad infinitum. Jedesmal, wenn Achilles die Schildkröte scheinbar erreicht hat, ist diese wieder ein (wenn auch allmählich kleiner werdendes) Stückchen vorangekommen. Daher ist die Schildkröte anscheinend uneinholbar.

Das ist jedoch ein Trugschluss. Denn nach der $(n+1)$-ten Momentaufnahme des Geschehens sind gerade $t_n = 10+1+\frac{1}{10}+\cdots+\frac{1}{10^n} = 11,\underbrace{111\cdots1}_{n}$ Sekunden vergangen. Alle diese Zeitpunkte liegen vor dem Zeitpunkt $11{,}111\cdots$, dem Grenzwert der Folge $\langle t_n\rangle$. Nun ist aber gerade dieser der Treffzeitpunkt τ. Denn bezeichnet s_A den von Achilles bis zur Treffzeit zurückgelegten Weg, sowie s_B den Weg der Schildkröte einschließlich der Vorgabe, so erhält man τ auf bekannte Weise aus der Gleichung $10\tau = s_A = s_B = \tau + 100$ zu $\tau = \frac{100}{9} = 11{,}111\cdots$.

Das Paradoxe an dem geistreichen Gedankenexperiment der Griechen erklärt sich also allein daraus, dass eine Folge von Zeitpunkten ins Auge gefasst wird, die sämtlich von der tatsächlichen Treffzeit τ liegen, und die auch noch gegen diese konvergiert. Damit erhält man dann auch eine Zerlegung des gesamten Zeitintervalls von der Start- bis zur Treffzeit in die unendliche Reihe $10 + 1 + \frac{1}{10} + \frac{1}{100} + \ldots$

Die hier auftretende Reihe $1 + \frac{1}{10} + (\frac{1}{10})^2 + \ldots$ ist, ebenso wie die Reihe (2), Spezialfall der so genannten *geometrischen Reihe*

(5) $\quad \sum_i c^i = 1 + c + c^2 + \ldots \qquad (0 < c < 1).$

Gleichung (4) in **4.3** besagt $s_n := 1+c+\ldots+c^n = \frac{1-c^{n+1}}{1-c}$. Wegen der hieraus unmittelbar folgenden Abschätzung $s_n < \frac{1}{1-c}$ konvergiert die geometrische Reihe. Ihr Grenzwert ist $\frac{1}{1-c}$. Um letzteres zu bestätigen, beachte man zunächst, dass $\langle c^n\rangle$ nach Übung 6.1 eine Nullfolge ist, und damit sicher auch die Restgliedfolge $\langle\frac{c^{n+1}}{1-c}\rangle = \langle\frac{1}{1-c} - s_n\rangle$. Nach dem Nullfolgenkriterium ist also in der Tat $\sum_i c^i = \frac{1}{1-c}$. Ferner ist für gegebene Werte b und $c < 1$ auch $\sum_i bc^i$ konvergent und es ist $\sum_i bc^i = \frac{b}{1-c}$. Auch die Reihe $\sum_i a_i c^i$ ist für $c < 1$ konvergent, wenn nur die a_i alle durch ein $b \in \mathbb{D}$ beschränkt sind. Es ist dann $\sum_i bc^i$ nämlich eine Majorante. Speziell ist jede der unendlichen Reihen

(6) $\quad z_0 + \frac{z_1}{g} + \frac{z_2}{g^2} + \ldots \qquad (2 \leq g \in \mathbb{N},\quad z_i \in \mathbb{N},\quad z_i < g \text{ für alle } i > 0)$

konvergent. Für $g = 10$ ist dies ohnehin klar gemäß (0), sofern $z_i < 9$ für unendlich viele i, d.h. falls $z_0{,}z_1\cdots z_n$ zulässig ist. Falls für alle $i > m$ aber $z_i = 9$ (oder $z_i = g-1$ im allgemeinen Falle), ist (6) eben auch konvergent. Falls $g = 10$ sieht man sehr leicht, dass dann gerade $z_0{,}z_1\cdots z_m + \varepsilon_m$ die Summe von (6) ist. Den Fall einer beliebigen Grundzahl $g \geq 2$ betrachten wir im nächsten Abschnitt.

Dieses Ergebnis wirft neues Licht auf unsere Ausgangsposition, nach der reelle Zahlen formal als Dezimalzahlen erklärt wurden; es verdeutlicht, dass wir von Anfang an statt $g = 10$ eine beliebige andere Grundzahl $g \geq 2$ hätten wählen können. Auch wird vollkommen klar, dass wir z.B. $0{,}999\cdots$ nicht unbedingt als unzulässig disqualifizieren mussten, sondern mit 1 von Anfang an hätten identifizieren können.

Für die Reihe (3) gilt mit $\mathbf{e} = \lim\langle(1+\frac{1}{n})^n\rangle$ (**6.1**), wie schon von EULER bemerkt,

(7) $\quad 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots = \mathbf{e}.$

Die gemäß (7) von EULER in [9] erstmals vorgenommene, auf 23 Dezimalen genaue Berechnung von $\mathbf{e}$ liefert $\mathbf{e} = 2{,}71828\,18284\,59045\,23536\,068\cdots$. Die Reihe (7) ist in der Tat sehr geeignet für die numerische Berechnung von $\mathbf{e}$ mit vorgeschriebener Stellenzahl. Denn sei $s := \sum_k \frac{1}{k!}$ und $s_n := \sum_{k\le n} \frac{1}{k!}$. Dann ergibt sich für die Summe der „Restreihe" $r_n := \frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \ldots = s - s_n$ für $n \ge 1$ die Abschätzung

$$\begin{aligned} r_n &= \tfrac{1}{(n+1)!}\cdot\left(1 + \tfrac{1}{n+2} + \tfrac{1}{(n+2)(n+3)} + \ldots\right) \\ &< \tfrac{1}{(n+1)!}\cdot\left(1 + \tfrac{1}{n+2} + \tfrac{1}{(n+2)^2} + \ldots\right) \le \tfrac{1}{(n+1)!}\cdot\left(1 + \tfrac{1}{3} + \tfrac{1}{3^2} + \ldots\right) = \tfrac{1{,}5}{(n+1)!}. \end{aligned}$$

Einfaches Ausrechnen liefert z.B. $r_{13} < \frac{1{,}5}{14!} < 2\varepsilon_{11}$. Also gibt bereits die Partialsumme s_{13} von (7) die Zahl $\mathbf{e}$ auf mindestens 10 Kommastellen genau an.

Es folgt nun der Beweis von (7). In **6.1** wurde für $e_n = (1+\frac{1}{n})^n$ bereits die Ungleichung $e_n \le s_n$ $(= \sum_{k\le n} \frac{1}{k!})$ gezeigt. Also $\mathbf{e} \le s = \sum_k \frac{1}{k!}$. Folglich genügt der Nachweis, dass $\langle s - e_n\rangle$ eine Nullfolge ist. Sei $\left[{n \atop k}\right] := \frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{n^k}$. Dann gilt für $1 \le k \le n$ wegen $1 \ge \left[{n \atop k}\right] \ge \frac{(n-k+1)^k}{n^k} = (1 - \frac{k-1}{n})^k$ und der Ungleichung (11) in **2.4** offenbar

$$1 - \left[{n \atop k}\right] \le 1 - (1 - \tfrac{k-1}{n})^k \le k \cdot \tfrac{k-1}{n} = \tfrac{1}{n}k(k-1).$$

Weil $\binom{n}{k} = \frac{n(n-1)\cdot\ldots\cdot(n-k+1)}{k!}$ und $e_n = \sum_{k\le n}\binom{n}{k}\frac{1}{n^k} = 1 + \sum_{k=1}^n \left[{n \atop k}\right]\frac{1}{k!}$, folgt somit für $n \ge 2$

$$\begin{aligned} s - e_n = \textstyle\sum_{k=1}^n \frac{1}{k!}(1 - \left[{n \atop k}\right]) + \sum_{k>n}\frac{1}{k!} &\le \tfrac{1}{n}\textstyle\sum_{k=1}^n \frac{k(k-1)}{k!} + \frac{1}{n}\sum_{k>n}\frac{n}{k}\cdot\frac{1}{(k-1)!} \\ &< \tfrac{1}{n}\left(\textstyle\sum_{k<n}\frac{1}{k!} + \sum_{k\ge n}\frac{1}{k!}\right) = \tfrac{s}{n}. \end{aligned}$$

Damit ist $\langle s - e_n\rangle$ in der Tat Nullfolge und $\mathbf{e} = \sum_k \frac{1}{k!}$ ist bewiesen. Es gelten nun sogar

$$\text{(e1)}\quad \mathbf{e} - e_n < \tfrac{\mathbf{e}}{2n+1} \quad ; \quad \text{(e2)}\quad \mathbf{e} - e_n > \tfrac{\mathbf{e}}{2n+2},$$

Übung 3. Die Ungleichung (e2) verursacht ein überaus träges Konvergenzverhalten der strikt wachsenden Folge $\langle e_n\rangle$. Während z.B. bereits $\mathbf{e} - \sum_{i\le 8}\frac{1}{i!} < \varepsilon_5$, ist nach (e2) immer noch $\mathbf{e} - e_{100\,000} > \varepsilon_5$, also wird $\mathbf{e}$ von $e_{100\,000}$ bis höchstens zur 4. Dezimalen genau approximiert. Eine kleine Umformung von (e1) und (e2) ergibt die Ungleichungen

$$\text{(e3)}\quad \mathbf{e}\tfrac{2n}{2n+1} < e_n < \mathbf{e}\tfrac{2n+1}{2n+2}.$$

Dies liefert unter Verwendung der (ausreichenden) Näherung 2,71828 für $\mathbf{e}$ das mit den angegebenen Ziffern genaue Ergebnis $e_{100\,000} = (1 + \frac{1}{100\,000})^{100\,000} = 2{,}71826\cdots$.

Wir haben hier die Irrationalzahl $\mathbf{e}$ benutzt, um eine Abschätzung über e_{10^5} zu gewinnen. Tatsächlich ist es ohne theoretische Vorbereitung einfacher, $\mathbf{e}$ mit hoher Genauigkeit auszurechnen als die scheinbar harmlose, abbrechende Dezimalzahl e_{10^5} mit ihren 500 000 Dezimalen (deren letzte eine 1 ist). Für $t_n := \frac{2n}{2n+1}\mathbf{e}$ folgt aus (e3)

$$\text{(e4)}\quad 0 < e_n - t_n < \tfrac{\mathbf{e}}{(2n+1)(2n+2)} < \tfrac{1}{n^2}.$$

Eine noch bessere Approximation der Größenordnung $\frac{1}{n^3}$ wird nach [31] gegeben durch den Term $t_n := \frac{12n+5}{12n+11}\mathbf{e}$. Für diesen gilt sogar $0 < t_n - e_n < \frac{1}{10n^3}$, so dass e_{10^5} durch t_{10^5} auf mindestens 15 Dezimalen genau angegeben wird.

Reihe (4) hat die Summe 1. Denn nach Übung 6.4 ist $\frac{1}{h_0} = \sum_{i<n} \frac{1}{h_i+1} + \frac{1}{h_n}$. Das ergibt

$$(8) \quad \tfrac{1}{h_0} = \textstyle\sum_i \frac{1}{h_i+1} \qquad \big(h_0 \in \mathbb{N}_+, \quad h_{i+1} = h_i(h_i+1)\big),$$

weil $h_n \geq 2^n$ und $\langle \frac{1}{h_n} \rangle$ daher Nullfolge ist. (4) ist der Sonderfall von (8) mit $h_0 = 1$ und hat deshalb die Summe 1. Setzt man LEONARDOs Algorithmus aus **6.3** auf ein $r < 1$ an, so erhält man nach (4) in **6.3** $r - \sum_{i<n} \frac{1}{g_i+1} < \frac{1}{g_n} \leq \frac{1}{2^n}$, was in $r = \sum_i \frac{1}{g_i+1}$ resultiert. Das ergibt für beliebiges $a \in \mathbb{D}_+$ wegen $r := a - \operatorname{Int} a < 1$ eine Darstellung

$$(9) \quad a = \operatorname{Int} a + \textstyle\sum_i \frac{1}{g_i+1} \qquad \big(g_{i+1} \geq g_i(g_i+1)\big).$$

Dabei steht in der Nebenbedingung statt $\geq$ das $=$-Zeichen ab einer gewissen Stelle n genau dann, wenn a rational ist. Denn sei etwa $g_{i+1} = g_i(g_i+1)$ für alle $i \geq n$. (9) und (8) mit $h_0 = g_n$ ergeben dann $a = \operatorname{Int} a + \sum_{i<n} \frac{1}{g_i+1} + \sum_{i\geq n} \frac{1}{g_i+1} = \operatorname{Int} a + \sum_{i<n} \frac{1}{g_i+1} + \frac{1}{g_n}$ und a ist damit rational. Sei umgekehrt a rational. Dann liefert die nach (8) mögliche Entwicklung des letzten Gliedes $\frac{1}{g_k+1}$ der LEONARDOschen Stammbruchdarstellung von $r = a - \operatorname{Int} a$ in eine unendliche Reihe auch eine Stammbruchreihe der Gestalt (9), in deren Nebenbedingung ab einer gewissen Stelle das $=$-Zeichen steht. Bei irrationalem a steht in der Nebenbedingung folglich $g_{i+1} > g_i(g_i+1)$ für unendlich viele i.

(9) heißt auch die SYLVESTER*sche Entwicklung* von a. Diese ist wie LEONARDOs Darstellung eindeutig, was sich ähnlich ergibt wie in Übung 6.5.

7.2 Der Zwischenwertsatz und erste Anwendungen

Die Lösbarkeit der Gleichung $x^n = a$ läuft auf die Frage hinaus, ob die Funktion $x \mapsto x^n$ den vorgegebenen Wert a annimmt. Diese lässt sich z.B. mit dem Hinweis auf den Zwischenwertsatz für stetige Funktionen positiv beantworten. Nun kann man hier auch einen etwas einfacheren Begriff verwenden, der vor allem für die numerische Analysis bedeutsam ist, die intervallweise definierte Lipschitz-Stetigkeit, siehe auch **8.4**. Die elementaren Funktionen sind alle Lipschitz-stetig in geeignet gewählten Intervallen ihres Definitionsbereichs, ein bedeutender Vorteil für ihre numerische Berechnung. Außerdem genügt es für unsere Zwecke, sich auf monoton wachsende Funktionen dieser Art zu beschränken. Das führt zu folgender Definition, wobei für $u, v \in \mathbb{D}$ mit $u < v$ das Intervall $\{x \in \mathbb{D} \mid u \leq x \leq v\}$ wie üblich mit $[u, v]$ bezeichnet wird.

Definition. Sei $X \subseteq \mathbb{D}$. Eine Funktion $f : X \to \mathbb{D}$ heiße *schlicht* in $[u, v]$, wenn f monoton wächst und ein $C \in \mathbb{D}$ existiert, so dass für alle h mit $x, x+h \in X \cap [u, v]$

$$(10) \quad f(x+h) - f(x) \leq hC,$$

d.h. wenn f dort Lipschitz-stetig ist. f heiße schlicht schlechthin, wenn je zwei Punkte des Definitionsbereichs von f einem Intervall angehören, in welchem f schlicht ist.

Man beachte, dass f nicht in ganz $[u,v]$ definiert sein muss, auch nicht an den Intervallgrenzen. In den Anwendungen wird meist $X = \mathbb{D}, \mathbb{E}, \mathbb{Q}$ oder ein bestimmtes Intervall hiervon sein. Um z.B. eine auf $\mathbb{E}$ überall definierte Funktion als schlicht (schlechthin) nachzuweisen, genügt es, die Schlichtheit in jedem Intervall $[0,n]$ zu bestätigen.

(10) ist offenbar gleichwertig mit $|fy - fx| \leq C|y-x|$ für alle $x, y \in X \cap [u,v]$. Setzt man $\Delta f := f(x+\Delta x) - f(x)$ mit $\Delta x := h$, so lässt sich (10) für $\Delta x \neq 0$ auch notieren als $\frac{\Delta f}{\Delta x} \leq C$. Schlichtheit umfasst also die Beschränktheit der *Differenzenquotienten* $\frac{\Delta f}{\Delta x}$ im gesamten Intervall $[u,v]$. Es ist anschaulich klar, dass diese für $f : x \mapsto x^2$ in jedem Intervall beschränkt sind. In der Tat, für $x, x+h \in [u,v]$ ist wegen $x + \frac{h}{2} < x + h \leq v$

$$f(x+h) - f(x) = (x+h)^2 - x^2 = 2h(x + \tfrac{h}{2}) < 2hv = Ch, \text{ mit } C = 2v.$$

Ferner ist für eine auf X definierte, in $I = [u,v]$ schlichte Funktion f mit jeder schlichten Folge $\langle x_n \rangle$ aus $X \cap I$ auch $\langle fx_n \rangle$ schlicht. Mehr noch, falls $\lim\langle x_n \rangle \in X$, gilt

$$(11) \qquad \lim\langle fx_n \rangle = f\xi \quad (\xi := \lim\langle x_n \rangle).$$

Denn mit $\langle \xi - x_n \rangle$ ist $\langle C(\xi - x_n) \rangle$ und daher sicher auch $\langle f\xi - fx_n \rangle$ eine Nullfolge. (11) besagt, dass f im Punkte ξ linksseitig stetig ist. f ist in ξ auch stetig schlechthin, und sogar gleichmäßig stetig in I, was hier nur erwähnt sei. Dabei heißt $f : X \to \mathbb{D}$ *stetig* im Punkte $\xi \in X$, wenn (11) für jede konvergente Folge $\langle x_n \rangle$ aus X erfüllt ist.

Bedingung (10) muss nicht für alle h geprüft werden. Es genügt, sich auf h-Werte < 1 zu beschränken. Denn jedes Intervall I lässt sich in endlich viele Teilintervalle (mit gemeinsamen Randpunkten) zerlegen, die eine Länge < 1 haben, und die Schlichtheit in jedem dieser Teilintervalle impliziert offenbar die in I. Es genügt auch, (10) nur für die Werte $h = \varepsilon_n$ zu verifizieren, was häufig erheblich einfacher ist.

Satz 7.1 (Zwischenwertsatz). *Es sei f eine im gesamten Intervall $[u,v] \subseteq \mathbb{D}$ definierte und dort schlichte reelle Funktion, sowie $fu < fv$. Dann nimmt f jeden Wert c mit $fu < c < fv$ an; kurz, die Gleichung $fx = c$ hat mindestens eine Lösung.*

Beweis. Sei U die gewiss beschränkte nichtleere Menge $\{x \in [u,v] \mid fx \leq c\}$ und sei $s = \sup U$. Sicher ist $u \leq s \leq v$. Wir behaupten, $fs = c$. Gemäß Voraussetzung gibt es ein C, so dass $x, x+\varepsilon_n \in [u,v] \;\Rightarrow\; f(x+\varepsilon_n) - fx \leq C\varepsilon_n$. Angenommen $fs < c$. Dann ist notwendig $s < v$, und es gibt ein n mit $s + \varepsilon_n \leq v$ und $C\varepsilon_n + fs \leq c$. Also ist $f(s+\varepsilon_n) \leq C\varepsilon_n + fs \leq c$, und damit $s + \varepsilon_n \in U$, im Widerspruch zu $s = \sup U$. Sei nun $fs > c$ angenommen, so dass sicher $u < s$, und sei ε_n so gewählt, dass $u + \varepsilon_n \leq s$ und $C\varepsilon_n < fs - c$. Für $a = s - \varepsilon_n$ $(\geq u)$ ergibt sich dann

$$fs - f(s-\varepsilon_n) = f(a+\varepsilon_n) - fa \leq C\varepsilon_n < fs - c,$$

also $c < f(s-\varepsilon_n)$. Sei $x \in U$, mit $s - \varepsilon_n \leq x$. Dann ist $c < f(s-\varepsilon_n) \leq fx$, was $x \in U$ widerspricht. Also verbleibt in der Tat nur die Möglichkeit $fs = c$. ∎

Es ist einfach nachzuweisen, dass Summe und Produkt schlichter Funktionen (mit demselben Definitionsbereich) wieder schlicht sind. Deswegen sind z.B. alle Funktionen $x \mapsto a_0 + a_1 x + \ldots + a_n x^n$ schlicht. Satz 7.1 ergibt ferner: Ist $f : \mathbb{D} \to \mathbb{D}$ schlicht,

strikt wachsend und unbeschränkt, so gibt es zu jedem $y > f0$ genau ein $x \in \mathbb{D}$ mit $fx = y$. Zum Beispiel erfüllt $x \mapsto ax$ für jedes $a > 0$ diese Bedingungen, womit nochmals die Existenz des Quotienten bewiesen wurde (dessen Existenz wird im Beweis von Satz 7.1 nicht benötigt!). Aber auch $x \mapsto x^n$ erfüllt für jedes $n \geq 1$ die genannten Bedingungen. Nach Satz 7.1 hat also $x^n = c$ für jedes reelle $c \geq 0$ genau eine Lösung. Es ist dies die mit $\sqrt[n]{c}$ bezeichnete *n-te Wurzel* aus c, für $n = 2$ auch *Quadratwurzel* aus c genannt. $x \mapsto \sqrt[n]{x}$ ist demnach gerade die Umkehrfunktion von $x \mapsto x^n$.

Aus den in Abschnitt **2** genannten Regeln der Potenzrechnung mit Exponenten aus $\mathbb{N}$ ergeben sich leicht die drei wichtigsten Gesetze der Wurzelrechnung, nämlich

$$\mathsf{R}_1: \ \sqrt[n]{ab} = \sqrt[n]{a} \cdot \sqrt[n]{b} \quad ; \quad \mathsf{R}_2: \ (\sqrt[n]{a})^m = \sqrt[n]{a^m} \quad ; \quad \mathsf{R}_3: \ a < b \ \Leftrightarrow \ \sqrt[n]{a} < \sqrt[n]{b}.$$

Zum Beweis von R_1 beachte man $(\sqrt[n]{a} \cdot \sqrt[n]{b})^n = (\sqrt[n]{a})^n \cdot (\sqrt[n]{b})^n = ab$. Also löst $\sqrt[n]{a} \cdot \sqrt[n]{b}$ die Gleichung $x^n = ab$, woraus wegen der Eindeutigkeit der Lösung R_1 folgt. R_2 ergibt sich aus $((\sqrt[n]{a})^m)^n = ((\sqrt[n]{a})^n)^m = a^m$ durch beidseitiges Ziehen der n-ten Wurzel. Schließlich ist R_3 eine Folge des strikten Wachstums von $x \mapsto x^n$. Denn eine in ihrem Definitionsbereich X strikt wachsende Funktion f bildet X bijektiv auf ihren Wertebereich Y ab, und die Umkehrfunktion $f^{-1}: Y \to X$ ist stets wieder strikt wachsend.

Auch zeigt man leicht: Ist f schlicht in I und $\frac{\Delta f}{\Delta x} \geq D$ für ein $D > 0$ (d.h. werden die $\frac{\Delta f}{\Delta x}$ nicht zu klein), ist auch f^{-1} im Bildintervall von f schlicht, mit der Konstanten $\frac{1}{D}$. Für $f: x \mapsto x^2$ z.B. ist $\frac{\Delta f}{\Delta x} = 2x \geq 2u$ mit $x \in [u, v]$. Also ist $f^{-1}: x \mapsto \sqrt{x}$ schlicht in jedem Intervall $[u', v']$ $(= [fu, fv])$, mit der Konstanten $\frac{1}{2\sqrt{u'}} = \frac{1}{2u}$ $(u > 0)$.

Sei nun f eine zunächst nur auf $\mathbb{E}$ definierte schlichte Funktion mit Werten aus $\mathbb{D}$. Ist $a \in \mathbb{D}$, so ist mit $\langle a_{.n} \rangle$ offenbar auch $\langle f a_{.n} \rangle$ wieder eine schlichte Folge, hat also einen wohlbestimmten Grenzwert $\lim \langle f a_{.n} \rangle$. Schon in Übung 3.5 wurde die durch

$$\bar{f}a = \lim \langle f a_{.n} \rangle \quad (= \sup\{f a_{.n} \mid n \in \mathbb{N}\}) \tag{12}$$

auf ganz $\mathbb{D}$ erklärte Funktion die *natürliche Fortsetzung* von f genannt. Der folgende überaus nützliche Satz zeigt, dass durch natürliche Fortsetzung von f auch dann keine anderen Funktionswerte entstehen, falls f von vornherein schon einen größeren Definitionsbereich $X \supseteq \mathbb{E}$ besitzt (etwa $X = \mathbb{Q}$), solange f dort monoton wächst.

Satz 7.2. *Sei $f: \mathbb{E} \to \mathbb{D}$ schlicht. Dann ist auch die natürliche Fortsetzung $\bar{f}$ schlicht auf ganz $\mathbb{D}$. Darüber hinaus ist $\bar{f}$ die einzige monotone Fortsetzung von f auf ganz $\mathbb{D}$.*

Beweis. $\bar{f}$ ist wachsende Fortsetzung von f nach Übung 3.5. Sei nun $\tilde{f}$ irgendeine wachsende Fortsetzung von f auf $\mathbb{D}$. Wir zeigen zuerst, $\tilde{f}$ ist schlicht in $[0, n]$. Sei $a < b \leq n$. Man wähle $a', b' \in \mathbb{E} \cap [0, n]$ so, dass $a' \leq a$, $b \leq b'$ und $a - a', b' - b \leq \frac{b-a}{2}$. Dann ist $b' - a' = b' - b + b - a + a - a' \leq 2(b - a)$, und $fb' - fa' \leq C(b' - a')$ ergibt

$$\tilde{f}b - \tilde{f}a \leq \tilde{f}b' - \tilde{f}a' = fb' - fa' \leq C(b' - a') \leq 2C(b - a).$$

Daher ist $\tilde{f}$ schlicht in $[0, n]$, für alle n. Folglich gilt $\tilde{f}a = \lim \langle \tilde{f} a_{.n} \rangle$ gemäß (11). Also ist $\tilde{f}a = \lim \langle \tilde{f} a_{.n} \rangle = \lim \langle f a_{.n} \rangle = \bar{f}a$ nach (12), für alle $a \in \mathbb{D}$. Damit ist die Eindeutigkeit einer monotonen Fortsetzung von f und zugleich deren Schlichtheit gezeigt. ∎

7.3 Potenzrechnung und Exponentialfunktion

Bisher kennen wir lediglich die Potenz b^n ($b \in \mathbb{D}$) als n-fach wiederholte Multiplikation. Nun pflegt man $\sqrt[n]{b}$ auch in der Form $b^{\frac{1}{n}}$ zu schreiben, also z.B. $2^{\frac{1}{2}}$ für $\sqrt{2}$. Wie kommt man zu dieser Schreibweise? Die Antwort ist einfach: will man, dass auch nur einige der unten aufgelisteteten Potenzregeln für andere Exponenten als nur natürliche Zahlen gültig bleiben, dann wird die Gleichung $b^{\frac{1}{n}} = \sqrt[n]{b}$ schlichtweg erzwungen. Denn

$$b = b^1 = b^{\frac{n}{n}} = b^{\overbrace{\frac{1}{n}+\ldots+\frac{1}{n}}^{n}} = \underbrace{b^{\frac{1}{n}} \cdot \ldots \cdot b^{\frac{1}{n}}}_{n} = (b^{\frac{1}{n}})^n,$$

also $b^{\frac{1}{n}} = \sqrt[n]{b}$. Dabei haben wir lediglich die Minimalforderungen P^1 und P^+ unten benutzt. Diese implizieren darüber hinaus $b^{\frac{m}{n}} = \underbrace{b^{\frac{1}{n}} \cdot \ldots \cdot b^{\frac{1}{n}}}_{m} = (\sqrt[n]{b})^m = \sqrt[n]{b^m}$.

