Wolfgang K. Seiler

Computeralgebra

Vorlesung im Frühjahrssemester 2023
an der Universität Mannheim

Dieses Skriptum entsteht parallel zur Vorlesung und soll mit möglichst geringer Verzögerung erscheinen. Es ist daher in seiner Qualität auf keinen Fall mit einem Lehrbuch zu vergleichen; insbesondere sind Fehler bei dieser Entstehensweise nicht nur möglich, sondern **sicher**. Dabei handelt es sich wohl leider nicht immer nur um harmlose Tippfehler, sondern auch um Fehler bei den mathematischen Aussagen. Da mehrere Teile aus anderen Skripten für Hörerkreise der verschiedensten Niveaus übernommen sind, ist die Präsentation auch teilweise ziemlich inhomogen.

Das Skriptum sollte daher mit Sorgfalt und einem gewissen Mißtrauen gegen seinen Inhalt gelesen werden. Falls Sie Fehler finden, teilen Sie mir dies bitte persönlich oder per e-mail (seiler@math.uni-mannheim.de) mit. Auch wenn Sie Teile des Skriptums unverständlich finden, bin ich für entsprechende Hinweise dankbar.

Falls genügend viele Hinweise eingehen, werde ich von Zeit zu Zeit Listen mit Berichtigungen und Verbesserungen zusammenstellen. In der online Version werden natürlich alle bekannten Fehler korrigiert.

Biographische Angaben von Mathematikern beruhen größtenteils auf den entsprechenden Artikeln im *MacTutor History of Mathematics archive* (`www-history.mcs.st-andrews.ac.uk/history/`), von wo auch die meisten abgedruckten Bilder stammen. Bei noch lebenden Mathematikern bezog ich mich, soweit möglich, auf deren eigenen Internetauftritt.

Kapitel 0
Einführung

§1: Was ist Computeralgebra?

Das Wort *Computeralgebra* legt nahe, daß es sich hier um die Anwendung von Computern in der Algebra geht. Das ist nicht ganz richtig, denn die französische Bezeichung *calcul formel* (formales Rechnen) beschreibt den Inhalt des Gebiets genauso gut. Unbestreitbar ist aber, daß Computeralgebra etwas mit Algebra zu tun hat, so daß wir uns zunächst fragen sollten, was Algebra ist.

Bei anderen Teilgebieten der Mathematik sagt uns der Name sofort worum es geht: Die Zahlentheorie beschäftigt sich mit den (ganzen) Zahlen, die Geometrie von γεωμετρία mit dem Vermessen der Erde, die Wahrscheinlichkeitstheorie mit Wahrscheinlichkeiten, die Statistik von *statisticum*, den Staat betreffend, ursprünglich vor allem mit Daten über den Staat. Die Algebra hat ihren Namen vom Titel eines Buchs:

Um 830 legte der arabische Gelehrte ABU DSCHA'FAR MUḤAMMAD IBN MŪSĀ AL-CHWĀRIZMĪ sein zweites Buch *Al-Kitāb al-muchtasar fi hisab al-dschabr wa-'l-muqābala* oder kurz *Kitāb al-dschabr wa-'l-muqābala* vor; *al-dschabr* gab der Algebra ihren Namen, und der Autorenname AL-CHWĀRIZMĪ führte zum Wort Algorithmus. In deutscher Übersetzung heißt der volle Titel etwa *Kurzgefaßtes Buch über das Rechnen durch Ergänzen und Ausgleichen. Al-dschabr,* das Ergänzen oder Vervollständigen, besteht darin, negative Terme in einer Gleichung auf die andere Seite zu bringen; in einem Beispiel aus dem Buch wird etwa aus (in moderner Schreibweise) $x^2 = 40x - 4x^2$ durch *al-dschabr* die Gleichung $5x^2 = 40x$. *Al-muqābala,* das Ausgleichen, besteht darin,

von zwei positiven Termen auf den beiden Seiten der Gleichung den einen auf Null zu reduzieren; aus $x^2+3x+5 = 7x+2$ wird also zunächst $x^2 + 5 = 4x + 2$ und dann $x^2 + 3 = 4x$.

ABU DSCHA'FAR MUHAMMAD IBN MŪSĀ AL-CHWĀRIZMĪ wurde um 780 geboren und arbeitete die meiste Zeit seines Lebens in Bagdad, insbesondere auch im *Haus der Weisheit,* das AL-MA'MŪM, der siebte Kalif, als wissenschaftliches Zentrum seines Reichs gegründet hatte. Eine der Aufgaben dieses Zentrums bestand darin, Texte klassischer griechischer Wissenschaftler ins Arabische zu übersetzen; viele Texte sind heute nur noch über diese Übersetzungen bekannt. Arbeitsgebiete am *Haus der Weisheit* waren vor allem Mathematik und Astronomie. Außer einem weiteren mathematischen Buch, das sich mit den indischen Ziffern befaßte, schrieb AL-CHWĀRIZMĪ auch Bücher über Geographie und Kartographie.

Aus heutiger Sicht besteht kaum ein Unterschied zwischen *al-dschabr* und *al-muqābala;* wir sagen einfach, daß wir einen Term auf die andere Seite bringen. Im neunten Jahrhundert waren die beiden Methoden noch grundverschieden, denn negative Zahlen begannen außerhalb Indiens erst im 16. Jahrhundert langsam in der Mathematik aufzutauchen. Auch die Null fing gerade erst an verwendet zu werden; davon handelt AL-CHWĀRIZMĪs erstes Buch, in dem er die indische Zahlenschrift in die arabische Welt brachte. Die Null wurde aber nicht als *Zahl* eingeführt, sondern nur als *Ziffer.* Dieses Wort kommt vom arabischen Wort für Null *ṣifr,* was von *ṣafira = leer sein* kommt, und das wiederum kommt vom Sanskrit-Wort *śūnya,* das Nichts oder die Leere. Wenn wir den Titel des Buchs aus heutiger Sicht interpretieren, ging es also um die Lehre vom auf die andere Seite bringen. Genauer gesagt handelt das Buch von der Lösung linearer und quadratischer Gleichungen.

In den darauffolgenden Jahrhunderten verstand man unter Algebra allgemeiner das Lösen von Polynomgleichungen in einer Variablen; gelegentlich wurden auch Systeme von Gleichungen in mehreren Variablen betrachtet.

Erst im 19. Jahrhundert begannen, vor allem im angelsächsischen Raum, erste Entwicklungen, die schließlich zu Beginn des zwanzigsten Jahrhunderts zur sogenannten *abstrakten Algebra* führten, mit der sich die meisten heutigen Vorlesungen und Lehrbücher über Algebra beschäftigen.

Sobald kurz nach dem zweiten Weltkrieg die ersten Computer an Universitäten auftauchten, wurden sie von Mathematikern nicht nur zum numerischen Rechnen eingesetzt, sondern auch für alle anderen Arten mathematischer Routinearbeiten, genau wie auch schon früher alle zur Verfügung stehenden Mittel benutzt wurden: Beispielsweise konstruierte D.H. LEHMER bereits vor rund achtzig Jahren, lange vor den ersten Computern, mit Fahrradketten Maschinen, die (große) natürliche Zahlen in ihre Primfaktoren zerlegen konnten.

Computer manipulieren Bitfolgen; von den meisten Anwendern wurden diese zur Zeit der ersten Computer zwar als Zahlen interpretiert, aber wie wenig später selbst die Buchhalter bemerkten, können sie natürlich auch Informationen ganz anderer Art darstellen. Deshalb wurden bereits auf den ersten Computern (deren Leistungsfähigkeit nach heutigen Standards nicht einmal der eines programmierbaren Taschenrechners entspricht) algebraische, zahlentheoretische und andere abstrakt mathematische Berechnungen durchgeführt wurden.

Programmiert wurde meist in Assembler, da die gängigen höhere Programmiersprachen der damaligen Zeit (FORTRAN, ALGOL 60, COBOL, ...) vor allem mit Blick auf numerische *bzw.,* im Fall von COBOL, betriebswirtschaftliche Anwendungen konzipiert worden waren.

Eine Ausnahme bildete die 1958 von JOHN MCCARTHY entwickelte Programmiersprache LISP, die speziell für symbolische Manipulation entwickelt wurde, vor allem solche im Bereich der künstlichen Intelligenz. In dieser Sprache wurden Ende der Sechzigerjahre die ersten Computeralgebrasysteme geschrieben: MACSYMA ab 1968 ebenfalls am M.I.T. zunächst vor allem für alle Arten von symbolischen Rechnungen in Forschungsprojekten des M.I.T., REDUCE ungefähr gleichzeitig von ANTHONY C. HEARN vor allem für Berechnungen in der Hochenergiephysik.

Beide Systeme verbreiteten sich schnell an den Universitäten und wurden bald auch schon für eine Vielzahl anderer Anwendungen benutzt; dies wiederum führte zur Weiterentwicklung der Systeme sowohl durch die ursprünglichen Autoren als auch durch Benutzer, die neue Pakete hinzufügten, und es führte auch dazu, daß anderswo neue Computer-

algebrasysteme entwickelt wurden, wie beispielsweise Maple an der University of Waterloo (einer der Partneruniversitäten von Mannheim). Mit der zunehmenden Nachfrage lohnte es sich auch, deutlich mehr Arbeit in die Entwicklung der Systeme zu stecken, so daß die neuen Systeme oft nicht mehr in LISP geschrieben waren, sondern in klassischen Programmiersprachen wie MODULA oder C *bzw.* später C++, die zwar für das symbolische Rechnen einen erheblich höheren Programmieraufwand erfordern als LISP, die dafür aber auch zu deutlich schnelleren Programmen führen.

Eine gewisse Zäsur bedeutete das Auftreten von *Mathematica* im Jahr 1988. Dies ist das erste System, das von Anfang an rein kommerziell entwickelt wurde. Der Firmengründer und Initiator STEVE WOLFRAM kommt zwar aus dem Universitätsbereich (bevor er seine Firma gründete, forschte er am *Institute for Advanced Studies* in Princeton über zelluläre Automaten), aber *Mathematica* war von Anfang an gedacht als ein Produkt, das an Naturwissenschaftler, Ingenieure und Mathematiker *verkauft* werden sollte. Ein wesentlicher Aspekt, der aus Sicht dieser Zielgruppe den Kauf von *Mathematica* attraktiv machte, obwohl zumindest damals noch eine ganze Reihe anderer Systeme frei oder gegen nominale Gebühr erhältlich waren, bestand in der Möglichkeit, auf einfache Weise Graphiken zu erzeugen. Bei den ersten Systemen hatte dies nie eine Rolle gespielt, da Graphik damals nur über teure Plotter und (zumindest in Universitätsrechenzentrum) mit Wartezeiten von rund einem Tag erstellt werden konnte. 1988 gab es bereits PCs mit (damals noch sehr schwachen) graphikfähigen Bildschirmen, und Visualisierung spielte plötzlich in allen Wissenschaften eine erheblich größere Rolle als zuvor.

