Zahlen, Operationen, Funktionen

H.R. Schneebeli

31. Mai 2021

Zusammenfassung

Wenn Analysis von den Eigenschaften der reellen Zahlen und reellwertiger Funktionen handelt, so verfolgt Numerik das Ziel, quantitative Ergebnisse zu liefern, deren Existenz die Analysis beweisen kann. Genauer: Numerik kann durch konkretes Rechnen wichtige Grundbegriffe der Analysis bloss imitieren. Approximationen sind zentral und wesentlich. Beide Bereiche verwenden gleiche oder ähnliche Namen für Begriffe, die eigentlich unterschieden werden müssen. In Beispielen werden hier typische Unterschiede zwischen den Zahlbegriffen, den arithmetischen Operationen und dem Funktionsbegriff herausgearbeitet. Diskretisierung, konkretes Rechnen, Rundung und Näherungsverfahren sind das Geschäft der Numerik. Jede konkrete numerische Operation kann bloss endlich viele Bits verändern – und einige davon eher zufällig. Der Bedarf an Energie, Speicherplatz und Zeit steht jeder exakten Arithmetik mit typischen reellen Zahlen im Wege.

Ein intuitives Verständnis für Zusammenhänge und Unterschiede zwischen Analysis und Numerik gehört zu den grundlegenden Erfahrungen, welche Elementarmathematik als Beitrag zur Allgemeinbildung leisten soll. Deshalb werden einfache Muster numerischer Probleme, Arbeitsweisen, Erfolge und Grenzen der Methoden vorgeführt.

Ziele Es wird gezeigt, wie Numerik, also konkretes Rechnen mit Fliesskommazahlen, abstrakte Existenzaussagen aus der Analysis in konkreten Anwendungen zu ergänzen vermag, wenn das Ziel kontrollierbare, quantitative Näherungen sind.

 $Klassische \ Analysis \ handelt von theoretischen \ Begriffen \ wie reelle Zahlen <math>\mathbb{R}$ oder komplexe Zahlen \mathbb{C} , Funktionen, Stetigkeit, lokale Veränderung, Ableitung als Funktion, verschiedene Integralbegriffe, Differentialgleichungen

Oft befasst man sich in der Analysis mit Fragen zur Existenz oder Eindeutigkeit von Lösungen gewisser Probleme, auch wenn deren Lösungen nicht explizit – etwa durch Lösungsformeln oder Angabe einer konkreten Zahl – beschreibbar sind. Formal exakte Lösungen sind im Bereich der Analysis seltene Ausnahmen. Dennoch gibt es analytisch beweisbare Sätze, welche die Existenz von Lösungen abstrakt garantieren. Das ist eine notwendige Voraussetzung vor jedem Einsatz von Numerik.

Numerik befasst sich mit quantitativen Näherungen für Lösungen, welche durch konkretes Rechnen mit Maschinenzahlen bei gegebenen Toleranzen angenähert werden und in vielen Anwendungen unverzichtbar sind.

1 Allerlei über Zahlen

Im 19. Jahrhundert wurde erkannt, dass es eine mathematische Konstruktion der reellen Zahlen braucht als Grundlage für logisch beweisbare Aussagen in der Analysis, vgl. [3].

Die rationalen Zahlen Q genügen den Bedürfnissen der Analysis nicht. Aber sie bilden ein Gerüst, mit dessen Hilfe sich im Prinzip jede reelle Zahl beliebig gut annähern lässt.

Wer numerisch rechnet, muss sich damit abfinden, dass in endlicher Zeit nur endlich viele Ziffern erzeugbar sind. Auch auf dem schnellsten Rechner lassen sich in endlicher Zeit nur endlich viele Bits erzeugen und jede Bitoperation kostet Energie und Zeit. Die grosse Mehrzahl der Vertreter aus dem Kontinuum der reellen Zahlen sind so unerreichbar. Diskretisierung und Näherungen sind beim praktischen Rechnen unumgänglich. Die Kunst besteht meist darin, Näherungen an die Kontinuumsmathematik im Rahmen des technisch Machbaren hinreichend effizient zu erzielen und die unumgänglichen Restfehler unter Kontrolle zu halten.

1.1 Im Paradies der Analysis, reelle Zahlen, formale Begriffe und Operationen

Die Zahlmengen $\mathbb Q$ und $\mathbb R$ mit ihren arithmetischen Operationen sind Beispiele für Zahlkörper. Ferner sind beide linear angeordnet. Es gibt eine Ordnungsbeziehung \leq so, dass für je zwei verschiedene Elemente a,b entweder a < b oder b < a gilt. Die Zahlen von $\mathbb R$, die in $\mathbb Q$ fehlen, heissen irrational. Das systematische Auffüllen der Lücken zwischen den rationalen Zahlen heisst Vervollständigung. Alle grundlegenden Begriffe der Analysis wie Grenzwert, Stetigkeit, Ableitung und Integral lassen sich nur dank der Vervollständigung zweckmässig definieren und handhaben. Für die klassischen Sätze der Analysis reichen die Eigenschaften von $\mathbb Q$ nicht aus. Die Vollständigkeit von $\mathbb R$ ist wesentlich. Sie wird beim Beweis wichtiger Existenzaussagen in der klassischen Analysis typischerweise benötigt. Für die Ziele der Analysis als Theorie schafft $\mathbb R$ somit ideale Verhältnisse.

Allerdings gibt es einen Preis. Die Menge $\mathbb R$ hat eine der Intuition schwer zugängliche Eigenschaft. Sie enthält wesentlich mehr Zahlen als die Mengen $\mathbb N$ oder $\mathbb Q$. Die rationalen Zahlen lassen sich mit $\mathbb N$ abzählen, die reellen Zahlen aber nicht. Man sagt daher, die Menge $\mathbb R$ ist *überabzählbar*. Diese mengentheoretische Eigenschaft bewirkt, dass wir eine typische reelle Zahl auf keinerlei Art in einem endlichen Kontext erfassen können, weder mit Worten noch mit Ziffern. In einigen Fällen haben sich besondere Symbole wie π als Bezeichnung eingebürgert, aber auch sie lösen das Dilemma nicht: Die allermeisten reellen Zahlen müssen aus prinzipiellen Gründen namenlos bleiben.