Das bedeutet aber noch nicht, dass die Erklärung $b^{\frac{m}{n}} = \sqrt[n]{b^m}$ legal ist; denn dazu muss $b^{\frac{m}{n}} = b^{\frac{p}{q}}$ für $\frac{m}{n} = \frac{p}{q}$ gesichert sein. Bevor wir darauf eingehen, formulieren wir zuerst die Potenzgesetze, jetzt aber gleich für beliebige Exponenten $x, y \in \mathbb{D}$ mit einer beliebigen positiven reellen Basis b. Wir betrachten diese Gesetze oder wenigstens einige davon zunächst nur als Forderungen, die an eine sinnvolle Verallgemeinerung der Potenz zu stellen sind. Dass sich diese sogar alle erfüllen lassen, und zwar auf genau eine Weise, wird in Satz 7.4 bewiesen. Für Exponenten aus $\mathbb{N}$ gelten alle diese Regeln nach **2.3**.

Regeln der allgemeinen Potenzrechnung

$$\begin{array}{lll}
\mathsf{P}^1 :\ b^1 = b, & \mathsf{P}^+ :\ b^{x+y} = b^x \cdot b^y, & \mathsf{P}_{\geq 1} :\ b^x \geq 1 \text{ für } b \geq 1, \\
\mathsf{P}_1 :\ 1^x = 1, & \mathsf{P}^\times :\ b^{x \cdot y} = (b^x)^y, & \mathsf{P}_{\leq 1} :\ b^x \leq 1 \text{ für } b \leq 1, \\
\mathsf{P}^0 :\ b^0 = 1, & \mathsf{P}_\times :\ (a \cdot b)^x = a^x \cdot b^x, & \mathsf{P}^< :\ x < y \ \Rightarrow\ b^x < b^y \quad (b > 1).
\end{array}$$

In der linken Spalte stehen Randbedingungen, in der Mitte Gleichungen, und rechts Monotoniebedingungen. Stets ist $b^x \neq 0$, weil $b^x \cdot (\frac{1}{b})^x = 1$ nach $\mathsf{P}_\times$ und P_1.

Zwischen obigen Regeln bestehen zahlreiche Abhängigkeiten. Zum Beispiel folgt P_1 sofort aus $\mathsf{P}_{\geq 1}$ und $\mathsf{P}_{\leq 1}$. Diese aus besonderem Grunde hervorgehobenen Regeln sind unter Beachtung von P^1, P^+ und der aus $b^x(\frac{1}{b})^x = 1$ folgenden Gleichung $(\frac{1}{b})^x = \frac{1}{b^x}$ nur Spezialfälle von $\mathsf{P}^<$. Auch folgt P^0 leicht aus $\mathsf{P}^1, \mathsf{P}^+$. Denn $b = b^{1+0} = b^1 \cdot b^0 = b \cdot b^0$, also $b^0 = 1$. Aus den angegebenen folgen leicht einige weitere Regeln, insbesondere

$$\mathsf{P}^- :\ b^{x-y} = \tfrac{b^x}{b^y} \quad (x \geq y), \quad \mathsf{P}_\div :\ (\tfrac{a}{b})^x = \tfrac{a^x}{b^x}, \quad \mathsf{P}_< :\ a < b \ \Rightarrow\ a^x < b^x \quad (x \neq 0),$$

$$\mathsf{P}_= :\ a^x = b^x \ \Rightarrow\ a = b \quad (x \neq 0), \quad \mathsf{P}^= :\ b^x = b^y \ \Rightarrow\ x = y \quad (b \neq 1).$$

P^- folgt aus $b^{x-y}b^x = b^x$ gemäß P^+, und ähnlich folgt $\mathsf{P}_\div$. Zum Beweis von $\mathsf{P}_<$ sei $0 < a < b$. Dann gibt es ein $c > 1$ mit $ac = b$. Also $a^x c^x = b^x$ und damit $a^x < b^x$, denn $c^x > 1$ für $x \neq 0$ gemäß $\mathsf{P}^<$ und P_1. Für $b < 1$ ist $x \mapsto b^x$ strikt fallend, weil nämlich $\left(\frac{1}{b}\right)^x = \frac{1}{b^x}$, und weil $x \mapsto \left(\frac{1}{b}\right)^x$ nach $\mathsf{P}^<$ strikt wächst.

In Satz 7.4 wird die bemerkenswerte und keineswegs offensichtliche Tatsache bewiesen, dass alle erwähnten Potenzgesetze schon eine Folge sind aus $\mathsf{P}^1, \mathsf{P}^+, \mathsf{P}_{\geq 1}, \mathsf{P}_{\leq 1}$. Diese vier Gesetze seien daher die *Basisregeln* der Potenzrechnung genannt.

Bemerkung 1. Aus den Basisregeln folgt unmittelbar nur, dass $f : x \mapsto b^x$ für $b \geq 1$ monoton wächst: Sei $x < y$, also $x+t = y$ für ein gewisses t. Es ist $1 \leq b^t$ gemäß $\mathsf{P}_{\geq 1}$, also $b^x \leq b^x \cdot b^t = b^y$. Erst Satz 7.4 wird zeigen, dass die Basisregeln auch das strikte Wachstum von f implizieren. Analoges gilt auch für $x \mapsto \frac{1}{b^x}$ im Falle $b < 1$, wegen $\mathsf{P}_{\leq 1}$.

Bemerkung 2. Schaut man sich die obigen Ausführungen genauer an, so sieht man leicht, dass sich der Beweis aller erwähnten Potenzgesetze aus den Basisregeln auf den Nachweis von $\mathsf{P}^\times$, $\mathsf{P}_\times$ und $\mathsf{P}^<$ reduziert.

Wir hatten anfänglich schon erkannt, dass a^r für $r = \frac{m}{n} \in \mathbb{Q}$ auf höchstens eine Weise sinnvoll erklärt werden kann. Nun steht auch einer Definition von a^r als $\sqrt[n]{a^m}$ auch insofern nichts im Wege, als diese nicht von der Darstellung von r abhängt. Denn für $\frac{m}{n} = \frac{p}{q}$ ist $mq = np$. Die Potenzgesetze in **2.3** und R_2 ergeben dann

$$\sqrt[n]{a^m} = \sqrt[n]{\sqrt[q]{(a^m)^q}} = \sqrt[n]{\sqrt[q]{a^{m\cdot q}}} = \sqrt[n]{\sqrt[q]{a^{n\cdot p}}} = \sqrt[n]{\sqrt[q]{(a^p)^n}} = \sqrt[n]{(\sqrt[q]{a^p})^n} = \sqrt[q]{a^p}.$$

Erst diese Betrachtungen rechtfertigen die folgende

Definition. Für $a \in \mathbb{D}$ und $r = \frac{m}{n} \in \mathbb{Q}$ sei $a^r = \sqrt[n]{a^m}$.

Demnach ist z.B. $2^{0,2} = 2^{\frac{2}{10}} = 2^{\frac{1}{5}} = \sqrt[5]{2} = 1{,}148\cdots$. Mit ausreichender Geduld lassen sich mittels R_1 – R_3 alle anfänglich aufgelisteten Potenzgesetze für rationale Exponenten nachrechnen. Doch verwenden wir in Satz 7.3 eine Überlegung, die uns – bis auf den Nachweis von P^+ – jede Rechenarbeit erspart. In diesem Satz bezeichnen r, s ausschließlich Elemente aus $\mathbb{Q}$. Wir benötigen dort außerdem

(13) $(1+a)^{\varepsilon_n} \leq 1 + a\varepsilon_n$.

Nach der Bernoullischen Ungleichung ist nämlich $1 + a = 1 + 10^n a\varepsilon_n \leq (1 + a\varepsilon_n)^{10^n}$. Hieraus folgt (13) durch Ziehen der 10^n-ten Wurzel auf beiden Seiten.

Satz 7.3. *Zu jeder positiven reellen Zahl b gibt es genau eine Funktion $f : \mathbb{Q} \to \mathbb{D}$ mit den Eigenschaften*

$$\mathsf{P}^1: \quad f1 = b, \qquad \mathsf{P}^+: \quad f(r+s) = f(r) \cdot f(s),$$

nämlich $f_b : r \mapsto b^r$ mit dem Definitionsbereich $\mathbb{Q}$. Darüber hinaus erfüllen die Funktionen f_b alle weiteren Potenzgesetze. Schließlich ist f_b für $b > 1$ schlicht und strikt wachsend; für $b < 1$ hat die Funktion $\frac{1}{f_b}$ diese Eigenschaften.

Beweis. f erfülle $\mathsf{P}^1, \mathsf{P}^+$. Dann ist, wie zu Beginn dieses Teilabschnitts schon festgestellt wurde, notwendigerweise $f\frac{m}{n} = \sqrt[n]{b^m}$. Denn man hätte dort doch nur überall fr statt b^r zu schreiben brauchen. Folglich ist $f = f_b$, und die Eindeutigkeit ist gezeigt. Gewiss gilt P^1 für f_b. Sei nun $r = \frac{m}{n}$, $s = \frac{k}{n}$. Dann folgt P^+ mit $\mathsf{R}_1, \mathsf{R}_2$ wie folgt:

$$b^{r+s} = b^{\frac{m}{n}+\frac{k}{n}} = b^{\frac{m+k}{n}} = \sqrt[n]{b^{m+k}} = \sqrt[n]{b^m \cdot b^k} = \sqrt[n]{b^m} \cdot \sqrt[n]{b^k} = b^r \cdot b^s.$$

Auch folgt $\mathsf{P}^<$ unmittelbar aus R_3. Aufgrund von Bemerkung 2 verbleiben daher nur die Beweise von $\mathsf{P}_\times$ und $\mathsf{P}^\times$. Seien $a, b > 0$. Man betrachte $f : r \mapsto (ab)^r$ und $g : r \mapsto a^r \cdot b^r$. Beide Funktionen erfüllen offenbar P^1, mit dem Wert ab an der Stelle $r = 1$. Sie erfüllen auch P^+. Für f ist das klar, und nur für g muss dies wirklich verifiziert werden:

$$g(r+s) = a^{r+s} \cdot b^{r+s} = a^r a^s b^r b^s = a^r b^r a^s b^s = g(r) \cdot g(s).$$

Nach der bereits bewiesenen Eindeutigkeitsaussage ist also notwendigerweise $f = g$ und $\mathsf{P}_\times$ ist bewiesen. Völlig analog zeigt man $\mathsf{P}^\times$ durch Vergleich von $f : s \mapsto (b^r)^s$ und $g : s \mapsto b^{rs}$ für festes $r \in \mathbb{Q}$, die bei $s = 1$ beide den Wert b^r haben.

Schlichtheit folgt, wenn f_b nur in jedem Intervall $[0, k]$ schlicht ist. Sei $b = 1 + a$ für ein $a \geq 0$, also $b^{\varepsilon_n} - 1 = (1+a)^{\varepsilon_n} - 1 \leq a\varepsilon_n$ gemäß (13). Für $r \in \mathbb{Q} \cap [0, k]$ ist dann

$$\begin{aligned} b^{r+\varepsilon_n} - b^r \;=\; b^r b^{\varepsilon_n} - b^r \;&=\; b^r(b^{\varepsilon_n} - 1) \\ &\leq\; b^r a\varepsilon_n \\ &<\; b^k b\varepsilon_n \;=\; b^{k+1}\varepsilon_n \quad \text{(wegen } r \leq k \text{ und } a < b\text{)}. \end{aligned}$$

Folglich ist f_b schlicht in $[0, k]$. Striktes Wachstum gilt wegen $\mathsf{P}^<$. Die Behauptung im Falle $b < 1$ folgt wegen $\frac{1}{b^r} = \left(\frac{1}{b}\right)^r$ in ganz entsprechender Weise. ∎

Dieser Satz besagt zugleich, dass für rationale Exponenten alle Potenzregeln schon aus P^1 und P^+ folgen, d.h. werden $\mathsf{P}^1, \mathsf{P}^+$ von $f : \mathbb{Q} \to \mathbb{D}$ erfüllt, so erfüllt f notwendigerweise auch alle übrigen Potenzgesetze.

Sei $b > 0$ eine beliebige reelle Zahl. Wir werden nun b^x auch für irrationale x erklären, und zwar einfach durch natürliche Fortsetzung. Nach der Eindeutigkeitsaussage von Satz 7.5 ist diese Art der Erklärung ohnehin die einzige, welche die Gültigkeit aller Potenzgesetze, ja nur der Basisregeln, für beliebige Exponenten zu sichern imstande wäre. Etwas genauer, wir erklären die Funktion $x \mapsto b^x$ für beliebige $x \geq 0$ durch natürliche Fortsetzung der entsprechenden, auf $\mathbb{E}$ eingeschränkten Funktion. Man benötigt dafür nur die Werte von b^r für $r \in \mathbb{E}$, obwohl dieser Ausdruck für alle rationalen Exponenten eigentlich schon festgelegt wurde. Sie umfasst sozusagen noch einmal alle rationalen Exponenten $\frac{m}{n}$, die nicht zu $\mathbb{E}$ gehören. Nach einer Bemerkung vor Satz 7.2 folgt aber die Übereinstimmung von $a^{\frac{m}{n}}$ mit $\sqrt[n]{a^m}$ auch für $\frac{m}{n} \notin \mathbb{E}$. Man beachte, auch die automatischen Rechner bestimmen b^x näherungsweise durch die Berechnung von $b^{x.n}$ mit einer durch den Rechner gegebenen Stellenzahl n solange sie numerisch rechnen, unabhängig davon ob x rational ist oder nicht. Sie müssen b^r nur für $r \in \mathbb{E}$ berechnen können.

Definition. Für $b \geq 1$ und $x \in \mathbb{D}$ sei $b^x = \lim\langle b^{x.n}\rangle$. Für $0 < b < 1$ sei $b^x = \frac{1}{a^x}$, mit $a := \frac{1}{b}$. Ferner sei $0^x = 0$ für alle $x \neq 0$. Für $b > 0$ heißt $\exp_b : \mathbb{D} \to \mathbb{D}$ mit $\exp_b x = b^x$ die *Exponentialfunktion mit der Basis b.*

Die Definition liefert für den Fall $x \in \mathbb{E}$ (d.h. $x = x.n$ für gewisses n) im Falle $b \geq 1$, und wegen $b^{x.n} = \frac{1}{a^{x.n}}$ nach Satz 7.3 auch im Falle $b < 1$ nichts neues. Daher gilt $b^x = \frac{1}{a^x}$ für alle $x \in \mathbb{D}$. Also steht uns bereits folgender Spezialfall von $\mathsf{P}_\times$ zur Verfügung:

(14) $\exp_{\frac{1}{b}} x = \frac{1}{\exp_b x}$ für alle $x \in \mathbb{D}$.

Weil ferner $g : \mathbb{E} \mapsto \mathbb{D}$ mit $gx = b^x$ nach Satz 7.3 für $b > 1$ schlicht und strikt monoton ist, erhalten wir nach Satz 7.2 dasselbe für deren natürliche Fortsetzung $\exp_b$, also das

Korollar. $\exp_b$ *ist für* $b > 1$ *schlicht und strikt monoton; ebenso* $\frac{1}{\exp_b}$ *für* $b < 1$.

Die Funktionenschar $\exp_b$ spielt in der Analysis eine fundamentale Rolle, wobei die Exponentialfunktion mit $b = \mathbf{e}$ aus mehreren Gründen besonders ausgezeichnet ist. Zum Beispiel ist sie die einzige durch den Punkt (0;1) verlaufene differenzierbare Funktion, die mit ihrer eigenen Ableitung übereinstimmt. Man schreibt meist nur exp für $\exp_{\mathbf{e}}$, so dass also $\exp(x) = \mathbf{e}^x$. Figur 2 gibt eine räumliche Darstellung dieser Funktionenschar für Werte der Basis zwischen $\frac{1}{\mathbf{e}} = 0{,}36\cdots$ und $\mathbf{e} = 2{,}71\cdots$ einschließlich, und dem Variablenbereich $0 \le x \le 2$. Es ist dies die auf der Fläche von links vorn nach rechts hinten gezeichnete Kurvenschar. Der rechte Rand des Flächenstücks zeigt gerade exp. Jede der Funktionen $\exp_b$ wächst für $b > 1$ mit größer werdendem x unbeschränkt und nimmt jeden Wert ≥ 1 genau einmal an. $\exp_b$ fällt anderseits für $b < 1$ sehr schnell auf nahezu 0 ab, wie dies die linke Randkurve des Flächenstücks verdeutlicht.

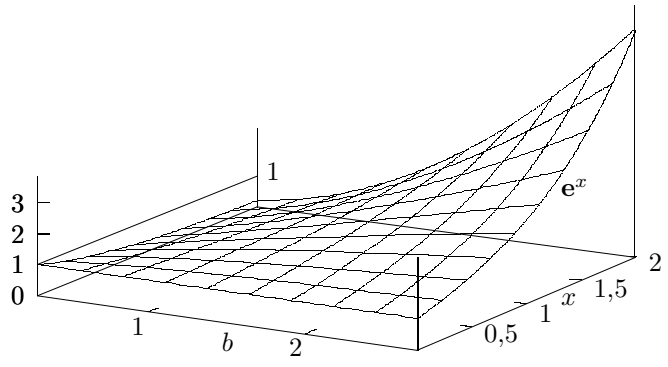


Fig. 2 Die Schar der Exponentialfunktionen $\exp_b$ *für* $\frac{1}{\mathbf{e}} \le b \le \mathbf{e}$

Eine maßstabgetreue Abbildung zeigt auch Figur 7 in **9.3**, wo diese Funktion auf beliebige reelle Argumente unter Einschluss negativer Exponenten erweitert wird. Es genügt, aus der Schar der Funktionen $\exp_b$ nur eine, etwa exp, zu kennen; die anderen kann man daraus leicht gewinnen. Denn offenbar ist

$$\exp(cx) = \mathbf{e}^{cx} = (\mathbf{e}^c)^x = b^x = \exp_b(x), \text{ mit } b = \mathbf{e}^c.$$

Für vorgegebenes $b > 1$ ist also $\exp_b$ für $b > 1$ durch exp gegeben; es ist nur das nach dem Zwischenwertsatz existierende, eindeutig bestimmte c mit $b = \mathbf{e}^c$ zu berechnen, welches auch mit $\ln b$ bezeichnet wird, siehe **7.4**. Dieser enge Zusammenhang ist der Grund, warum einfach nur von *der* Exponentialfunktion gesprochen wird und womit in der Regel exp gemeint ist. Wegen (12) lässt sich auch die Exponentialfunktion mit einer Basis < 1 durch eine rationale Operation gewinnen. Nach Einführung der negativen Zahlen darf man auch schreiben $\exp_{\frac{1}{b}}(x) = \exp_b(-x)$, d.h. man wählt im Term $\exp(cx)$ einfach $c = -1$. Mit anderen Worten, $x \mapsto \mathbf{e}^{cx}$ mit $c \in \mathbb{R}$ ist die allgemeine Form einer Exponentialfunktion. Diese hat viele Darstellungsformen. Wir erwähnen vor allem

(15) $\mathbf{e}^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots,$

(16) $\mathbf{e}^x = \lim\langle(1 + \frac{x}{n})^n\rangle,$

wobei die erste für die Analysis besonders wichtig ist. Die unendliche Reihe rechts in (15) ist Beispiel einer so genannten *Potenzreihe* und für jedes x konvergent; denn deren Partialsummenfolge ist schlicht gemäß Übung 6.3. Dort wurde auch die Folge rechts in (16) als schlicht und damit konvergent erkannt.

Die Gleichung $\lim\langle(1 + \frac{x}{n})^n\rangle = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots$ beweist man mittel der binomischen Formel fast genau so, wie dies in **7.1** für den Fall $x = 1$ geschah. Es genügt also, (16) auf elementarem Wege zu beweisen, um auch (15) ohne Differentialkalkül oder Potenzreihentheorie zu erhalten, siehe dazu Übung 4. Jede der beiden Gleichungen (15) und (16) eröffnet eine weitere Erklärungsmöglichkeit der Exponentialfunktion.

Mit einer bloßen Erklärung von $\exp_b$ oder b^x ist allerdings noch nicht viel gewonnen. Worauf es vor allem ankommt ist der Nachweis der uneingeschränkten Gültigkeit aller Potenzgesetze. Wir werden diesen Nachweis nunmehr lückenlos erbringen und darüber hinaus beweisen, dass die Definition von b^x durch natürliche Fortsetzung die *einzige* Erklärungsmöglichkeit darstellt, um auch nur einige der Potenzgesetze – nämlich die Basisregeln – zu erhalten. Letzteres ist nicht nur eine an sich interessante Tatsache, sondern sie wird uns helfen, ohne viel Rechnen auch die übrigen Potenzgesetze nachzuweisen. Dass diese dann eine notwendige Folge der Basisregeln sind, ergibt sich ganz nebenbei. Anders als in Satz 7.3 erscheinen in Satz 7.4 gewisse Ungleichungen. Wir erwähnen, dass diese oder äquivalente Bedingungen (z.B. die Monotonie oder Stetigkeit) für Satz 7.4 unverzichtbar sind. In Abschnitt **10** werden wir diesen Satz auf eine ganz andere Weise erhalten, nämlich als Nebenprodukt des Maßzahlensatzes.

Satz 7.4. *Zu jedem $b \in \mathbb{D}_+$ gibt es eine und nur eine Funktion $f : \mathbb{D} \to \mathbb{D}$ mit den Eigenschaften $\mathsf{P}_{\geq 1}$: $fx \geq 1$ für $b \geq 1$ und $\mathsf{P}_{\leq 1}$: $fx \leq 1$ für $b \leq 1$, sowie*

$$\mathsf{P}^1 : \ f1 = b, \qquad \mathsf{P}^+ : \ f(x + y) = f(x) \cdot f(y) \text{ für alle } x, y \in \mathbb{D},$$

nämlich $f = \exp_b$. Darüber hinaus erfüllen die Funktionen $\exp_b$ alle Potenzgesetze.

Beweis. Wir beweisen Eindeutigkeit zuerst und betrachten zunächst den Fall $b \geq 1$. f erfülle die genannten Voraussetzungen. Die Einschränkung von f auf $\mathbb{Q}$ genügt den Voraussetzungen von Satz 7.3. Also ist $f(\frac{m}{n}) = b^{\frac{m}{n}}$. Auch ist diese Einschränkung schlicht, und dasselbe gilt natürlich auch für die mit g bezeichnete Einschränkung von f auf $\mathbb{E}$. Weil die (gewöhnliche) Monotonie von f nach Bemerkung 1 aus den Basisregeln direkt folgt – und nur hier benötigen wir die Eigenschaft $\mathsf{P}_{\geq 1}$ – ist f monoton, und somit nach Satz 7.2 gerade die natürliche Fortsetzung von g; kurzum, $f = \exp_b$. Damit ist die Eindeutigkeit bewiesen und auch $\exp_b(\frac{m}{n}) = b^{\frac{m}{n}}$ ist gezeigt. Dasselbe gilt im Falle $b < 1$; denn $\frac{1}{\exp_b}$, und gemäß $\mathsf{P}_{\leq 1}$ auch $\frac{1}{f}$, sind monoton wachsende Fortsetzungen der nach Satz 7.3 schlichten Funktion $\frac{1}{g}$, und damit identisch.

Es verbleibt der Nachweis, dass die $\exp_b$ tatsächlich alle Potenzgesetze erfüllen. P^1 ist klar. P^+ ergibt sich im Falle einer Basis $b \geq 1$ wie folgt:

$$\begin{aligned} b^x \cdot b^y &= \lim\langle b^{x_{.n}}\rangle \cdot \lim\langle b^{y_{.n}}\rangle && \text{(Definition)}\\ &= \lim\langle b^{x_{.n}} \cdot b^{y_{.n}}\rangle && \text{(Spezialfall von Übung 5.3)}\\ &= \lim\langle b^{x_{.n}+y_{.n}}\rangle && \text{(Satz 7.3)}\\ &= b^{x+y} && \text{(wegen (11) und weil } x+y = \lim\langle x_{.n}+y_{.n}\rangle). \end{aligned}$$

Der Fall $b < 1$ lässt sich auf den behandelten sofort zurückführen; denn ist $a = \frac{1}{b}$, so gilt $b^x = \frac{1}{a^x}$ nach (14), und weil $a > 1$, folgt $b^x \cdot b^y = \frac{1}{a^x} \cdot \frac{1}{a^y} = \frac{1}{a^{x+y}} = b^{x+y}$.

Nach Satz 7.3 gilt für die Einschränkung g von $\exp_b$ auf $\mathbb{E}$ sicher $gx > 1$ für $b > 1$, und $gx < 1$ für $b < 1$, wenn immer $x \in \mathbb{E}_+$. Dasselbe gilt dann offenbar auch für die natürliche Fortsetzung von g. Damit erhalten wir nicht nur $P_{\geq 1}, P_{\leq 1}$ sondern leicht auch $P^<$. Es verbleiben nach Bemerkung 2 also lediglich die Nachweise von $\mathsf{P}_\times$ und $\mathsf{P}^\times$. Hierfür verwenden wir ähnlich wie in Satz 7.3 die schon bewiesene Eindeutigkeitsaussage. Für $\mathsf{P}_\times$ betrachte man im Falle $a, b \geq 1$ oder $a, b < 1$ die Funktionen $f : x \mapsto (ab)^x$ und $g : x \mapsto a^x \cdot b^x$ zuerst für $a, b \geq 1$. Sowohl f als auch g erfüllen sicher P^1 (mit ab für b). Beide Funktionen erfüllen offenbar auch P^+, sowie die Bedingungen $P_{\geq 1}$ und $P_{\leq 1}$. Damit folgt $f = g$, und $\mathsf{P}_\times$ ist für diesen Fall gezeigt. Die verbleibenden Fälle $a < 1 \leq ab \leq b$ und $a \leq ab < 1 \leq b$ lassen sich auf den behandelten leicht zurückführen. Zum Beispiel ist im ersten Fall nach dem Bewiesenen $b^x = (ab)^x(\frac{1}{a})^x = (ab)^x \frac{1}{a^x}$. In derselben Weise zeigt man $\mathsf{P}^\times$ durch Betrachtung von $f : x \mapsto (b^y)^x$ und $g : x \mapsto b^{y \cdot x}$. ∎

7.4 Logarithmen

In der Vergangenheit war das logarithmische Rechnen, d.h. die Verwandlung von Multiplikationen und Divisionen in Additionen und Subtraktionen, eine der wesentlichen Anwendungen der Logarithmen. Heute ist diese Art des Rechnens kaum noch üblich. Aber das Vorhandensein der Potenzen und Logarithmen auf Taschenrechnern deutet darauf hin, dass sie unabhängig davon ein breites Anwendungsspektrum haben. Das rührt vor allem daher, dass in Wissenschaft und Technik häufig Formeln verwendet werden, die diese Funktionen explizit enthalten. Uns geht es hier weniger um eine Auflistung von Anwendungen, sondern um die Klärung des Begriffs selbst. Ein beliebtes Verfahren ist z.B. die Einführung der logarithmischen Funktion ln durch

$$\ln x = \int_1^x \frac{dt}{t} \qquad (x > 0).$$

Das setzt natürlich voraus, dass der Integralkalkül voll entwickelt ist, weil aus dieser Definition die Eigenschaften von ln herzuleiten sind. Die wohl eleganteste Einführung von ln ist ihre Definition als gewisser Isomorphismus von der multiplikativen Gruppe der positiven Zahlen in die additive Gruppe der reellen Zahlen, siehe hierzu **10.6**.