Der Nachteil der ersten *Mathematica*-Versionen war eine im Vergleich zur Konkurrenz ziemlich hohe Fehlerquote bei den mathematischen Berechnungen. (Perfekt ist in diesem Punkt auch heute noch kein Computeralgebrasystem.) Der große Vorteil der einfachen Erzeugung von Graphiken sowie das sehr gute Begleitbuch von STEVE WOLFRAM, das deutlich über dem Qualitätsniveau auch heute üblicher Softwaredokumentation lag, bescherte *Mathematica* einen großen Erfolg. Da auch Systeme wie MACSYMA und MAPLE mittlerweile in selbständi-

ge Unternehmen ausgegliedert worden waren, führte die Konkurrenz am Markt schnell dazu, daß Graphik auch ein wesentlicher Bestandteil anderer Computeralgebrasysteme wurde und daß *Mathematica* etwas vorsichtiger mit den Regeln der Mathematik umging. Heute unterscheiden sich die beiden kommerziell dominanten Systeme Maple und *Mathematica* nicht mehr wesentlich in ihren Graphikfähigkeiten und ihrer (geringen, aber bemerkbaren) Häufigkeit mathematischer Fehler. Hinzu kam der Markt der Schüler und Studenten, so daß ein am Markt erfolgreiches Computeralgebrasystem auch in der Lage sein muß, die Grundaufgaben der Schulmathematik und der Mathematikausbildung zumindest der ersten Semester der gefragtesten Studiengänge zu lösen.

Da die meisten, die mit dem Begriff *Computeralgebra* überhaupt etwas anfangen können, an Computeralgebrasysteme denken, hat sich dadurch auch die Bedeutung des Worts *Computeralgebrasystem* verändert: Gemeinhin versteht man darunter nicht mehr nur ein Programm, das symbolische Berechnungen ermöglicht, sondern eines, das über ernstzunehmende Graphikfähigkeiten verfügt und viele gängige Aufgabentypen lösen kann, ohne daß der Benutzer notwendigerweise versteht, wie man solche Aufgaben löst.

In der Computeralgebra als Teilgebiet der Mathematik geht es allerdings nicht um Computeralgebrasystemen, auch wenn diese größtenteils von Computeralgebraikern stammen und deren Algorithmen implementieren. Vielmehr geht es darum, Algorithmen zu finden, mit denen sich algebraische Probleme möglichst effizient lösen lassen. Viele dieser Algorithmen, die teilweise schon aus dem 19. Jahrhundert stammen, sind so aufwendig, daß sie sich nur mit Computerhilfe auf interessante Probleme anwenden lassen, was sicherlich mit zu dem Namen *Computeralgebra* geführt hat und natürlich auch dazu, daß Computeralgebraiker intensiv sowohl gängige Computeralgebrasysteme als auch selbst entwickelte spezielle Systeme benutzen.

Die Computeralgebra befasst sich sowohl mit dem klassischen Problem der Algebra, der Lösung von nichtlinearen Gleichungen und Gleichungssystemen, als auch mit konstruktiven Verfahren der abstrakten Algebra wie dem Rechnen in abstrakten Gruppen oder der Berechnung von GALOIS-Gruppen und ähnlichen Problemen. Auch analytische Proble-

me lassen sich zumindest teilweise mit Methoden der Computeralgebra behandeln, beispielsweise gibt es algebraische Algorithmen zur symbolischen Integration und auch zur Lösung von Differentialgleichungen.

In dieser Vorlesung, die keine Algebra-Vorlesung voraussetzt, sollen vor allem Algorithmen zur Lösung klassischer Probleme betrachtet werden, insbesondere das der Lösung von Gleichungen und Gleichungssystemen. Da schon bei Gleichungen in einer Variablen die Schwierigkeit sehr schnell mit dem Grad der Gleichung ansteigt, gehören dazu auch Verfahren zur Zerlegung eines Polynoms in ein Produkt von Polynomen kleineren Grades – sofern dies möglich ist. Auch der EUKLIDische Algorithmus, sowohl für Zahlen als auch für Polynome, wird uns immer wieder begegnen.

§2: Numerisches, exaktes und symbolisches Rechnen

Mit vielen Fragestellungen der Computeralgebra wie etwa der Lösung von Polynomgleichungen oder Systemen solcher Gleichungen beschäftigt sich auch die numerische Mathematik; um die unterschiedlichen Ansätze beider Gebiete zu verstehen, müssen wir uns die Unterschiede zwischen numerischem Rechnen, exaktem Rechnen und symbolischem Rechnen klar machen.

Numerisches Rechnen gilt gemeinhin als *das* Rechnen mit reellen Zahlen. Kurzes Nachdenken zeigt, daß wirkliches Rechnen mit reellen Zahlen weder mit Papier und Bleistift noch per Computer möglich ist: Die Menge $\mathbb{R}$ der reellen Zahlen ist schließlich überabzählbar, aber sowohl unsere Gehirne als auch unsere Computer sind endlich. Der Datentyp **real** oder **float** oder auch **double** einer Programmiersprache kann daher unmöglich das Rechnen mit reellen Zahlen exakt wiedergeben.

Tatsächlich genügt das Rechnen mit reellen Zahlen per Computer völlig anderen Regeln als denen, die wir vom Körper der reellen Zahlen gewohnt sind. Zunächst einmal müssen wir uns notgedrungen auf eine endliche Teilmenge von $\mathbb{R}$ beschränken; in der Numerik sind dies traditionellerweise die sogenannten Gleitkommazahlen.

Eine Gleitkommazahl wird dargestellt in der Form $x = \pm m \cdot b^{\pm e}$, wobei die *Mantisse* m zwischen 0 und 1 liegt und der *Exponent* e eine ganze Zahl aus einem gewissen vorgegebenen Bereich ist. Die Basis b ist in heutigen Computern gleich zwei, in einigen alten Mainframe Computern sowie in vielen Taschenrechnern wird auch $b = 10$ verwendet.

Praktisch alle heute gebräuchliche CPUs für Computer richten sich beim Format für m und e nach dem IEEE-Standard 754 von 1985. Hier ist $b = 2$, und einfach genaue Zahlen werden in einem Wort aus 32 Bit gespeichert. Das erste dieser Bits steht für das Vorzeichen, 0 für positive, eins für negative Zahlen. Danach folgen acht Bit für den Exponenten e und 23 Bit für die Mantisse m.

Die acht Exponentenbit können interpretiert werden als eine ganze Zahl n zwischen 0 und 255; wenn n keinen der beiden Extremwerte 0 und 255 annimmt, wird das Bitmuster interpretiert als die Gleitkommazahl (Mantisse im Zweiersystem)

$$\pm 1{,}m_1 \ldots m_{23} \times 2^{n-127}\,.$$

Die Zahlen, die in obiger Form dargestellt werden können, liegen somit zwischen $2^{-126} \approx 1{,}175 \cdot 10^{-37}$ und $(2 - 2^{-23}) \cdot 2^{127} \approx 3{,}403 \cdot 10^{38}$. Das führende Bit der Mantisse ist stets gleich eins (sogenannte normalisierte Darstellung) und wird deshalb gleich gar nicht erst abgespeichert. Der Grund liegt natürlich darin, daß man ein führendes Bit Null durch Erniedrigung des Exponenten zum Verschwinden bringen kann – es sei denn, man hat bereits den niedrigstmöglichen Exponenten $n = 0$, entsprechend $e = -127$.

Für $n = 0$ gilt daher eine andere Konvention: Jetzt wird die Zahl interpretiert als

$$\pm 0{,}m_1 \ldots m_{23} \times 2^{-126}\,;$$

man hat somit einen (unter Numerikern nicht unumstrittenen) *Unterlaufbereich* aus sogenannten *subnormalen* Zahlen, in dem mit immer weniger geltenden Ziffern Zahlen auch noch positive Werte bis hinunter zu $2^{-23} \times 2^{-126} = 2^{-149} \approx 1{,}401 \cdot 10^{-44}$ dargestellt werden können, außerdem natürlich die Null, bei der sämtliche 32 Bit gleich Null sind.

Auch der andere Extremwert n = 255 hat eine Sonderbedeutung: Falls alle 23 Mantissenbit gleich Null sind, steht dies je nach Vorzeichenbit für $\pm\infty$, andernfalls für NAN *(not a number),* d.h das Ergebnis einer illegalen Rechenoperation wie $\sqrt{-1}$ oder $0/0$. Das Ergebnis von $1/0$ dagegen ist nicht NAN, sondern $+\infty$, und $-1/0 = -\infty$.

Doppeltgenaue Gleitkommazahlen werden entsprechend dargestellt; hier stehen insgesamt 64 Bit zur Verfügung, eines für das Vorzeichen, elf für den Exponenten und 52 für die Mantisse. Durch die elf Exponentenbit können ganze Zahlen zwischen Null und 2047 dargestellt werden; abgesehen von den beiden Extremfällen entspricht dies dem Exponenten $e = n - 1023$.

Der Exponent e sorgt dafür, daß Zahlen aus einem relativ großen Bereich dargestellt werden können, er hat aber auch zur Folge, daß die Dichte der darstellbaren Zahlen in den verschiedenen Größenordnung stark variiert: Am dichtesten liegen die Zahlen in der Umgebung der Null, und mit steigendem Betrag werden die Abstände benachbarter Zahlen immer größer.

Um dies anschaulich zu sehen, betrachten wir ein IEEE-ähnliches Gleitkommasystem mit nur sieben Bit, einem für das Vorzeichen und je drei für Exponent und Mantisse. Das folgende Bild zeigt die Verteilung der so darstellbaren Zahlen (mit Ausnahme von NAN):

Um ein Gefühl dafür zu bekommen, was dies für das praktische Rechnen mit Gleitkommazahlen bedeutet, betrachten wir ein analoges System mit der uns besser vertrauten Dezimaldarstellung von Zahlen (für die es einen eigenen IEEE-Standard 854 von 1987 gibt), und zwar nehmen wir an, daß wir eine dreistellige dezimale Mantisse haben und Exponenten zwischen -3 und 3. Da es bei einer von zwei verschiedenen Basis keine Möglichkeit gibt, bei einer normalisierten Mantisse die erste Ziffer einzusparen, schreiben wir die Zahlen in der Form $\pm 0{,}m_1m_2m_3 \cdot 10^e$.