Wir können uns denken, dass ein Orakel oder der Zufall sie uns zuspielen kann, wenn wir sagen 'sei x eine reelle Zahl'. Eine namenlose reelle Zahl muss mit einem generischen Symbol, beispielsweise x, bezeichnet werden. Wenn nun das Symbol uns befähigt, in endlich vielen Worten ein wohlbestimmtes Individuum auszuzeichnen, so erweist es sich durch diese Tatsache im Nachhinein als nicht mehr namenlos. Dieser Fall tritt ein bei der Definition der Zahl π als Flächeninhalt des Einheitskreises. Aber damit ist π als Zahlwert noch nicht bestimmt. Wer sagt, $\pi = 22/7$, irrt sich. Wer behauptet, $\pi \approx 22/7$, macht eine für grobe Anwendungen nützliche numerische Näherung.

Jede reelle Zahl x lässt sich als Grenzwert einer Folge von Dezimalzahlen (mit je endlich vielen Dezimalziffern) darstellen. Damit wird auch klar, welche tragende Funktion dem Gerüst $\mathbb{Q} \subset \mathbb{R}$ in allen konkreten Rechnungen zukommt.

Oft handelt die Analysis von Beziehungen zwischen Zahlen eher als von den Zahlen selbst. Die Zahl π lässt sich wie alle irrationalen Zahlen als Grenzwert einer unendlichen Folge von rationalen Zahlen beschreiben. Oft dienen die reellen Zahlen dazu, Relationen zwischen Grössen zu beschreiben, beispielsweise in der Kreisgeometrie $U=\pi d$. Dies bedeutet aber nur: Wenn wir den Umfang U und den Durchmesser d eines Kreises ganz exakt kennen würden, so liesse sich $\pi=U/d$ durch eine einzige Division berechnen. Die Verwendung des Konjunktiv als Möglichkeit ist typisch, wenn es um konkrete Interpretationen der Analysis geht. Niemand kann gleichzeitig die absolut exakten Zahlwerte von U und d im Falle eines Kreises angeben.

Analysis lehrt uns, dass π eine transzendente Zahl ist. Der absolut exakte numerische Wert von π ist also niemals in endlich vielen Worten oder Bits erfassbar.

Da \mathbb{R} linear angeordnet ist, liegt es nahe, eine Gerade zu verwenden, um \mathbb{R} bildlich darzustellen. Dabei geht man davon aus, dass es eine umkehrbar eindeutige Zuordnung zwischen allen Zahlen in \mathbb{R} und allen Punkten einer Geraden gibt. Diese Beziehung kann durch ein Koordinatensystem hergestellt werden. Wer der Sache auf den Grund geht, bemerkt, dass die Geometrie uns die reellen Zahlen nicht automatisch zur Verfügung stellt. Es gibt eine Zahlgerade, die nur rationale Koordinaten kennt oder eine solche, die nur aus Punkten besteht, die sich mit Zirkel und Lineal aus den rationalen Zahlen konstruieren lassen. Die Darstellung von \mathbb{R} als Zahlgerade ist eine bequeme Metapher. Anschauliche, naive Geometrie klärt die Grundlagen der Analysis jedoch nicht hinreichend. Hilbert durchleuchtet in [1] Zusammenhänge zwischen Körpertheorie und Geometrie.

Die mathematische Analysis geht davon aus, dass Rechenoperationen auch im Bereich der reellen Zahlen mit unbegrenzter Genauigkeit, absoluter Zuverlässigkeit und ohne Zeitaufwand nach den Grundgesetzen der arithmetischen Operationen in \mathbb{R} möglich sei. In der numerischen Analysis sind realistischere Annahmen die Regel: Ob wir im Kopf rechnen oder mit einem Computer, Rechnen setzt biologische, chemische oder physikalische Prozesse voraus, die mit endlich viel Materie und Energie je eine gewisse endliche Zeit beanspruchen und insgesamt in endlicher Zeit beendet werden müssen. Praktisches Rechnen hat in diesem Sinne eine experimentelle Komponente und schliesst die Möglichkeit nicht aus, dass das Experiment misslingen kann. Die Numerik lässt sich nur im Kontext der verfügbaren Technologien richtig einschätzen.

Eine analoge Situation gibt es in der Geometrie. Es gibt eine abstrakten Theorie zur Möglichkeit idealer Konstruktionen mit idealen Werkzeugen Zirkel und Lineal, beziehungsweise idealen Kreisen und Geraden. Daneben stehen konkret ausführbare Konstruktionen, die mit den materiellen Werkzeugen Zirkel und Lineal und deren begrenzten Möglichkeiten konkret ausführbar sind. Mehr dazu folgt in den Aufgaben.

1.2 In der Wirklichkeit der Numerik, Maschinenzahlen

Die Numerik beackert im Vergleich zur Analysis einen kargen Boden. Die idealen Verhältnisse, die in der Analysis angenommen werden, kennt die Rechenpraxis nicht. Man könnte vom Standpunkt des konkreten Rechnens aus von einer virtuellen Realität der reellen Zahlen sprechen.

Alle, die praktisch rechnen müssen, sind mit einer anderen Wirklichkeit konfrontiert. In der Regel sind schon die Daten, die in die Rechnungen eingehen, mit Unsicherheiten behaftet. Die Frage, ob eine gemessene Zahl rational oder irrational sei, macht keinen Sinn. Alle Messergebnisse lassen sich hinreichend gut durch Dezimalzahlen mit endlich vielen Ziffern approximieren. Tatsache ist, dass Menschen oder Maschinen in endlicher Zeit nur endlich viele Zahlen, Ziffern oder Bits lesen, schreiben oder bearbeiten können. Welches System dabei benutzt wird, ist eher eine Frage der technischen Umsetzung. Wir bleiben hier beim Dezimalsystem, das vielen Menschen näher liegt als das Dualsystem aus der Computertechnik.

Im Hintergrund steht jedoch auch bei den Mathematikanwendern sehr oft die Analysis. Sie bietet einen idealisierten, theoretischen Rahmen, in welchem die numerischen Berechnungen konzipiert und die Ergebnisse interpretiert oder gar auf Plausibilität geprüft werden.

Beim konkreten Rechnen mit dem Computer wird in einer festen, endlichen Auswahl von rationalen Zahlen operiert. Dies sind die Maschinenzahlen. Oft gibt es gar zwei Arten, solche, die diskrete Zahlmengen, also $\mathbb N$ oder $\mathbb Z$ imitieren und solche, die sich möglichst wie ein

Kontinuum verhalten sollten und als Ersatz für $\mathbb R$ oder $\mathbb C$ dienen müssen. Wir behandeln nur zwei Fälle, und wir beschränken uns auf eine vereinfachende, untechnische Beschreibung und ausgewählte Eigenschaften, die für die jeweiligen Konstruktionen typisch sind. Insbesondere werden Beispiele betrachtet, in denen sich Maschinenzahlen gerade *nicht so verhalten* wie ihre Vorbilder aus der theoretischen Mathematik. Maschinenzahlen sind untrennbar mit der Binärdarstellung der Zahlen und der arithmetischen Operationen verknüpft, die im Prozessor verwendet werden. Ferner gehört die Übersetzung zwischen der Binärdarstellung und der Dezimaldarstellung zum Konzept der Maschinenzahlen.