Nach den bisherigen Vorbereitungen können wir indes die Logarithmusfunktion direkt als Umkehrung der Exponentialfunktion erklären. Für $b > 1$ ist $\exp_b$ gemäß dem Korollar in **7.3** schlicht, strikt monoton wachsend und unbeschränkt. Nach dem Zwischenwertsatz existiert also zu jedem $y \geq 1$ genau ein $x \geq 0$ mit $b^x = y$. Ist aber $0 < b < 1$,

so hat $x \mapsto \frac{1}{\exp_b x}$ diese Eigenschaften. Daraus folgt offensichtlich, dass $\exp_b$ jetzt jeden Wert $y \leq 1$ genau einmal annimmt.

Definition. Sei $b > 1$ und $x \geq 1$ oder aber $b < 1$ und $x \leq 1$. Dann sei $\log_b x$ der eindeutig bestimmte Wert y derart, dass $b^y = x$. Die Funktion $x \mapsto \log_b x$ heißt die *Logarithmusfunktion zur Basis b.*

Für $\log_{10}$ schreibt man häufig lg (die BRIGG*schen Logarithmen*). Für $\log_2$ schreibt man meist ld (***l**ogarithmus **d**ualis*). Im Falle $b = \mathbf{e}$ spricht man vom *natürlichen Logarithmus* (***l**ogarithmus **n**aturalis*) und schreibt in der Regel ln für $\log_{\mathbf{e}}$.

Nach Definition gilt stets $\log_b 1 = 0$, $\log_b b = 1$, und allgemeiner $\log_b b^n = n$. Konkrete Zahlenwerte sind z.B. $\ln 10 = 2{,}30258\cdots$ und $\lg 2 = 0{,}30103\cdots$. Man beachte, dass immer $\log_b a \cdot \log_a b = \log_b(b) = 1$ gemäß L_3 unten. Figur 3 am Schluss veranschaulicht die Funktion ld. Ihr Bild entsteht aus dem von $\exp_2$ durch Spiegelung an der Achse $x = y$. In **9** werden beide Funktionen erweitert. Dabei wird sich herausstellen, dass $\operatorname{ld} x$ für $0 < x < 1$ sinnvoll nur als eine negative Zahl erklärt werden kann.

Aus der Definitionsidentität $b^{\log_b x} = x$ ergeben sich mit Hilfe der Potenzgesetze einige für das Rechnen mit Logarithmen maßgebliche Gleichungen, nämlich

L_1 : $\log_b xy = \log_b x + \log_b y$,
L_2 : $\log_b x^y = y \cdot \log_b x$,
L_3 : $\log_a x = \frac{\log_b x}{\log_b a}$ (insbesondere $\log_a b = \frac{1}{\log_b a}$).

Es folgt L_1 aus $b^{\log_b xy} = xy = b^{\log_b x} \cdot b^{\log_b y} = b^{\log_b x + \log_b y}$ durch Vergleich der Exponenten, und L_2 aus $b^{\log_b x^y} = x^y = (b^{\log_b x})^y = b^{y \cdot \log_b x}$. Schließlich ergibt sich L_3 aus

$$b^{\log_b x} = x = a^{\log_a x} = \left(b^{\log_b a}\right)^{\log_a x} = b^{\log_b a \,\cdot\, \log_a x}.$$

Gleichung L_1 ist die Grundlage des logarithmischen Rechnens mittels einer Tafel. Um gegebene Werte x, y zu multiplizieren, entnimmt man der Tafel die Werte $\log_b x$ und $\log_b y$, addiert diese und erhält $\log_b xy$. Sodann liest man aus der Tafel den Wert $x \cdot y$ ab. L_2 sorgt dafür, dass Logarithmen nur in einem engeren Bereich tabuliert werden müssen, etwa für Werte x mit $1 \leq x < 10$ bei $b = 10$; denn jedes $x \geq 10$ lässt sich für geeignetes x' als $x = x' \cdot 10^n$ mit $1 \leq x' < 10$ schreiben, was auf bloße Kommaverschiebung in x hinausläuft. L_3 schließlich dient der Umrechnung der Logarithmen von einer Basis in die andere. Gemäß L_3 ist z.B. $\lg x = \frac{1}{\ln 10} \cdot \ln x = \frac{1}{2{,}30\cdots} \cdot \ln x$ für alle $x \geq 1$.

Bei obigen Formeln und den Beweisen ist vorerst darauf zu achten, dass die Argumente jeweils im Definitionsbereich von $\log_b$ liegen, also z.B. $x, y \geq 1$ für eine Basis $b > 1$ in L_1. Das bedeutet natürlich eine erhebliche Behinderung. Denn sobald ein Faktor < 1 ist, wäre man mit der Kunst des logarithmischen Multiplizierens am Ende, und die Wahl einer Basis < 1 schafft entsprechende Probleme mit Faktoren > 1. Hier also tritt die Notwendigkeit der Einbeziehung negativer Zahlen in den Kalkül mit aller Deutlichkeit hervor. Wären diese nicht schon da gewesen, hätte man sie nun erfinden müssen. Historisch gesehen waren es in der Tat die Logarithmen, durch die alle Vorbehalte gegen negative Zahlen endgültig schwanden. Allerdings definierte man bis ins 19. Jahrhundert

hinein man negative Zahlen vorwiegend geometrisch und nur teils algebraisch, aber im Grunde nur intuitiv. Deswegen ist die Ablehnung negativer Zahlen durch den Erfinder der „Buchstabenrechnung“ VIETA (1540 – 1603) durchaus verzeihlich.

Taschenrechner ermitteln $\log_b x$ bekanntlich in Sekundenbruchteilen. Auch die Erfinder der ersten Logarithmentafel haben weniger gerechnet als man zunächst glaubt. Das beruht wesentlich auf einer geschickten Basiswahl. In allen Fällen lag und liegt diese auch in modernen Taschenrechnern in der Nähe von $\mathbf{e}$ oder dem Reziproken $\frac{1}{\mathbf{e}}$. Sieht man von Unerheblichkeiten ab, so hatte die erste Logarithmentafel von NAPIER im Jahre 1614 die Basis $(1-\frac{1}{10^7})^{10^7}$, welche näherungsweise $=\frac{1}{\mathbf{e}}$ ist, Übung 6.2. Für die Tafel von BÜRGI (1620) war sie $(1+\frac{1}{10^4})^{10^4}$ $(=2{,}71814\cdots$ gemäß (e4) in **7.1**). Diese stimmt immerhin in den ersten drei Kommastellen mit $\mathbf{e}$ überein.

Die Basis $b_n := e_{10^n} = (1+\frac{1}{10^n})^{10^n} = (1+\varepsilon_n)^{10^n}$ ist für die Berechnung einer n-stelligen Logarithmentafel vor allem deswegen sehr geeignet, weil $\Delta\log_{b_n} x$ nahezu gleich $\frac{\Delta x}{x}$ ist für kleine Schrittweiten Δx von x, Übung 5. Um diesen Vorteil etwas zu erläutern, setze man $x_k := (1+\varepsilon_n)^k$. Weil $b_n^{k\varepsilon_n} = (1+\varepsilon_n)^{10^n\cdot k\varepsilon_n} = x_k$, folgt $\log_{b_n} x_k = k\varepsilon_n$ und die Schrittweite dieser Werte für $k = 0, 1, 2, \ldots$ ist gerade ε_n. Man erhält so den durch die Tabelle unten beschriebenen Anfang einer Logarithmentafel. Die Zahlenwerte in Klammern entsprechen der Wahl $n = 4$, also denen einer 4-stelligen Tafel.

x	$(n=4)$	$\log_{b_n} x$	$(\log_{b_4} x)$
1	(1,0000)	0	0
$1+\varepsilon_n$	(1,0001)	ε_n	(0,0001)
$(1+\varepsilon_n)^2$	(1,0002)	$2\varepsilon_n$	(0,0002)
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(1+\varepsilon_n)^{100}$	(1,0100)	$100\varepsilon_n$	(0,0100)
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(1+\varepsilon_n)^{141}$	(1,0141)	$141\varepsilon_n$	(0,0141)
$(1+\varepsilon_n)^{142}$	(1,0143)	$142\varepsilon_n$	(0,0142)
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Logarithmentafel zur Basis e_{10^n}

Wie aus den Zahlenwerten der Tafel ersichtlich, beträgt für kleine k die Schrittweite $x_{k+1} - x_k$ des Arguments (oder *Numerus*) x auch nur ε_n. Denn $x_k = (1+\varepsilon_n)^k$ ist für nicht zu große k nahezu gleich $1 + k\varepsilon_n$. Für $k^2 \le 10^n$ ist x_k gleich $1 + k\varepsilon_n$ sogar bis auf n Stellen genau. Denn für $\varepsilon := \varepsilon_n$ ist dann $k^{1+i}\varepsilon^i \le 1$ für alle $i \ge 1$, und wir erhalten

$$(1+\varepsilon)^k - (1+k\varepsilon) = \tbinom{k}{2}\varepsilon^2 + \tbinom{k}{3}\varepsilon^3 + \ldots + \tbinom{k}{n}\varepsilon^n \le \varepsilon(\tfrac{k^2\varepsilon}{2!} + \tfrac{k^3\varepsilon^2}{3!} + \ldots + \tfrac{k^n\varepsilon^{n-1}}{n!})$$
$$\le \varepsilon(\tfrac{1}{2!} + \ldots + \tfrac{1}{n!}) < \varepsilon(\mathbf{e} - 2) < \varepsilon.$$

So ist z.B. noch $k^2\varepsilon_4 \le 1$ für $k \le 100$, also ist $(1+\varepsilon_4)^{100} = 1 + 100\varepsilon_4$ in den Grenzen 4-stelliger Genauigkeit. Dies gilt sogar noch für $k = 141$, denn $(1+\varepsilon_4)^{141} = 1,0141991\cdots$.

Erst ab $k = 142$ beginnt der Numerus beim Erstellen der Tafel von rechts nach links allmählich zu springen. Aufgrund dieser Verhältnisse ist es bequemer, nicht den Logarithmus auszurechnen, sondern man schreibt diesen schrittweise hin und berechnet den zugehörigen Numerus, wobei sich in der angegebenen Weise und mit anderen Tricks der Rechenaufwand ziemlich gering halten lässt. Auch moderne Rechner gehen etwa in dieser Weise vor; durch blitzschnelles Ausrechnen der Exponentialfunktion suchen sie den der Eingabe am nächsten gelegenen Numerus.

Obige Tafel entspricht derjenigen von Bürgi. Napiers Tafel hingegen kann (mit umgekehrtem Vorzeichen) näherungsweise als eine solche zur Basis **e** angesehen werden, weil $\log_{\frac{1}{b}} x = -\log_b x$, **9.3**. Es ist deshalb unerheblich, ob $(1+\varepsilon_n)^{10^n}$ oder $(1-\varepsilon_n)^{10^n}$ die Berechnungsbasis einer n-stelligen Tafel darstellt. Beide Werte sind in den Grenzen n-stelliger Genauigkeit reziprok zueinander; denn $(1+\varepsilon_n)^{10^n}(1-\varepsilon_n)^{10^n} = (1-\varepsilon_n^2)^{10^n}$, was unter Beachtung von (5) in **3.5** die behauptete Genauigkeits-Abschätzung ergibt:

$$1-\varepsilon_n = 1-10^n\varepsilon_n^2 < (1-\varepsilon_n^2)^{10^n} < 1.$$

Hat man keine Tafel oder Rechner zur Hand, hilft oft eine Überschlagsrechnung zur Ermittlung des Logarithmus einer Zahl bis auf etwa eine Kommastelle. Will man z.B. $\lg 2$ ermitteln, gehe man aus von $2^9 < 10^3 < 2^{10}$. Das ergibt offenbar $9\lg 2 < 3 < 10\lg 2$, also $\frac{9}{10}\lg 2 < 0{,}3 < \lg 2$, und somit $0{,}3 < \lg 2 < \frac{10}{9}\cdot 0{,}3 = 0{,}333\cdots$. Damit wurde die erste Kommastelle von $\lg 2$ präzise ermittelt.

Sei $k \in \mathbb{N}$. Wir wollen die Anzahl $\ell_2 k$ der Binärziffern von k abschätzen, siehe **1.2**. Dies ist wichtig für die Planung des Umfangs von Speicherzellen von Rechnern, die mit vorgeschriebener, z.B. 10-stelliger Genauigkeit rechnen sollen. Setzt man $m := \operatorname{Int}\operatorname{ld} k$, so folgt $2^m \le 2^{\operatorname{ld} k} \ (= k) < 2^{m+1}$. Weil nun 2^m genau $m+1$ Binärziffern hat und 2^{m+1} kleinste natürliche Zahl ist mit einer Binärziffer mehr, ergibt sich offenbar $\ell_2 k = m+1$, also $\ell_2 k = \operatorname{Int}\operatorname{ld} k + 1$. Alles eben Gesagte gilt nun ganz analog für jede Grundzahl $g \ge 2$. Damit erhalten wir für die g-adische Länge von k die exakte Formel

$$(17)\quad \ell_g k = \operatorname{Int}\log_g k + 1 \qquad (g \ge 2,\ k > 0).$$

Gesetzt, wir verfügen im Moment über keine Rechenhilfsmittel und sind genötigt, die binäre Länge von 10^{10} überschlagsmäßig zu ermitteln. Soviel Bitplätze benötigt man gerade, um eine beliebige der 10^{10} natürlichen Zahlen $< 10^{10}$, wie sie in der Anzeige eines Taschenrechners mit 10-stelligem Display erscheinen, intern zu speichern. Wegen $2^3 < 10 < 2^4$ ist $3 < \operatorname{ld} 10 < 4$. Das aber genügt nicht, um $\ell_2 10^{10}$ zu bestimmen, denn dazu benötigen wir nach (17) die natürliche Zahle $\operatorname{Int}\operatorname{ld} 10^{10} = \operatorname{Int}(10\cdot\operatorname{ld} 10)$, also mindestens eine Kommastelle von $\operatorname{ld} 10$. Diese ermitteln wir durch *lineare Interpolation*: Der Verlauf von ld zwischen den Punkten (8;3) und (16;4) unterscheidet sich nur wenig von dem einer diese Punkte verbindenden Geraden g, der *Interpolationsgeraden*, Figur 3. Eine elementare Überlegung zeigt, dass der Punkt (10;3,25) auf g liegt. ld verläuft in geringem Abstand oberhalb von g, so dass $\operatorname{ld} 10 \approx 3{,}3$ angenommen werden kann. Es ist $\ell_2 10^{10} = \operatorname{Int}(10\cdot\operatorname{ld} 10) + 1$ und dies ist gemäß Schätzung der Wert $10\cdot 3{,}3 + 1 = 34$. Tatsächlich ist $\operatorname{ld} 10 = \frac{1}{\lg 2} = 3{,}32192\cdots$, so dass diese Schätzung nahezu exakt ist.

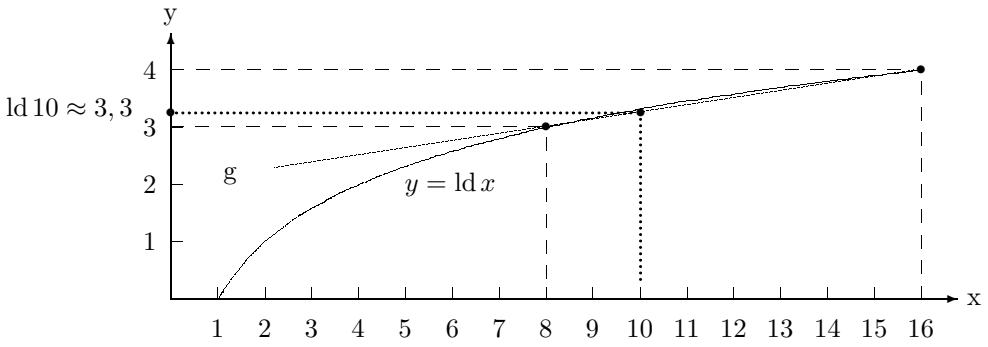


Fig. 3 Die Funktion ld *mit Interpolationsgerade durch* (8;3) *und* (16;4).

Wir haben hier stillschweigend benutzt, dass ld *konvex* ist, d.h. ist $1 \leq a < x < b$ und h die Interpolationsgerade durch die Punkte $(a;\mathrm{ld}\,a)$, $(b;\mathrm{ld}\,b)$, so ist $hx < \mathrm{ld}\,x$. Es genügt, dies für ln nachzuweisen (Übung 6); denn dasselbe gilt dann für alle Funktionen $\log_b$ mit $b > 1$, weil $\log_b x = c \cdot \ln x$ mit $c = \log_b \mathbf{e}$. Die Beobachtung, dass ld oder ln auf dem Bildschirm eines Rechners konvex erscheinen, ist nur ein Hinweis, aber noch lange kein Beweis, dass diese Funktionen tatsächlich konvex sind.

7.5 Übungen

1. Man zeige $\sqrt{2 \cdot \sqrt{2 \cdot \sqrt{2}}} \ldots = 2$, d.h. es ist zu zeigen, dass die (monotone) Folge $a_1 = \sqrt{2},\ a_2 = \sqrt{2\sqrt{2}},\ a_3 = \sqrt{2\sqrt{2\sqrt{2}}}, \ldots$ konvergiert und den Grenzwert 2 hat.

2. Sei $e_0 = 1,\ a_0 = 3$, $e_n = (1 + \frac{1}{n})^n$, $a_n = e_n(1 + \frac{1}{2n})$ $(n > 0)$ und $b_n = e_n(1 + \frac{1}{2n+1})$. Man beweise: $\langle a_n \rangle$ konvergiert strikt fallend und $\langle b_n \rangle$ strikt wachsend gegen $\mathbf{e}$. Dies verdeutlicht, wie empfindlich das Monotonieverhalten gewisser Folgen ist.

3. Man beweise mit Hilfe von Übung 2 die Ungleichungen (e1) und (e2) in **7.1**.

4. Sei $f_n : x \mapsto (1 + \frac{x}{n})^n$ und $f : x \mapsto \lim \langle f_n x \rangle$. Man beweise auf elementarem Wege $fx = \mathbf{e}^x$, womit zugleich (15) und (16) in **7.3** bewiesen sind. (15) ergibt insbesondere $1 + x < \mathbf{e}^x$, oder gleichwertig, $\ln(1 + x) < x$, für alle $x \geq 0$.

5. Man beweise (a) $x - \frac{x^2}{2} < \ln(1 + x)$ für $x < 1$. Damit zeige man, für alle $x \geq 1$ und $0 < \Delta x < 1$ gilt (b) $0 < \frac{1}{x} - \frac{\Delta \ln x}{\Delta x} < \frac{\Delta x}{2x^2}$. Also ist $\frac{\Delta \ln x}{\Delta x} \approx \frac{1}{x}$, mit einem Fehler der Größenordnung Δx.

6. Sei $c > 0$. Man zeige (a) $f_c : x \mapsto (1 + \frac{c}{x})^x$ wächst monoton, oder gleichwertig, $x \mapsto (1 + cx)^{\frac{1}{x}}$ fällt monoton. (b) ln ist konvex. (c) Ist $\langle a_n \rangle$ eine (fallende) Nullfolge $(a_n \neq 0)$, so konvergiert $\langle \frac{\ln(1+a_n)}{a_n} \rangle$ wachsend gegen 1.

7. Sei f die überall in $\mathbb{D}$ erklärte Funktion $x \mapsto \frac{2ax}{a+x^2}$ $(a > 0)$. Man zeige: f ist schlicht im Intervall $[0, \sqrt{a}\,]$, und $x \leq fx \leq f(\sqrt{a}) = \sqrt{a}$ für alle $x \leq \sqrt{a}$.

Abschnitt 8

Elementare Rechenverfahren

Dieser Abschnitt dient dem Verständnis praktisch wichtiger numerischer Algorithmen. Einige davon gehören in ihrer Grundstruktur zur Hardware automatischer Rechner, wie etwa der in **8.1** ausführlich behandelte Divisionsalgorithmus, eines der wichtigsten Rechenverfahren überhaupt. Bei seiner Ausführung muss nur addiert, subtrahiert und multipliziert werden. Mit demselben Verfahren bewerkstelligen die Computer auch die Umrechnung von einer Zahlendarstellung in die andere. Bekanntlich rechnen die meisten Computer binär. Um mit dem Nutzer kommunizieren zu können, müssen sie imstande sein, Zahldarstellungen in verschiedenen Systemen blitzschnell ineinander umzurechnen. Dies geschieht mittels des in **8.2** behandelten g-adischen Algorithmus, der nichts anderes ist als der Divisionsalgorithmus mit speziellen Parameterwerten.

Eine elegante Verallgemeinerung des g-adischen Algorithmus ist der in **8.3** vorgestellte CANTORsche Algorithmus, bei dem die g-adischen Darstellungen als Sonderfälle eines allgemeineren Darstellungskonzepts erscheinen. Diese Algorithmen sind Beispiele so genannter *Iterationsverfahren*. Ein solches Verfahren ist grob dadurch gekennzeichnet, dass die Ausgangswerte einer oder mehrerer gegebener Operationen zu Eingangswerten derselben Operationen gemacht werden und die Schleife der wiederholten Anwendung derselben Operationen erst dann verlassen wird, wenn gewisse Bedingungen erfüllt sind, die bei jedem Durchlauf der Schleife vom Rechner zu testen sind. Auf diese Weise können z.B. gewisse Irrationalzahlen wie $\sqrt{2}$, $\boldsymbol{\pi}$ oder $\mathbf{e}$ heute problemlos und in kurzer Zeit auf Milliarden von Dezimalstellen berechnet, Gleichungen mit einer riesigen Anzahl von Unbekannten gelöst und Differentialgleichungen näherungsweise integriert werden, deren praktische Lösung früher hoffnungslos war.

Mit den komplizierteren Rechenverfahren befasst sich die numerische Analysis, die durch die immer leistungsfähigeren Rechner zu einem faszinierenden Gebiet geworden ist. Die Begleitung der Theorie durch das Experiment am Computer kann unvermutete Sachverhalte offenbaren, auf die man sich allerdings erst dann wirklich verlassen kann, wenn sie mathematisch bewiesen sind. Diese Bemerkung betrifft natürlich auch andere Disziplinen, darunter nicht nur solche von mathematischem Charakter.

8.1 Der Divisionsalgorithmus

Es handelt sich hierbei um eines der grundlegenden Rechenverfahren, dessen Ausgangsobjekte in der einfachsten Version zwei natürliche Zahlen a und b sind, der *Dividend* und der *Divisor* und dessen Ergebnis eine abbrechende oder auch eine nichtabbrechende Dezimalzahl $u_0{,}u_1u_2\cdots$ ist, kurz notiert

$$a : b = u_0{,}u_1u_2\cdots.$$

Wir werden rigoros beweisen, und zwar ohne jeden Rückgriff auf unendliche Reihen, dass $u_0{,}u_1u_2\cdots$ der Quotient $\frac{a}{b}$ ist, also $\frac{a}{b} = u_0{,}u_1u_2\cdots$. Dies wird sogleich für beliebige Operanden $a, b \in \mathbb{D}$ nachgewiesen, mit der einzigen Einschränkung $b \neq 0$.

Für die Beschreibung des Algorithmus verwenden wir die Funktion Int. Mit ihrer Hilfe lässt sich für gegebene $a, b \in \mathbb{D}$ die größte natürliche Zahl n bestimmen, so dass $nb \leq a$. Und zwar ist $n = \operatorname{Int}\frac{a}{b}$; denn $nb \leq a < (n+1)b$ ergibt nach Division durch b die Ungleichung $n \leq \frac{a}{b} < n+1$, also $n = \operatorname{Int}\frac{a}{b}$. Im Hinblick auf eine im nächsten Teilabschnitt diskutierte Verallgemeinerung des Divisionsalgorithmus sei die Zahl 10 vorläufig mit g bezeichnet. In **8.2** wird in diesem Algorithmus die Grundzahl $g = 10$ nämlich durch eine beliebige andere natürliche Zahl ≥ 2 ersetzt werden.

Das Schema des schriftlichen Divisionsalgorithmus ist bekanntlich das folgende:

$$\begin{array}{rl} a \;:\; b \;=\; u_0{,}u_1u_2\cdots & \\ -\underline{bu_0} & \\ \quad r_0 & \mid\cdot g \\ \quad -\underline{bu_1} & \\ \qquad r_1 & \mid\cdot g \\ \qquad\quad \ddots & \end{array}$$

Das bedeutet im einzelnen: Zuerst wird die größte natürliche Zahl u_0 bestimmt, so dass gerade noch $bu_0 \leq a$. Mit anderen Worten, $u_0 = \operatorname{Int}\frac{a}{b}$. Weil $c - \operatorname{Int} c < 1$ für alle $c \in \mathbb{D}$, ist $r_0 := a - bu_0 = a - b\operatorname{Int}\frac{a}{b} = b(\frac{a}{b} - \operatorname{Int}\frac{a}{b}) < b$. Sodann wird die größte natürliche Zahl u_1 bestimmt mit $bu_1 \leq r_0 g$. Weil $\frac{r_0}{b} < 1$, ist $\frac{gr_0}{b} < g\ (= 10)$, also ist $u_1 = \operatorname{Int}\frac{gr_0}{b}$ eine Ziffer. Analog werden $r_1 := gr_0 - bu_1\ (< b)$, $u_2\ (< g)$, $r_2\ (< b)$ bestimmt, usw. Formal lautet diese rekursive Definition der u_n und r_n wie folgt, wobei die Reste r_n im Ergebnis nicht erscheinen und nur eine Hilfsrolle spielen:

$$(1)\qquad \begin{array}{lllllll} u_0 & = & \operatorname{Int}\frac{a}{b}, & r_0 & = & a - bu_0 & \big(= b(\frac{a}{b} - \operatorname{Int}\frac{a}{b})\big); \\ u_{n+1} & = & \operatorname{Int}\frac{gr_n}{b}, & r_{n+1} & = & r_n g - bu_{n+1} & \big(= b(\frac{gr_n}{b} - \operatorname{Int}\frac{gr_n}{b})\big). \end{array}$$

Das nach (1) berechnete u_{n+1} ist tatsächlich Dezimalziffer, also $u_{n+1} < g\ (= 10)$; denn $r_0 < b$ und ebenso $r_{n+1} = b(\frac{gr_n}{b} - \operatorname{Int}\frac{gr_n}{b}) < b$. Daher ist $\frac{r_n}{b} < 1$ für alle n, und folglich

$$u_{n+1} = \operatorname{Int}\frac{gr_n}{b} \leq \frac{gr_n}{b} = g\frac{r_n}{b} < g.$$

Offenbar ist eine schriftliche Division von Hand nichts anderes als ein ökonomisch geschriebenes Protokoll einer Berechnung gemäß den Rekursionsgleichungen (1).