Zunächst einmal ist klar, daß die Summe zweier Gleitkommazahlen aus diesem System nicht immer als Gleitkommazahl im selben System darstellbar ist: Ein einfaches Gegenbeispiel wäre die Addition der

größten darstellbaren Zahl $0{,}999 \cdot 10^3 = 999$ zu $5 = 0{,}5 \cdot 10^1$: Natürlich ist das Ergebnis 1004 nicht mehr im System darstellbar. Der IEEE-Standard sieht vor, daß in so einem Fall eine *overflow*-Bedingung gesetzt wird und das Ergebnis gleich $+\infty$ wird. Wenn man (wie es die meisten Compiler standardmäßig tun) die *overflow*-Bedingung ignoriert und mit dem Ergebnis $+\infty$ weiter rechnet, kann dies zu akzeptablen Ergebnissen führen: Beispielsweise wäre die Rundung von $1/(999 + 5)$ auf die Null für viele Anwendungen kein gar zu großer Fehler, auch wenn es dafür in unserem System die sehr viel genauere Darstellung $0{,}996 \cdot 10^{-3}$ gibt. Spätestens wenn man das Ergebnis mit 999 multipliziert, um den Wert von $999/(999 + 5)$ zu berechnen, sind die Konsequenzen aber katastrophal: Nun bekommen wir eine Null anstelle von $0{,}996 \cdot 10^0$. Ähnlich sieht es auch aus, wenn wir anschließend 500 subtrahieren: $\infty - 500 = \infty$, aber $(999 + 5) - 500 = 504$ ist eine Zahl, die sich in unserem System sogar exakt darstellen ließe!

Auch ohne Bereichsüberschreitung kann es Probleme geben: Beispielsweise ist

$$123 + 0{,}0456 = 0{,}123 \cdot 10^3 + 0{,}456 \cdot 10^{-1} = 123{,}0456$$

mit einer nur dreistelligen Mantisse nicht exakt darstellbar. Hier sieht der Standard vor, daß das Ergebnis zu einer darstellbaren Zahl gerundet wird, wobei mehrere Rundungsvorschriften zur Auswahl stehen. Voreingestellt ist üblicherweise eine Rundung zur nächsten Maschinenzahl; wer etwas anderes möchte, kann dies durch spezielle Bits in einem Prozessorstatusregister spezifizieren. Im Beispiel würde man also $123 + 0{,}0456 = 123$ oder (bei Rundung nach oben) 124 setzen und dabei zwangsläufig einen Rundungsfehler machen.

Wegen solcher unvermeidlicher Rundungsfehler gilt das Assoziativgesetz selbst dann nicht, wenn es keine Bereichsüberschreitung gibt: Bei Rundung zur nächsten Maschinenzahl ist beispielsweise

$$(0{,}456 \cdot 10^0 + 0{,}3 \cdot 10^{-3}) + 0{,}4 \cdot 10^{-3} = 0{,}456 \cdot 10^0 + 0{,}4 \cdot 10^{-3} = 0{,}456 \cdot 10^0,$$

aber

$$0{,}456 \cdot 10^0 + (0{,}3 \cdot 10^{-3} + 0{,}4 \cdot 10^{-3}) = 0{,}456 \cdot 10^0 + 0{,}7 \cdot 10^{-3} = 0{,}457 \cdot 10^0.$$

Ein mathematischer Algorithmus, dessen Korrektheit unter Voraussetzung der Körperaxiome für $\mathbb{R}$ bewiesen wurde, muß daher bei Gleitkommarechnung kein korrektes oder auch nur annähernd korrektes Ergebnis mehr liefern – ein Problem, das keinesfalls nur theoretische Bedeutung hat.

In der numerischen Mathematik ist dieses Problem natürlich schon seit Jahrzehnten bekannt; das erste Buch, das sich ausschließlich damit beschäftigte, war

J.H. WILKINSON: Rounding errors in algebraic processes, *Prentice Hall,* 1963; Nachdruck bei *Dover,* 1994.

Heute enthält fast jedes Lehrbuch der Numerischen Mathematik entsprechende Abschnitte; zwei Bücher in denen es speziell um diese Probleme, ihr theoretisches Verständnis und praktische Algorithmen geht, sind

FRANÇOISE CHAITIN-CHATELIN, VALÉRIE FRAYSSÉ: Lectures on finite precision computations, SIAM, 1996

sowie das sehr ausführlichen Buch

NICHOLAS J. HIGHAM: Accuracy and stability of numerical algorithms, SIAM, 1996.

Eine ausführliche und elementare Darstellung der IEEE-Arithmetik und des Umgangs damit findet man in

MICHAEL L. OVERTON: Numerical Computing with IEEE Floating Point Arithmetic – *Including One Theorem, One Rule of Thumb and One Hundred and One Exercises,* SIAM, 2001.

Um zu sehen, wie sich Probleme mit Rundungsfehlern bei algebraischen Fragestellungen auswirken können, wollen wir zum Abschluß dieses Paragraphen ein Beispiel aus WILKINSONs Buch betrachten. Er geht aus vom Polynom zwanzigsten Grades

$$f = (X-1)(X-2)(X-3)\cdots(X-18)(X-19)(X-20)$$

mit den Nullstellen $1, 2, \ldots, 20$. In ausmultiplizierter Form würde es mehrere Zeilen benötigen: Der größte Koeffizient, der von X^2, hat

zwanzig Dezimalstellen, und die meisten anderen haben nicht viel weniger.

Der Koeffizient von X^{19} ist allerdings noch überschaubar: Wie man sich leicht überlegt, ist er gleich der negativen Summe der Zahlen von eins bis zwanzig, also -210.

WILKINSON stört nun diesen Koeffizienten um einen kleinen Betrag und berechnet die Nullstellen des so modifizierten Polynoms. Betrachten wir etwa die Nullstellen von $g = f - 10^{-9}X^{19}$. Wir ersetzen in f also den Koeffizienten -210 durch $-210{,}000000001$. Die neuen Nullstellen sind, auf fünf Nachkommastellen gerundet,

$$\begin{aligned}
&1{,}0000, \quad 2{,}0000, \quad 3{,}0000, \quad 4{,}0000, \quad 5{,}0000,\\
&6{,}0000, \quad 7{,}0000, \quad 8{,}0001, \quad 8{,}9992, \quad 10{,}008,\\
&10{,}957, \quad 12{,}383 \pm 0{,}10867i, \quad 14{,}374 \pm 0{,}77316i,\\
&16{,}572 \pm 0{,}88332i, \quad 18{,}670 \pm 0{,}35064i, \quad 20{,}039\,.
\end{aligned}$$

Durch kleinste Veränderungen an einem einzigen Koeffizienten, wie sie beispielsweise jederzeit durch Rundungen entstehen können, kann sich also selbst das qualitative Bild ändern: Hier etwa reduziert sich die Anzahl der (für viele Anwendungen einzig relevanten) reellen Nullstellen von zwanzig auf zwölf. Schon wenn wir verläßliche Aussagen über die Anzahl reeller Nullstellen brauchen, können wir uns also nicht allein auf numerische Berechnungen verlassen, sondern brauchen alternative Methoden wie zum Beispiel explizite Lösungsformeln, mit denen wir auch theoretisch arbeiten können.

§3: Unentscheidbarkeitsprobleme

Ein auch nur moderat komplizierter symbolischer Ausdruck läßt sich praktisch immer auf eine Vielzahl von Arten darstellen, die teils offensichtlich gleich sind, teils aber auch auf den ersten Blick nichts miteinander zu tun haben. Einige Beispiele:

$$\frac{10}{15} = \frac{2}{3}\,, \quad \sqrt{8} = 2\sqrt{2}\,, \quad \sqrt{4+2\sqrt{3}} = 1+\sqrt{3}$$

$$(a+b)^2 = a^2 + 2ab + b^2\,,\quad \frac{X^5-1}{X-1} = 1 + X + X^2 + X^3 + X^4\,,$$

$$X^5 - 15X^4 + 85X^3 - 225X^2 + 274X - 120$$
$$= (X-1)(X-2)(X-3)(X-4)(X-5)\,,$$

$$\sin x \cos x = \frac{\sin 2x}{2}\,,\quad 1 + \tan^2 x = \frac{1}{\cos^2 x}$$

Nur in wenigen dieser Fälle ist eine der beiden Darstellung für alle Arten von Anwendungen der anderen vorzuziehen; meist hat mal die eine, mal die andere Form ihre Vorteile.

Andererseits gehört es zu den Grundaufgaben jeglicher Art des Rechnens, daß man entscheiden muß, ob zwei Ausdrücke gleich sind. Dies ist dann am einfachsten, wenn jeder Ausdruck intern durch eine eindeutig bestimmte kanonische Form dargestellt wird. In einem System, daß alle Ergebnisse auf eine solche kanonische Form bringt, lassen sich zwei Ausdrücke einfach dadurch auf Gleichheit testen, daß man ihre Differenz berechnet; die Ausdrücke sind genau dann gleich, wenn das Ergebnis die kanonische Darstellung der Null ist.

Gegen eine solche Darstellung sprechen sowohl theoretische als auch praktische Gründe: Wenn beispielsweise Polynome stets in ausmultiplizierter Form dargestellt werden, läuft man Gefahr, ein als Produkt von Linearfaktoren gegebenes Polynom zunächst auszumultiplizieren, um dann anschließend mit großer Mühe seine Nullstellen zu bestimmen. Stellt man Polynome dagegen in faktorisierter Form da, so kann es passieren, daß ein als Summe von Potenzen gegebenes Polynom zunächst mit großem Aufwand faktorisiert wird, und wir anschließend beispielsweise eine Stammfunktion suchen, wofür diese Faktorisierung wieder rückgängig gemacht werden muß. Das Ergebnis müßte dann wieder faktorisiert werden, wobei je nach Wahl der Integrationskonstanten sehr verschiedene Ergebnisse entstehen können.

In älteren Computeralgebrasystemen wie REDUCE war es üblich, alles auszumultiplizieren; in den heute gebräuchlichen Systemen wie MAPLE und MATHEMATICA werden Umformungen nur noch durchgeführt, wenn es entweder für die jeweilige Rechnung notwendig ist (Zur Berechnung der Stammfunktion eines Polynoms muß dieses in ausmultiplizierter

Form vorliegen) oder wenn es der Anwender explizit verlangt. Lediglich in einigen offensichtlichen Fällen bemühen sich auch diese Systeme um Normalisierung: Beispielsweise werden Brüche stets in gekürzter Form dargestellt und bei Summen werden gleichartige Terme zusammengefaßt.