Technische Definitionen mit allen Einzelheiten finden sich in den einschlägigen IEEE-Normen. (Vgl. [4])

Der Name Maschinenzahlen weckt die Erwartung, dass die Hardware die Arithmetik vollständig bestimmt, weil ein Arithmetikprozessor diese Zahlen bearbeitet. Diese Sicht der Dinge ist heute nicht mehr immer zutreffend. Bei Rechnern, deren Prozessoren mehrfache Kerne besitzen, braucht es auch die entsprechende Software, um die volle Rechenpotenz der Hardware gezielt zu nutzen. Anderseits können Programme wie Computeralgebrasysteme arithmetische Operationen mit mehrfacher Genauigkeit ausführen, falls dies erforderlich ist. Das geht allerdings auf Kosten der Rechengeschwindigkeit. Letztlich wird die Speichergrösse oder die Laufzeit der Programme die Genauigkeit von Berechnungen auch mit ganzen Zahlen begrenzen. Ferner ist zu beachten, dass der menschliche Leser nicht in der Lage ist, etwa eine formal exakte Darstellung eines Bruches zu interpretieren, wenn Zähler und Nenner auch nur einige hundert Ziffern enthalten. Der Ausweg führt über die Vereinfachungen und die Reduktion auf das menschliche Mass. Das wird in der Form von Dezimalzahlen mit begrenzter Stellenanzahl in Mantisse und Exponent erreicht [7].

Zur Verdeutlichung sei an die hochgenauen Approximationen an die Zahl π erinnert. Es sind Billionen von Stellen dieser Zahl berechnet worden. Es macht aber keinen Sinn, Jahre eines Menschenlebens mit der naiven Lektüre dieser Ziffernfolge zu vergeuden. Die Analyse des Ergebnisses selbst erfordert wieder einen Computereinsatz, zum Beispiel, um statistische Eigenschaften der Ziffernhäufigkeit zu untersuchen. Aber auch hier stossen wir an Grenzen, weil nur ein verschwindend kleines Anfangsstück konkret berechnet wurde, ein Nichts im Vergleich zu den fehlenden Stellen, die nie berechnet werden können.

1.2.1 Imitation der ganzen Zahlen

Manche Programmierumgebungen zeichnen ganzzahlige Maschinenzahlen aus. Ihre technische Spezifikation nimmt Rücksicht auf die intern verwendeten Binärzahlen. Wir nehmen hier vereinfachend die Sicht der Nutzer ein, die mit ganzen Zahlen im Dezimalsystem rechnen. Die ganzen Rechnerzahlen $\mathbb{M}_{\mathbb{Z}}$ umfassen alle ganzen Zahlen k aus einem Intervall $-I_{\max} \leq k \leq I_{\max}$ von \mathbb{Z} . Auf $\mathbb{M}_{\mathbb{Z}}$ wird die Anordnung $k \leq m$ von \mathbb{Z} unverändert vererbt. Insbesondere ist in $\mathbb{M}_{\mathbb{Z}}$ immer exakt entscheidbar, ob die Aussage $a \leq b$ für die beiden ganzen Maschinenzahlen a, b gilt. Die arithmetischen Operationen Summe, Differenz, Produkt, Division mit Rest lassen sich nur beschränkt von \mathbb{Z} auf $\mathbb{M}_{\mathbb{Z}}$ übertragen. Notwendig und hinreichend für die erfolgreiche Übertragung ist, dass das in \mathbb{Z} berechnete Ergebnis der Operation zum Bereich der ganzen Maschinenzahlen $\mathbb{M}_{\mathbb{Z}}$ gehört. Trifft diese Bedingung zu, so gelten beim Rechnen in $\mathbb{M}_{\mathbb{Z}}$ genau dieselben Regeln wie in \mathbb{Z} . Andernfalls ist eine Fehlermeldung overflow zu erwarten.

1.2.2 Imitation der reellen Zahlen, Fliesskommazahlen $M_{\mathbb{R}}$

Weit wichtiger als $\mathbb{M}_{\mathbb{Z}}$ sind für die Numerik die Maschinenzahlen $\mathbb{M}_{\mathbb{R}} \subset \mathbb{Q}$. Wiederum stellen wir uns auf den vereinfachenden Standpunkt der Nutzer, welche die Elemente von $\mathbb{M}_{\mathbb{R}}$ als De-

zimalzahlen erfahren. Genauer: jede Maschinenzahl $x \neq 0$ aus $\mathbb{M}_{\mathbb{R}}$ lässt sich mit Zehnerziffern d_i darstellen in der Form

$$x = \sigma * d_1 d_2 \cdots d_n * 10^k$$
 mit ganzen Maschinenzahlen k und n aus $\mathbb{M}_{\mathbb{Z}}$

Die Zahl k heisst Zehnerexponent von x. Der Faktor $d_1 \dots d_n$ ist die Zehnermantisse oder Ziffernfolge von x. Ihre maximale Länge n ist für alle Maschinenzahlen begrenzt. Das Vorzeichen σ hat einen der beiden Werte 1 oder -1. Für x=0 ist kein Vorzeichen definiert. Ein Grund ist folgender: Das Vorzeichen definiert einen Homomorphismus von der multiplikativen Gruppe \mathbb{R}^* [reelle Zahlen ohne Null] in die multiplikative zyklische Gruppe mit zwei Elementen realisiert als $C_2 = \{1, -1\}$. Deshalb wird für 0 gar kein Vorzeichen benötigt.

Es gibt in $\mathbb{M}_{\mathbb{R}}$ auch zwei *Vorzeichen* $\sigma \in \{-1,1\}$. Aber diese Vorzeichen verlieren die typische Eigenschaft eines Homomorphismus in der Rechnerarithmetik, weil es in $\mathbb{M}_{\mathbb{R}}$ *Nullteiler* gibt. Das sind Zahlen $p,q \neq 0$ mit $p \cdot q \doteq 0$, wobei $a \doteq b$ bedeutet, dass zwei Maschinenzahlen a,b wegen Rundungseffekten *ununterscheidbar* geworden sind.