Es ist leicht zu erkennen, dass die Ziffernfolge $\langle u_i \rangle$ des Ergebnisses effektiv bestimmt werden kann, wenn die Ziffern von a schrittweise berechenbar sind und wenn der Divisor eine endliche Dezimalzahl ist. Hingegen ergeben sich bei einem beliebigen Divisor b unter Umständen gewisse praktische Probleme, weil dann die Int-Operation in (1) möglicherweise nicht präzise genug ausgeführt werden kann. Schon um $\operatorname{Int} \frac{a}{b}$ zu bestimmen, ist dasjenige n mit $nb \leq a < (n+1)b$ zu finden. Das gelingt im Falle $a > b$ durch schrittweise Subtraktion mit absoluter Sicherheit i.a. nur für abbrechende Operanden. Inwieweit man in der Praxis mit abbrechenden Näherungen von b (oder auch a) rechnen darf, hängt natürlich von der Aufgabenstellung ab. Dies alles ist jedoch für die Theorie und den Beweis von Satz 8.1 unten völlig unerheblich.

Es seien nun a und $b \neq 0$ gegeben und $\frac{a}{b} = z_0{,}z_1z_2\cdots$. Wir beweisen als erstes, dass dieser Algorithmus genau die Ziffern von $\frac{a}{b}$ liefert. Damit wird übrigens zugleich gezeigt, dass das Ergebnis des Algorithmus stets eine zulässige Dezimalzahl ist.

Satz 8.1. *Sei $a : b = u_0{,}u_1u_2\cdots$ das Ergebnis des Divisionsalgorithmus angewandt auf die Operanden $a \in \mathbb{D}$, $b \in \mathbb{D}_+$ und sei $\frac{a}{b} = z_0{,}z_1z_2\cdots$. Dann ist $u_n = z_n$ für alle n.*

Beweis. Wir verifizieren induktiv über n die Gleichungen

$$(*) \qquad u_n = z_n \;\; ; \;\; r_n = b \cdot 0{,}z_{n+1}z_{n+2}\cdots$$

was die Behauptung des Satzes mit einschließt. Gewiss ist $u_0 = \operatorname{Int} \frac{a}{b} = z_0$ und daher

$$r_0 = b \cdot (\tfrac{a}{b} - u_0) = b \cdot (z_0{,}z_1z_2\cdots - z_0) = b \cdot 0{,}z_1z_2\cdots.$$

Sei $(*)$ für n vorausgesetzt. Dann ist $\frac{gr_n}{b} = g \cdot 0{,}z_{n+1}z_{n+2}\cdots = z_{n+1}{,}z_{n+2}z_{n+3}\cdots$ und folglich $u_{n+1} = \operatorname{Int} \frac{gr_n}{b} = z_{n+1}$. Hieraus ergibt sich mit $r_n = b \cdot 0{,}z_{n+1}z_{n+2}\cdots$ dann auch

$$\begin{aligned} r_{n+1} &= gr_n - b \cdot u_{n+1} = b \cdot g \cdot 0{,}z_{n+1}z_{n+2}\cdots - b \cdot u_{n+1} \\ &= b(z_{n+1}{,}z_{n+2}z_{n+3}\cdots - u_{n+1}) \\ &= b \cdot 0{,}z_{n+2}z_{n+3}\cdots \qquad (\text{wegen } u_{n+1} = z_{n+1}). \quad \blacksquare \end{aligned}$$

Es ist klar, dass $\frac{a}{b} = \sup X$ mit $X = \{x \in \mathbb{D} \mid bx \leq a\}$. Man bestätigt unschwer, dass der Divisionsalgorithmus im Grunde nur eine Spezialisierung des Berechnungsverfahrens der Ziffern von $\sup X$ gemäß dem Beweis von Satz 3.2 darstellt. Z.B. ist $u_0 = \operatorname{Int} \frac{a}{b}$ nichts anderes als die größte natürliche Zahl k, so dass gerade noch $k \leq x$ für ein $x \in X$, ganz entsprechend der Definition von z_0 in Satz 3.2.

Der Divisionsalgorithmus kann auch in Gestalt eines so genannten *Flussdiagramms* veranschaulicht werden (Figur 4), einer nützlichen Vorstufe für dessen Programmierung auf einem Taschenrechner oder in irgendeiner Programmiersprache. Ein entsprechend dem Diagramm programmierter Rechner druckt die Ziffern von $\frac{a}{b} = u_0{,}u_1u_2{,}\cdots$ schrittweise aus (PRINT-*Operation*). Er bricht die Rechnung ab, falls $\frac{a}{b} \in \mathbb{E}$, also nur endlich viele Ziffern zu bestimmen sind. Dies ist der Fall genau dann, wenn einer der Reste r_n und damit alle nachfolgenden verschwinden, ebenso wie die Ziffern $u_{n+1}, u_{n+2}, \ldots$ Alle nicht durch INPUT belegten Variablen des Diagramms, nämlich u, r, n haben anfänglich den Wert 0. Die Variable g erhält zu Beginn den Wert 10. Sie wurde nur zwecks Nutzung

des Algorithmus für die g-adische Entwicklung als Parameter eingeführt.

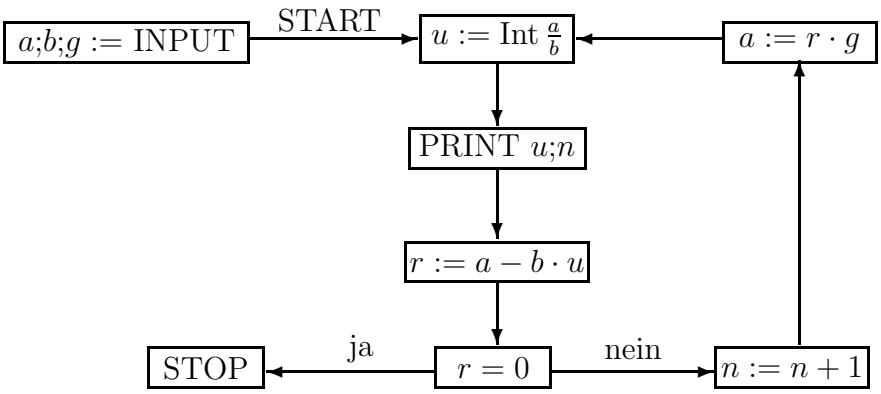


Fig. 4 Flußdiagramm des Divisionsalgorithmus

Rechtecke symbolisieren Operationsschritte. $\boxed{x := t}$ bedeutet, dass der Variablen x der jeweilige Werte des Terms t zugewiesen wird, wobei x durchaus in t vorkommen kann, wie etwa in $\boxed{n := n + 1}$. Das Diagramm enthält die Testoperation $\boxed{r = 0\ ?}$, d.h. es ist die Frage ob r im jeweiligen Rechenschritt den Wert 0 hat mit ja oder nein zu beantworten. Der Rechenfluss verzweigt sich entsprechend der Antwort. Im Verlaufe der Rechnung erhalten die Variablen i.a. neue Werte. Der Ablauf der Rechnung gemäß diesem Flussdiagramm ist einfach zu verfolgen und entspricht offenbar vollkommen dem Divisionsalgorithmus mit Abbruch der Rechnung, falls die Division „aufgeht“.

Man übersetzt das Diagramm sehr einfach in eine Programmiersprache (auch moderner Taschenrechner), z.B. durch Benutzung einer DO...UNTIL...END-Struktur und erhält so bei umsichtiger Formulierung des Programms die Dezimalen des Quotienten weit über die Stellenkapazität des Rechners hinaus. Das Diagramm nutzt zur Bestimmung von $\mathrm{Int}\,\frac{a}{b}$ der Form nach die interne Rechnerdivision. Dies könnte wegen der Rundungsorganisation des Rechners zu Fehlern führen. Berücksichtigt man, dass $\mathrm{Int}\,\frac{a}{b}$ nichts anderes ist als der Quotient bei der so genannten Division mit Rest, kann die besonders empfindliche Operation $\boxed{u := \mathrm{Int}\,\frac{a}{b}}$ des Diagramms auch ohne Divisionen in einer Schleife fortlaufender Subtraktionen mit hoher Präzision ausgeführt werden.

Nun lässt sich der Divisionsalgorithmus natürlich auch so programmieren, dass das Verfahren bei überflüssigen Rechnungen überhaupt abbricht. Überflüssig wird weiteres Rechnen nicht nur wenn ein Rest r gleich 0 wird, sondern auch dann, wenn einer der Reste einem früher berechneten gleicht. Sei nämlich $r_{q+p} = r_q$ für gewisse (minimal gewählte) p, q. Dann ist auch $r_{q+p+1} = r_{q+1}$, $r_{q+p+2} = r_{q+2}$ usw. Parallel dazu ergibt sich $u_{q+p+i} = u_{q+i}$ für $i = 1, 2, \ldots$ Kurzum, der Rechenprozess ist in eine Periode eingetreten und liefert die periodische Dezimalzahl $u_0{,}u_1 \cdots u_q\overline{u_{q+1} \cdots u_{q+p}}$ mit der Periodenlänge p. Es lässt sich durch eine auf der Hand liegende Erweiterung des obigen Diagramms problemlos erreichen, dass der Programmablauf bei Eintritt in eine Periode beendet wird und zwecks vollständiger Information zugleich die Periodenlänge ausgedruckt wird.

Periodizität tritt nun offenbar immer ein, wenn die Operanden a, b natürliche Zahlen sind. Denn dann sind alle Reste r_i natürliche Zahlen $< b$. Daher ist entweder einer der Reste $r_1, \ldots, r_b$ (und damit alle nachfolgenden) gleich 0, oder aber alle diese Reste sind von 0 verschieden und mindestens zwei dieser Reste sind identisch (dies ist ein einfaches Beispiel eines so genannten Schubfach-Arguments). Damit haben wir nebenbei folgenden Satz bewiesen, der den Satz 6.2 gewissermaßen vervollständigt:

Satz 8.2. *Jede rationale Zahl $\frac{m}{n}$ ist periodisch. Dabei gilt $p + q \leq n$ und $p < n$, wobei p die Periodenlänge, q der Periodenindex ist.*

Übrigens folgt aus dem ersten Teil des Satzes mittels Formel (1) in Satz 6.2: Jede positive rationale Zahl r hat mit geeignetem m eine Darstellung

$$r = \frac{m}{(10^p-1)10^q} = \frac{m}{\underbrace{9\cdots9}_{p}\,\underbrace{0\cdots0}_{q}} \quad (p > 0).$$

Zum Beispiel sind $\frac{1}{13} = \frac{76923}{999999}$ und $\frac{1}{14} = \frac{714285}{9999990}$ solche Darstellungen, und zwar die kürzesten dieser Art. Es folgt hieraus offenbar gänzlich ohne Zahlentheorie, dass eine zu 10 teilerfremde natürliche Zahl Teiler ist von $10^p - 1$ für geeignetes p.

Wir illustrieren die Überlegenheit der programmierten Division an folgendem

Beispiel. Sei $a = 1{,}000\,009$ und $b = 17$. Division auf einem Rechner mit 13-stelligem Display liefert $1{,}000\,009 \div 17 = 0{,}058824\,058824\cdots$. Man könnte daher vermuten, dass die Periodenlänge des Quotienten 6 beträgt. Die genaue Berechnung nach einem Programm (oder eine Hobby-Rechnung von Hand) ergibt jedoch

$$\tfrac{1{,}000\,009}{17} = 0{,}0588240\overline{588235294117647}$$

mit der beim Nenner 17 maximalen Periodenlänge 16. Hierzu beachte man, dass die Periodenlängen von $r \in \mathbb{Q}$ und $r \xrightarrow{n} (= 10^n r)$ übereinstimmen. Für $\frac{1000009}{17}$ ist diese aber ≤ 16 gemäß Satz 8.2. Dass sie genau 16 ist, kann man mit etwas Zahlentheorie auch im voraus bestimmen: 1000009 und 17 sind teilerfremd und $p = 16$ ist das kleinste p, so dass $10^p - 1$ durch 17 teilbar ist; siehe hierzu Übung 2.

Eine Bemerkung zur Bestimmung des Quotienten $\frac{a}{b}$ für den Fall nichtabbrechender Operanden a, b. Dieser kann jedenfalls dann (näherungsweise) berechnet werden, wenn a, b in dem Sinne berechenbar sind, dass durch ein Verfahren zu jedem n Näherungen $a_n, b_n \in \mathbb{E}$ mit $|a - a_n|, |b - b_n| < \varepsilon_n$ effektiv bestimmt werden können, siehe auch **8.4**. Es lässt sich ähnlich wie für Summe und Produkt unschwer ausrechnen, wie groß n sein muss, damit $|\frac{a}{b} - \frac{a_n}{b_n}| < \varepsilon_m$, wobei ε_m die vorgeschriebene Genauigkeit sein soll. Allerdings handelt es sich hierbei nicht um die ziffernweise Berechenbarkeit. Denn ziffernweise Berechenbarkeit der Operanden überträgt sich i.a. auf keine der arithmetischen Operationen. Eine Zahl a kann im erwähnten etwas allgemeineren Sinne berechenbar sein, ohne dass einem konkreten Berechnungsverfahren auch nur eine Information über den letztgültigen Wert von $\mathrm{Int}\, a$ entnommen werden kann, siehe **5.4**. Dieses Phänomen kann sehr störend sein bei Computerberechnungen mit empfindlichen Tests.

8.2 Der g-adische Algorithmus

Wie in **1.2** schon erwähnt wurde, hätte man sich auch auch auf ein g-adisches System mit einer Grundzahl $g \neq 10$ einigen können, und die Wahl $g = 8$ wäre in vieler Hinsicht sogar vorteilhafter gewesen. Unabhängig davon haben Computer-Experten bei komplexen Installationen und auch Nutzer (bei Änderungen der Konfiguration) häufig Veranlassung, Zahlen verschiedener g-adischer Systeme ineinander umzurechnen. Das betrifft vor allem die Grundzahlen $g = 2,\ g = 8,\ g = 10$ und $g = 16$. Im *Hexadezimalsystem* ($g = 16$) benutzt man neben den zehn Dezimalziffern die Buchstaben A, . . . ,F zur Bezeichnung der sechs fehlenden Ziffern. Diese Umrechnungen sind so von einer bloßen Spielerei zu einer praktisch wichtigen Angelegenheit geworden.

Jede positive natürliche Zahl k hat im g-adischen System eine eindeutige Darstellung

$$k = z_0 g^n + z_1 g^{n-1} + \ldots + z_{n-1} g_0 + z_n = (z_0 z_1 \cdots z_n)_g \qquad (z_i < g,\ z_0 \neq 0),$$

die der dezimalen Darstellung vollkommen entspricht, siehe **11.6**. Wie man diese mit dem g-adischen Divisionsalgorithmus schnell errechnet, sehen wir weiter unten. Zuvor stellen wir die Frage, was entspricht den abbrechenden und nichtabbrechenden Dezimalzahlen im g-adischen System? Die Antwort ist im Prinzip einfach. Denn die in den Abschnitten **3** - **5** entwickelte Theorie der Dezimalzahlen kann ohne jede Änderung auf das g-adische System übertragen werden. Summe und Produkt lassen sich in derselben Weise definieren. Diese Bemerkung zeigt, dass nicht nur bezüglich der natürlichen Zahlen, sondern auch bezüglich der reellen Zahlen die Wahl von $g = 10$ nur die Wahl eines bestimmten Koordinatensystems ist. Alle wesentlichen Dinge bleiben bezüglich einer Transformation des Koordinatensystems invariant.

Als Zugeständnis an die Tradition wollen wir jedoch an die Theorie der unendlichen Reihen anknüpfen und definieren g-adische reelle Zahlen durch unendliche Reihen der Gestalt $\sum_i \frac{z_i}{g^i} = z_0 + \frac{z_1}{g} + \frac{z_1}{g^2} + \ldots$ Dabei seien die z_i für $i > 0$ Ziffern des g-adischen Systems, also natürliche Zahlen $< g$. Auch z_0 sei g-adisch dargestellt. Obige Reihe ist nach den Ergebnissen in **7.1** konvergent und hat eine wohlbestimmte Summe a. Wir schreiben nun $z_0{,}z_1 z_2 \cdots_g$ für $z_0 + \frac{z_1}{g} + \frac{z_2}{g^2} + \ldots$ und nennen $z_0{,}z_1 z_2 \cdots_g$ eine *g-adische Darstellung* (oder Entwicklung) von a. Fragen, die deren Eindeutigkeit betreffen, erörtern wir am Schluss. In jedem Falle gilt $z_0{,}z_1 z_2 \cdots_g \overset{n}{\rightarrow} = z_0{,}z_1 z_2 \cdots_g \cdot g^n$, Übung 4.

Die erste sich hier erhebende Frage ist, ob nun auch jede positive reelle Zahl eine g-adische Darstellung besitzt. Angesichts der Vorbemerkung über das „g-adische Koordinatensystem“ ist dies zwar klar, doch wollen wir uns auf eine formale g-adische Darstellung ja nicht berufen. Daher muss diese Darstellbarkeit bewiesen werden.

Die dazu erforderlichen Hilfsmittel stehen aber zur Verfügung. Wir betrachten den Divisionsalgorithmus mit folgender Änderung: $g = 10$ wird durch die Basis g des gewählten g-adischen Systems ersetzt. a sei irgend eine zunächst in Dezimaldarstellung gegebene reelle Zahl und nunmehr sei von Beginn an $b = 1$. Sonst bleibt alles beim alten. Die Formeln (1) zur Berechnung der Folgen u_n und r_n spezialisieren sich dann zu

$$(2)\qquad \begin{array}{rclcrcl} u_0 & = & \operatorname{Int} a, & & r_0 & = & a - u_0 \qquad (= a - \operatorname{Int} a); \\ u_{n+1} & = & \operatorname{Int}(gr_n), & & r_{n+1} & = & gr_n - u_{n+1} \quad (= gr_n - \operatorname{Int} gr_n). \end{array}$$

Es ist $r_n < 1$ für alle n, wei $r - \operatorname{Int} r < 1$ für alle $r \in \mathbb{D}$. Also $gr_n < g$, so dass $u_{n+1} = \operatorname{Int} gr_n \leq gr_n < g$ eine g-adische Ziffer ist. Damit ist klar, dass der so genutzte Algorithmus, angewandt auf a, nach Umrechnung von $u_0 = \operatorname{Int} a$ in das g-adische System eine wohlbestimmte reelle Zahl $u_0{,}u_1u_2\cdots_g = \sum_i \frac{u_i}{g^i}$ liefert. Wir werden nun gleich sehen, dass dieses Ergebnis mit a identisch ist, also gerade eine g-adische Darstellung von a liefert. Deshalb spricht man nunmehr vom *g-adischen Algorithmus*.

Satz 8.3. *Sei $u_0{,}u_1u_2\cdots_g$ das Ergebnis des g-adischen Algorithmus, angewandt auf eine reelle Zahl $a \geq 0$. Dann ist $a = u_0{,}u_1u_2\cdots_g$ $(= \sum_i \frac{u_i}{g^i})$.*

Beweis. Zuerst beweisen wir durch Induktion über n die Formel

$$(*)\qquad a = u_0 + \tfrac{u_1}{g^1} + \ldots + \tfrac{u_n}{g^n} + \tfrac{r_n}{g^n}.$$

Für $n = 0$ heißt dies $a = u_0 + r_0$, und dies gilt gemäß der ersten Zeile von (2). Sei $(*)$ für n vorausgesetzt. Gemäß (2) ist $gr_n = u_{n+1} + r_{n+1}$, also $\frac{r_n}{g^n} = \frac{u_{n+1}}{g^{n+1}} + \frac{r_{n+1}}{g^{n+1}}$. Daher gilt $(*)$ auch für $n+1$, und folglich für alle n. Um nun die Richtigkeit von $a = \sum_i \frac{u_i}{g^i}$ einzusehen, hat man sich nur davon zu überzeugen, dass $\langle \frac{r_n}{g^n} \rangle$ eine Nullfolge ist. Das aber ist klar; denn wie oben schon festgestellt wurde, ist $r_n < 1$, also $\frac{r_n}{g^n} < \frac{1}{g^n}$. ∎

Beispiel. Wir stellen $\boldsymbol{\pi} = 3{,}14159\,26535\,89\cdots$ [1)] mit dem 8-adischen Algorithmus als Oktalzahl dar. Statt die Reste wie bei üblicher Division mit 10, werden diese nunmehr mit 8 multipliziert. Wegen der Pünktchen bei $\boldsymbol{\pi}$ kennen wir nur die ausgeschriebenen Dezimalen von $8r_0, 8r_1, \ldots$ genau. Damit sind alle Vorkommaziffern von $8r_0$ bis $8r_{10}$ korrekt, und damit auch die angegebenen 11 Kommastellen des Ergebnisses.

$$\begin{array}{ll} \underline{3{,}141592653589\cdots} & = \; 3{,}11037552421\cdots_8 \\ \underline{1{,}13274122871\cdots} & (= 8r_0) \\ \underline{1{,}061929829\cdots} & (= 8r_1) \\ \vdots & \vdots \end{array}$$

Die Umrechnung einer Dezimalzahl in das Oktalsystem ist sehr nützlich als Zwischenstufe für ihre Umrechnung in das Binärsystem. Man braucht nämlich deren oktale Darstellung nur ziffernweise umzuschreiben, wobei den Ziffern $0, 1, 2, 3, \ldots, 7$ der Reihe nach die Tripel $000, 001, 010, 011, \ldots, 111$ von Binärziffern entsprechen, Übung 3. Für $\boldsymbol{\pi}$ erhält man deshalb ohne jede weitere Rechnung aus dem obigen Beispiel

$$\boldsymbol{\pi} = 3{,}11037552421\cdots_8 = 11{,}001\,001\,000\,011\,111\,101\,101\,010\,100\,010\,001\cdots_2\,.$$

Damit erhält man zugleich auch die Umrechnung einer Dezimalzahl in das Hexadezimalsystem. Jede Hexadezimalziffer $0, \ldots, 9, \mathrm{A}, \ldots, \mathrm{F}$ entspricht nämlich genau einem

[1)] $\boldsymbol{\pi}$ wurde inzwischen (2005) auf über 200 Milliarden Dezimalziffern berechnet. Eines ihrer vielen Geheimnisse ist, ob die Folge dieser Ziffern, oder besser, der Ziffern ihrer Binärdarstellung (unten) den Kriterien einer Zufallsfolge genügt, und wenn nicht, welches Muster sich in der Ziffernfolge verbirgt.

Quadrupel von Binärziffern. Man erhält so insbesondere

$$\boldsymbol{\pi} = 11{,}0010\,0100\,0011\,1111\,0110\,1010\,1000\,1000\,1\cdots_2 = 3{,}243\mathrm{F6A88}\cdots_{16}\,.$$

Betrachten wir nun den Divisionsalgorithmus mit gegebener Grundzahl $g \geq 2$ und heben die oben gemachte Einschränkung $b = 1$ wieder auf. Man spricht dann vom *g-adischen Divisionsalgorithmus*, denn das Resultat $u_0{,}u_1u_2\cdots_g$ ist eine g-adische Entwicklung von $\frac{a}{b}$. Dies beweist man wörtlich wie Satz 1, ausgehend von einer durch Satz 8.3 garantierten Darstellung $\frac{a}{b} = z_0{,}z_1z_2\cdots_g$, welche die für den Beweis benötigten Eigenschaften (z.B. $\operatorname{Int}\frac{a}{b} = z_0$) tatsächlich auch besitzt. Eine g-adische Darstellung von $\frac{a}{b}$ bei dezimal gegebenen Operanden lässt sich also direkt ermitteln, ohne $\frac{a}{b}$ zuvor dezimal auszurechnen und dann mit dem g-adischen Algorithmus umzuwandeln. Noch wichtiger aber ist, dass der g-adische Divisionsalgorithmus es erlaubt, zugleich auch die natürliche Zahl $\operatorname{Int} a$ (speziell also jedes $a \in \mathbb{N}$) in das g-adische System umzurechnen. Denn sei $a < g^{n+1}$, also $\frac{a}{g^n} < g$, und sei $a : g^n = u_0{,}u_1u_2\cdots_g$ Ergebnis der Anwendung des g-adischen Divisionsalgorithmus, so dass $a = g^n \cdot u_0{,}u_1u_2\cdots_g = u_0{,}u_1u_2\cdots_g \overset{n}{\rightarrow}$. Wegen $\frac{a}{g^n} < g$ ist nun auch $u_0 = \operatorname{Int}\frac{a}{g^n}$ eine g-adische Ziffer und wir erhalten

$$a = u_0{,}u_1u_2\cdots_g \overset{n}{\rightarrow} = u_0\cdots u_n{,}u_{n+1}u_{n+2}\cdots_g\,.$$

Folglich ist $\operatorname{Int} a = (u_0\cdots u_n)_g$. Das alles gilt speziell für $a = \operatorname{Int} a$, d.h. $a \in \mathbb{N}$.

Beispiel. Es ist $100{,}6 < 8^3$, also $\frac{100{,}6}{8^2} < 8$. Der 8-adische Divisionsalgorithmus liefert

$$\begin{array}{rcl}
100{,}6 \;:\; 64 & = & 1{,}44\,4631\,4631\cdots_8 \\
\underline{-64}\;\; (=-u_0\cdot 64) & & \\
36{,}6 \quad \cdot 8 & = & 292{,}8 \quad (=8r_0) \\
 & & \underline{-256} \quad\;\; (=-u_1\cdot 64) \\
 & & 36{,}8 \quad \cdot 8 = \underline{294{,}4}\;\; (=8r_1) \\
 & & \ddots
\end{array}$$

Es ergibt sich also $100{,}6 = 1{,}44\overline{4631}_8 \overset{2}{\rightarrow} = 144{,}\overline{4631}_8 = 1\cdot 8^2 + 4\cdot 8 + 4 + \frac{4}{8} + \frac{6}{8^2} + \ldots$ Satz 8.2 gilt völlig unabhängig von der gewählten Basis. Deswegen wundert es nicht, daß 100,6 auch im Oktalsystem periodisch ist. Natürlich hängt u.a. die Periodenlänge wesentlich von der Basis ab. Im Binärsystem z.B. hat 100,6 die Periodenlänge 12.

Wie sieht es nun mit der Eindeutigkeit der g-adischen Darstellung aus? Für $g = 10$ ist z.B. $\frac{1}{10} + \frac{0}{10^2} + \frac{0}{10^3} + \ldots = \frac{1}{10} = \frac{0}{10} + \frac{9}{10^2} + \frac{9}{10^3} + \ldots$ Eindeutigkeit kann daher sicher nur unter zusätzlichen Bedingungen erwartet werden. Verlangt man $z_i \neq 9$ für unendlich viele i in einer Darstellung $a = \sum_i \frac{z_i}{10^i}$, dann sind die Ziffern z_i in der Tat eindeutig bestimmt. Denn die Zusatzbedingung besagt für $g = 10$ nichts anderes als $z_0{,}z_1z_2\cdots$ ist zulässig; und weil $a = z_0{,}z_1z_2\cdots$, sind die z_i eindeutig bestimmt. Für beliebiges $g \geq 2$ gilt ein völlig analoger Sachverhalt. Den Beweis hierfür liefern wir in **8.3**.