Das theoretische Argument gegen kanonische Darstellungen ist, daß es solche Darstellungen nur für sehr eingeschränkte Klassen von Zahlen und Funktionen gibt: Wie wir gleich sehen werden, ist selbst für reelle Zahlen im allgemeinen unentscheidbar, wann zwei auf unterschiedliche Weise dargestellte Zahlen gleich sind.

Dieses negative Ergebnis kam hat seinen Ausgangspunkt in einem positiv formulierten Problem von DAVID HILBERT. Dieser stellte auf dem Internationalen Mathematikerkongress 1900 in Paris 23 Probleme vor, von denen er glaubte, daß sie für die Mathematik des 20. Jahrhunderts wichtig sein sollten. Die Probleme kamen aus allen Teilgebieten der Mathematik und hatten auch sehr unterschiedlichen Schwierigkeitsgrad: Einige wurden schon sehr bald gelöst, andere sind auch ein Jahrhundert später noch ungelöst. Das zehnte Problem lautete:

Man gebe ein Verfahren an, das für eine beliebige diophantische Gleichung entscheidet, ob sie lösbar ist.

Wie sich zeigte, war HILBERT hier zu optimistisch: 1970 bewies YURI V. MATIYASEVICH, daß es kein solches Verfahren geben kann, da sich jedes sogenannte rekursiv aufzählbare Problem auf die Frage nach der Lösbarkeit eines diophantischen Gleichung zurückführen läßt. Da zu den rekursiv aufzählbaren Problem auch unlösbare wie das Halteproblem für TURING-Maschinen gehören, folgte daraus die Unmöglichkeit des von HILBERT geforderten Verfahrens.

Da reelle Zahlen $x_1, \ldots, x_n$ genau dann ganz sind, wenn $\sum_{i=1}^{n} \sin^2 \pi x_i$ verschwindet, übersetzte DANIEL RICHARDSON dies in den folgenden Unmöglichkeitssatz für reeller Zahlen:

Satz von Richardson: Es gibt kein Verfahren, das in endlich vielen Schritten entscheidet, ob ein beliebig vorgegebener Ausdruck bestehend

aus rationalen Zahlen, π, einer Variablen x sowie den Funktionen +, $\cdot$, Sinus und Betrag gleich Null ist.

Tatsächlich bewies RICHARDSON ein etwas schwächeres Resultat, denn seine Arbeit erschien bereits 1969, also ein Jahr vor der von MATIYASEVICH, so daß er nur ein schwächeres Resultat verwenden konnte. Zusammen mit dem Resultat von MATIYASEVICH zeigt seine Methode aber sofort den angegebenen Satz.

Mehr zum zehnten HILBERTschen Problem und seinen Konsequenzen findet man bei

YURI V. MATIYASEVICH: Hilbert's Tenth Problem, *MIT Press,* 1993

Kapitel 1
Polynomgleichungen in einer Veränderlichen

Wir beginnen mit dem klassischen Grundproblem der Algebra, dem Lösen von Polynomgleichungen in einer Variablen, d.h. Gleichungen der Form

$$a_d x^d + a_{d-1} x^{d-1} + \cdots + a_1 x + a_0 = 0 \,.$$

Über den Zahlbereich, in dem die Koeffizienten liegen, wollen wir uns dabei im Augenblick noch keine großen Gedanken machen. In den meisten Beispielen werden die Koeffizienten ganze, rationale, reelle oder komplexe Zahlen sein; der Zahlbereich könnte aber auch einfach irgendein Körper oder Ring sein. Um Lösungen zu finden, müssen wir oft auch in einem größeren Zahlbereich suchen: Die Gleichung $2x - 3 = 0$ hat beispielsweise ganzzahlige Koeffizienten, aber keine ganzzahligen Lösungen, sondern nur die rationale Lösung $x = \frac{2}{3}$.

Wir werden stets annehmen, daß a_d nicht verschwindet und bezeichnen dann d als den *Grad* der Gleichung. Gleichungen vom Grad eins oder lineare Gleichungen sind problemlos zu lösen: Da gemäß unserer Annahme in $ax + b = 0$ der Koeffizient a von x nicht verschwindet, können wir (eventuell erst nach Übergang zu einem größeren Zahlbereich) b auf die andere Seite bringen (je nach Vorzeichen von b ist das *al-dschabr* oder *al-muqābala*) und dann beide Seiten durch a dividieren, um die Lösung $x = -\frac{b}{a}$ zu erhalten.

§1: Quadratische Gleichungen

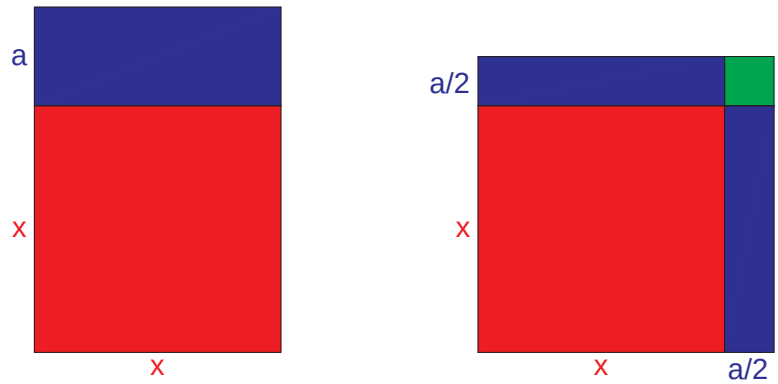
Gleichungen vom Grad zwei werden üblicherweise als quadratische Gleichungen bezeichnet; Verfahren zur ihrer Lösung waren in allen

frühen Hochkulturen bekannt. Die ältesten erhaltenen Hinweise deuten darauf hin, daß die Babylonier schon vor rund vier Jahrtausenden damit vertraut waren.

Der Ansatz zur Lösung der Gleichung $x^2 + ax = b$ läßt sich am einfachsten geometrisch verstehen: Wir suchen nach einem Quadrat mit unbekannter Seitenlänge x derart, daß die Fläche des Quadrats zusammen mit der des Rechtecks mit Seiten x und a gleich b ist.

Die linke unter den beiden folgenden Zeichnungen zeigt dieses Quadrat und darüber das Rechteck; auf der rechten Seite ist die Hälfte des Rechtecks neben das Quadrat gewandert, so daß abgesehen von dem kleinen Quadrat rechts oben nun ein Quadrat mit Seitenlänge $x + \frac{a}{2}$ entstanden ist. Die Größe des kleinen Quadrats ist bekannt: Seine Seitenlänge ist $\frac{a}{2}$. Wir suchen somit eine Zahl x derart, daß das Quadrat mit Seitenlänge $x+\frac{a}{2}$ die Fläche $b+\frac{a^2}{4}$ hat; das Problem ist also zurückgeführt auf das Ziehen einer Quadratwurzel:

$$x = -\frac{a}{2} \pm \sqrt{b + \frac{a^2}{4}}\,.$$



(Eine ähnliche Zeichnung befindet sich übrigens auch im Buch von AL-CHWĀRIZMĪ; er teilt das Rechteck mit Seiten a und x allerdings auf in vier Rechtecke mit Seiten $a/4$ und x und setzt diese an die vier Seiten des Quadrats. Das gibt eine etwas schönere Zeichnung, dafür muß er vier Quadrate mit Seitenlänge $a/4$ hinzufügen, um auf ein Quadrat mit Seitenlänge $x + a/2$ zu kommen.)

Wie die Babylonier auf diese Lösungsformel kamen, ist nicht bekannt; in den überlieferten Schriften wird nur der fertige Lösungsweg anhand von

Beispielen präsentiert. Sie wußten aber auf jeden Fall, daß die Summe der Lösungen der Gleichung $x^2 - ax + b = 0$ gleich a ist und ihr Produkt gleich b. In der Tat: Sind x_1 und x_2 die beiden Nullstellen des Polynoms $X^2 - aX + b$, so ist

$$X^2 - aX + b = (X - x_1)(X - x_2) = X^2 - (x_1 + x_2)X + x_1x_2\,,$$

woraus die Behauptung durch Koeffizientenvergleich folgt. (Ein formaler Beweis für das erste Gleichheitszeichen folgt, in allgemeinerem Zusammenhang, in §4.)

Damit ist das Lösen der Gleichung $x^2 - ax + b = 0$ äquivalent dazu, zwei Zahlen x_1 und x_2 zu finden mit

$$x_1 + x_2 = a \quad \text{und} \quad x_1x_2 = b\,.$$

Die führt zu einer alternativen Herleitung der Lösungsformel: Wir machen den Ansatz

$$x_{1/2} = \frac{a}{2} \pm u$$

mit einer neuen Unbekannten u. Damit ist die erste Gleichung automatisch erfüllt. Für die zweite erhalten wir nach der den Babyloniern bekannten dritten binomischen Formel

$$b = x_1x_2 = \left(\frac{a}{2} + u\right)\left(\frac{a}{2} - u\right) = \frac{a^2}{4} - u^2\,, \text{ also } \quad u = \sqrt{\frac{a^2}{4} - b}\,.$$

Somit ist $x_{1/2} = \dfrac{a}{2} \pm \sqrt{\dfrac{a^2}{4} - b}$. (Vor $\frac{a}{2}$ steht hier kein Minuszeichen, weil in der Gleichung eines vor a steht.)

Falls $\frac{a^2}{4} - b > 0$ ist, liefert uns das zwei reelle Lösungen; falls der Radikand Null ist, fallen beide zusammen.