Wegen der fest vorgegebenen Mantissenlänge enthält jede der in $\mathbb{M}_{\mathbb{R}}$ verfügbaren Dekaden die gleiche Anzahl Maschinenzahlen. In $\mathbb{M}_{\mathbb{R}}$ ist jede Dekade eine um den Faktor 10 vergrösserte Kopie der vorangehenden Dekade. Daher bleibt die *relative Auflösung* in allen Dekaden gleich.

Der Übergang von \mathbb{R} auf $\mathbb{M}_{\mathbb{R}}$ erfolgt in zwei Schritten. Jeder der beiden Schritte zerstört einen Teil der Eigenschaften, die \mathbb{R} als Zahlkörper auszeichnen.

- 1. Begrenzung auf ein Intervall $I = \{x \mid -X_{\max} \le x \le X_{\max}\}$
- 2. Rundung jeder Zahl x aus I auf die nächstgelegene Maschinenzahl rd(x) aus $\mathbb{M}_{\mathbb{R}}$. Zahlen exakt in der Mitte zwischen zwei benachbarten Maschinenzahlen müssen nach einer zusätzlichen Konvention behandelt werden.

In $\mathbb{M}_{\mathbb{R}}$ gibt es eine kleinste positive Zahl $R_{\min} > 0$, aber nicht jede Zahl in $\mathbb{M}_{\mathbb{R}}$ ist ein ganzzahliges Vielfaches von R_{\min} . Sonst wäre ja die absolute Auflösung zwischen zwei Maschinenzahlen immer R_{\min} . Typischerweise ist $1 + R_{\min}$ keine Maschinenzahl, Runden ergibt $\mathrm{rd}(1 + R_{\min}) = 1$. Es ergeben sich weitere ungewohnte Folgerungen: Die üblichen Rechenregeln aus \mathbb{R} lassen sich in $\mathbb{M}_{\mathbb{R}}$ nicht exakt nachbilden. Zudem wird die Anordnung der reellen Zahlen beim Übergang zu $\mathbb{M}_{\mathbb{R}}$ höchstens näherungsweise respektiert. Aus der Beziehung a < b in \mathbb{R} muss nicht folgen $\mathrm{rd}(a) < \mathrm{rd}(b)$ in $\mathbb{M}_{\mathbb{R}}$. Aus $\mathrm{rd}(a) = \mathrm{rd}(b)$ folgt $a \approx b$ mit einigem Interpretationsspielraum für die Bedeutung des \approx -Zeichens.

Die Maschinengenauigkeit ist die kleinste Zahl $\varepsilon>0$, für die gilt $(1+\varepsilon)-1>0$. Die Maschinengenauigkeit ist ein Mass für die relative Genauigkeit bei der Zahldarstellung. Die relative Genauigkeit von Näherungen kann nie besser sein als die Maschinengenauigkeit. Die kleinste positive Maschinenzahl R_{\min} charakterisiert die Lücke zwischen 0 und den positiven Zahlen in $\mathbb{M}_{\mathbb{R}}$. Das ist ein Mass für eine absolute Genauigkeit, die nur in der kleinsten Dekade erreicht wird, die in $\mathbb{M}_{\mathbb{R}}$ darstellbar ist. In der Regel ist $\varepsilon\gg R_{\min}$.

Tatsächlich gibt es verschiedene Realisierungen von $\mathbb{M}_{\mathbb{R}}$. Eine davon trägt das Label IEEE einer US-Organisation von Ingenieuren und gilt seit 1985 als Quasistandard.

Es folgen einige Beispiele im Zusammenhang mit Genauigkeit, Rundung und Rechengesetzen:

- 1. Da $\epsilon \ll 1$ gilt, ist $(1+\varepsilon)\cdot(1-\varepsilon)=1-\varepsilon^2\doteq 1$. In der Rechnerarithmetik gilt dann also $1-\varepsilon\doteq(1+\varepsilon)^{-1}$.
- 2. Aus $a \cdot b = 0$ in $\mathbb{M}_{\mathbb{R}}$ folgt nicht a = 0 oder b = 0, wie das Beispiel $a = R_{\min} > 0$ und b = 0.6 zeigt. In $\mathbb{M}_{\mathbb{R}}$ gibt es *Nullteiler*.

3. Allgemein gilt in $\mathbb{M}_{\mathbb{R}}$ das Distributivgesetz nicht, denn

$$0.6R_{\min} + 0.6R_{\min} = 0 + 0 = 0 \quad \text{und} \quad (0.6 + 0.6) \cdot R_{\min} \ge R_{\min} > 0$$

4. Allgemein gilt in $\mathbb{M}_{\mathbb{R}}$ das Assoziativgesetz der Multiplikation nicht, denn

$$(R_{\min} \cdot \sqrt{R_{\min}}) \cdot \frac{2}{\sqrt{R_{\min}}} = 0 \quad \text{und} \quad R_{\min} \cdot (\sqrt{R_{\min}} \cdot \frac{2}{\sqrt{R_{\min}}}) > R_{\min} > 0$$

1.3 Das Kontinuum $\mathbb R$ und die endliche Menge $\mathbb M_{\mathbb R}$ als Diskretisierung mit Unschärfen

Die endliche Menge $\mathbb{M}_{\mathbb{R}}$ ist diskret. Wie kann sie Eigenschaften des Kontinuums \mathbb{R} simulieren? Diese Imitation gelingt nicht alleine mit Eigenschaften der Menge $\mathbb{M}_{\mathbb{R}}$, es braucht dazu noch eine Besonderheit der Operationen. Eine mahnende Faustregel lautet: Jede Operation in $\mathbb{M}_{\mathbb{R}}$ ist ein möglicher Grund für Unterschiede zu den idealisierten Operationen in \mathbb{R} . Das Runden einer Zahl x auf eine Maschinenzahl $\tilde{x} = rd(x)$ und andere Approximationen erzeugten Differenzen zu den Operationen in R. Es ist sinnvoll, sich die Auswirkung als 'Lärm' oder zufälliges Rauschen vorzustellen, das Rundungsrauschen, das die idealen Operationen stört. Es überdeckt Lücken zwischen benachbarten Elementen von $\mathbb{M}_{\mathbb{R}}$ in einem stochastischen Sinne. Natürlich erzeugt der Computer keine wirklich zufälligen Schwankungen. Aber wir sind nicht in der Lage, die Rundungsfehler und die Approximationsfehler in einer langen Rechnung deterministisch zu verfolgen, ebenso wie wir beim Wurf einer Münze ihren Flug nicht so genau kontrollieren können, dass das Ergebnis sicher vorhersagbar wird. So wie sich das überabzählbare Kontinuum der individuellen Benennung seiner Elemente in jeder konkreten Sprache entzieht und wir uns die unbenennbaren Zahlen durch eine Art Orakel geben lassen müssen, so scheint die Unschärfe in den Operationen in $\mathbb{M}_{\mathbb{R}}$ eine geradezu unverzichtbare stochastische Komponente zu sein, eine Notwendigkeit, um mit $\mathbb{M}_{\mathbb{R}}$ das Kontinuum \mathbb{R} zu imitieren, [5].