Man kann auch Eindeutigkeit erzwingen, indem man verlangt, dass in einer Darstellung $a = \sum_i \frac{z_i}{g^i}$ alle Ziffern z_i ab einer gewissen Stelle identisch $g-1$ sind. Eine solche Darstellung existiert für eine beliebige Grundzahl $g \geq 2$. Z.B. ist $0{,}7_8 = 0{,}6777\cdots_8$.

8.3 Der CANTORsche Algorithmus

G. CANTOR hat einige wichtige Beiträge zur Theorie der unendlichen Reihen geliefert und unter anderem eine interessante Verallgemeinerung des g-adischen Algorithmus angegeben. Die folgenden Darlegungen hierüber verwenden nur einige Grundtatsachen über unendliche Reihen mit positiven Gliedern. Negative Zahlen treten hierbei nach wie vor nicht in Erscheinung.

Es sei $g_1, g_2, \ldots$ eine Folge natürlicher Zahlen ≥ 2, die wir uns im folgenden beliebig, aber fest gewählt denken und eine CANTORsche *Basis* nennen. Der CANTORsche Algorithmus verläuft wie der g-adische Algorithmus, nur wird im n-ten Schritt die Rolle der Zahl g von g_n wahrgenommen. Im Flussdiagramm der Figur 4 lässt sich dies durch Einfügung der Operation $\boxed{g := g_n}$ nach Ausführung von $\boxed{n := n+1}$ und Einfügung eines Unterprogramms zur Berechnung der g_n problemlos berücksichtigen. Die Rekursionsformeln (2) zur Berechnung von u_n und r_n nehmen dann für den CANTORschen Algorithmus die folgende Gestalt an, wobei $a \in \mathbb{D}$ beliebig vorgegeben ist:

$$(3)\qquad \begin{array}{rclcrcll} u_0 & = & \operatorname{Int} a, & & r_0 & = & a - u_0 & \big(= a - \operatorname{Int} a\big); \\ u_{n+1} & = & \operatorname{Int}(g_{n+1} r_n), & & r_{n+1} & = & g_{n+1} r_n - u_{n+1} & \big(= g_{n+1} r_n - \operatorname{Int}(g_{n+1} r_n)\big). \end{array}$$

Diese produzieren eine Folge $\langle u_n \rangle$ natürlicher Zahlen mit $u_n < g_n$ für $n > 0$. Denn offenbar ist $r_n < 1$ für jedes n, und daraus folgt $u_{n+1} \leq g_{n+1} r_n < g_{n+1}$. Der g-adische Algorithmus ist demnach der Spezialfall des CANTORschen für $g_1 = g_2 = \ldots = g$. Natürlich kann im CANTOR'schen Algorithmus von den u_n für $n > 0$ nur noch in sehr allgemeinem Sinne als den „Ziffern" gesprochen werden. (3) impliziert offenbar $r_n = \frac{u_{n+1}}{g_{n+1}} + \frac{r_{n+1}}{g_{n+1}}$, also $a = u_0 + r_0 = u_0 + \frac{u_1}{g_1} + \frac{r_1}{g_1} = u_0 + \frac{u_1}{g_1} + \frac{u_2}{g_1 \cdot g_2} + \frac{r_2}{g_1 \cdot g_2}$ und allgemein

$$a = u_0 + \frac{u_1}{g_1} + \frac{u_2}{g_1 \cdot g_2} + \ldots + \frac{u_n}{g_1 \cdot \ldots \cdot g_n} + \frac{r_n}{g_1 \cdot \ldots \cdot g_n} \qquad (r_n < 1).$$

Setzt man zusätzlich $g_0 := 1$ und $G_n := g_0 \cdot \ldots \cdot g_n$, lässt sich dies auch schreiben als

$$(4)\qquad a = \textstyle\sum_{i \leq n} \frac{u_i}{G_i} + \frac{r_n}{G_n} \qquad (r_n < 1).$$

Daraus entnimmt man, dass die unendliche Reihe $\sum_i \frac{u_i}{G_i} = u_0 + \frac{u_1}{G_1} + \frac{u_2}{G_2} + \ldots$ konvergiert. Sie hat außerdem gerade den Wert a. Denn die jeweils letzten Summanden $\frac{r_n}{G_n}$ in (4) bilden eine Nullfolge, weil $G_n \geq 2^n$ für alle n, und damit $\frac{r_n}{G_n} < \frac{1}{G_n} \leq \frac{1}{2^n}$.

Die Darstellung $a = \sum_i \frac{u_i}{G_i}$ heiße die CANTORsche *Entwicklung von* a. Wir wollen zuerst zeigen, dass diese noch die zusätzliche Eigenschaft $u_i \neq g_i - 1$ für unendlich viele Indizes i besitzt, also im übertragenen Sinne zulässig ist. Denn angenommen, dies ist nicht der Fall, etwa $u_i = g_i - 1$ für alle $i > n$. Dann gilt $r_{i+1} = g_{i+1} \cdot r_i - (g_{i+1} - 1)$ für $i \geq n$ gemäß (3). Eine leichte Umformung hiervon ergibt $1 - r_i = \frac{1 - r_{i+1}}{g_{i+1}}$ und daher

$$1 - r_n = \frac{1 - r_{n+1}}{g_{n+1}} = \frac{1 - r_{n+2}}{g_{n+1} \cdot g_{n+2}} = \ldots = \frac{1 - r_{n+k}}{g_{n+1} \cdot \ldots \cdot g_{n+k}} \leq \frac{1}{g_{n+1} \cdot \ldots \cdot g_{n+k}} \leq \frac{1}{2^k}.$$

Diese Ungleichung gilt für beliebige k, woraus offenbar $1 - r_n = 0$ folgt, d.h. $r_n = 1$. Dies aber steht im Widerspruch zu $r_n < 1$. Es muss in der Tat also unendlich viele Indizes i geben mit $u_i < g_i - 1$. Man beachte, dass damit auch eine Aussage über das Ergebnis

$\sum_i \frac{u_i}{g^i}$ des g-adischen Algorithmus gemacht wird, nämlich $u_i \neq g-1$ für unendlich viele Indizes i. Für $g = 10$ ist dies natürlich nach Satz 8.1 schon bekannt.

Unter einer CANTOR*schen Reihe* bezüglich einer gegebenen CANTORschen Basis $\langle g_n \rangle$ sei nun eine unendliche Reihe der Gestalt verstanden, wie sie sich beim CANTORschen Algorithmus ergibt, also eine unendliche Reihe folgender Gestalt, wobei die z_i sämtlich natürliche Zahlen sind, und wie oben $G_n = \prod_{i \leq n} g_i$ und $G_0 = g_0 = 1$ gesetzt sei:

(5) $z_0 + \frac{z_1}{G_1} + \frac{z_2}{G_2} + \ldots$ $(z_i < g_i$ für $i > 0$; $z_i \neq g_i - 1$ für unendlich viele $i)$.

Bisher wurde nur bewiesen, dass eine derartige Reihe konvergiert, sofern sie Ergebnis des CANTORschen Algorithmus ist. Ferner zeigt dieser Algorithmus, dass jede reelle Zahl $a \geq 0$ auf wenigstens eine Weise als CANTORsche Reihe darstellbar ist. Nun ist aber *jede* CANTORsche Reihe konvergent; denn wegen $\frac{z_{i+1}}{g_{i+1}} < 1$ und $G_n \geq 2^n$ ist

$$\begin{aligned} \sum_{i \leq n} \frac{z_i}{G_i} &= z_0 + \frac{z_1}{g_1} + \frac{1}{G_1} \cdot \frac{z_2}{g_2} + \frac{1}{G_2} \cdot \frac{z_3}{g_3} + \ldots + \frac{1}{G_{n-1}} \cdot \frac{z_n}{g_n} \\ &< z_0 + 1 + \frac{1}{G_1} + \frac{1}{G_2} + \ldots + \frac{1}{G_{n-1}} \\ &\leq z_0 + 1 + \frac{1}{2} + \frac{1}{2^2} + \ldots + \frac{1}{2^{n-1}} < z_0 + 2. \end{aligned}$$

$\langle \sum_i \frac{z_i}{G_i} \rangle$ ist also schlicht. Wir benötigen im folgenden eine schärfere Abschätzung. Dazu betrachten wir die unendlichen Reihen $\frac{g_{n+1}-1}{G_{n+1}} + \frac{g_{n+2}-1}{G_{n+2}} + \ldots$ und zeigen zuerst

(6) $\frac{g_{n+1}-1}{G_{n+1}} + \frac{g_{n+2}-1}{G_{n+2}} + \ldots = \frac{1}{G_n}$ $\qquad (n = 0, 1, \ldots)$.

In der Tat, die Konvergenz der Reihe links in (6) folgt für beliebiges n aus

$$\begin{aligned} \frac{g_{n+1}-1}{G_{n+1}} + \ldots + \frac{g_{n+i}-1}{G_{n+i}} &= (\frac{1}{G_n} - \frac{1}{G_{n+1}}) + (\frac{1}{G_{n+1}} - \frac{1}{G_{n+2}}) + \ldots + (\frac{1}{G_{n+i-1}} - \frac{1}{G_{n+i}}) \\ &= \frac{1}{G_n} - \frac{1}{G_{n+i}} = \frac{1}{G_n}(1 - \frac{1}{g_{n+1} \cdot \ldots \cdot g_{n+i}}) < \frac{1}{G_n}, \end{aligned}$$

und Gleichung (6) sodann aus $\frac{g_{n+1}-1}{G_{n+1}} + \frac{g_{n+2}-1}{G_{n+2}} + \ldots = \lim_{i \to \infty} \frac{1}{G_n}(1 - \frac{1}{g_{n+1} \cdot \ldots \cdot g_{n+i}}) = \frac{1}{G_n}$. (6) ergibt nun für die „Restsummen“ der Reihe $\sum_i \frac{z_i}{G_i}$ die Abschätzung

(7) $\sum_{i>n} \frac{z_i}{G_i} = \frac{z_{n+1}}{G_{n+1}} + \frac{z_{n+2}}{G_{n+2}} + \ldots < \frac{g_{n+1}-1}{G_{n+1}} + \frac{g_{n+2}-1}{G_{n+2}} + \ldots = \frac{1}{G_n}$ $\qquad (n \geq 0)$.

Das $<$-Zeichen in (7) ist richtig; denn obwohl i.a. nur $z_i \leq g_i - 1$, gilt für wenigstens ein $i \geq n$ tatsächlich $z_i < g_i - 1$, und dies genügt um die Ungleichung (7) zu sichern.

Wir beweisen nunmehr die folgende, alles weitere entscheidende Behauptung

(a) Das Ergebnis $\langle u_i \rangle$ der Anwendung des CANTORschen Algorithmus auf die Summe $a = \sum_i \frac{z_i}{G_i}$ einer CANTORsche Reihe ist gerade die Folge $\langle z_i \rangle$.

Mit dieser Behauptung hätten wir gleich zweierlei bewiesen, und zwar die folgenden beiden Tatsachen:

(b) Jede Folge $\langle z_i \rangle$ natürlicher Zahlen mit $z_i < g_i$ für alle $i > 0$ und $z_i \neq g_i - 1$ für unendlich viele i kommt als Ergebnis des CANTORschen Algorithmus wirklich vor, nämlich als Ergebnis der Anwendung auf die reelle Zahl $a = \sum_i \frac{z_i}{G_i}$,

(c) Die Darstellung einer reellen Zahl a als CANTORsche Reihe ist eindeutig; denn ist $a = \sum_i \frac{z_i}{G_i} = \sum_i \frac{z'_i}{G_i}$, so ist $u_i = z_i$ und $u_i = z'_i$, also $z_i = z'_i$ für alle i. Dies betrifft insbesondere die g-adische Entwicklung, also den Fall $g_1 = g_2 = \ldots = g$.

Der Nachweis von (a) ist nun nach unseren Vorbereitungen nicht schwierig und erinnert an den Beweis von Satz 8.1. Sei $a = \sum_i \frac{z_i}{G_i} = z_0 + \sum_{i>0} \frac{z_i}{G_i}$ und $u_0, u_1, \ldots$ das Ergebnis des CANTORschen Algorithmus angewandt auf a. Wir behaupten, für jedes n gilt

$$(*) \quad u_n = z_n \quad ; \quad r_n = G_n \cdot \sum_{i>n} \frac{z_i}{G_i} \; (= \tfrac{z_{n+1}}{g_{n+1}} + \tfrac{z_{n+2}}{g_{n+1} \cdot g_{n+2}} + \ldots).$$

In der Tat, es ist $u_0 = \operatorname{Int} a$ gleich z_0, weil $z_0 \leq a$ und $a - z_0 = \sum_{i>0} \frac{z_i}{G_i} < \frac{1}{G_0} = 1$ gemäß (7). Damit ist auch $r_0 = a - \operatorname{Int} a = \sum_{i>0} \frac{z_i}{G_i} = G_0 \cdot \sum_{i>0} \frac{z_i}{G_i}$, d.h. $(*)$ gilt für $n = 0$. Seien diese Formeln richtig für n. Dann ergibt sich wegen $g_{n+1} G_n = G_{n+1}$

$$g_{n+1} r_n = G_{n+1} \cdot \sum_{i>n} \frac{z_i}{G_i} = z_{n+1} + G_{n+1} \cdot \sum_{i>n+1} \frac{z_i}{G_i}.$$

Nach (7) ist $G_{n+1} \cdot \sum_{i>n+1} \frac{z_i}{G_i} < 1$. Also ist $\operatorname{Int}(g_{n+1} r_n) = z_{n+1}$, d.h. $u_{n+1} = z_{n+1}$, und folglich auch $r_{n+1} = g_{n+1} r_n - u_{n+1} = G_{n+1} \cdot \sum_{i>n+1} \frac{z_i}{G_i}$. Damit ist (a) bewiesen.

Wir betrachten nunmehr den Sonderfall CANTORscher Reihen für die Basis $\langle g_n \rangle$ mit $g_n = n + 1$, also $g_0 = 1, g_1 = 2$, $g_2 = 3$, usw. Dann erhalten wir eine der erstaunlichsten Unterscheidungen rationaler und irrationaler Zahlen, nämlich den folgenden

Satz 8.4. *Jede reelle Zahl $a \geq 0$ hat genau eine Darstellung der Gestalt*

$$(8) \quad a = z_0 + \tfrac{z_1}{2!} + \tfrac{z_2}{3!} + \tfrac{z_3}{4!} + \ldots$$

mit $z_i \in \mathbb{N}$ und $z_i \leq i$ für $i > 0$, sowie $z_i \neq i$ für unendlich viele i. Darüber hinaus ist a rational genau dann, wenn ihre Darstellung (8) *abbricht.*

Beweis. Der CANTORsche Algorithmus ergibt die Existenz einer Darstellung (8) mit $z_i < g_i = i + 1$, oder gleichwertig $z_i \leq i$, sowie $z_i \neq g_i - 1$ oder gleichwertig $z_i \neq i$ für unendlich viele i. Und (c) sichert deren Eindeutigkeit. Gewiss ist a rational, wenn ihre Darstellung (8) abbricht. Sei anderseits $a = \frac{m}{n}$ gegeben. Dann ergibt (4) mit $z_i = u_i$

$$\tfrac{m}{n} = \tfrac{z_0}{1!} + \tfrac{z_1}{2!} + \ldots + \tfrac{z_{n-1}}{n!} + \tfrac{r_{n-1}}{n!} \qquad (0 \leq r_{n-1} < 1).$$

Multiplikation dieser Gleichung mit $n!$ ergibt

$$\tfrac{m \cdot n!}{n} = z_0 \cdot n! + \tfrac{z_1 \cdot n!}{2!} + \ldots + \tfrac{z_{n-1} \cdot n!}{n!} + r_{n-1}.$$

Offenbar lassen sich in allen Bruchtermen dieser Formel die Nenner wegkürzen, woraus sich schließen lässt, dass r_{n-1} eine natürliche Zahl ist. Damit verbleibt – weil $r_{n-1} < 1$ – nur die Möglichkeit $r_{n-1} = 0$. Es sind dann gemäß (3) alle nachfolgenden Reste, und damit auch die natürlichen Zahlen $z_n, z_{n+1}, \ldots$ in (8) identisch gleich 0. ∎

Für $\mathbf{e} = 1 + \frac{1}{1!} + \frac{1}{2!} + \ldots = 2 + \frac{1}{2!} + \frac{1}{3!} + \ldots$ erhält man in einer Darstellung (8) gerade $z_0 = 2$, während die „Ziffern“ z_i für $i > 0$ sämtlich gleich 1 sind. Also erhalten wir das

Korollar. *Die* EULERsche *Zahl* $\mathbf{e}$ *ist irrational.*

Man beachte, dass dieser Beweis der Irrationalität von $\mathbf{e}$ wesentlich auf der Eindeutigkeit

der Darstellung (8) beruht. Ohne die Nebenbedingung $z_i \neq i$ für unendlich viele i hat man gemäß (6) für $n = 0$ und mit $g_i = i + 1$ (also $G_i = (i + 1)!$) neben der trivialen Gleichung $1 = 1 + \frac{0}{2!} + \frac{0}{3!} + \ldots$ auch die erwähnenswerte Gleichung

$$1 = \tfrac{1}{2!} + \tfrac{2}{3!} + \tfrac{3}{4!} + \ldots$$

Es gibt noch manche andere Reihendarstellungen reeller Zahlen, siehe [27, 11]. Auch kann man unendliche Produkte und Quotienten, z.B. die numerisch interessanten Kettenbruchfolgen betrachten. Diese sind von der Gestalt $a_0 + \frac{1}{a_1},\ a_0 + \frac{1}{a_1 + \frac{1}{a_2}}, \ldots$ $(a_i > 0)$, und deren Limes bezeichnet man im Konvergenzfalle mit $a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots}}}$.

8.4 Berechenbarkeit und Iteration

Wir geben im folgenden einen kurzen Einblick in einen Problemkreis, der von erheblichem Interesse für die Theorie und die Praxis mathematischer Anwendungen ist, den der expliziten Berechenbarkeit. Eine reelle Zahl a (≥ 0) ist im intuitiven Sinne berechenbar, wenn zu jedem n eine n-stellige Näherung a_n mit $|a - a_n| < \varepsilon_n$ explizit angegeben werden kann. Das ist i.a. etwas weniger als die explizite Berechenbarkeit ihrer Dezimalziffern, worauf in **5.4** schon hingewiesen wurde. Es lässt sich auch schreiben $a_n = N_a(n)\overleftarrow{n}$, mit $N_a(n) \in \mathbb{N}$ und einer gewissen berechenbaren oder allgemein-rekursiven Funktion $n \mapsto N_a(n)$. Auf diese Weise lässt sich reelle, und ähnlich auch komplexe Berechenbarkeit innerhalb der weit entwickelten Theorie der rekursiven (zahlentheoretischen) Funktionen präzisieren. Intuitiv ist klar und dies lässt sich streng beweisen, dass die rationalen Operationen $+, -, \times, \div$, aber auch nichtrationale Operationen wie Potenzierung und Logarithmierung, von in diesem Sinne berechenbaren Argumenten wieder zu berechenbaren Werten führen.

Ist $p(x) = c$ eine lösbare Polynomgleichung, so kann eine Lösung mit beliebig hoher Genauigkeit auch berechnet werden. Dies gilt allgemeiner für Gleichungen der Gestalt $f(x) = c$, wenn z.B. gewisse Differenzierbarkeitsforderungen an die Funktion f gestellt werden, und legt die Frage nach der Formulierung möglichst allgemeiner Bedingungen über die Berechenbarkeit einer Lösung von $f(x) = c$ für berechenbare Zahlen c und Funktionen f nahe, falls eine Lösung überhaupt existiert.

Dazu muss offenbar gesagt werden, was die Berechenbarkeit einer reellen Funktion f bedeuten soll. Hier bietet sich folgende Erklärung an. f heiße *berechenbar*, wenn es ein effektives Verfahren[2)] gibt, mit dessen Hilfe sich für jedes berechenbare x aus dem Definitionsbereich den Wert fx beliebig genau berechnen lässt. Es sei zum Beispiel f eine im Intervall $I = [0, v]$ schlichte und berechenbare Funktion, und c eine berechenbare

[2)] Dieser Begriff wird in der Rekursionstheorie präzisiert. Mit den berechenbaren reellen Funktion befasst sich die so genannte *Rekursive Analysis*, wo Elemente der Analysis, der Numerik und der Rekursionstheorie zusammenfließen. Nach der Erörterung in 5.4 ist plausibel, dass z.B. die simple Funktion $a \mapsto \operatorname{Int} a$ nicht berechenbar ist. Es lässt sich zeigen, dass eine berechenbare Funktion immer stetig ist.

Zahl mit $f0 < c < fv$. Dann hat die Gleichung $fx = c$ nach Satz 7.1 die Lösung $\xi = \sup\{x \in I \mid fx \leq c\}$, deren Berechnung sich leicht programmieren lässt. Es kommt nur darauf an, ob der Test $\boxed{fx \leq c}$ mit $x = \xi_n + i\varepsilon_{n+1}$ für $i = 1, \ldots, 9$ effektiv ausführbar ist, wobei $\xi_n \leq \xi \leq \xi_n + \varepsilon_n$ ein schon berechneter Näherungswert ist. Dieser ist z.B. für $x \mapsto x^n$ präzise ausführbar, weil $x^n \in \mathbb{E}$ für $x \in \mathbb{E}$. Daher lässt sich $\sqrt[n]{c}$ sogar ziffernweise berechnen, falls die Information vorliegt ob $c \in \mathbb{E}$ oder nicht, und c im letzteren Falle selbst auch ziffernweise berechenbar ist. Die Ausführbarkeit des obigen Tests beruht hier wesentlich darauf, dass mit $x \in \mathbb{E}$ auch $x^n \in \mathbb{E}$. Bei nichtrationalen berechenbaren Funktionen f ist i.a. nur die übliche näherungsweise Berechenbarkeit einer Lösung von $fx = c$ garantiert. Das gilt auch dann noch, wenn f in I nur stetig und berechenbar ist, für jedes berechenbare c zwischen fu und fv.

Im Einzelfall werden den Gegebenheiten besonders angepasste Verfahren zur Berechnung von Zwischenwerten gewählt. So rechnen die Computer die n-te Wurzel meist logarithmisch und natürlich nur näherungsweise aus, es sei denn, man schreibt ein eigenes Programm. Die Möglichkeiten hierfür sind vielfältig. Am beliebtesten sind Iterationsverfahren, weil deren Programmierung besonders bequem ist. Beispiele behandeln wir unten. Sie beruhen auf folgendem Satz, einem von zahlreichen ähnlichlautenden Konvergenzsätzen der numerischen Analysis.

Satz 8.5 (Fixpunktsatz). *Sei $I \subseteq \mathbb{D}$ ein abgeschlossenes Intervall, sowie $f : I \to I$ eine in I schlichte Funktion, und $\langle x_n \rangle$ eine beliebige Folge mit $x_0 \in I$* und

$$x_{n+1} = fx_n \quad \text{für alle } n. \tag{9}$$

Dann ist $\langle x_n \rangle$ konvergent, und $\lim\langle x_n \rangle$ ist Fixpunkt von f. Ist f zudem schlicht in I mit $C < 1$, hat f in I nur einen einzigen Fixpunkt.

Beweis. $\langle x_n \rangle$ falle monoton und ist dann sicher konvergent; wegen $x_{n+1} = fx_n$ und $\xi := \lim\langle x_n \rangle = \lim\langle x_{n+1} \rangle$ ist auch $\xi = \lim\langle fx_n \rangle$. Weil $\lim\langle fx_n \rangle = f\xi$ nach (11) in **7.2**, ergibt sich $\xi = f\xi$. Sei nun $x_k < x_{k+1}$ $(= fx_k)$ für ein gewisses k. Ein simpler Induktionsschluss zeigt $x_n \leq x_{n+1}$ für alle $n \geq k$. Auch ist $\langle x_n \rangle_{n \geq k}$ beschränkt, weil diese Folge ganz in I verbleibt. Also ist sie schlicht und damit konvergent, und es gilt $\xi := \lim\langle x_n \rangle_{n \geq k} = \lim\langle x_n \rangle$. Die Gleichung $f\xi = \xi$ erschließt man wie eben. Sei nun $fx - fy \leq C(x - y)$ mit $C < 1$ für alle $x, y \in I,\ x > y$. Ist $\xi_1 = f\xi_1,\ \xi_2 = f\xi_2$, und wäre etwa $\xi_1 < \xi_2$, so folgt $\xi_2 - \xi_1 \leq C(\xi_2 - \xi_1)$, also der Widerspruch $1 \leq C$. ∎

Dass eine den Voraussetzungen dieses Satzes genügende Funktion f überhaupt einen Fixpunkt hat, ist keine Neuheit. Dazu genügt nach Beispiel 3 in **3.3** bereits monotones Wachstum von f. Auch konvergiert dann immer noch die Folge $\langle x_n \rangle$, nur muss deren Limes nicht notwendig auch ein Fixpunkt von f sein. Auch lässt sich die Voraussetzung $x_0 \in I$ offenbar noch abschwächen zu $x_i \in I$ für ein gewisses i, weil $\langle x_n \rangle$ und $\langle x_n \rangle_{n \geq i}$ dasselbe Konvergenzverhalten haben. Für Anwendungen von Satz 8.5 ist hauptsächlich der Fall interessant, dass f genau einen Fixpunkt ξ in I hat. Jede gemäß (9) definierte Folge – eine so genannte *Iterationsfolge* – konvergiert dann automatisch gegen ξ. Oft weiß man aus anderen Gründen, dass f in I höchstens einen Fixpunkt hat, so dass die

dafür hinreichende (aber nicht notwendige) Bedingung $C < 1$ gar keine Rolle spielt. Betrachten wir als Beispiel die bereits in Übung 7.7 im Intervall $[0, \sqrt{a}]$ als schlicht nachgewiesene Funktion $f : x \mapsto \frac{2ax}{a+x^2}$ für $a = 10$ nun in einem Intervall $I = [u, v]$ mit $0 < u < v := \sqrt{a}$, siehe hierzu Figur 5. f bildet I in sich ab und hat dort nur einen Fixpunkt ξ, nämlich am oberen Intervallende; denn die Gleichung $x = \frac{2ax}{a+x^2}$ hat nur die Lösungen $\xi = 0$ und $\xi = \sqrt{a}$. Die untere Intervallgrenze u (= 0,5 in der Figur) ist nicht 0, um den Fixpunkt 0 aus der Betrachtung fernzuhalten und darf innerhalb des Spielraums $0 < u < \sqrt{a}$ frei gewählt werden. Nach Satz 8.5 konvergiert jede Iterationsfolge $\langle x_n \rangle$ mit $x_0 \in I$ gegen den einzigen Fixpunkt $\xi = \sqrt{a}$ von f in I. Damit hat man ein bequemes iteratives Verfahren zur Berechnung von $\sqrt{a}$ mittels der rationalen Operationen. Man kann die Berechnung von $\sqrt{10}$ nach diesem Iterationsverfahren mit dem Eingangswert $x_0 = 1$ in der Figur längs der gestrichelten Linien deutlich verfolgen. Je größer n ist, desto näher liegt x_n bei $\sqrt{a}$. Gewisse Informationen über die Konvergenzgeschwindigkeit dieses Verfahrens, welche wesentlich durch die Anzahl der Durchläufe durch die Iterationschleife bestimmt wird, enthält die Bemerkung weiter unten.