Für die Babylonier, die ihre Mathematik benutzten, um Größen aus der realen Welt zu berechnen, ging es nur um reelle Lösungen; heute interessieren wir uns auch für Gleichungen, bei denen unter der Wurzel eine negative oder sogar eine komplexe Zahl steht. Im Falle einer negativen Zahl ist die Wurzel rein imaginär und somit problemlos; für eine komplexe Zahl allerdings stellt sich die Frage, wie wir die Wurzel aus

$c + di$ mit $c, d \in \mathbb{R}$ und $d \neq 0$ in der Form $u + iv$ mit $u, v \in \mathbb{R}$ darstellen können. Die Gleichung

$$(u + iv)^2 = u^2 - v^2 + 2iuv = c + id$$

führt auf die beiden reellen Gleichungen

$$u^2 - v^2 = c \quad \text{und} \quad 2uv = d\,.$$

Wegen $d \neq 0$ können auch u und v nicht verschwinden; daher können wir die zweite Gleichung umformen zu $v = d/2u$ und das in die erste Gleichung einsetzen:

$$u^2 - \frac{d^2}{4u^2} = c\,.$$

Multiplikation mit $4u^2$ macht daraus

$$4u^4 - d^2 = 4cu^2 \quad \text{oder} \quad u^4 - cu^2 - \frac{d^2}{4} = 0\,.$$

Dies ist eine quadratische Gleichung für u^2 mit den beiden Lösungen

$$u^2 = \frac{c}{2} \pm \frac{1}{2}\sqrt{c^2 + d^2}\,.$$

Als Quadrat einer reellen Zahl muß $u^2 \geq 0$ sein; die Lösung mit dem Minuszeichen kommt daher nicht in Frage: Für negatives $c \in \mathbb{R}$ ist sie offensichtlich negativ, und für positives c auch, denn wegen $d \neq 0$ ist $\sqrt{c^2 + d^2}$ größer als der Betrag von c. Somit sind

$$u = \pm\sqrt{\frac{c}{2} + \frac{1}{2}\sqrt{c^2 + d^2}} \quad \text{und} \quad v = \frac{d}{2u}$$

problemlos berechenbar.

§2: Kubische Gleichungen

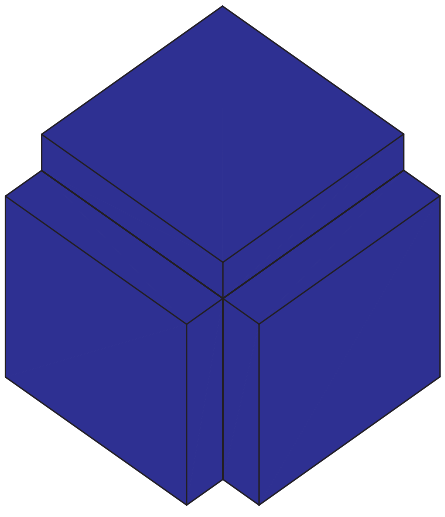
Geometrische Probleme, die (analytisch betrachtet) auf kubische Gleichungen führen, waren bereits den Griechen bekannt; ein Beispiel ist etwa die Dreiteilung eines beliebigen Winkels oder das Problem der Würfelverdoppelung. Verfahren zur Lösung solcher Probleme waren

durchaus bekannt, allerdings keine, die mit Zirkel und Lineal auskommen. Da alles außer diesen beiden einfachsten Hilfsmitteln als unrein galt, spielten solche Verfahren in der griechischen Mathematik nur eine sehr untergeordnete Rolle und fanden kein weites Interesse.

In der Algebra tauchten vor gut 500 Jahren, genauer ab 1515, Verfahren zur Lösung spezielle kubischer Gleichungen auf.

Wenn wir versuchen, für die Gleichungen $x^3 + ax^2 = b$ eine ähnlich Strategie zu finden wie im Fall der Gleichung $x^2 + ax = b$, müssen wir ins Dreidimensionale gehen und auf den Würfel mit Kantenlänge x eine quadratische Säule mit Basisquadrat der Seitenlänge x und Höhe a stellen. Um sie so zu verteilen, daß wir möglichst nahe an einen neuen Würfel kommen, müssen wir jeweils ein Drittel davon auf drei der Seitenflächen des Würfels platzieren.

Leider fehlt hier nun nicht nur ein Würfel der Kantenlänge $\frac{a}{3}$, sondern auch noch drei quadratische Säulen der Höhe x auf Grundflächen mit Seitenlänge $\frac{a}{3}$. Wir können das Volumen des Würfels mit Seitenlänge $x + \frac{a}{3}$ also nicht einfach durch die bekannten Größen a, b ausdrücken, sondern haben auch noch einen Term mit der Unbekannten x.



Trotzdem ist diese Idee nützlich, denn sie erlaubt es uns, in einer allgemeinen kubischen Gleichung den x^2-Term zu eliminieren. Wenn wir Gleichungen über einem Körper betrachten, können wir durch den höchsten Koeffizienten dividieren und erhalten dann eine Gleichung

der Form $x^3 + ax^2 + bx + c = 0$. Mit dem Ansatz $y = x + a/3$ können wir diese schreiben als

$$\left(y - \frac{a}{3}\right)^3 + a\left(y - \frac{a}{3}\right)^2 + b\left(y - \frac{a}{3}\right) + c$$
$$= y^3 + \left(b - \frac{a^2}{3}\right) y + c - \frac{ab}{3} + \frac{2a^3}{27} = 0\,.$$

Es reicht daher, wenn wir Gleichungen der Form

$$y^3 + py + q = 0$$

lösen können. Die Tatsache, daß die ersten Ansätze zur Lösung solcher Gleichungen erst im 16. Jahrhundert auftauchten, legt allerdings nahe, daß die Lösung wohl nicht ganz einfach sein wird.

Der Trick, der schließlich zum Erfolg führte, ist folgender: Wir schreiben y als Summe zweier neuer Zahlen u und v und machen dadurch das Problem auf den ersten Blick nur schwieriger. Andererseits ist diese Summendarstellung natürlich alles andere als eindeutig; wir können daher hoffen, daß es auch dann noch Lösungen gibt, wenn wir an u und v zusätzliche Forderungen stellen und dadurch das Problem vielleicht vereinfachen.

Einsetzen von $y = u + v$ führt auf die Bedingung

$$(u+v)^3 + p(u+v) + q = u^3 + 3u^2v + 3uv^2 + v^3 + p(u+v) + q = 0\,.$$

Dies können wir auch anders zusammenfassen als

$$(u^3 + v^3 + q) + (3uv + p)(u + v) = 0\,,$$

und natürlich verschwindet diese Summe insbesondere dann, wenn beide Summanden einzeln verschwinden. Falls es uns also gelingt, zwei Zahlen u, v zu finden mit

$$u^3 + v^3 = -q \quad \text{und} \quad 3uv = -p\,,$$

haben wir eine Lösung gefunden.

Zwei solche Zahlen u, v erfüllen erst recht die schwächere Bedingung

$$u^3 + v^3 = -q \quad \text{und} \quad u^3 \cdot v^3 = -\frac{p^3}{27}\,;$$

wir kennen also die Summe und das Produkt ihrer dritten Potenzen. Damit kennen wir aber, wie wir schon im Zusammenhang mit quadratischen Gleichungen gesehen haben, auch u^3 und v^3 als Lösungen der quadratischen Gleichung $x^2 + qx - \frac{1}{27}p^3 = 0$. Somit ist

$$u^3 = -\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}} = -\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}$$

$$\text{und} \qquad v^3 = -\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3},$$

wobei es auf die Reihenfolge natürlich nicht ankommt.

Damit kennen wir u^3 und v^3. Für u und v selbst gibt es dann jeweils drei Möglichkeiten, allerdings führen nicht alle neun Kombinationen dieser Möglichkeiten zu Lösungen, denn für eine Lösung muß ja die Bedingung $3uv = -p$ erfüllt sein, nicht nur $u^3 \cdot v^3 = -\frac{1}{27}p^3$.

Dies läßt sich am besten dadurch gewährleisten, daß wir für u *irgendeine* der drei Kubikwurzeln von u^3 nehmen und dann $v = -p/3u$ setzen. Die drei Lösungen der kubischen Gleichung $y^3 + py + q = 0$ sind also

$$y = u - \frac{p}{3u} \quad \text{mit} \quad u = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}},$$

wobei für u nacheinander jede der drei Kubikwurzeln eingesetzt werden muß. (Es spielt keine Rolle, welche der beiden Quadratwurzeln wir nehmen, denn ersetzen wir die eine durch die andere, vertauschen wir dadurch einfach u und v.)

Da selbst von den drei Kubikwurzeln einer reellen Zahl nur eine reell ist, müssen wir zur Bestimmung aller drei Lösungen einer kubischen Gleichung mit reellen Koeffizienten *immer* auch mit komplexen Zahlen rechnen, selbst wenn alle Lösungen reell sind.

Wenn wir eine Kubikwurzel w_0 einer komplexen Zahl kennen, lassen sich die beiden anderen leicht bestimmen: Ist w eine von ihnen, so ist $(w/w_0)^3 = 1$, d.h. w/w_0 ist eine Nullstelle des Polynoms $X^3 - 1$. Dieses hat die Eins als Nullstelle; wenn wir durch $X - 1$ dividieren erhalten wir

den Quotienten X^2+X+1, der nach der Lösungsformel für quadratische Gleichungen die beiden Nullstellen

$$\rho = -\frac{1}{2} + \frac{i}{2}\sqrt{3} \quad \text{und} \quad \bar{\rho} = -\frac{1}{2} - \frac{i}{2}\sqrt{3}$$

hat. Die beiden anderen Kubikwurzeln sind also $w_0\rho$ und $w_0\bar{\rho}$. Ihr Produkt $\rho\bar{\rho}$ ist $|\rho|^2 = 1$, d.h. die beiden sind invers zueinander, so daß die Division durch eine der beiden Wurzeln äquivalent ist zur Multiplikation mit der anderen. Ist in der Lösungsformel für die kubische Gleichung daher u irgendeine feste dritte Wurzel, so sind die drei Lösungen gleich

$$u - \frac{p}{3u}, \quad u\rho - \frac{p}{3u\rho} = u\rho - \frac{p}{3u}\bar{\rho} \quad \text{und} \quad u\bar{\rho} - \frac{p}{3u\bar{\rho}} = u\bar{\rho} - \frac{p}{3u}\rho\,.$$

Die erste Lösung einer kubischen Gleichung geht wohl aus SCIPIONE DEL FERRO (1465–1526) zurück, der von 1496 bis zu seinem Tod an der Universität Bologna lehrte. 1515 fand er eine Methode, um die Nullstellen von $x^3 + px = q$ für *positive* Werte von p und q zu bestimmen (Negative Zahlen waren damals in Europa noch nicht im Gebrauch). Er veröffentlichte diese jedoch nie, so daß NICCOLO FONTANA (1499–1557, oberes Bild), genannt TARTAGLIA (der Stotterer), dieselbe Methode 1535 noch einmal entdeckte und gleichzeitig auch noch eine Modifikation, um einen leicht verschiedenen Typ kubischer Gleichungen zu lösen. TARTAGLIA war mathematischer Autodidakt, war aber schnell als Fachmann anerkannt und konnte seinen Lebensunterhalt als Mathematiklehrer in Verona und Venedig verdienen.