Diskretisierung und Rundung sind zwei wichtige Gründe für Diskrepanzen zwischen analytischen und numerischen Ergebnissen in Fällen, die einen direkten Vergleich zulassen. Bekanntlich gibt es Fragen, die sich prinzipiell nicht durch formal exakte algebraische oder analytische Methoden beantworten lassen [6]. Numerische Näherungen können in solchen Fällen weiterführen als die rein formalen Methoden, falls man sich mit angenäherten Antworten zufrieden geben kann. Für Grundlagen vgl. [2].

2 Funktionen

2.1 Ideale, Träume, Wirklichkeit

Wir sind uns gewohnt, Funktionen mit Hilfe von Termen zu definieren, etwa so:

$$g: x \mapsto \sin(x)/\sqrt{1+x^2}$$

Genau genommen müssten noch ein Definitionsbereich und ein Wertebereich angegeben sein, damit alle Eigenschaften der Funktion mit der Definition festgelegt werden. Wichtig ist aber, dass es uns auch nicht gelingt, mit Termdarstellungen alle denkbaren reellwertigen Funktionen auf einem Intervall $I \neq \emptyset$, $f: I \to \mathbb{R}$ mit einer Zuordnungsformel zu definieren. Das misslingt ja schon bei der überabzählbaren Menge der Konstanten aus \mathbb{R} . Die Menge aller Funktionen, die auf einem reellen Intervall definiert sind und Werte in den reellen Zahlen annehmen, ist

nochmals viel umfangreicher als die Menge \mathbb{R} selbst. Es fehlen uns Worte unhd Bezeichnungen in einem endlichen Alphabeth, um das konkret zu schreiben, was wir abstrakt in Gedanken konzipieren.

Typische Funktionen in der Analysis treten als Ergebnisse von unendlichen Prozessen in der Form von Grenzwerten oder als Idealisierungen auf, mit denen sich gewisse geforderte Bedingungen erst erfüllen lassen. Funktionen, die den Bedürfnissen der Analysis entsprechen, sind nur ausnahmsweise termdefiniert und noch viel seltener effektiv und exakt berechenbar. Dennoch wird in der Analysis angenommen, dass zu jedem Input aus dem Definitionsbereich einer Funktion genau ein Output existiert. Allerdings bedeutet 'existiert' nicht, dass es uns gelingt, mit elementaren Operationen diesen Funktionswert in endlich vielen Rechenschritten zu erzeugen. Sogar bei der wichtigsten elementaren Funktion, der Exponentialfunktion, ist dies ein unerfüllbarer Traum.

2.2 Standardfunktionen, praktische Realisierungen und Näherungen

Schon ganz einfache Taschenrechner sind mit einem Satz von Standardfunktionen ausgestattet, Wurzeln und Potenzen, Exponentialfunktion, Logarithmen, trigonometrische Funktionen und ihre Umkehrungen. Jede technisch-wissenschaftliche Software verfügt über einen Satz von Standardfunktionen, die den gleichen Namen tragen, wie ihre Vorbilder aus der Algebra und Analysis.

Die einfachsten Funktionen sind die Konstanten und die identische Funktion, welche den Input zurückgibt. Wer mit den Rechnerzahlen $\mathbb{M}_{\mathbb{R}}$ auskommen muss, bemerkt, dass es natürlich wiederum beachtliche Unterschiede gibt zwischen den analytischen Vorbildern und den entsprechenden numerischen Realisierungen von Standardfunktionen allein unter Benutzung der Rechnerzahlen. Mit den arithmetischen Operationen in $\mathbb{M}_{\mathbb{R}}$ lassen sich in endlich vielen Schritten nur rationale Funktionen berechnen und das in der Praxis bloss für endlich viele [besser endlich wenige!] Rechnerzahlen und der vorhandenen Rechnerarithmetik.

Jede numerische Realisierung einer reellen Funktion $f:x\mapsto f(x)$ erzeugt letzlich bloss eine endliche Liste von Zahlenpaaren der Form $(\tilde{x},\tilde{f}(\tilde{x}))$, wobei \tilde{x} eine Maschinenzahl aus dem Definitionsbereich von f sein muss und \tilde{f} für eine numerische Berechnung steht, die in endlich vielen Anwendungen von Grundoperationen des Rechners mindestens eine Näherung für $f(\tilde{x})$ liefert. Jede rationalen Funktionen q mit Koeffizienten aus $\mathbb{M}_{\mathbb{R}}$ lässt sich im Definitionsbereich $D_q \cap \mathbb{M}_{\mathbb{R}}$ in endlich vielen Schritten bis auf Rundungsfehler berechnen. In der Nähe von Polstellen müssen mögliche overflow-Probleme behandelt werden, was voraussetzt, dass die Nennernullstellen von q hinreichend genau berechenbar sind. Wenn wir davon ausgehen, dass bei allen Rechnerzahlen x das Vorzeichen eindeutig und richtig erkannt wird, so sind beispielsweise auch die Vorzeichenfunktion oder die Betragsfunktion in diesem Sinne berechenbar.

Ein extremes Beispiel für die Diskrepanz zwischen der Realisierung auf einem Rechner und der abstrakten Definition einer Funktion zeigt die Funktion r, die für eine rationale Zahl x den Funktionswert r(x) = 0 ausgibt, für irrationale x den Funktionswert 1. Wegen $\mathbb{M}_{\mathbb{R}} \subset \mathbb{Q}$ ist \tilde{r} die Konstante 0. Allerdings ist r auch ein Beispiel für eine Funktion, deren Berechnung prinzipielle Schwierigektien bereitet, wenn man für jeden Input x eine Antwort verlangt. Das läuft darauf hinaus, effektiv entscheiden zu können, ob eine beliebige reelle Zahl x rational sei, obwohl nur ein endliches Anfangsstück der Darstellung als Ziffernfolge zur Verfügung steht.