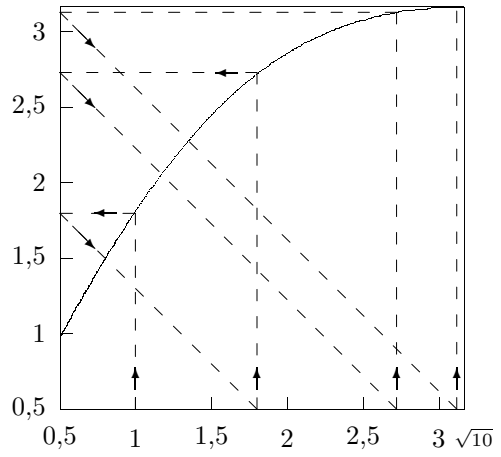
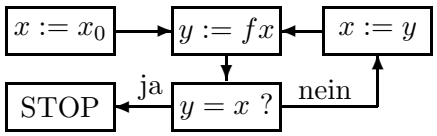


Fig. 5 Iterative Berechnung des Fixpunktes $\sqrt{a}$ von $x \mapsto \frac{2ax}{a+x^2}$ für $a = 10$.



Die Programmierung des Verfahrens entsprechend nebenstehendem, leicht durchschaubaren Flussdiagramm verwendet den Test $\boxed{y = x\ ?}$, der angibt, ob die Berechnung nahe genug am Konvergenzpunkt liegt. Die Rechnung endet, wenn die Werte von x und y nach einer gewissen Anzahl von Schleifendurchläufen im Rahmen der Genauigkeit eines gegebenen Rechners ununterscheidbar sind, oder, wie man auch sagt, der Rechenprozess in den Bereich des „Rundungsrauschens“ gelangt. Auf dem programmierbaren Taschenrechner erkennt man dies bei einigen Iterationsverfahren nach Ersetzung des Abbruchtests durch $\boxed{\text{DISPLAY } x, y}$ an einem periodischen Wechsel der letzten Ziffer in der Anzeige. Hier die Berechnung von $\sqrt{10}$ nach dem obigen Verfahren mit der 0-ten Näherung $x_0 = 1$ auf einem Taschenrechner mit 12-stelligem Display:

$$x_1 = 1{,}81818181818, \quad x_2 = 2{,}73291925465, \quad x_3 = 3{,}12890616374,$$
$$x_4 = 3{,}16209970755, \quad x_5 = 3{,}16227765516, \quad x_6 = 3{,}16227766017.$$

Bereits nach dem 6. Iterationsschritt ändern sich die Folgenglieder nicht mehr und $\sqrt{10} = 3{,}16227\,76601\,6838\cdots$ ist so genau berechnet, wie es die Anzeige des Rechners zulässt. Für $x_0 = 3$ wären sogar nur 3 Iterationsschritte nötig. Übrigens muss man nicht auf $x_0 \leq \sqrt{a}$ bestehen. x_0 muss positiv, darf sonst aber beliebig sein. Falls $x_0 > \sqrt{a}$, liegt bereits $x_1 = fx_0$ wieder in I, wie man sich leicht überlegt.

Weitere Beispiele. (a) Man sieht leicht, dass $f : x \mapsto \sqrt{2x}$ in $[u, v]$ schlicht ist, wenn nur $u > 0$. Für $I := [1, 2]$ ist auch $f : I \to I$. Daher konvergiert die Iterationsfolge $\langle x_n \rangle$ mit $x_0 = 1$, also $x_1 = \sqrt{2}$, $x_2 = \sqrt{2\sqrt{2}}$, ... – siehe Übung 7.1 – gegen den einzigen Fixpunkt von f in I, nämlich die Lösung $\xi = 2$ der Fixpunktgleichung $x = fx$, also von $x = \sqrt{2x}$. Damit haben wir einen leicht überschaubaren Beweis für $\lim\langle x_n \rangle = 2$.
(b) Sei $x_0 = 0$ und $x_{n+1} = \sqrt{2 + x_n}$, also $x_1 = \sqrt{2}$, $x_2 = \sqrt{2 + \sqrt{2}}$, ... Auch $\langle x_n \rangle$ ist konvergent mit $\lim\langle x_n \rangle = 2$. Denn $f : x \mapsto \sqrt{2 + x}$ ist schlicht in $I = [0, 2]$ mit $C = \frac{1}{\sqrt{8}} < 1$; ferner gilt $f : I \to I$, und $\xi = 2$ ist nach Satz 8.5 einziger Fixpunkt von f in I.
(c) Manche Computer vermeiden direkte Divisionen und ermitteln $\frac{1}{a}$ für $a \in \mathbb{D}_+$ mit dem folgendem Iterationsverfahren: Sie suchen schnellstmöglich einen Wert x_0 mit $0 < x_0 \cdot a \leq 1$ und berechnen dann iterativ $x_{n+1} = 2x_n - ax_n^2$ bis zum Eintritt des Rundungsrauschens. Auch hier sichert Satz 8.5 die Konvergenz von $\langle x_n \rangle$ gegen den Fixpunkt $\frac{1}{a}$ der in $[0, \frac{1}{a}]$ leicht als schlicht nachweisbaren Funktion $x \mapsto 2x - ax^2$.

Bemerkung. Eine konvergente Folge $\langle x_n \rangle$ mit $\xi = \lim\langle x_n \rangle$ konvergiert *linear*, *quadratisch* bzw. *kubisch*, wenn es eine Konstante c gibt mit $d_{n+1} \leq c \cdot d_n^{\,k}$ für alle n, wobei $d_n = |\xi - x_n|$ und $k = 1, 2$ bzw. 3, sowie $c < 1$ für $k = 1$. Grob gesprochen, die Anzahl der genauen Dezimalen der Näherung x_n von x wächst linear, quadratisch bzw. kubisch mit n. So konvergiert die in den Beispielen oben zuletzt genannte Folge quadratisch, Übung 7. Auch die durch $x_{n+1} = \frac{2ax_n}{a+x_n^2}$ definierte Iterationsfolge konvergiert quadratisch, ebenso wie die des so genannten „Babylonischen“ Verfahrens in Übung 6, ein Spezialfall des NEWTONschen Näherungsverfahrens. Es geht noch besser: Die durch $x_{n+1} = \frac{x_n(x_n^2+3a)}{3x_n^2+a}$ definierte schlichte Folge $\langle x_n \rangle$ konvergiert sogar kubisch gegen $\sqrt{a}$. Beginnend mit $x_0 = 3$, hat man für $a = 10$ im ersten Iterationsschritt $\sqrt{10}$ bereits auf 3 Dezimalen, und im zweiten auf alle Dezimalen des Displays (in Wahrheit auf sogar 13 Dezimalen) genau. Hingegen konvergiert die Folge $\langle e_n \rangle$ aus **6.2** nicht einmal linear, wie mittels der Ungleichungen (e1),(e2) aus **7.1** unschwer erschlossen werden kann.

Eine in einem Intervall I definierte Funktion f heißt *Lipschitz-stetig in* I, wenn

(10) Es gibt eine Konstante C mit $|fx - fy| \leq C|x - y|$ für alle $x, y \in I$

erfüllt ist. Eine in I monoton wachsende Funktion ist dort Lipschitz-stetig genau dann, wenn sie schlicht ist. Falls in (10) $C < 1$ gewählt werden kann, heißt f eine *Kontraktion*. Satz 8.5 ähnelt dem BANACH*schen Fixpunktsatz*, der behauptet, dass eine Kontraktion f von I genau einen Fixpunkt ξ in I hat und dass *jede* Iterationsfolge $\langle x_n \rangle$ mit $x_k \in I$ für ein k gegen ξ konvergiert. Der Beweis ist ähnlich dem von Satz 8.5 weshalb er hier übergangen sei. Eine von vielen Anwendungen dieses Satzes ist folgende:

In [22] wird am Beispiel $a = 47$ ein iteratives Verfahren zur Berechnung von $\sqrt[3]{a}$ für $a \in \mathbb{N}_+$ suggeriert, welches das LEONARDO*sche Iterationsverfahren* genannt sei. Es ist vermutlich das erste, schriftlich fixierte numerische Iterationsverfahren im modernen Sinne und funktioniert für reelle $a \geq 1$ genauso wie in LEONARDOs Beispiel: Man errate zuerst ein $u \geq 1$ mit $u^3 \leq a < (u+1)^3$ – ein solches u findet man stets auch in $\mathbb{N}$ – setze $x_0 := u$ und berechne sodann iterativ $x_{n+1} = x_n + \frac{a - x_n^3}{3(u+1)x_n}$. Tatsächlich konvergiert $\langle x_n \rangle$ nach dem BANACHschen Fixpunktsatz gegen die Lösung $\sqrt[3]{a}$ der Fixpunktgleichung $x = Fx$ mit $F \mapsto x + \frac{a - x^3}{3(u+1)x}$. Denn F ist Kontraktion eines geeigneten Intervalls

$[u,v]$ für $v \geq u+1$, Übung 8. Das Verfahren konvergiert nicht sehr schnell. Beginnend mit $x_0 = 3$, benötigt man 11 Schritte, um z.B. $\sqrt[3]{47} = 3{,}608\cdots$ auf 10 Dezimalen zu berechnen. Aber für Handrechnungen ist es durchaus brauchbar. Zum Beispiel ist schon $x_1 = 3{,}555\cdots$. Hier noch eine weitere einfache Anwendung:

Die berühmte, durch $\varphi_0 = 1,\ \varphi_1 = 1$ und (a) $\varphi_{n+1} = \varphi_n + \varphi_{n-1}$ für $n > 0$ definierte Folge heißt die FIBONACCI–Folge. Also $\varphi_2 = 2,\ \varphi_3 = 3,\ \varphi_4 = 5$, aber z.B. ist bereits $\varphi_{58} = 956\,722\,026\,041$. Sei $q_n := \frac{\varphi_{n+1}}{\varphi_n}$. Wir beweisen, die Folge $\langle q_n\rangle$ konvergiert gegen $\frac{1+\sqrt{5}}{2} = 1{,}618\cdots$, der Verhältniszahl des Goldenen Schnitts. (a) ergibt $q_{n+1} = 1 + \frac{1}{q_n}$. Nun ist $f : x \mapsto 1 + \frac{1}{x}$ Kontraktion von $I := [\frac{3}{2}, 2]$; denn (10) ist mit $C = \frac{2}{3}$ erfüllt, wie man leicht sieht. Weil $q_2 \in I$, ist $\langle q_n\rangle$ nach dem BANACHschen Fixpunktsatz konvergent, und weil $\frac{1+\sqrt{5}}{2}$ die Fixpunktgleichung $x = 1 + \frac{1}{x}$ in I löst, gilt $\lim\langle q_n\rangle = \frac{1+\sqrt{5}}{2}$. Auf q_0 (also auf φ_0 und φ_1) kommt es demnach gar nicht an. Das erkennt man auch deshalb, weil $q_1 = 1 + \frac{1}{q_0},\ q_2 = 1 + \frac{1}{1+\frac{1}{q_0}}, \ldots$, und damit $\frac{1+\sqrt{5}}{2} = \lim\langle q_n\rangle = 1 + \cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{\ddots}}}$.

8.5 Übungen

1. Man zeige: eine positive rationale Zahl a ist im Dezimalsystem reinperiodisch genau dann, wenn es eine Darstellung $a = \frac{m}{n}$ gibt, so dass n und 10 teilerfremd sind. Damit sind Summe und Differenz (im Falle ihrer Existenz), sowie das Produkt reinperiodischer Zahlen wieder reinperiodisch. Diese bilden also einen (in $\mathbb{D}$ dichten) $\mathcal{E}$-Bereich reeller Zahlen, der die Elemente aus $\mathbb{E}\backslash N$ sozusagen auslässt.

2. Seien m, n teilerfremd und sei $a = \frac{m}{n}$ reinperiodisch mit Periodenlänge p. Man zeige (a) n teilt $10^p - 1$, (b) p ist die kleinste Zahl aus $\mathbb{N}$ derart, dass n Teiler ist von $10^p - 1$. (Stets ist $p \leq$ Anzahl der zu n teilerfremden $k \leq n$, EULER.)

3. Sei $a = z_0{,}z_1z_2\cdots_8$ eine Oktalzahl, also $0 \leq z_i < 8$ für $i > 0$, und auch z_0 sei im Oktalsystem geschrieben. Man zeige: die binäre Entwicklung von a erhält man durch ziffernweise Umschrift der Oktalziffern von a wie im Text geschildert.

4. Man betrachte die Kommaverschiebung für g-adische reelle Zahlen a und beweise $a\overset{n}{\rightarrow} = a \cdot g^n$, sowie $a\overset{n}{\leftarrow} = a \cdot \frac{1}{g^n}$.

5. Man beweise, $a \in \mathbb{D}$ ist rational genau dann, wenn für jedes $g \geq 2$ die g-adische Entwicklung von a periodisch ist. Diese Eigenschaft ist also basisinvariant.

6. Häufig wird folgendes Iterationsverfahren zur Berechnung von $\sqrt{a}$ für $a > 0$ betrachtet: $x_0 > 0$ sei beliebig, und $x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n})$. Man beweise, dass $\langle x_n\rangle_{n>0}$ eine fallende, gegen $\sqrt{a}$ konvergierende Folge ist.

7. Sei $a > 0$ und $0 < ax_0 \leq 1$, sowie $x_{n+1} = 2x_n - ax_n^2$. Man zeige: $\langle x_n\rangle$ ist eine schlichte Folge und diese konvergiert quadratisch gegen ihren Grenzwert $\frac{1}{a}$.

8. Sei $u \geq 1$. Man zeige: $F : x \mapsto x + \frac{a-x^3}{3(u+1)x}$ kontrahiert $[u,v]$ für ein geeignetes $v > u+1$. Dies sichert die Konvergenz des LEONARDOschen Verfahrens.

Abschnitt 9

Negative und komplexe Zahlen

Die bisherigen Ausführungen zeigen, dass man interessante Teile der Analysis auch ohne negative Zahlen bequem entwickeln kann. Dennoch macht sich an einigen Stellen deren Fehlen empfindlich bemerkbar, z.B. beim logarithmischen Rechnen, und auch viele Iterationsverfahren wären nur beschwerlich zu formulieren. Nun gibt es keineswegs nur innermathematische Gründe für den Gebrauch negativer Zahlen, die diesen die Rolle eines reinen Rechenhilfsmittels zuweisen würden. Auch im Alltag treten sie in Erscheinung. Man denke an Beispiele wie Kontostand, Temperatur-Skalen, usw. Trotz dieses Wirklichkeitsbezuges deutet die Bezeichnung „negativ“ darauf hin, dass man derartigen Zahlen zunächst mit Skepsis begegnete.

Die Konstruktion negativer Elemente geht für alle $\mathcal{E}$-Bereiche kanonisch (in derselben Weise) vonstatten. Zwar interessiert man sich meist nur für die Hinzufügung negativer Zahlen für spezielle Bereiche wie $\mathbb{N}$ und $\mathbb{D}$, aber es erschwert weder die Aufgabenstellung noch deren Durchführung, wenn diese sich auf einen beliebigen Elemenarbereich bezieht; im Gegenteil, die Allgemeinheit der Betrachtung macht den Vorgang durchsichtiger. Freilich darf man sich unter einem $\mathcal{E}$-Bereich stets einen konkreten Bereich, etwa den der natürlichen oder der nichtnegativen reellen Zahlen vorstellen.

Negative Zahlen werden hier nicht mit in Universitätsvorlesungen beliebten Verfahren der differenzengleichen Paare, sondern direkt eingeführt. Ein Hauptvorteil dieser Konstruktion ist der enge Bezug zur Realisierung des Rechnens in Computern. Bei der Projektierung der Rechnerchips wird nämlich so verfahren wie in **9.2** beschrieben. Ein anderer, wenn auch unbedeutender Vorteil ist, dass die schon vorhandenen Zahlen „bleiben was sie sind“. Wir beweisen ausführlich, dass die Konstruktion negativer Zahlen unter Beachtung natürlicher Forderungen auf nur eine einzige Weise zum Erfolg führen kann. Dieser Aspekt wird mitunter etwas vernachlässigt.

Die Konstruktion negativer Zahlen hat wesentliche Gemeinsamkeiten mit derjenigen komplexer Zahlen, trotz der Unterschiede in den Details. Zu jedem Körper, in welchem die Gleichung $x^2 = -1$ unlösbar ist, lässt sich kanonisch ein Erweiterungskörper $K[\mathbf{i}]$ so konstruieren, dass diese Gleichung durch $\pm\mathbf{i}$ gelöst wird und auch dies gelingt unter Beachtung natürlicher Forderungen nur auf eine Weise wie in **9.4** gezeigt wird.

9.1 Ringerweiterungen und Konsequenzen aus der Existenzannahme

Werfen wir zunächst noch einmal einen Blick auf das Axiomensystem in Abschnitt **2**. Nur das Axiom E verursacht die Einschränkung hinsichtlich der Subtraktion, obwohl E anderseits nützliche Konsequenzen hat wie etwa die Monotoniegesetze. Die Aufgabenstellung besteht also darin, einen gegebenen $\mathcal{E}$-Bereich A durch Hinzufügen neuer Elemente so zu einem Bereich $B \supseteq A$ zu erweitern, dass die Subtraktion in B keinen Beschränkungen mehr unterliegt. Es versteht sich von selbst, dass die Grundoperationen von B Fortsetzungen derjenigen von A sein sollen. Man erwartet, dass auch in B Addition und Multiplikation wieder assoziativ und kommutativ sind und dass 0 und 1 weiterhin neutrale Elemente bleiben. Schließlich sollte B nach Möglichkeit wieder geordnet sein, und zwar so, dass $a < b$ für $a, b \in A$ in B und in A dieselbe Bedeutung hat. Natürlich soll dies möglichst so geschehen, dass B keine neuen positiven Elemente enthält. Alle diese Forderungen sind tatsächlich erfüllbar, und zwar nur auf eine einzige Art und Weise, wobei Einzigkeit hier natürlich nur Einzigkeit bis auf Isomorphie bedeuten kann.

Die erwähnte Aufgabenstellung lässt sich wie folgt präzisieren, wobei wir die Frage der Anordnung vorerst außer acht lassen. Gegeben ist ein $\mathcal{E}$-Bereich A. Gesucht ist ein Bereich $B \supseteq A$, so dass folgenden Forderungen (I) und (II) Genüge getan wird:

(I) B erfüllt die Rechengesetze $\mathsf{N}^+, N^\times, K^+, K^\times, A^+, A^\times$ und D, sowie die Aussage

$$\mathsf{L}: \quad \textit{Für alle } a, b \textit{ gibt es ein } x \textit{ mit } b + x = a.$$

(II) B enthält keine überflüssigen Elemente; genauer, ist $A \subseteq B' \subseteq B$ und erfüllt auch B' die Forderung (I), so ist $B' = B$.

Eine diesen Forderungen genügende Erweiterung B von A heiße eine *Ringerweiterung* von A (genau genommen eine *minimale* Ringerweiterung). Allgemein heißt ein Bereich B mit Addition, Multiplikation und neutralen Elementen 0 und 1, welcher den unter (I) genannten Rechengesetzen genügt, ein *Ring*, genauer, ein *Ring mit* 1 [1)].

Es werde nun im folgenden angenommen, es existiere eine Ringerweiterung B unseres vorgegebenen $\mathcal{E}$-Bereichs A. Daraus ergeben sich eine Reihe von Informationen, die uns für die explizite Konstruktion in **9.2** den genauen Weg weisen werden. Man beachte dabei stets den hypothetischen Charakter der Ausführungen, die von der Existenz einer Ringerweiterung für A ausgehen. Es ist im Prinzip dasselbe Vorgehen wie bei der Gleichungslösung: zuerst leitet man eine oder mehrere Bedingungen aus der Annahme her, eine Lösung existiere; dann erst wird die Lösung explizit angegeben.

[1)] noch genauer, ein *kommutativer* Ring mit 1, denn $\mathsf{K}^\times$ wird in der Definition von *Ring* nicht gefordert. In diesem Abschnitt meint *Ring* stets *kommutativer Ring*. Wird auch auf die Forderung L verzichtet, spricht man von einem (kommutativen) *Halbring*. Ein $\mathcal{E}$-Bereich ist ein solcher Halbring, genauer ein *geordneter* Halbring, der wegen des Axioms E auch ein *natürlich geordneter Halbring* genannt wird. $\mathcal{E}$-Bereiche sind demnach natürlich geordnete Halbringe.

Von grundsätzlicher Bedeutung ist, dass in einem Ring B die Streichungsregel und damit die eindeutige Lösbarkeit von $a + x = b$ für $a, b \in B$ – anders als in **2** – gänzlich ohne Monotoniegesetze erschlossen werden kann. Gemäß L existiert nämlich zu jedem $a \in B$ ein $a' \in B$ mit $a' + a = a + a' = 0$. Ist daher $a + x_1 = a + x_2$, so ergibt sich $x_1 = 0 + x_1 = a' + a + x_1 = a' + a + x_2 = 0 + x_2 = x_2$. Folglich gibt es zu allen $a, b \in B$ genau ein d mit $b + d = a$, das allgemein mit $a - b$ bezeichnet wird. Damit ist die Subtraktion überall auf B erklärt und es gilt

(1) $\quad a - b = d \;\Leftrightarrow\; b + d = a.$

Speziell hat damit die Gleichung $a + x = 0$ eine eindeutige Lösung, die man mit $-a$ bezeichnet und *das additive Inverse* von a nennt. Demnach gilt $a + -a = 0$. Das Minuszeichen tritt also zugleich als ein- und als zweistelliges Operationssymbol auf. Man beachte $- - a = a$; denn wegen K^+ gilt $-a + a = 0$. Aus $-a = -b$ ergibt sich $a = - -a = - -b = b$. Die Operation $\mu : x \mapsto -x$ von B, welche auch die *Minusfunktion* von B heißt, vermittelt daher eine Bijektion von B auf sich selbst.

Wir beweisen zuerst für einen beliebigen Ring B für alle $a, b \in B$ die Gleichungen

(2) $a + -b = a - b$, (3) $a + -b = -(b - a)$, (4) $-a + -b = -(a + b)$,
(5) $a \cdot 0 = 0 \cdot a = 0$, (6) $a \cdot -b = -(ab)$, (7) $- a \cdot -b = ab$

Nach $\mathsf{A}^+, \mathsf{K}^+$ und N^+ ist $b + (a + -b) = a + b + -b = a + 0 = a$. Also gilt (2) gemäß (1). Damit folgt $(b - a) + (a + -b) = b + -a + a + -b = 0$. Also ist $a + -b$ mit dem additiven Inversen von $b - a$ identisch, was (3) gerade behauptet. (4) folgt analog aus $(a + b) + (-a + -b) = 0$. (5) beweist man wie in **2**. (6) ergibt sich mit (1) und D aus $a \cdot b + a \cdot -b = a(b + -b) = a0 = 0$. (7) ergibt sich durch zweimaliges Anwenden von (6) aus $-a \cdot -b = -(-a \cdot b) = -(b \cdot -a) = - - (ba) = ba = ab$.

Sei nun B unsere hypothetische Ringerweiterung von A. Auch die Subtraktion von B erweitert die auf A partiell bereits gegebene, weil (1) für $a, b, d \in A$ im Falle $a \geq b$ ja auch in A gilt. Wie in **2.1** sei $A_+ = A \backslash \{0\}$. Nach (2) - (7) gilt speziell für $a, b \in A$

$$(8) \qquad \begin{array}{l} a + -b = \begin{cases} a - b & \text{für} \quad a \geq b \\ -(b - a) & \text{für} \quad b > a \end{cases} \quad ; \quad - a + -b = -(a + b); \\ a \cdot -b = -(ab) \qquad\qquad ; \qquad - a \cdot -b = ab. \end{array}$$

Sei $A_- := \{-a \mid a \in A_+\}$ $(\subseteq B)$. Es ist $A_+ \cap A_- = \emptyset$. Denn sei $a \in A_+$, also $-a \in A_-$. Wäre auch $-a \in A_+$, so folgt wegen $a + -a = 0$ und E in **2.1** offenbar $a < 0$, im Widerspruch zu $a > 0$.

Satz 9.1. *Sei B eine Ringerweiterung von A. Dann gelten die Gleichungen* (8), *sowie*

(9) $\quad B = A \cup A_- = A_- \cup \{0\} \cup A_+.$

Ferner gibt es (bis auf Isomorphie) höchstens eine Ringerweiterung von A. Schließlich kann B auf genau eine Weise so geordnet werden, dass B auch das Monotoniegesetz M^+ erfüllt; es gilt dann auch $\mathsf{M}^\times$ in B, und B hat bezüglich dieser Ordnung dieselben positiven Elemente wie A.

Beweis. (8) wurde schon gezeigt. Sei $B' := A \cup A_-$ ($\subseteq B$). Die rechten Seiten von (8) liegen für $a, b \in A$ offenbar in B', also auch die linken. Hieraus folgt, dass B' gegenüber $+, \cdot$ abgeschlossen ist. Außerdem enthält B' mit jedem Element auch dessen additives Inverse: Für $a \in A_+$ ist $-a \in A_-$; für $b \in A_-$ sei $b = -a$ mit $a \in A_+$, so dass dann $-b = --a = a \in A_+$. Mit a, b liegt also auch $a - b = a + -b$ in B'. Damit erfüllt $B' \subseteq B$ alle Axiome für Ringe wie man leicht prüft. Also $B' = B$ nach (II). Das beweist (9). Sei nun nebst B auch $\tilde{B}$ Ringerweiterung von A, sagen wir mit der Minusfunktion $\tilde{\mu}$. Wegen (9) gilt dann $\tilde{B} = A \cup \{\tilde{\mu}a \mid a \in A_+\}$, und $\varphi : B \mapsto \tilde{B}$ mit $\varphi(a) = a$ und $\varphi(\mu a) = \tilde{\mu}a$ für $a \in A$ ist ein Isomorphismus von B auf $\tilde{B}$, wie man leicht bestätigt. Nun zeigen wir, dass B unter Wahrung von M^+ auch geordnet werden kann. Wir erklären

$$(10) \quad a < b \ :\Leftrightarrow \ b - a \in A_+ \qquad (a, b \in B)$$

und behaupten, B wird durch $<$ geordnet. Sicher ist $a \not< a$ für alle $a \in B$. Die Transitivität T gilt wegen $b - a, c - b \in A_+ \ \Rightarrow \ c - a = (c - b) + (b - c) \in A_+$. Auch ist $a < b$ oder $b < a$ für alle $a, b \in B$ mit $a \neq b$, weil nach (9) entweder $b - a \in A_+$ oder $a - b = -(b - a) \in A_+$. Ferner ist für $a, b \in A$ offenbar $a < b$ im alten Sinne genau dann, wenn $a < b$ im neuen Sinne. Für $a < b$ liefert (10) deswegen $(b+c)-(a+c) = b-a \in A_+$. Also ist $a + c < b + c$ und M^+ ist bewiesen. Nach (10) ist speziell $0 < b \Leftrightarrow b \in A_+$, so dass A_+ genau die Menge der in B bezüglich $<$ positiven Elemente ist. Deshalb folgt im Falle $c > 0$ aus $a < b$ sicher $bc - ac = (b - a)c > 0$, und daher $ac < bc$. Damit ist $\mathsf{M}^\times$ bewiesen. Für eine M^+ erfüllende Anordnung $<$ von B gilt nun andererseits auch (10). Denn $a < b$ ergibt $0 = a - a < b - a$ nach M^+, was dasselbe bedeutet wie $b - a \in A_+$. Analog folgt $b - a \in A_+ \ \Rightarrow \ a < b$. Damit ist auch die Eindeutigkeitsbehauptung über die Ordnung der Ringerweiterung gezeigt worden. ∎

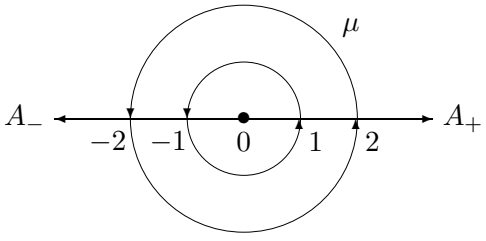


Fig. 6 Ringerweiterung von A

Figur 6 gibt eine einprägsame Veranschaulichung einer Ringerweiterung [2] $B \supseteq A$. Es lässt sich B als eine durch 0 verlaufene Gerade veranschaulichen, wobei links von 0 die negativen, rechts von 0 die positiven Elemente von B liegen. Die Minusfunktion μ vermittelt eine Bijektion von A_+ auf A_- und von A_- auf A_+. M^+ ergibt $-b < -a \Leftrightarrow a < b$, speziell für $a, b \in A$, womit die Figur auch die Anordnung von B in anschaulich korrekter Weise wiedergibt. Diese ist allein durch diejenige von A und die auf A eingeschränkte Minusfunktion bestimmt. Ferner zeigen die Gleichungen (8), dass auch die Grundoperationen in $B = A \cup A_-$ einzig durch Addition, Subtraktion und Multiplikation in A, sowie die auf A eingeschränkte Minusfunktion bestimmt sind. Damit ist nunmehr vollständig klar, dass die Konstruktion einer Ringerweiterung von A nur eine einzige Chance auf Erfolg hat.