Die Lösung allgemeiner kubischer Gleichungen geht auf den Mathematiker, Arzt und Naturforscher GIROLAMO CARDANO (1501–1576, unteres Bild) zurück, dem TARTAGLIA nach langem Drängen und unter dem Siegel der Verschwiegenheit seine Methode mitgeteilt hatte. LODOVICO FERRARI (1522–1565) kam 14-jährig als Diener zu CARDANO; als dieser merkte, daß FERRARI schreiben konnte, machte er ihn zu seinem Sekretär. 1540 fand FERRARI die Lösungsmethode für biquadratische Gleichungen; 1545 veröffentlichte CARDANO trotz seines Schweigeversprechens gegenüber TARTAGLIA die Lösungsmethoden für kubische und biquadratische Gleichungen in seinem Buch *Ars magna.*

Im sechzehnten Jahrhundert wurde das natürlich nicht so formuliert: Die mathematische Formelschreibweise führte schließlich erst VIÈTE einige Jahrzehnte später ein. TARTAGLIA, der die Lösungsmethode für die Gleichung $x^3 + px = q$ für positive Werte p, q fand, arbeite im übrigen auch nicht mit den Größen u und v, sondern mit deren dritten Potenzen. Er beschrieb seine Methode gegenüber CARDANO in einem Gedicht:

Quando chel cubo con le cose appresso
Se agguaglia à qualche numero discreto
Trouan dui altri differenti in esso.

Depoi terrai questo per consueto
Che'llor produtto sempre sta eguale
Al terzo cubo delle cose neto.

El residuo poi suo generale
Delli lor lati cubi ben sottratto
Varra la tua cosa principale.

Frei übersetzt: Wenn der Kubus zusammen mit dem Produkt mit einer Sache eine gewisse Zahl ergibt, drücke diese aus als eine Differenz zweier anderen. Danach stelle sicher, daß das Produkt dieser beiden immer gleich dem Kubus eines Drittels der Sache ist. Die Lösung ist dann die Differenz der Kubikwurzeln der beiden.

In heutiger mathematischer Sprechweise: Zur Lösung der Gleichung $x^3 + px = q$ schreibe q als eine Differenz $q = U - V$. Stelle sicher, daß $UV = \left(\frac{p}{3}\right)^3$ ist. Dann ist $x = \sqrt[3]{U} - \sqrt[3]{V}$.

Betrachten wir als einfaches Beispiel die Gleichung

$$(x-1)(x-2)(x-3) = x^3 - 6x^2 + 11x - 6 = 0\,;$$

sie hat nach Konstruktion die drei Lösungen $1, 2$ und 3.

Falls wir das nicht wüßten, würden wir als erstes durch die Substitution $y = x - 2$ den quadratischen Term eliminieren. Einsetzen von $x = y + 2$ liefert

$$(y+2)^3 - 6(y+2)^2 + 11(y+2) - 6$$
$$= y^3 + 6y^2 + 12y + 8 - 6y^2 - 24y - 24 + 11y + 22 - 6 = y^3 - y\,,$$

wir müssen also zunächst die Gleichung $y^3 - y = 0$ lösen. Hierzu brauchen wir selbstverständlich keine Lösungstheorie kubischer Gleichungen: Ausklammern von y und die dritte binomische Formel zeigen sofort, daß

$$y^3 - y = y(y^2 - 1) = y(y+1)(y-1)$$

genau an den Stellen $y = -1, 0, 1$ verschwindet, und da $x = y + 2$ ist, hat die Ausgangsgleichung die Lösungen $x = 1, 2, 3$.

Wenden wir trotzdem unsere Lösungsformel an: Bei dieser Gleichung ist $p = -1$ und $q = 0$, also

$$u_1 = \sqrt[3]{\sqrt{\frac{-1}{27}}} = \sqrt[6]{\frac{-1}{27}} = \sqrt{\frac{-1}{3}}$$

für die rein imaginäre Kubikwurzel. Das zugehörige v_1 muß die Gleichung $u_1 v_1 = \frac{1}{3}$ erfüllen, also ist $v_1 = -u_1$, und wir erhalten als erste Lösung $y_1 = u_1 + v_1 = 0$.

Die beiden anderen Kubikwurzeln erhalten wir, indem wir die bekannte Kubikwurzel mit einer der beiden komplexen dritten Einheitswurzeln multiplizieren, d.h. also mit ρ und mit $\bar{\rho}$.

$$u_2 = \sqrt{\frac{-1}{3}}\rho = \frac{\sqrt{3}}{3} i \left(-\frac{1}{2} + \frac{\sqrt{3}\, i}{2} \right) = -\frac{1}{2} - \frac{\sqrt{3}}{6} i$$

und

$$v_2 = \frac{1}{3u_2} = \frac{-2}{3 + \sqrt{3}\, i} = \frac{-2(3 - \sqrt{3}\, i)}{3^2 + (\sqrt{3})^2} = -\frac{1}{2} + \frac{\sqrt{3}}{6} i\,;$$

wir erhalten somit die Lösung $y_2 = u_2 + v_2 = -1$.

Die dritte Kubikwurzel

$$u_3 = \sqrt{\frac{-1}{3}}\bar{\rho} = \frac{\sqrt{3}}{3} i \left(-\frac{1}{2} - \frac{\sqrt{3}\, i}{2} \right) = \frac{1}{2} - \frac{\sqrt{3}}{6} i$$

schließlich führt auf

$$v_3 = \frac{1}{3u_3} = \frac{2}{3 - \sqrt{3}\, i} = \frac{2(3 + \sqrt{3}\, i)}{3^2 + (\sqrt{3})^2} = \frac{1}{2} + \frac{\sqrt{3}}{6} i$$

und liefert so die Lösung $y_3 = u_3 + v_3 = 1$.

Etwas komplizierter wird es bei der Gleichung

$$x^3 - 7x + 6 = 0\,.$$

Da sie keinen x^2-Term hat, können wir gleich $p = -7$ und $q = 6$ in die Formel einsetzen und erhalten

$$u = \sqrt[3]{-3 + \sqrt{\frac{6^2}{4} - \frac{7^3}{27}}} = \sqrt[3]{-3 + \sqrt{-\frac{400}{4 \cdot 27}}} = \sqrt[3]{-3 + \frac{10}{9}\sqrt{3}\,i}\,.$$

Was nun? Wenn wir einen Ansatz der Form $u = r + is$ machen, kommen wir auf ein System von zwei kubischen Gleichungen in zwei Unbekannten, also ein schwierigeres Problem als unsere Ausgangsgleichung.

Eine Alternative ist die Polarkoordinatendarstellung komplexer Zahlen: Eine komplexe Zahl $z = x + iy$ läßt sich bekanntlich auch darstellen als $z = re^{i\varphi}$ mit $r = |z| = \sqrt{x^2 + y^2}$ und $x = r\cos\varphi$, $y = r\sin\varphi$. Da $e^{i\varphi} \cdot e^{i\psi} = e^{i(\varphi+\psi)}$ ist, werden bei der Multiplikation zweier komplexer Zahlen in Polarkoordinatendarstellung die Beträge miteinander multipliziert und die Winkel addiert. Daher ist $\sqrt[3]{|z|}\big(\cos\frac{\varphi}{3} + i\sin\frac{\varphi}{3}\big)$ eine dritte Wurzel von z. Leider gibt es aber keine einfache Formel, die Sinus und Kosinus von $\frac{\varphi}{3}$ durch $\cos\varphi$ und $\sin\varphi$ ausdrückt. Aus den Additionstheoremen können wir uns natürlich leicht Formeln für $\cos 3\varphi$ verschaffen; wir erhalten

$$\cos 3\varphi = 4\cos^3\varphi - 3\cos\varphi\,.$$

Um $x = \cos\frac{\varphi}{3}$ zu berechnen, müssen wir also die kubische Gleichung $4x^3 - 3x = \cos\varphi$ lösen, was uns wiederum auf die Berechnung einer Kubikwurzel führt, *usw.*

Trotzdem ist die obige Darstellung der Lösung nicht völlig nutzlos: Sie gibt uns immerhin Formeln für den Real- und den Imaginärteil der Lösung, und diese Formeln können wir numerisch auswerten.

Für den hier interessierenden Radikanden $z = -3 + \frac{10}{9}\sqrt{3}\,i$ ist

$$|z| = \sqrt{(-3)^2 + \left(\frac{10}{9}\sqrt{3}\right)^2} = \sqrt{\frac{9 + 300}{81}} = \sqrt{\frac{9 \cdot 27 + 100}{27}}$$
$$= \sqrt{\frac{343}{27}} = \sqrt{\frac{7^3}{3^3}} = \frac{7}{9}\sqrt{21}\,.$$

Somit ist

$$\cos\varphi = \frac{x}{|z|} = -\frac{9}{49}\sqrt{21} \approx -0{,}84169975767$$

und

$$\sin\varphi = \frac{y}{|z|} = \frac{10}{49}\sqrt{7} \approx 0{,}5399492473\,.$$

Der Arkuskosinus des ersten Werts ist ungefähr 2,571215844, der Arkussinus des zweiten 0,5703768102. Wenn wir φ im Intervall $(-\pi,\ \pi]$ suchen, folgt aus der Negativität von $\cos\varphi$, daß $|\varphi| > \frac{\pi}{2}$ sein muß; daher ist φ ungefähr gleich dem ersten der beiden Werte. (Der zweite ist natürlich $\pi - \varphi$, was den gleichen Sinus hat.)

Damit können wir eine dritte Wurzel näherungsweise bestimmen; wir erhalten

$$u_1 = \sqrt[3]{|z|}\left(\cos\frac{\varphi}{3} + i\sin\frac{\varphi}{3}\right) \approx 0{,}9999999994 + 1{,}154700538i$$

(Je nach Taschenrechner oder Programm kann das Ergebnis auch leicht verschieden sein.) und damit als erste Lösung

$$x_1 = u_1 + \frac{7}{3u_1} \approx 1{,}9999999999 - 10^{-9}i\,.$$

Wie jedes numerische Ergebnis stimmt diese Zahl natürlich nur näherungsweise und hängt im übrigen auch sowohl von der Stellenzahl als auch der Rundung ab. Zumindest in diesem Fall ist die Hypothese, daß es sich hier um eine durch Rundungsfehler verfälschte Zwei handeln könnte, eine Überlegung wert. Einsetzen zeigt, daß die Zwei tatsächlich eine Lösung ist. Für die beiden anderen dividieren am besten das Polynom durch $X-2$ und lösen dann die Quotientengleichung $x^2+2x-3 = 0$.

Obwohl die drei Lösungen $1, 2$ und -3 unserer Gleichung allesamt ganzzahlig sind, konnten wir dies also durch bloßes Einsetzen in unsere Formel nicht erkennen und konnten insbesondere die Kubikwurzel nur durch Erraten und Nachprüfen in einer einfachen Form darstellen.