Im Beispiel der Funktion $q: x \mapsto x^2 - 2$ fehlt die Nullstelle $\sqrt{2}$ in $\mathbb{M}_{\mathbb{R}}$. Die nächstliegende Rechnerzahl ist $w = \mathrm{rd}(\sqrt{2})$. Die Einsetzprobe mit w ergibt exakt gerechnet $q(w) \neq 0$ und in

 $\mathbb{M}_{\mathbb{R}}$ eine Näherung $\tilde{q}(w) \approx 0$ jedoch keine verlässliche Aussage (allenfalls ist die Näherung in $\mathbb{M}_{\mathbb{R}}$ nicht von 0 unterscheidbar, was wir mit $\tilde{q}(w) \doteq 0$ bezeichnen). Natürlich wissen wir in diesem Beispiel, dass w eine gute Näherung ist.

Die folgende Überlegung zeigt den Mehraufwand bei der Einsetzprobe in der Numerik und die Rolle des Zwischenwertsatzes aus der Analysis: Eine kleine Verschiebung von w zu \tilde{w} wird in einigen Fällen $\tilde{q}(\tilde{w}_+) > \delta$, in anderen $\tilde{q}(\tilde{w}_-) < \delta$ ergeben, wobei $\delta > 0$ eine kleine Zahl ist, die garantiert, dass die Vorzeichen bei $\tilde{q}(\tilde{w})$ trotz der verschiedenen Näherungen richtig erkannt werden. Weil wir aber wissen, dass q eine stetige Funktion ist, so garantiert der Zwischenwertsatz mindestens eine echte Nullstelle von q zwischen w_- und w_+ . Numerische Nullstellensucher geben neben einer Näherung gelegentlich eine Meldung no sign change aus. Das deutet darauf hin, dass der Algorithmus so eingerichtet ist, dass mindestens zwei Näherungen für die Lösung gesucht werden, in denen der Funktionswert sicher unterschiedliche Vorzeichen aufweist. Die Meldung ist eine Warnung, die im Falle von Nullstellen mit einseitiger Berührung der Koordinatenachse sinnvoll ist. Dann kann nämlich geschehen, dass der Nullstellensucher mit Kandidaten für Annäherungen an eine Nullstelle die Einsetzprobe durchführt und angenäherte Funktionswerte findet, die zwar ganz in der Nähe von 0 liegen, aber alle das gleiche Vorzeichen aufweisen. Die numerische Evidenz reicht dann nicht aus, um sicher auf die Existenz einer Nullstelle in der Nähe der untersuchten Kandidaten zu schliessen.

Es gibt zwei wesentliche Kontrollfragen, die immer angebracht sind, wenn in einer Überlegung eine Computergraphik wesentlich ist:

- Weiss ich, was ich sehe?
- Sehe ich, was ich weiss?

3 Aufgaben

3.1 Zum Rechnen mit Zirkel und Lineal

Die Geometrie Euklids ist ein theoretischer Rahmen für Konstruktionen mit Zirkel und Lineal. Diese beiden Werkzeuge werden freilich idealisiert. Daneben gibt es die tatsächlich mit materiellen Werkzeugen ausgeführten geometrischen Konstruktionen. Zwischen praktischen Konstruktionen und der Geometrie Euklids gibt es eine Beziehung, die jener zwischen der Numerik und der Analysis gleicht. Jede Zeichenfläche ist begrenzt, jede gezeichnete Linie hat eine gewisse nicht zu vernachlässigende Strichdicke, dadurch entstehen beim Zeichnen keine exakten Schnittpunkte, sondern Durchschnitte von Linien endlicher Breite, die im Falle von schleifenden Schnitten die Konstruktionsgüte stark beeinträchtigen.

Mit Zirkel und Lineal lässt sich im Prinzip aus einer beliebig gegebenen Einheitsstrecke (OE) mit $O \neq E$ in der Ebene eine Zahlgerade konstruieren. Durch Abtragen orientierter Strecken lässt sich die Addition nachbilden, Orientierungswechsel entspricht dem Vorzeichenwechsel. Mit den Strahlensätzen lassen sich Multiplikation und Division geometrisch realisieren. Der Höhensatz kann benutzt werden, um die Quadratwurzel zu beliebigen positiven Strecken zu konstruieren. Interessanterweise bilden die aus der Einheitsstrecke (OE) konstruierbaren Punkte auf der Geraden OE einen Zahlkörper $\mathbb K$. Er enthält die rationalen Zahlen, aber auch gewisse irrationale, da alle Lösungen von quadratischen Gleichungen mit Koeffizienten in $\mathbb K$ in $\mathbb K$ liegen. Da jeder Punkt von $\mathbb K$ in endlich vielen elementaren Konstruktionsschritten aus (OE) gewonnen wird, ist $\mathbb K$ abzählbar. Die Operationen auf $\mathbb K$ sind stetig. Sie lassen sich also eindeutig auf $\mathbb R$ erweitern. Die Realisierungen der arithmetischen Grundoperationen und des Wurzelziehens als Konstruktionen mit Zirkel und Lineal ergeben

im Rahmen der elementaren Geometrie finite Algorithmen für arithmetische Operatioen ohne Rücksicht auf die Zahldarstellung. Insbesondere ist dabei die Einteilung in rationale oder irrationale Zahlen belanglos.

- 1. Wie lassen sich auf der Zahlgeraden Addition und Subtraktion als Operationen mit Zirkel und Lineal realisieren?
- 2. Wie lassen sich Multiplikation und Division planimetrisch auf einer Zahlgeraden realisieren?
- 3. Wie lässt sich die Quadratwurzel aus einer positiven Zahl planimetrisch mit Zirkel und Lineal konstruieren?
- 4. Warum sind die reellen Lösungen einer quadratischen Gleichung $x^2 + p \cdot x + q = 0$ mit Zirkel und Lineal konstruierbar?
- 5. Lösen Sie das Gleichungssystem

$$y = \frac{1}{5}x$$
$$y = \frac{1}{7} + \frac{1}{6}x$$

grafisch. Wie stark unterscheiden sich die Antworten der grafischen Lösung von der algebraischen?

- 6. Welche Probleme treten bei allen praktischen Realisierungen von Zahloperationen im Rahmen der Planimetrie auf? Inwiefern ergeben sich ähnliche Probleme bei der Verwendung von Maschinenzahlen?
- 7. Wie lassen sich die arithmetischen Operationen und die Quadratwurzel in \mathbb{C} , gedacht als Zeichenebene, mit Zirkel und Lineal realisieren? Ist jede quadratische Gleichung in \mathbb{C} mit Zirkel und Lineal lösbar?
- 8. Weshalb lassen sich dritte Wurzeln in C nur in Ausnahmefällen mit Zirkel und Lineal konstruieren? Warum gelingt die Konstruktion von Quadratwurzeln oder von vierten Wurzeln in C?
- 9. Angenommen, wir können beliebige Winkel ω durch Markierungen auf dem Einheitskreis ablesen. Wie lassen sich dann die trigonometrischen Funktionen cos, sin, tan und ihre Umkehrungen durch Konstruktionen mit Zirkel und Lineal in der Zeichenebene durch endliche Konstruktionen realisieren?
- 10. Angenommen, die Exponentialfunktion $\exp: x \mapsto \exp(x)$ und ihre Umkehrung, die Logarithmusfunktion $\ln: t \mapsto \ln(t)$ seien durch endliche Konstruktionen mit Zirkel und Lineal realisierbar. Wie liesse sich dann $\sqrt[3]{2}$ aus der Einheitsstrecke mit Zirkel und Lineal konstruieren? Warum kommt diese Hypothese mit der Nicht-Konstruierbarkeit der Würfelverdoppelung in Konflikt?