[2] Der Name stammt nicht von diesem Bild. Vielmehr entstammen die im 19. Jahrhundert geprägten Begriffe *Gruppe, Ring, Körper, Verband* dem Gesellschaftsleben und sollten zum Ausdruck bringen, dass die Elemente des betreffenden Bereichs miteinander in wohldefinierten Beziehungen stehen.

9.2 Konstruktion der Ringerweiterung

Nach den ausführlichen Darlegungen des letzten Abschnitts ist die Konstruktion einer Ringerweiterung für einen gegebenen $\mathcal{E}$-Bereich A eine klar vorgezeichnete Aufgabe. Für deren explizite Durchführung gibt es nun ganz unterschiedliche Möglichkeiten. Am bekanntesten ist die Methode der differenzengleichen Paare. Diese entspricht zwar mehr dem Geiste der modernen Algebra, nicht aber der genetischen Entwicklungsgeschichte negativer Zahlen. Wir wollen indes wenigstens deren Grundidee kurz beschreiben.

Man geht aus von einer Klasseneinteilung aller geordneten Paare $(a, a') \in A \times A$ nach der folgendermaßen erklärten Äquivalenzrelation $\sim$, der „Differenzengleichheit“:

$$(a, a') \sim (b, b') \;\Leftrightarrow\; a + b' = a' + b \qquad (a, a', b, b' \in A).$$

Sei nun B die Menge aller Äquivalenzklassen nach dieser Relation (die sich auch als „Verschiebungen“ von A veranschaulichen lassen). Auf B werden dann in bestimmter Weise Rechenoperationen $+, \cdot$ erklärt und es wird nachgewiesen, dass B ein Ring ist. Als nächstes wird gezeigt, dass $a \mapsto (b, b')$ mit $b - b' = a$ eine Einbettung[3)] von A in B ist, und schließlich noch, dass B in der Tat Ringerweiterung von A ist.

Wir schreiten nun aber zur direkten Konstruktion der Ringerweiterung eines gegebenen $\mathcal{E}$-Bereichs A. Bei dieser Methode entfällt die Klassenbildung und die schon vorhandenen Elemente bleiben was sie waren. Es wird, grob gesprochen, zu jedem Element $a \in A_+$ ein entsprechendes negatives Element „erschaffen“. Natürlich darf man sich unter A einen konkreten Zahlenbereich vorstellen. Dies ändert oder vereinfacht nichts an den erforderlichen Betrachtungen.

Ist $a \in A_+$, so heiße das geordnete Paar $(-, a)$ mit dem Minuszeichen in der ersten Komponente *das dem Element a entsprechende negative Element*. Statt $(-, a)$ schreiben wir $-a$. Sei $A_- := \{-a \mid a \in A_+\}$ und $B := A \cup A_-$. Es ist dann natürlich $-a \neq -b$ für $a \neq b$, und auch $A \cap A_- = \emptyset$.

Wir erweitern nun die arithmetischen Operationen von A auf B. Es soll $-a$ das additive Inverse von $a \in A_+$ werden. Also können gemäß Satz 9.1 nur die folgenden Definitionen zum Erfolg führen; darin bezeichnen a, b beliebige Elemente aus A mit $b \neq 0$:

$$(11)\qquad \begin{array}{lll} a + -b & := & \begin{cases} a - b, & \text{falls} \quad a \geq b \\ -(b - a) & \text{falls} \quad b > a \end{cases} \; ; \qquad -a + -b := -(a + b); \\ a \cdot -b & := & -(a \cdot b) \; ; \quad -a \cdot -b := a \cdot b \; ; \quad 0 \cdot -b := -b \cdot 0 := 0. \end{array}$$

Außerdem sei $-b + a = a + -b$ und $-b \cdot a = a \cdot -b$ gesetzt, so dass K^+ und $\mathsf{K}^\times$ bereits gesichert sind. Definition (11) ist im Prinzip nur eine Abschrift von (8) in **9.1**. Genau so sind auch die Rechenoperationen in der Hardware eines Rechners geschaltet.

$\mathsf{N}^+, \mathsf{N}^\times$ gelten jetzt in ganz B, denn $-a + 0 = -(a - 0) = -a$ und $-a \cdot 1 = -(a \cdot 1) = -a$ für $a \in A$. Ferner ist $a + -a = a - a = 0$, so dass $-a$ tatsächlich ein additives Inverses

[3)] Allgemein eine Injektion $\eta : A \to B$, so dass das Bild ηA eine isomorphe Kopie von A ist.

von $a \in A_+$ ist. Auch ist a wegen K^+ gerade ein additives Inverses von $-a$, also hat jedes Element von B ein additives Inverses. Dessen eindeutige Bestimmtheit folgt jedoch erst durch den Nachweis von A^+. Dieser ist entscheidend. Er ist wegen zahlreicher Fallunterscheidungen aber ein nur etwas Geduld erforderndes Puzzlespiel. Hierbei treten einige der etwas subtileren Rechengesetze aus Abschnitt **2** noch einmal in Aktion.

Beweis von A^+. In der folgenden Tabelle sind die Fälle aufgeführt, die einer Betrachtung bedürfen, wobei die Variablen im Term $x + (y + z)$ der Reihe nach als der erste, zweite und dritte Operand bezeichnet werden. Die Zeichen + und − zeigen an, ob der betreffende Operand in A_+ bzw A_- liegt. Nicht aufgeführt sind die Fälle in denen einer der Operanden $= 0$ ist. Diese erledigen sich nämlich von selbst, denn wegen des bereits bewiesenen Gesetzes N^+ ist z.B. $x + (y + 0) = x + y = (x + y) + 0$. Bis zum Abschluss der Konstruktion bezeichnen a, b, c ausschließlich Elemente aus A_+.

Fall	1. Operand	2. Operand	3. Operand
1	+	+	−
2	+	−	+
3	−	+	+
4	+	−	−
5	−	+	−
6	−	−	+
7	−	−	−

Zunächst lassen sich die Fälle 2 und 3 auf Fall 1 reduzieren, und zwar wegen K^+. Denn

$$\begin{aligned}
a + (-b + c) &= a + (c + -b) && (\mathsf{K}^+ \text{ wurde schon gezeigt}) \\
&= (a + c) + -b && (\text{gemäß Fall 1}) \\
&= (c + a) + -b \\
&= c + (a + -b) && (\text{gemäß Fall 1}) \\
&= (a + -b) + c. \\
-a + (b + c) &= (b + c) + -a \\
&= (c + b) + -a \\
&= c + (b + -a) && (\text{gemäß Fall 1}) \\
&= c + (-a + b) &= (-a + b) + c.
\end{aligned}$$

Analog reduzieren sich die Fälle 5 und 6 auf Fall 4. Fall 7 folgt mit (4) in **9.1** aus

$$-a + (-b + -c) = -a + -(b + c) = -(a + b + c) = -(a + b) + -c = (-a + -b) + -c.$$

Fall 1a $a + b > b \geq c$.

$$\begin{aligned}
a + (b + -c) &= a + (b - c) && (\text{Definition}) \\
&= (a + b) - c && (\text{(5) in } \mathbf{2.4}) \\
&= (a + b) + -c && (\text{Definition})
\end{aligned}$$

Fall 1b $a + b \geq c > b$.

$$\begin{aligned}
a + (b + -c) &= a + -(c - b) && (\text{Definition}) \\
&= a - (c - b) && (\text{weil } a \geq c - b) \\
&= (a + b) - c && (\text{(9) in } \mathbf{2.4}) \\
&= (a + b) + -c && (\text{Definition})
\end{aligned}$$

Fall 1c $c > a + b > b$. $\quad a + (b + -c) = a + -(c - b)$ (Definition)
$= -((c - b) - a)$ (weil $c - b > a$)
$= -(c - (a + b))$ ((8) in **2.4**)
$= (a + b) + -c$ (Definition).

Weil stets $a + b \geq b$, sind damit offenbar alle Unterfälle des Falles 1 erschöpft.

Völlig analog behandelt man nun Fall 4 durch Unterscheidung der Unterfälle $a \geq b + c$, $b + c > a \geq b$ und $b + c \geq b > a$. Zum Beispiel ist in erstgenannten Unterfall

$$\begin{aligned} a + (-b + -c) = a + -(b + c) &= a - (b + c) && \text{(Definition)} \\ &= (a - b) - c && (\text{(8) in } \mathbf{2.4}) \\ &= (a + -b) + -c && (\text{weil } a \geq b,\ a - b \geq c). \end{aligned}$$

Beweis von $\mathsf{A}^\times$. Auch hier sind die Fälle 1 – 7 zu betrachten, wobei sich wie eben die Fälle 2,3 auf 1, und 5,6 auf 4 reduzieren. Dabei lässt sich von vornherein ausschließen, dass einer der Operanden $= 0$ ist, weil $\mathsf{A}^\times$ sich in diesem Falle von selbst erledigt. Im Falle 1 gilt nun die Behauptung nach Definition (11) wie folgt:

$$a \cdot (b \cdot -c) = a \cdot -(b \cdot c) = -(abc) = (ab) \cdot -c.$$

Fall 4 erledigt sich wegen $a(-b \cdot -c) = a(bc) = (ab)c = -(ab) \cdot -c = (a \cdot -b) \cdot -c$, und Fall 7 wegen $-a \cdot (-b \cdot -c) = -a \cdot (bc) = -(abc) = ab \cdot -c = (-a \cdot -b) \cdot -c$.

Nach diesen Ausführungen bereitet auch der Nachweis von D keinerlei Schwierigkeiten. Z.B. ist im Fall 1 – und ähnlich erledigt man die verbleibenden Fälle – nach Definition

$$(a + b) \cdot -c = -((a + b)c) = -(ac + bc) = -ac + -bc = a \cdot -c + b \cdot -c.$$

Damit ist $B = A \cup A_-$ ein Ring, der auch Bedingung (II) für Ringerweiterungen erfüllt; denn für jeden Ring B' mit $A \subseteq B' \subseteq B$ gilt $B' \supseteq A \cup A_-$, also $B' = B$. Damit ist die Konstruktion abgeschlossen.

Ein Ring mit geordneter Trägermenge, in welchem die beiden Monotoniegesetze gelten, heißt ein *geordneter Ring*. Nach Satz 9.1 kann die Ringerweiterung eines $\mathcal{E}$-Bereichs A immer auch als ein geordneter Ring verstanden werden. So ist die Ringerweiterung von $\mathbb{N}$ nichts anderes als der (geordnete) Ring der ganzen Zahlen. Ein geordneter Ring R ist stets *nullteilerfrei*, d.h. $ab \neq 0$ für alle $a, b \in R\backslash\{0\}$. Dies ist nach $\mathsf{M}^\times$ klar für $a, b > 0$, und wegen (6),(7) damit auch in den Fällen $a < 0 < b$ bzw. $a, b < 0$.

Es ist ferner klar, dass ein Bezeichnungssystem für die Elemente eines $\mathcal{E}$-Bereiches A auch ein solches für die Elemente der Ringerweiterung von A ergibt. Betrachten wir etwa den Bereich $\mathbb{E}$ [4)] und nennen dessen Ringerweiterung $\mathbb{F}$. Die Elemente aus $\mathbb{E}$ sind bereits (in der üblichen Weise) bezeichnet. Unter Verwendung des Minussymbols lassen sich dann auch alle Elemente von $\mathbb{F}$ bezeichnen: Für $a = z_0{,}z_1 \cdots z_n \in \mathbb{E}_+$ bezeichne $-z_0{,}z_1 \cdots z_n$ das Element $-a$ $(\in \mathbb{E}_-)$. Wegen $\mathbb{F} = \mathbb{E} \cup \mathbb{E}_-$ sind damit alle Elemente von $\mathbb{F}$ auch schon bezeichnet. 0 darf wahlweise auch mit -0 oder $+0$ bezeichnet werden.

[4)] Ganz analoge Betrachtungen lassen sich auch für die Zahlenbereiche $\mathbb{N}, \mathbb{Q}$ und $\mathbb{D}$, oder jeden anderen $\mathcal{E}$-Bereich durchführen, für dessen Elemente ein Bezeichnungssystem vorliegt.

Der Vorteil dieses Bezeichnungssystems ist, dass die Minusfunktion gerade durch den so genannten Vorzeichenwechsel repräsentiert wird. Damit wird die Ausführung der arithmetischen Operationen auf elektronischen Rechnern vollständig auf das Manipulieren mit nichtnegativen Operanden und der Minusfunktion μ reduziert. Hierin liegt nun auch der Grund, warum Computer das Rechnen unter Einschluss negativer Zahlen so problemlos beherrschen. Das Operieren mit der einstelligen Minusfunktion ist dabei besonders einfach, weil dies formal nur auf ein „Hinzufügen eines Minuszeichens“ und – falls ein solches schon vorhanden ist – auf dessen Weglassen hinausläuft, also nur ein einziges Bit Information kostet.

9.3 Der Körper der reellen Zahlen

Man vergegenwärtige sich noch einmal, dass sich die Konstruktion einer Ringerweiterung im besonderen auf die Bereiche $\mathbb{N}$, $\mathbb{E}$, $\mathbb{Q}$ und $\mathbb{D}$ anwenden lässt. In den letzten beiden Fällen liegt der Sonderfall eines $\mathcal{E}$-Bereichs A vor, in welchem *die Division ausführbar ist*, d.h. in A gilt die Aussage

$\mathsf{D}iv$: *Für alle* a, b *mit* $b \neq 0$ *gibt es ein* x *mit* $(*)$ $b \cdot x = a$.

In einem solchen Falle gilt $\mathsf{D}iv$ auch in der Ringerweiterung B von A. Denn für $b \in A_+$ gibt es ein b' mit $b \cdot b' = b' \cdot b = 1$, auch ein *multiplikatives Inverses*, genannt. Dasselbe gilt wegen (7) auch für $b \in B_-$. Dies hat offensichtlich die Gültigkeit von $\mathsf{D}iv$ in ganz B zur Folge, weil $(*)$ durch $x = b'a$ gelöst wird. Auch ist die Lösung wegen der aus $\mathsf{M}^\times$ folgenden Kürzungsregel eindeutig.

Falls in einem Ring die Division ausführbar ist, heißt dieser ein *Körper*. Ein geordneter Ring, in welchem die Division ausführbar ist, wird dann auch ein *geordneter Körper* genannt. Die Ringerweiterungen von $\mathbb{Q}$ und von $\mathbb{D}$ sind also Körper, und darüber hinaus sogar geordnete Körper. Durch Ringerweiterung erhalten wir also

- Aus $\mathbb{N}$ den Ring $\mathbb{Z}$ der ganzen (oder ganz-rationalen) Zahlen,
- Aus $\mathbb{E}$ den Ring $\mathbb{F}$ der endlichen (oder abbrechenden) Dezimalzahlen,
- Aus $\mathbb{Q}$ den (ebenfalls mit $\mathbb{Q}$ bezeichneten) Körper der rationalen Zahlen,
- Aus $\mathbb{D}$ den Körper $\mathbb{R}$ der reellen Zahlen.

Bemerkung. Der Ring $\mathbb{Z}$ ist natürlich ein besonders ausgezeichneter Ring. Auch die Körper $\mathbb{Q}$ und $\mathbb{R}$ haben zahlreiche Besonderheiten, abgesehen davon, dass sie zugleich auch geordnete Ringe sind. Für diese Bereiche gibt es auch eine Vielzahl vollständiger Charakterisierungen, d.h. eine jeweilige Sammlung von Eigenschaften, durch welche der betreffende Zahlenbereich bis auf Isomorphie eindeutig gekennzeichnet ist. So ist $(\mathbb{Z}, 0, 1, +, <)$ bis auf Isomorphie der einzige diskret geordnete Ring mit Einselement, und $(\mathbb{Z}, 0, +)$ die einzige diskret geordnete abelsche Gruppe, siehe dazu **10.4**. Dabei heißt eine beliebige geordnete Menge *diskret geordnet*, wenn jeder ihrer Schnitte ein Sprung im Sinne von **3.4**) ist. Auf vollständige Beschreibungen des Körpers $\mathbb{R}$ werden wir in **10.5** eingehen.

Die positiven Zahlen in $\mathbb{R}$ sind von der Form $z_0{,}z_1z_2\cdots$ und nach den Ausführungen in **9.2** sind die negativen von der Form $-z_0{,}z_1z_2\cdots$. Das Analogon zur n-ten kanonischen Näherung, nämlich $-z_0{,}z_1\cdots z_n$ für $a=-z_0{,}z_1z_2\cdots$, ist nunmehr $\geq a$. Denn die Ordnungsverhältnisse im Negativbereich sind denen des Positivbereichs in der durch Figur 6 illustrierten Weise entgegensetzt. Wesentlich ist vor allem, dass auch in $\mathbb{R}$ der Satz von der oberen Grenze gilt. Denn enthält $X\subseteq\mathbb{R}$ ein Element aus $\mathbb{D}$, so ist das Supremum von X in $\mathbb{D}$ zugleich Supremum von X in $\mathbb{R}$. Andernfalls ist $\{-x\mid x\in X\}\subseteq\mathbb{D}$. Hat diese Menge in $\mathbb{D}$ das Infimum s, ist anschaulich klar und leicht zu bestätigen, dass $-s$ das Supremum von X in $\mathbb{R}$ ist.

$\mathbb{R}$ ist also ein *lückenlos geordneter Körper*. Es lassen sich nun die wesentlichen Begriffe des Abschnitts **7** fast ohne Änderungen auch auf $\mathbb{R}$ übertragen. Wir erwähnen etwa den der schlichten Funktion in einem Intervall, das nunmehr teilweise oder auch ganz aus negativen Zahlen bestehen kann. Insbesondere überträgt sich der Zwischenwertsatz 7.2 ohne jede Änderung. Auch übertragen sich alle relevanten Gleichungen, wie etwa die binomische Formel, von $\mathcal{E}$-Bereichen auf deren Ringerweiterungen, und das gilt für die meisten Ungleichungen (etwa die Bernoullische), die durch den Wegfall hinderlicher Einschränkungen in der Regel einen erweiterten Geltungsbereich erhalten. Wir erwähnen, dass in $\mathbb{R}$, ja schon in jedem geordneten Ring, jedes Quadrat positiv ist. Denn für $a>0$ folgt dies aus $\mathsf{M}^\times$, und für $a<0$ aus $a^2=(-a)^2$ gemäß (7) in **9.1**.

Blickt man auf die Potenzrechnung in **7.3** zurück, drängt sich die Erklärung der Potenz auch für negative Werte des Exponenten fast von selbst auf. Man setzt $b^{-x}=\frac{1}{b^x}$ für $x>0$, insbesondere $b^{-1}=\frac{1}{b}$ $(b>0)$. Einzig diese Definition ist sinnvoll. Denn wenn P^+ gültig bleiben soll, ist $b^x\cdot b^{-x}=b^{x-x}=b^0=1$. Die Definitionsgleichung $b^{-x}=\frac{1}{b^x}$ ist also eine notwendige Folge hiervon. Die angegebene und nur diese Definition bewahrt die Potenzgesetze; genauer, letztere gelten nunmehr für beliebige reelle Exponenten. Um dies zu bestätigen, könnte man mit der erweiterten Definition alle Potenzgesetze der Reihe nach verifizieren. Man kann sich das Leben aber erleichtern. Denn da $\mathsf{P}_{\geq 1}$ und $\mathsf{P}_{\leq 1}$ offensichtlich ihre Gültigkeit bewahren, genügt es, lediglich das Basisgesetz P^+ zu bestätigen, weil sich die Art und Weise der in **7.3** ausführlich erläuterten Rückführung der Potenzregeln auf die Basisregeln nicht ändert. Zum Beweis von P^+ sind entsprechend der Definition (11) in **9.2** nur drei Fallunterscheidungen zu treffen. Zum Beispiel ist wie gefordert $a^{x+-y}=a^{x-y}=\frac{a^x}{a^y}=a^x\cdot a^{-y}$ im Falle $x\geq y$ (≥ 0). Völlig analog behandelt man die beiden verbleibenden Fälle der Definition (11). Entsprechend der erweiterten Definition von b^x erweitern sich die exponentialen und logarithmischen Funktionen und Satz 7.4 gilt mit der gleichen Formulierung auch für $\mathbb{R}$ anstelle von $\mathbb{D}$.

Die Definition von $\log_b x$ als derjenige Wert y mit $b^y=x$, wird beibehalten, und weil sich die Potenzgesetze nicht ändern, bleiben auch die darauf beruhenden logarithmischen Rechenregeln in **7.4** dieselben. Insbesondere ist $\log_b\frac{1}{x}=\log_b x^{-1}=-\log_b x$ gemäß L_2. Ebenso ist $\log_{b^{-1}}x=-\log_b x$. Dies folgt gemäß Definition des Logarithmus aus $b^{\log_b x}=x=(b^{-1})^{\log_{b^{-1}}x}=b^{-\log_{b^{-1}}x}$. Demnach unterscheiden sich $\log_b x$ und $\log_{b^{-1}}x$ nur durch das Vorzeichen, worauf in **7.4** schon hingewiesen wurde.

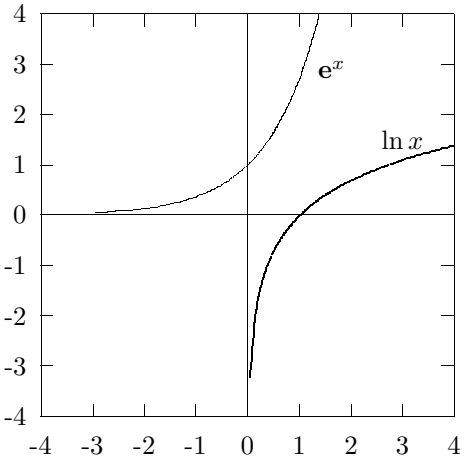


Fig. 7 Exponential- und Logarithmusfunktion

Die einzig verbleibende, im Reellen nicht aufhebbare Einschränkung ist, dass das Argument von $\log_b$ stets positiv sein muss, weil sich nämlich die Exponentialfunktionen $\exp_b$ im Reellen nur für positive b sinnvoll erklären lassen. Figur 7 zeigt die entsprechend der angegebenen Definition erweiterte Exponentialfunktion exp sowie ihre Umkehrfunktion ln im Intervall $[-4, +4]$. Der Graph von ln entsteht durch Spiegelung desjenigen von exp an der Achse $x = y$.

9.4 Der Körper der komplexen Zahlen

Von $\mathbb{R}$ ausgehend lässt sich nun leicht der Körper der komplexen Zahlen konstruieren. Dies ist eine vergleichsweise harmlose Angelegenheit, die sogar einfacher ist als die Konstruktion negativer Zahlen. Im Schoße der mittelalterlichen Algebra geboren, hatten komplexe Zahlen bereits zu Zeiten EULERs begonnen, eine Rolle in der Analysis zu spielen, die weit über ihren ursprünglichen Zweck, nämlich die unbeschränkte Auflösbarkeit von quadratischen und kubischen Gleichungen zu erreichen, hinausreichte. Einer der Gründe für ihre herausragende Bedeutung ist, dass durch ihre Benutzung ein Zusammenhang zwischen solchen auf den ersten Blick ganz verschiedenartigen Funktionen wie der Exponentialfunktion und den trigonometrischen Funktionen hergestellt wird, den eine im Reellen verbleibende Analysis gänzlich verbergen würde. Die Scheu, welche der Anfänger vor den komplexen Zahlen empfindet, entspringt allein mangelnder Gewohnheit; komplexe Zahlen kommen in Rechenaufgaben des Alltags eben selten vor, wohl aber bereits in der Ingenieurmathematik (etwa der Elektrotechnik). Prinzipiell wären komplexe Zahlen sogar überhaupt vermeidbar, und zwar in demselben Sinne wie negative Zahlen – alle diese Zahlen benützenden Sätze oder Beschreibungen physikalischer Vorgänge könnten nach einem gewissen Algorithmus so umformuliert werden, dass negative oder komplexe Zahlen darin nicht mehr erscheinen. Es widerspräche aber dem gesunden Menschenverstand, so zu verfahren, weil man sich dann freiwillig ergiebiger Quellen mathematischer Schöpferkraft berauben würde. So wenig wie rationale Zahlen rational und irrationale irrational im ursprünglichen Wortsinne sind, so wenig haben reelle Zahlen eine reale und imaginäre Zahlen eine nur eingebildete Existenz.

Ähnlich wie im Falle der Ringerweiterungen gibt es eine kanonische Konstruktion, die zu *jedem* Körper K, in welchem -1 kein Quadrat ist wie in $\mathbb{R}$ und in jedem geordneten Körper, in eindeutiger Weise einen Erweiterungskörper $K[\mathbf{i}] \supseteq K$ liefert, der diesen Mangel – falls er als solcher betrachtet wird – beseitigt. Diese Konstruktion ist wie

diejenige negativer Zahlen unabhängig von einer speziellen Darstellung der Elemente von K und wie dort bis auf Isomorphie eindeutig durch die Forderung bestimmt, dass der Erweiterungskörper keine überflüssigen, nicht dem Ziel der Konstruktion dienenden neuen Elemente enthält, Übung 3. $K[\mathbf{i}]$ heißt in der Algebra auch „K adjungiert $\mathbf{i}$“. In diesem Erweiterungskörper von K gilt $\mathbf{i}^2 = -1$, und natürlich auch $(-\mathbf{i})^2 = -1$. Die Konstruktion läuft z.B. für $K = \mathbb{Q}$ in genau der gleichen Weise ab [5)]. Auf eine lineare Anordnung der neuen Zahlen im herkömmlichen Sinne muss allerdings verzichtet werden; denn wäre $K[\mathbf{i}]$ geordnet, müsste -1 notwendig negativ, als Quadrat aber zugleich positiv sein. In historischer Rückschau ist wahrscheinlich dies der eigentliche Grund für die Bezeichnung „imaginär“ (für $\sqrt{-1}$); denn die Vorstellung über wirklichkeitsbezogene Zahlen erschien untrennbar von einer Anordnung. Aber dieser Verzicht ist nicht wirklich schwerwiegend. Er wird kompensiert durch gewisse topologische Eigenschaften des Erweiterungskörpers und durch gewonnene Vorteile bei weitem aufgewogen.