In manchen Fällen ist die Anwendung der Lösungsformel auch völlig problemlos. Für die Gleichung $x^3 + 6x + 6 = 0$ mit $p = -6$ und $q = 6$

erhalten wir

$$u_1 = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} = \sqrt[3]{-3 + \sqrt{9-8}} = \sqrt[3]{-2} = -\sqrt[3]{2}$$

für die reelle Wurzel; die erste Lösung ist also

$$x_1 = u_1 - \frac{p}{3u_1} = -\sqrt[3]{2} - \frac{2}{\sqrt[3]{2}} = -\sqrt[3]{2} - \sqrt[3]{4}\,.$$

Für die zweite und dritte Lösung müssen wir mit $u_2 = u_1\rho$ *bzw.* $u_3 = u_1\bar{\rho}$ anstelle von u_1 arbeiten und erhalten

$$x_2 = -\sqrt[3]{2}\,\rho - \frac{2}{\sqrt[3]{2}\,\rho} = -\sqrt[3]{2}\,\rho - \sqrt[3]{4}\,\bar{\rho} \qquad \text{und}$$

$$x_3 = -\sqrt[3]{2}\,\bar{\rho} - \frac{2}{\sqrt[3]{2}\,\bar{\rho}} = -\sqrt[3]{2}\,\bar{\rho} - \sqrt[3]{4}\,\rho\,,$$

was nach Einsetzen von $\rho = -\frac{1}{2} + \frac{1}{2}\sqrt{3}i$ und $\bar{\rho} = -\frac{1}{2} - \frac{1}{2}\sqrt{3}i$ auf die beiden komplexen Lösungen

$$x_{2/3} = -\sqrt[3]{2}\left(-\frac{1}{2} \pm \frac{\sqrt{3}}{2}i\right) - \sqrt[3]{4}\left(-\frac{1}{2} \mp \frac{\sqrt{3}}{2}i\right)$$

$$= \frac{\sqrt[3]{2} + 3\sqrt[3]{4}}{2} \pm \frac{\sqrt{3}\left(\sqrt[3]{2} - \sqrt[3]{4}\right)}{2}i$$

führt.

Leider kann es aber auch in Fällen, in denen wir eine reelle Kubikwurzel finden können, gelegentlich Schwierigkeiten geben. Betrachten wir etwa die Gleichung

$$x^3 - 3x^2 + 9x + 13 = 0\,.$$

Hier setzen wir $x = y + 1$ und erhalten die neue Gleichung

$$(y+1)^3 - 3(y+1)^2 + 9(y+1) + 13$$

$$= y^3 + 3y^2 + 3y + 1 - 3(y^2 + 2y + 1) + 9y + 9 + 13$$

$$= y^3 + 6y + 20 = 0$$

mit $p = 6$ und $q = 20$. Damit ist $\frac{p}{3} = 2$ und $\frac{q}{2} = 10$, also

$$u = \sqrt[3]{-10 + \sqrt{100+8}} = \sqrt[3]{-10 + \sqrt{108}} = \sqrt[3]{-10 + 6\sqrt{3}}$$

Da 108 größer ist als $(-10)^2 = 100$, gibt es eine positive reelle Wurzel u_1; wir rechnen zunächst mit dieser und erhalten als erste Lösung

$$y_1 = u_1 - \frac{p}{3u_1} = \sqrt[3]{-10+6\sqrt{3}} - \frac{2}{\sqrt[3]{-10+6\sqrt{3}}}\,.$$

Damit haben wir im Prinzip eine Lösung gefunden. Wenn wir sie allerdings numerisch auswerten, erhalten wir etwas wie -1,99999998, und damit drängt sich natürlich der Verdacht auf, daß dies gleich -2 sein könnte. Einsetzen von $y = -2$ in unsere kubische Gleichung zeigt in der Tat, daß

$$(-2)^3 + 6\cdot(-2) + 20 = -8 - 12 + 20 = 0$$

ist. Aber warum ist

$$\sqrt[3]{-10+6\sqrt{3}} - \frac{2}{\sqrt[3]{-10+6\sqrt{3}}} = -2\,,$$

und wie, vor allem, kann man das der linken Seite ansehen?

Wie die Erfahrung der Computeralgebra zeigt, kann es extrem schwierig sein, auch nur zu entscheiden, ob zwei Wurzelausdrücke gleich sind; direkte allgemeine Verfahren dazu gibt es nicht. Unsere Formel gibt uns daher zwar immer drei Wurzelausdrücke, die Lösungen der gegebenen Gleichung sind, aber diese können für Zahlen stehen, die sich auch sehr viel einfacher ausdrücken lassen.

Im vorliegenden Fall, wo die numerische Berechnung eine Vermutung nahelegt, können wir wieder versuchen, diese zu beweisen: Aus der vermuteten Gleichung

$$u_1 - \frac{2}{u_1} = -2 \quad \text{folgt} \quad u_1^2 - 2 = -2u_1\,.$$

Quadratische Ergänzung macht daraus $(u_1+1)^2 = 3$, also ist u_1 eine der beiden Zahlen $-1 \pm \sqrt{3}$. Die dritte Potenz davon ist

$$(-1\pm\sqrt{3})^3 = -1 \pm 3\sqrt{3} - 3\cdot 3 \pm 3\sqrt{3} = -10 \pm 6\sqrt{3}\,,$$

also ist tatsächlich $u_1 = -1 + \sqrt{3}$ und

$$y_1 = -1 + \sqrt{3} - \frac{2}{-1+\sqrt{3}} = -1 + \sqrt{3} - \frac{2(-1-\sqrt{3})}{(-1+\sqrt{3})(-1-\sqrt{3})}$$
$$= -1 + \sqrt{3} + \frac{2+2\sqrt{3}}{-2} = -2\,.$$

Nachdem wir u_1 in einfacher Form ausgedrückt haben, lassen sich auch die anderen beiden Lösungen berechnen:

$$u_2 = u_1\rho = (-1+\sqrt{3}) \cdot \frac{-1+\sqrt{3}\,i}{2} = \frac{(1-\sqrt{3})+(3-\sqrt{3})i}{2}$$

und

$$u_3 = u_1\bar{\rho} = (-1+\sqrt{3}) \cdot \frac{-1-\sqrt{3}\,i}{2} = \frac{(1-\sqrt{3})-(3-\sqrt{3})i}{2}$$

Damit ist

$$\frac{2}{u_2} = \frac{4\big((1-\sqrt{3})-(3-\sqrt{3})i\big)}{(1-\sqrt{3})^2+(3-\sqrt{3})^2} = \frac{4\big((1-\sqrt{3})-(3-\sqrt{3})i\big)}{16-8\sqrt{3}}$$
$$= \frac{\big((1-\sqrt{3})-(3-\sqrt{3})i\big)(2+\sqrt{3})}{2(2-\sqrt{3})(2+\sqrt{3})} = \frac{-(1+\sqrt{3})-(3+\sqrt{3})i}{2}\,,$$

also

$$y_2 = u_2 - \frac{2}{u_2} = \frac{(1-\sqrt{3})+(3-\sqrt{3})i}{2} + \frac{(1+\sqrt{3})+(3+\sqrt{3})i}{2} = 1+3i\,.$$

Entsprechend folgt $y_3 = u_3 - \dfrac{2}{u_3} = 1 - 3i$.

Die Mathematiker des sechzehnten Jahrhunderts, auf die die Lösungsformel für kubische Gleichungen zurückgeht, hatten natürlich weder Computer noch Taschenrechner noch komplexe Zahlen; auch Dezimalbrüche in ihrer heutigen Form kamen erst im siebzehnten Jahrhundert auf, als die ersten Tafeln trigonometrischer Funktionen und kurz später auch Logarithmen veröffentlicht wurden. Doch auch ohne diese Hilfsmittel konnten sie erstaunlich gut mit der Lösungsformel umgehen. In §3.2 des Buchs

TEO MORA: Solving Polynomial Equation Systems I: The Kronecker-Duval Philosophy, *Cambridge University Press,* 2003

sind zwei Beispiele für ihre Vorgehensweise zu finden:

Bei der Gleichung $x^3 + 3x - 14 = 0$ ist $p = 3$ und $q = -14$, also

$$u = \sqrt[3]{7 + \sqrt{7^2 + 1^3}} = \sqrt[3]{7 + 5\sqrt{2}}\,.$$

Beim vorigen Beispiel hatten wir gesehen, daß

$$\sqrt[3]{-10 + 6\sqrt{3}} = -1 + \sqrt{3}$$

ist; eine Zahl der Form $a + b\sqrt{3}$ hat in diesem speziellen Fall also eine Kubikwurzel derselben Form $c+d\sqrt{3}$. Das gilt natürlich nicht allgemein; die Kubikwurzel aus $\sqrt{3}$, die sechste Wurzel von drei also, läßt sich sicher nicht in der Form $c + d\sqrt{3}$ mit ganzen Zahlen c und d schreiben. Trotzdem können wir unser Glück versuchen.

Für den Radikanden $7 + 5\sqrt{2}$ machen wir natürlich einen Ansatz der Form $c + d\sqrt{2}$. Wir wollen, daß

$$(c + d\sqrt{2})^3 = c^3 + 3c^2 d\sqrt{2} + 6cd^2 + 2d^3\sqrt{2} = 7 + 5\sqrt{2}$$

ist mit $c, d \in \mathbb{Z}$, also

$$c^3 + 6cd^2 = 7 \quad \text{und} \quad 3c^2 d + 2d^3 = 5\,.$$

Damit haben wir, wie schon oben erwähnt, ein System von *zwei* kubischen Gleichungen anstelle von einer, jetzt allerdings suchen wir nur nach ganzzahligen Lösungen. Aus der ersten Gleichung können wir c ausklammern und erhalten $c(c^2 + 6d^2) = 7$. Somit muß c ein Teiler von sieben sein, d.h. $c = \pm 1$ oder $c = \pm 7$. Die negativen Zahlen scheiden aus, da die Klammer nicht negativ werden kann, und auch $c = 7$ ist nicht möglich, denn dann wäre die linke Seite mindestens gleich 7^3. Wenn es eine ganzzahlige Lösung gibt, muß daher $c = 1$ sein; durch Einsetzen folgt, daß dann mit $d = \pm 1$ die erste Gleichung in der Tat erfüllt ist. Die zweite Gleichung $d(3c^2 + 2d^2) = 5$ zeigt, daß auch d positiv sein muß und $c = d = 1$ beide Gleichungen erfüllt. Somit ist

$$u = \sqrt[3]{7 + 5\sqrt{2}} = 1 + \sqrt{2}$$

für die reelle unter den drei Kubikwurzeln. Da wir eine Gleichung mit reellen Koeffizienten haben, muß auch das zugehörige v reell sein und kann genauso wie u bestimmt werden:

$$v = \sqrt[3]{7 - 5\sqrt{2}} = 1 - \sqrt{2} \quad \text{und} \quad x = u + v = 2\,.$$

Damit war die Gleichung für die Zwecke des sechzehnten Jahrhunderts gelöst, denn da es noch keine komplexen Zahlen gab, suchte auch niemand nach komplexen Lösungen.