3.2 Über $\mathbb R$ und $\mathbb M_\mathbb R$

- 11. Oft wird in der Numerik ein Modell für $\mathbb{M}_{\mathbb{R}}$ verwendet mit einer 16-stelligen Mantisse.
 - (a) Es ist üblich, von einem Weltalter von rund $1.5 \cdot 10^{10}$ Jahren auszugehen. Wie viele Stunden dauert der kleinste Bruchteil Δt des Weltalters, der sich als Dezimalzahl mit Exponent 0 und einer 16-stelligen Mantisse angeben lässt?
 - (b) Die mittlere Entfernung der Erde zur Sonne, die Astronomische Einheit AU, beträgt rund $1.5 \cdot 10^8$ km. Wie viele Meter misst die kleinste Teilstrecke der AU, die sich als Dezimalzahl mit Exponent 0 und einer 16-stelligen Mantisse beschreiben lässt?
 - (c) Stimmt es, dass man die Entfernung der Erde zum Mond von rund 384000 km mit einer 12-stelligen Mantisse bis auf die Grösse eines Staubkornes auflösen kann?
- 12. Wie viele Stellen benötigt die Mantisse, um den besten heute bekannten Wert der folgenden physikalischen Konstanten anzugeben?
 - (a) Lichtgeschwindigkeit $c = 2,99792458 \cdot 10^8 \text{m} \cdot \text{s}^{-1}$ [1983]
 - (b) Gravitationskonstante $G = 6.67428(67) \cdot 10^{-11} \,\mathrm{N \cdot m^2 \cdot kg^{-2}}$
 - (c) Masse des Protons $m_p = 1.672621637(83) \cdot 10^{-27} \,\mathrm{kg}$
- 13. Eine Zahl z > 0 wird binär mit einer Mantisse von 12 Bit gespeichert. Anschliessend wird diese Binärzahl \tilde{z} als Dezimalzahl mit gerundeter 4-stelliger Mantisse angezeigt.
 - (a) Es sei z=1/3. Wie lautet die zwölfstellige Mantisse von \tilde{z} in der Binärdarstellung? Welche vierstellige Mantisse wird daraus beim Umwandeln als Dezimalzahl \tilde{z} ? Wie gross ist die vierstellige Dezimalzahl $3 \cdot \tilde{z}$?
 - (b) Es sei z=1/5. Wie lautet die zwölfstellige Mantisse \tilde{z} in der Binärdarstellung? Welche vierstellige Mantisse wird daraus beim Umwandeln als Dezimalzahl \tilde{z} ? Wie gross ist die vierstellige Dezimalzahl $5 \cdot \tilde{z}$?

3.3 Zu Funktionen, Graphen, Abtastungen und Pixelmustern

Die folgenden Aufgaben handeln von einer Funktion, zum Beispiel f, einer Abtastung s_f und einer Abtastung einer Näherung \tilde{s}_f . Der Funktionsgraph wird durch feine Linien symbolisiert. Abtastungen der Funktionen werden markiert mit *, Abtastungen der numerischen Näherungen mit \circ .

- 14. Die Abbildung 1 zeigt den Graphen einer Funktion f, den Graphen einer Abtastung s_f und den Graphen einer numerischen Näherung \tilde{s}_f . Kommentieren Sie Gemeinsamkeiten und Unterschiede der drei Graphen. Welche Probleme ergeben sich, wenn anhand des Graphen von \tilde{s}_f
 - (a) die Nullstellen von f gesucht werden?
 - (b) die Extrema von f zu bestimmen sind?
- 15. Die Abbildung 2 zeigt wiederum Graphen einer Funktion g, ihrer Abtastung s_g und einer numerischen Realisierung als Näherung \tilde{s}_g . Kommentieren Sie das lokale Verhalten in den gezeigten Fallbeispielen je für die Funktion, die Abtastung und die numerische Näherung. Was zeigt der Graph, was die Abtastung und was die Näherung? Welche

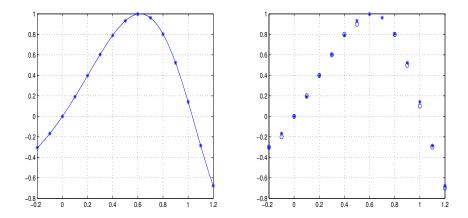


Abbildung 1: Graph einer Funktion, Abtastung und Abtastung einer Näherung

Missverständnisse oder Fehlschlüsse können sich ergeben, wenn nur die Abtastung oder Näherung interpretiert werden, ohne an den idealen Graphen zu denken? Welche korrigierende Rolle spielen die Vorkenntnisse aus der Analysis bei der Interpretation? Welche Zusatzinformation ist nötig, um diese Interpretation sachgerecht zu machen?

(a) einfache Nullstelle (transversaler Schnitt der horizontalen Achse)

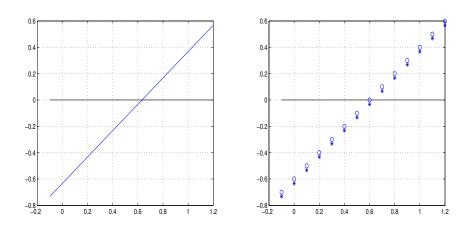


Abbildung 2: Nullstelle mit transversalem Schnitt (Nullstelle 1. Ordnung) Graph, Abtastung und Näherung von $g_1: x \mapsto x - \sqrt{0.4}$,

- (b) Die Abbildung 3 zeigt das typische Muster bei einer zweifachen Nullstelle (einseitige Berührung der horizontalen Achse).
- (c) Abbildung 4 stellt zwei Arten von Störungen dar, die dieses Muster zerstören. Zu beachten sind neben dem Verhalten der Funktion das Verhalten von Abtastung * und Annäherung o. Welche Unterschiede zeigen sich im Vergleich zum Verhalten einer leicht gestörten Nullstelle mit transversalem Schnitt? Welche Folgerungen ergeben sich für numerische Gleichungslöser?
- (d) Die Abbildung 5 zeigt das typische Verhalten von Graph, Abtastung und Näherung nahe bei einer dreifachen Nullstelle (zweiseitige Berührung der horizontalen