Definition. Für einen Körper K sei $K[\mathbf{i}]$ der Bereich $\{(a,b) \mid a,b \in K\}$ mit den folgendermaßen erklärten Operationen der Addition und Multiplikation:

$$(a,a') + (b,b') = (a+b, a'+b') \quad ; \quad (a,a') \cdot (b,b') = (ab - a'b', ab' + a'b).$$

Einfaches Nachrechnen zeigt, $K' := K[\mathbf{i}]$ ist ein (kommutativer) Ring mit Null $(0,0)$ und Eins $(1,0)$, und für $\mathbf{i} := (0,1)$ gilt $\mathbf{i}^2 = (-1,0)$. Die Gültigkeit fast aller Ringaxiome sieht man mit bloßem Auge; lediglich der Nachweis von $\mathsf{A}^\times$ erfordert Schreibstift und Papier. Ferner darf $a \in K$ mit $(a,0) \in K'$ identifiziert werden; denn $\eta : a \mapsto (a,0)$ ist injektiv und eine Einbettung von K in K', also $\eta(a+b) = \eta a + \eta b$ und $\eta(a \cdot b) = \eta a \cdot \eta b$, wie man unmittelbar sieht. Wir setzen also $(a,0) = a$, so dass speziell $\mathbf{i}^2 = -1$. Ferner ergibt sich $\mathbf{i}b = (0,1) \cdot (b,0) = (0,b)$. Deshalb hat jedes $(a,b) \in K'$ eine Darstellung

$$(a,b) = (a,0) + (0,b) = a + \mathbf{i}b.$$

In dieser Schreibweise ist $(a + \mathbf{i}a')(b + \mathbf{i}b') = ab - a'b' + \mathbf{i}(ab' + a'b)$. Sei nun -1 in K kein Quadrat. Dann (und auch nur dann) ist K' sogar ein Körper. Dies besagt

Satz 9.2 *$K[\mathbf{i}]$ ist ein Körper genau dann, wenn -1 in K kein Quadrat ist.*

Beweis. Sei -1 in K kein Quadrat, mit anderen Worten, sei die Gleichung $x^2 = -1$ in K unlösbar. Nach Übung 2 ist diese Eigenschaft gleichwertig mit $a^2 + b^2 \neq 0$, wenn immer $(a,b) \neq 0$ $(= (0,0))$, d.h. wenn a, b nicht beide 0 sind. Sei also $(a,b) \in K[\mathbf{i}] \backslash \{0\}$. Dann existieren $x, y \in K$ mit $(a,b) \cdot (x,y) = 1$ $(= (1,0))$, oder gleichwertig

$$(*) \qquad \begin{aligned} ax - by &= 1 \\ bx + ay &= 0, \end{aligned}$$

nämlich $x = \frac{a}{a^2+b^2}$, $y = \frac{-b}{a^2+b^2}$. Also ist K' ein Körper. Sei dies umgekehrt der Fall und $(a,b) \neq 0$ beliebig gewählt. Es gibt dann $x, y \in K$ mit $(a,b) \cdot (x,y) = 1$, also ist auch $(*)$ erfüllt. $(*)$ ergibt $(a^2 + b^2)x = a$, $(a^2 + b^2)y = -b$. Daraus folgt $a^2 + b^2 \neq 0$, sonst wäre doch $a = -b = 0$. Also ist nach Übung 2 die Gleichung $x^2 = -1$ in K unlösbar. ∎

[5)] Sie funktioniert für *jeden* Körper, auch wenn -1 dort bereits Quadrat ist; dann ergibt sich lediglich ein nicht besonders interessanter Ring, und es sollte z.B. besser $\mathbf{j}$ statt $\mathbf{i}$ geschrieben werden.

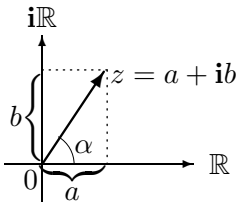


Fig. 8 Zahlenebene

Für $K = \mathbb{R}$ ist $K[\mathbf{i}]$ demnach ein Körper, der meist mit $\mathbb{C}$ bezeichnete *Körper der komplexen Zahlen*. Dessen Elemente sind von der Form $a + \mathbf{i}b$ mit $a, b \in \mathbb{R}$ und lassen sich als Punkte (oder Ortsvektoren) einer Ebene veranschaulichen, der *komplexen Zahlenebene* mit den Achsen $\mathbb{R}$ und $\mathbf{i}\mathbb{R} = \{\mathbf{i}r \mid r \in \mathbb{R}\}$, Figur 8. $|z| := \sqrt{a^2 + b^2}$ heißt *Betrag* oder *Norm* der komplexen Zahl $z = a + \mathbf{i}b$ [6)]. Der gerichtete Winkel α zwischen der Halbachse $\mathbb{R}_+$ und z ($\neq 0$) heißt das *Argument* von z und wird mit $\arg z$ bezeichnet. $\arg 0$ ist undefiniert. Weil $\arg z$ für $z \neq 0$ nur modulo $2\boldsymbol{\pi}$ eindeutig ist, normiert man $\arg z$ meist so, dass $-\boldsymbol{\pi} < \arg z \leq \boldsymbol{\pi}$, also $\arg 1 = 0$, $\arg -1 = \boldsymbol{\pi}$, $\arg \mathbf{i} = \frac{\boldsymbol{\pi}}{2}$ und $\arg -\mathbf{i} = -\frac{\boldsymbol{\pi}}{2}$.

Für $z = a + \mathbf{i}b \in \mathbb{C}$ sei $r := |z|$, $\alpha := \arg z$. Wie die Figur zeigt, hat z wegen $\cos\alpha = \frac{a}{r}$ und $\sin\alpha = \frac{b}{r}$ die Darstellung $z = r\cos\alpha + \mathbf{i}r\sin\alpha = r(\cos\alpha + \mathbf{i}\sin\alpha)$, genannt die *polare Darstellung* von z. Für $x = r(\cos\alpha + \mathbf{i}\sin\alpha)$ und $y = s(\cos\beta + \mathbf{i}\sin\beta)$ ist nach den ihrer geometrischen Herkunft wegen als bekannt vorausgesetzten Additionstheoremen für die trigonometrischen Funktionen

$$\begin{aligned} x \cdot y &= rs\big((\cos\alpha \cdot \cos\beta - \sin\alpha \cdot \sin\beta) + \mathbf{i}(\cos\alpha \cdot \sin\beta + \sin\alpha \cdot \cos\beta)\big) \\ &= rs\big(\cos(\alpha+\beta) + \mathbf{i}\sin(\alpha+\beta)\big) \end{aligned}$$

Damit erhält man eine einfache geometrische Veranschaulichung des Produktes komplexer Zahlen $\neq 0$: *Die Beträge werden multipliziert und die Argumente addiert.*

Das Rückgrat der komplexen Analysis ist die Erweiterungsfähigkeit von exp auf $\mathbb{C}$ so, dass P^+: $\exp(x+y) = \exp x \cdot \exp y$ für alle $x, y \in \mathbb{C}$ gilt, nämlich durch die Erklärung $\mathbf{e}^{a+\mathbf{i}b} := \mathbf{e}^a \cdot (\cos b + \mathbf{i}\sin b)$, Übung 5. Weil $z = |z| \cdot (\cos\alpha + \mathbf{i}\sin\alpha)$ mit $\alpha = \arg z$, darf $z = |z| \cdot \mathbf{e}^{\mathbf{i}\alpha}$ geschrieben werden. Wegen $\cos\boldsymbol{\pi} = -1$ und $\sin\boldsymbol{\pi} = 0$ erhält man

$$\mathbf{e}^{\mathbf{i}\boldsymbol{\pi}} = -1,$$

eine Gleichung, welche die vier berühmtesten Zahlen miteinander in Beziehung setzt.

Bemerkung. In der angegebenen Erweiterung der reellen Exponentialfunktion liegt keine Willkür; natürliche Zusatzbedingungen sichern, dass exp unter Erhalt des Basisgesetzes P^+ *eindeutig* auf ganz $\mathbb{C}$ fortgesetzt werden kann. Es genügt z.B. die komplexe Differenzierbarkeit im Nullpunkt, sowie die Wahl der ersten der beiden möglichen Alternativen $\mathbf{e}^{\mathbf{i}\frac{\boldsymbol{\pi}}{2}} = \mathbf{i}$ oder $\mathbf{e}^{\mathbf{i}\frac{\boldsymbol{\pi}}{2}} = -\mathbf{i}$. Denn grundsätzlich hat die durch Spiegelung einer komplexen Funktion f an der reellen Achse entstehende Funktion dieselben Eigenschaften wie f.

9.5 Interne Vorgänge in elektronischen Rechnern

Elektronenrechner führen bekanntlich nicht nur die elementar-arithmetischen Operationen aus, sondern es werden auch nichtrationale Funktionen, wie exp, sin und cos und deren Umkehrungen mit hoher Genauigkeit berechnet. Selbst Taschenrechner benötigen dazu nur Sekundenbruchteile. Wie ist das möglich?

[6)] Damit ist $|a|$ auch für $a \in \mathbb{R}$ wohldefiniert. Es ist $|a| = a$ für $a \geq 0$ und $|a| = -a$ für $a < 0$. So definiert man übrigens $|a|$ für $a \in R$ in jedem geordneten Ring R.

Obwohl die Hersteller von Chips und Computern verständlicherweise wenig über die innere Organisation ihrer Produkte mitteilen und sich im wesentlichen auf die Beschreibung von Leistungsdaten und Betriebssystemen beschränken, lässt sich einiges doch allein durch äußerliche Manipulation erraten.

Bei binärer Darstellung einer positiven reellen Zahl entspricht die Kommaverschiebung einer Multiplikation mit einer entsprechenden Potenz von 2 mit ganzzahligem Exponenten, Übung 8.3. Nun kann man sich die Kommaverschiebung immer so ausgeführt denken, dass vor dem Komma nur eine einzige von 0 verschiedene Ziffer, im Binärsystem also eine 1 zu stehen kommt. Mithin ist klar, dass es zu jeder positiven reellen Zahl a genau ein $k \in \mathbb{Z}$ und wohlbestimmte Binärziffern $z_1, z_2, z_3, \ldots$ gibt mit

$$(12) \quad a = 1{,}z_1z_2\cdots_2 \cdot 2^k = \begin{cases} 1{,}z_1z_2\cdots_2 \overset{k}{\rightarrow} & \text{für } a > 1 \text{ und damit } k \geq 0, \\ 1{,}z_1z_2\cdots_2 \overset{-k}{\leftarrow} & \text{für } 1 > a > 0, \text{ daher } k < 0. \end{cases}$$

Die negativen reellen Zahlen lassen sich analog darstellen, indem nur ein Minuszeichen hinzugefügt wird. $1{,}z_1z_2\cdots_2$ heiße die *Mantisse*, k der *Exponent* dieser schon in **4.1** erwähnten Gleitkommadarstellung. $a = 0$ habe die Mantisse und den Exponenten 0.

Es bezeichne $\mathbb{F}^2$ den Ring der abbrechenden Binärzahlen; dessen Elemente sind nach **8.2** zugleich darstellbar als abbrechende Oktalzahlen, Hexadezimalzahlen, usw. Für $a \in \mathbb{F}^2$ ist die Mantisse endlich. Z.B. ist $1001{,}101_2 = 1{,}001\,101_2 \overset{3}{\rightarrow} = 1{,}001\,101_2 \cdot 2^3$ (dies ist im Dezimalsystem die Zahl 9,625). Damit ist eine Zahl $a \in \mathbb{F}^2 \backslash \{0\}$ durch folgende Angaben vollständig beschrieben, und diese Beschreibung liegt ihrer Speicherung in den meisten Rechnern zugrunde; die Zahl 0 mit verschwindender Mantisse hat aus verschiedenen Gründen einen Sonderstatus und wird extra kodiert:

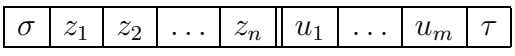
(a) Eine einzelne Ziffer σ zur Kodierung des Vorzeichens von a: Es sei $\sigma = 0$ für $a > 0$, und $\sigma = 1$ für $a < 0$ (Ziffer meint hier und in (b) bis (d) stets Binärziffer);

(b) Die Ziffernfolge $z_1, \ldots, z_n$ zur Darstellung der Mantisse $(1{,}z_1 \cdots z_n)_2$ von a;

(c) Die Ziffernfolge $u_1, \ldots, u_m$, wobei $(u_1 \cdots u_m)_2$ die binäre Darstellung von $|k|$ und k der (eventuell negative) Exponent von a ist;

(d) eine einzelne Ziffer τ zur Kodierung des Vorzeichens des Exponenten k oder der Kommaverschiebungsrichtung in (12): $\tau = 0$ für $k \geq 0$, $\tau = 1$ für $k < 0$.

Diese vier Informationen werden in der Regel in einer einzigen Speicherzelle eines Rechners untergebracht, etwa in der folgenden Weise; dabei repräsentiert jedes Kästchen einen *Bitplatz* (ein Platz, wo 0 oder 1 gespeichert werden kann).



Mantissenteil Exponententeil

Meistens sind nun die Zahlen n (der *Mantissenumfang*) und m (der *Exponentenumfang*) fest vorgegeben, weil eine Speicherzelle ihr Volumen nicht ohne weiteres ändern kann. Dem entspricht die Unmöglichkeit der Eingabe beliebig großer oder kleiner Zahlen.

Beispiel. Die Speicherung von $\boldsymbol{\pi} = 11{,}001\cdots_2 = 1{,}1001\cdots_2 \xrightarrow{1}$ (siehe **8.2**) lässt sich wie folgt beschreiben. Dabei beachte man, dass die 0 am Ende des Exponententeils die im vorliegenden Falle einfache Kommaverschiebung *nach rechts* kodiert:

0	100 100 100 001 111 110 110 101 010 001	000 000 001	0

Wir werden jetzt den Umfang einer Speicherzelle eines 10-stelligen Taschenrechners der Mittelklasse durch ein externes Experiment ungefähr bestimmen. Nach Umschaltung auf die Gleitkommadarstellung sehen wir, dass $\underbrace{9,9\cdots9}_{8}\cdot 10^{99}$ die größte im Display darstellbare Zahl ist. Diese ist zwar größer als die vermutete Anzahl der Atome im Weltall, aber der Abstand zur nächstkleineren darstellbaren Zahl beträgt 10^{92}, so dass mit Zahlen dieser Größenordnung nur noch sehr begrenzt operiert werden kann. Wir gehen von der berechtigten Annahme aus, dass die Gleitkommaanzeige der internen Speicherzellenaufteilung entspricht. Danach muss im Mantissenteil einer Speicherzelle die 8-stellige Zahl $k := 99\cdots9 = 10^8 - 1$ untergebracht werden können. Für die Länge der binären Darstellung dieser Zahl ergibt sich wegen $\operatorname{ld} 10^8 = 8\cdot\operatorname{ld}10 \approx 8\cdot 3{,}3 = 26{,}4$ nach (17) in **7.4** $\ell_2 k = \operatorname{Int} 26{,}4 + 1 = 27$. Also braucht man für den Mantissenteil 27 Bitplätze. Realistisch ist $n + 1 = 30$ für den Mantissenumfang einschließlich Speicherplatz des Vorzeichens, weil die interne Genauigkeit etwas größer ist als die der Anzeige. Wegen $\ell_2 99 = \operatorname{Int}\operatorname{ld}99 + 1 = \operatorname{Int}\frac{\lg 99}{\lg 2} + 1 = 7$ braucht man für die Speicherung des Exponenten 7 Bitplätze. Realistisch ist $m + 1 = 10$ einschließlich Vorzeichenkodierung. Das macht insgesamt 40 Bitplätze. Damit entspricht die schematische Darstellung einer Speicherzelle in etwa der des letzten Beispiels. Taschenrechner der Spitzenklasse verfügen über 8 Byte Speicherzellen für reelle Zahlen, die 64 Bitplätze enthalten.

Wie wird nun z.B. die Funktion ln berechnet? Hier bietet sich zunächst die Berechnung gemäß der unendlichen Reihe [7)]

$$(13)\quad \ln x = (x-1) - \tfrac{(x-1)^2}{2} + \tfrac{(x-1)^3}{3} - + \ldots \quad (0 < x < 2)$$

an, wobei nach einer gewissen Gliederanzahl die Rechnung abgebrochen wird. Leider ist dies nun noch nicht sehr vorteilhaft. Denn obwohl eine Potenzreihe $\sum_n a_n\cdot(x-c)^n$ in jedem abgeschlossenen Intervall ihres Konvergenzbereichs gleichmäßig konvergiert, ist die Konvergenz in unterschiedlichen Intervallteilen i.a. unterschiedlich schnell. Kurz, die Approximation durch eine gewählte Partialsumme ist ungleichmäßig. Deutlich erkennt man dies auch an der Reihenentwicklung der Exponentialfunktion (Formel (15) in **7.3**), bei der z.B. für kleinere Werte von $|x|$ wenig Glieder, bei größer werdenden Werten aber mehr Glieder benötigt werden, um die vorgeschriebene Genauigkeit zu erreichen. Dies alles könnte bei der Schaltung der Hardware durch interne Tests zwar berücksichtigt werden, aber eben dies würde die Rechenzeit beträchtlich in die Höhe treiben.

Dem erwähnten Effekt ungleichmäßiger Approximation durch Potenzreihen für $\exp, \ln$, und analog bei $\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - + \ldots$ und $\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - + \ldots$ überlagert

[7)] Für Reihenentwicklung von Funktionen und einige andere nachfolgend genannte Begriffe sei auf Lehrbücher der Analysis verwiesen, etwa [16], [28], [39].

sich noch eine zweite, computerbedingte Schwierigkeit. Für größere Werte von $|x|$ haben die Glieder der zuletzt angegebenen Reihen große Absolutwerte, so dass nach erfolgter Addition und Subtraktion der Reihenglieder nur noch wenige, oder gar keine Stellen des Ergebnisses mehr zuverlässig sind. Aus den genannten Gründen sind Potenzreihen für die computerinterne Berechnung der elementaren Funktionen unbrauchbar.

Nun bietet sich jedoch eine andere Berechnungsweise an. Zunächst ist klar, dass statt ln irgendeine andere Logarithmenfunktion zu einer passenden Basis berechnet werden kann. Tatsächlich wird häufig zuerst die Funktion ld (oder ihre Umkehrfunktion $\exp_2$) berechnet. Es kann dann wegen der Umrechnungsformel

$$\ln x = \ln 2 \cdot \operatorname{ld} x = 0{,}101100010111\cdots_2 \cdot \operatorname{ld} x$$

auch ln berechnet werden. Sei $a = (1{,}z_1\cdots z_n)_2 \cdot 2^k$ (alle z_i Binärziffern; $k \in \mathbb{Z}$). Dann ist $\operatorname{ld} a = \operatorname{ld}(1{,}z_1\cdots z_n)_2 + \operatorname{ld} 2^k = \operatorname{ld}(1{,}z_1\cdots z_n)_2 + k$. Es braucht also nur $\operatorname{ld}(1{,}z_1\cdots z_n)_2$ berechnet zu werden, denn k steht ja bereits im Exponententeil der Speicherzelle von a. Nun liegt die Mantisse $(1{,}z_1\cdots z_n)_2$ im Intervall $[1, 2]$. Damit läuft die Aufgabe auf die Berechnung von $\operatorname{ld} x$ in diesem Intervall hinaus.

Es ist nun eine wichtige Tatsache, dass jede in einem Intervall definierte Funktion, die sich dort „anständig“ verhält (z.B. dort mehrfach differenzierbar ist), in diesem Intervall durch ein gewisses Polynom gleichmäßig approximiert werden kann, z.B. durch ein so genanntes Tschebyscheff-Polynom. Nur dieses ist in Abhängigkeit von der gewünschten Genauigkeit zu bestimmen, und sodann ein Berechnungsverfahren intern zu verdrahten. Eine analoge Bemerkung gilt hinsichtlich der Exponentialfunktion und der trigonometrischen Funktionen. In all diesen Fällen rechnet der Elektronenrechner mit in der Hardware (genauer, im Zentralchip, dem „Gehirn“ des Rechners) implementierten gleichmäßigen polynomialen Approximationen in einem geeigneten Intervall. Das geht blitzschnell, weil bei der Wertberechnung der Approximationspolynome keine Testschleifen durchlaufen werden müssen [8)]. Angesichts der fortgeschrittenen Speichertechnologie in modernen Hochleistungs-Chips ist es heutzutage auch kein Problem mehr, genügend viele Stützstellen des Reduktionsintervalls einiger elementarer Funktionen direkt in der Hardware zu speichern. Die Berechnung dieser Funktionen für eingegebene Werte der Argumente erfolgt dann mit hoher Präzision durch Interpolation mittels geeigneter polynomialer Splines.

Nicht nur die Planung von internen Rechenprogrammen für Computer, sondern auch die Programmierung eigener Rechenverfahren durch den Nutzer erfordert große Erfahrung, will man die Gefahr rundungsbedingter Fehlergebnisse in kontrollierbaren Grenzen halten. Als ein instruktives Beispiel betrachten wir folgendes Berechnungsverfahren für $\ln x$, das sich für Taschenrechner ohne Logarithmus aber mit Wurzelfunktion geradezu anbietet. Man zeigt unschwer, dass die Folge $\langle n(\sqrt[n]{x} - 1)\rangle$ für $x > 0$ (fallend) gegen $\ln x$ konvergiert, Übung 7. (Dies wird gelegentlich zur Einführung der Funktion ln genutzt.)

[8)] Für Taschenrechner mit durchweg garantierter 10-stelliger Genauigkeit sind auch polynomiale Approximationen noch zu langsam und man verwendet die in **7.4** schon erwähnten alten und einfachen Algorithmen von Napier, Briggs und anderen, siehe z.B. [24].

Setzt man daher $u_0 = 1$, $u_{n+1} = 2u_n$ und $v_0 = x$, $v_{n+1} = \sqrt{v_n}$, so konvergiert auch das Produkt $u_n(v_n - 1) = 2^n(\sqrt[2^n]{x} - 1)$ gegen $\ln x$, sogar relativ schnell. Mit anderen Worten, $\ln x$ lässt sich mit den Grundrechenarten und mit Wurzelziehen auf scheinbar einfache Art iterativ berechnen. Nun nimmt aber der erste Faktor u_n für große n Werte von exorbitanter Größenordnung an, der zweite Faktor hingegen wird im Absolutbetrag sehr klein (für $x > 1$ konvergiert $\langle v_n \rangle$ fallend gegen 1 und für große n „löscht die Subtraktion $v_n - 1$ die Stellen aus“). Daher scheitert das angegebene Iterationsverfahren zur Berechnung von $\ln x$ beim Rechnen mit fester Stellenzahl.

Es sei noch erwähnt, dass die oben skizzierte Berechnungsmöglichkeit der elementaren Funktionen durch Reduktion auf ein geeignetes Intervall – die so genannte *Bereichsreduktion* – ebenfalls ihre Tücken hat. So weigern sich manche Computer, sin und cos für große Argumente auszurechnen. Das ist eine sinnvolle Vorsichtsmaßnahme. Ungehemmtes Rechnen kann nämlich leicht falsche Gegebenheiten vortäuschen. Denn nehmen wir an, auf einem n-stelligen Rechner soll in einer Zwischenrechnung $\sin(10^{n+1}a)$ ermittelt werden. Im allgemeinen entstand a durch Rundung, d.h. das Argument liegt günstigstenfalls im Intervall mit den Grenzen $10^{n+1}(a - 0{,}5 \cdot \varepsilon_n)$ und $10^{n+1}(a + 0{,}5 \cdot \varepsilon_n)$. Die Breite dieses Intervalls ist $10^{n+1}\varepsilon_n = 10$, also größer als die des üblichen Reduktionsintervalls $[0, \frac{\pi}{2}]$ der sin-Funktion, in welchem diese bereits jeden ihrer Werte annimmt. Das Resultat kann sozusagen beliebig falsch sein.

9.6 Übungen

1. Sei n ungerade. Man zeige: in einem geordneten Ring R (**9.2**) hat die Gleichung $(*)$: $x^n = c$ mit $c \in R$ höchstens eine, und für $R = \mathbb{R}$ genau eine Lösung.

2. Man beweise, für einen Körper K sind äquivalent: (i) -1 ist Quadrat, (ii) mit a ist auch $-a$ Quadrat, (iii) es gibt ein Paar $(a, b) \neq (0, 0)$ mit $a^2 + b^2 = 0$.

3. Sei K Körper, in welchem -1 kein Quadrat ist, und $K' \supseteq K$ Erweiterungskörper mit $\iota^2 = -1$ für ein gewisses $\iota \in K'$. Man zeige, es gibt eine Einbettung η von $K[\mathbf{i}]$ in K' mit $\eta\mathbf{i} = \iota$. Kurz, K' enthält einen zu $K[\mathbf{i}]$ isomorphen Unterkörper.

4. Man zeige mit Übung 3 ohne zu rechnen: $\overline{x + y} = \overline{x} + \overline{y}$ und $\overline{x \cdot y} = \overline{x} \cdot \overline{y}$, wobei für $z = a + \mathbf{i}b \in K[\mathbf{i}]$ das *konjugierte* Element $\overline{z}$ als $a - \mathbf{i}b$ erklärt ist.

5. Man beweise P^+: $\mathbf{e}^{x+y} = \mathbf{e}^x \cdot \mathbf{e}^y$ für alle $x, y \in \mathbb{C}$. Das ergibt speziell $\mathbf{e}^{\mathbf{i}\alpha n} = (\mathbf{e}^{\mathbf{i}\alpha})^n$ für alle $\alpha \in \mathbb{R}$, also $\cos n\alpha + \mathbf{i}\sin n\alpha = (\cos\alpha + \mathbf{i}\sin\alpha)^n$. Danach ist zum Beispiel
$$\cos 3\alpha = \cos^3\alpha - 3\cos\alpha \cdot \sin^2\alpha \quad ; \quad \sin 3\alpha = 3\cos^2\alpha \cdot \sin\alpha - \sin^3\alpha\,.$$

6. Man beweise: die Gleichung $(*)$: $x^n = c$ $(n \geq 1)$ hat für jedes $c \in \mathbb{C}\backslash\{0\}$, insbesondere für $c \in \mathbb{R}\backslash\{0\}$, in $\mathbb{C}$ genau n verschiedene Lösungen.

7. Man zeige: $\langle n(x^{\frac{1}{n}} - 1) \rangle$ konvergiert für jede reelle Zahl $x > 0$ fallend gegen $\ln x$.