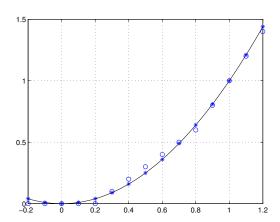


Abbildung 3: Nullstelle 2. Ordnung, einseitige Berührung der Achse, Bsp $g_2:x\,\mapsto\,x^2$

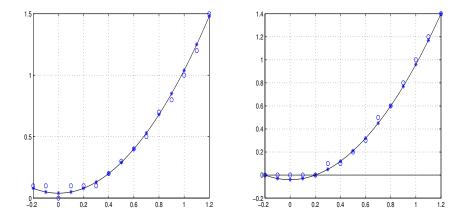


Abbildung 4: Zwei Funktionen, Störungen einer Nullstelle 2. Ordnung, Bsp $g_2^+:x\mapsto g_2(x)+0.04,\quad g_2^-:x\mapsto g_2(x)-0.04$

Achse). Leichte Variationen dieser Situation sind in Abbildung 6 dargestellt. Welcher Unterschied zum Verhalten bei der Nullstelle zweiter Ordnung zeigt sich? Wie unterscheiden sich Graph, Abtastung und Näherung in den drei gezeigten Fällen? $g_3: x \mapsto x^3$

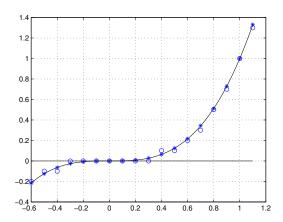


Abbildung 5: Nullstelle 3. Ordnung, zweiseitige Berührung der Achse, Bsp $g_3: x \mapsto x^3$

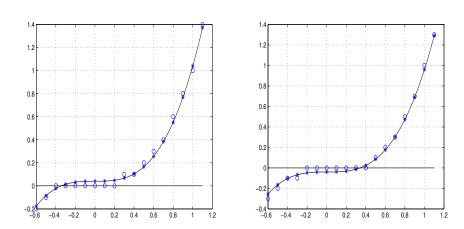


Abbildung 6: Störungen einer Nullstelle 3. Ordnung, Bsp $g_3^+: x \mapsto g_3(x) + 0.04, \quad g_3^-: x \mapsto g_3(x) - 0.04$

16. Die folgende Aufgabe zeigt, was geschehen kann, wenn eine hochfrequente Sinusfunktion auf dem Pixelmuster des Bildschirms dargestellt wird. Zur Illustration wird ein vereinfachtes Beispiel gewählt. Die Funktion $h_{10}: t \mapsto \sin(20\pi \cdot t)$ hat 10 volle Perioden auf dem Einheitsintervall. Diese Funktion wird abgestastet in den Zeitpunkten $t_r = \frac{r}{9}$ mit ganzzahligem r. Dieses Experiment lässt sich auch mit dem Abtasttheorem von Shannon in Verbindung bringen oder einfach mit einer kleinen Rechnung:

$$\sin(20\pi \cdot t_r) = \sin(\frac{10}{9}2\pi r) = \sin(2r\pi + \frac{2}{9}r\pi) = \sin(\frac{2}{9}r\pi), \text{ weil } r \in \mathbb{Z}.$$

Die Abtastung s_h in den Punkten t_r erscheint in Abbildung 7 zugleich mit dem Graphen von h_{10} und von $h_1: t \mapsto \sin(2\pi t)$. Was zeigt dieses Experiment? Wie erscheint

der Graph einer Sinusfunktion mit 1025 Schwingungen auf einem Bildschirm mit 1024 Pixeln in der Horizontalen? Wie lässt sich das Phänomen erklären?

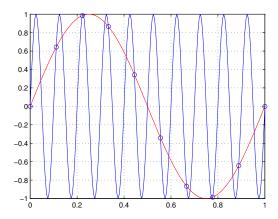


Abbildung 7: Abtastung s_h von $h_{10}: t \mapsto \sin(20\pi t)$ ist auch Abtastung zur Funktion $h_1: t \mapsto \sin(2\pi t)$

Welches allgemeine Problem spricht dieses Beispiel an? Wie lässt sich das Beispiel verallgemeinern? [Stichwort: Aliasing und Shannon's Abtasttheorem, undersampling]

- 17. Manchmal hört man im Zusammenhang mit computergenerierten Darstellungen von Funktionsgraphen die Empfehlung: Ein Graph ist nie genug! Was ist da dran?
 - (a) Welches Problem lässt sich mindern, wenn mehrere Graphen mit verschiedener Auflösung gezeichnet werden?
 - (b) Welche Vergröberungen verursacht das Pixelmuster des Bildschirms?

Eine Projektidee Angenommen, es steht eine Numeriksoftware zur Verfügung mit einer 12-stelligen Mantisse. Wie lässt sich mit diesem Programm

- 1. eine ganze Zahl in der Basis 10²⁰ simulieren?
- 2. eine Addition realisieren mit Dezimalzahlen, die eine 20-stellige Mantisse haben?
- 3. eine Multiplikation realisieren mit Dezimalzahlen, die eine 20-stellige Mantisse haben?
- 4. zu einer Dezimalzahl $a \neq 0$ mit 20-stelliger Mantisse die reziproke Zahl finden?
- 5. eine Arithmetik mit 10n-stelliger Mantisse konstruieren?

Literatur

- [1] HILBERT, D. Grundlagen der Geometrie, B.G. Teubner, Leipzig, Berlin, 1904
- [2] Higham N. J. Accuracy and Stability of Numerical Algorithms, 2nd edn. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics, 2002
- [3] Landau, E., Grundlagen der Analysis, Chelsea, New York, 1965

- [4] OVERTON, M.L., Numerical Computing with IEEE Floating Point Arithmetic, SIAM, Society for Industrial and Applied Mathematics, 2001
- [5] Poincaré, Henri, La Science et l'Hypothèse, Flammarion, Paris, 1904
- [6] RICHARDSON, DANIEL, Some undecidable problems involving elementary functions of a real variable, PhD thesis, University of Wisconsin Madison, 1968
- [7] SIAM, Society for Industrial and Applied Mathematics, *The 100 Digits Challenge*, 2001

Ich danke H.R. Vollmer für die Unterstützung in IATEX und beim Ausmerzen von Fehlern.