Wir interessieren uns allerdings für komplexe Lösungen; die beiden noch fehlenden Lösungen können wir entweder berechnen als $u\rho + v\bar{\rho}$ und $u\bar{\rho} + v\rho$, oder aber wir dividieren das Polynom $X^3 + 3X - 14$ durch $X - 2$ und erhalten das quadratische Polynom $X^2 + 2X + 7$ mit den Nullstellen $-1 \pm \sqrt{6}\,i$.

Bei Gleichungen mit drei reellen Nullstellen führt die Lösungsformel, wie wir in §9 sehen werden, *immer* übers Komplexe, aber auch damit wurden CARDANO und seine Zeitgenossen fertig. MORA betrachtet als Beispiel dafür die Gleichung $x^3 - 21x - 20 = 0$. Hier ist

$$u = \sqrt[3]{10 + \sqrt{10^2 - 7^3}} = \sqrt[3]{10 + \sqrt{-243}} = \sqrt[3]{10 + 9\sqrt{-3}}\,.$$

$\sqrt{-3}$ war für CARDANO im Gegensatz zu $\sqrt{2}$ keine Zahl; trotzdem rechnete er damit als mit einem abstrakten Symbol gemäß der Regel $\sqrt{-3} \cdot \sqrt{-3} = -3$.

Wenn wir wieder auf unser Glück vertrauen und einen Ansatz der Form $u = c + d\sqrt{-3}$ machen, kommen wir auf das Gleichungssystem

$$c^3 - 9cd^2 = 10 \quad \text{und} \quad 3c^2d - 3d^3 = 9\,.$$

Ausklammern von c *bzw.* d und Kürzen der zweiten Gleichung durch drei führt auf

$$c(c^2 - 9d^2) = 10 \quad \text{und} \quad d(c^2 - d^2) = 3\,.$$

Wenn es ganzzahlige Lösungen gibt, muß wegen der zweiten Gleichung $d = \pm 1$ oder $d = \pm 3$ sein. $d = \pm 1$ führt auf $c^2 - 1 = \pm 3$, also $d = 1$ und $c = \pm 2$; für $d = \pm 3$ läßt sich kein ganzzahliges c finden. Einsetzen in die erste Gleichung zeigt, daß $c = -2, d = 1$ das System löst, also ist

$u_1 = -2 + \sqrt{-3}$ eine der drei Wurzeln. Die erste Lösung der kubischen Gleichung ist also

$$\begin{aligned} x_1 &= -2 + \sqrt{-3} + \frac{7}{-2 + \sqrt{-3}} \\ &= -2 + \sqrt{-3} + \frac{7(-2 + \sqrt{-3})}{(-2 + \sqrt{-3})(-2 - \sqrt{-3})} \\ &= -2 + \sqrt{-3} + \frac{-14 + 7\sqrt{-3}}{7} = -4\,. \end{aligned}$$

Zur Bestimmung der beiden anderen Lösungen haben wir verschiedene Möglichkeiten: Wir können das Polynom $X^3 - 21X - 20$ durch $X - x_1$, also $X + 4$, dividieren und damit das Problem auf die Lösung einer quadratischen Gleichung reduzieren, oder wir können die beiden weiteren Werte für u als $u_2 = u_1\rho$ und $u_3 = u_1\bar{\rho}$ berechnen Zur Zeit CARDANOs gab es keine dieser Möglichkeiten: Mit Polynomen rechnete man erst im achtzehnten Jahrhundert, und die komplexen Zahlen wurden sogar erst im neunzehnten Jahrhundert eingeführt. CARDANO rechnete zwar mit „Symbolen“ wie $\sqrt{-3}$, aber die einzige Lösung der Gleichung $x^3 = 1$ war für ihn – wie für alle seiner Zeitgenossen – die Eins.2s

Wie die obige Rechnung zeigte, sind $c = -2$ und $d = 1$ die einzigen ganzzahligen Lösungen der Gleichung $(c + d\sqrt{-3})^3 = 10 + 9\sqrt{-3}$. Vielleicht gibt es aber weitere Lösungen, wenn wir für c und d auch rationale Zahlen zulassen. Nun haben wir allerdings bei der Suche nach ganzzahligen Lösungen viel mit Teilbarkeit argumentiert, und im Körper der rationalen Zahlen ist jedes Element durch jedes andere außer der Null teilbar. Wir müssen uns daher auf Zahlen mit einem festen Nenner beschränken; dann kommen wir wieder zu Beziehungen zwischen ganzen Zahlen und können versuchen, wie oben vorzugehen.

Der kleinstmögliche Nenner ist zwei; versuchen wir also unser Glück mit dem Ansatz

$$\left(\frac{c}{2} + \frac{d}{2}\sqrt{-3}\right)^3 = 10 + 9\sqrt{-}3\,,$$

wobei c und d wieder ganze Zahlen sein sollen. Ausmultiplizieren,

Multiplikation mit acht und Ausklammern führt auf die Gleichungen

$$c(c^2 - 9d^2) = 80 \quad \text{und} \quad d(c^2 - d^2) = 24\,,$$

wobei mindestens eine der Zahlen c und d ungerade sein muß, da wir ansonsten wieder eine Wurzel mit ganzzahligem Real- und Imaginärteil bekommen, also die bereits bekannte. Da rechts jeweils gerade Zahlen stehen, sieht man leicht, daß dann beide Zahlen ungerade sein müssen; damit bleiben also für c als Teiler von achtzig nur die Möglichkeiten ± 1 und ± 5. Für d als Teiler von 24 sind $d = \pm 1$ und $d = \pm 3$ möglich. Einsetzen zeigt, daß $c = -1, d = -3$ und $c = 5, d = 1$ die einzigen Lösungen sind. Die beiden verbleibenden Kubikwurzeln von $-10 + 9\sqrt{-3}$ sind somit

$$u_2 = \frac{5}{2} + \frac{1}{2}\sqrt{-3} \quad \text{und} \quad u_3 = -\frac{1}{2} - \frac{3}{2}\sqrt{-3}\,.$$

Damit lassen sich nun leicht

$$x_2 = u_2 + \frac{7}{u_2} = -1 \quad \text{und} \quad x_3 = u_3 + \frac{7}{u_3} = 5$$

berechnen. Die Gleichung $x^3 - 21x - 20 = 0$ hat also die drei ganzzahligen Lösungen $-1, -4$ und 5.

Wie die Beispiele in diesem Paragraphen zeigen, haben wir es beim exakten Lösen kubischer Gleichungen mit der hier betrachteten Formel oft mit komplizierten Ausdrücken zu tun, von denen sich nachher (nach teilweise recht trickreichen Ansätzen) herausstellt, daß sie sich tatsächlich sehr viel einfacher darstellen lassen. Dies ist ein allgemeines Problem der Computeralgebra, zu dem es leider keine allgemeine Lösung gibt, denn nach dem im vorigen Kapitel erwähnten Satz von RICHARDSON kann es keinen Algorithmus geben, der in allen Fällen entscheidet, ob zwei Ausdrücke in reellen Zahlen gleich sind oder nicht.

Die obigen Beispiele zeigen, daß das Lösung von kubischen Gleichung deutlich aufwendiger ist als das von quadratischen und daß die Lösungsformel hier keine problemlos anwendbare Mitternachtsformel ist. So ist es durchaus verständlich, daß CARDANO seinem Buch, in dem er die Lösung kubischer und biquadratischer Gleichungen behandelte, den Titel *Ars magna,* die „große Kunst“ gab.

§3: Biquadratische Gleichungen

1840 fand CARDANOs ehemaliger Diener und späterer Sekretär LODOVICO FERRARI eine Lösungsmethode für Gleichungen vom Grad vier. Hier wird zunächst der kubische Term von

$$x^4 + ax^3 + bx^2 + cx + d = 0$$

eliminiert durch die Substitution $x = y - \frac{a}{4}$; dies führt auf eine Gleichung der Form

$$y^4 + py^2 + qy + r = 0\,.$$

Zu deren Lösung benutzen wir (nach FERRARI) einen anderen Trick als im kubischen Fall: Wir versuchen, die Gleichung so zu modifizieren, daß wir ihre Lösungen als die Lösungen zweier quadratischer Gleichungen berechnen können.

Dazu nehmen wir an, wir hätten eine Lösung y der Gleichung und betrachten dazu für eine zunächst noch beliebige Zahl u die Zahl $y^2 + u$. Da $y^4 + py^2 + qy + r$ verschwindet, ist $y^4 = -py^2 - qy - r$, also

$$(y^2 + u)^2 = y^4 + 2uy^2 + u^2 = (2u - p)y^2 - qy + u^2 - r\,.$$

Falls rechts das Quadrat eines linearen Polynoms $sy + t$ steht, ist

$$(y^2 + u)^2 = (sy + t)^2 \Longrightarrow y^2 + u = \pm(sy + t)\,,$$

wir müssen also nur die beiden quadratischen Gleichungen

$$y^2 \mp (sy + t) + u = 0$$

lösen, um die Lösungen der biquadratischen Gleichung zu finden.

Natürlich läßt sich die rechte Seite $(2u-p)y^2 - qy + u^2 - r$ im allgemeinen nicht als ein Quadrat $(sy+t)^2$ schreiben; wir können aber hoffen, daß sie zumindest für gewisse spezielle Werte der bislang noch willkürlichen Konstanten u eines ist.

Ein quadratisches Polynom $\alpha Y^2 + \beta Y + \gamma$ ist genau dann Quadrat eines linearen, wenn die beiden Nullstellen der quadratischen Gleichung

$\alpha y^2 + \beta y + \gamma = 0$ übereinstimmen. Diese Nullstellen können wir nach der Formel aus §1 berechnen:

$$y^2+\frac{\beta}{\alpha}y+\frac{\gamma}{\alpha} = 0 \Longrightarrow y = -\frac{\beta}{2\alpha} \pm \sqrt{\frac{\beta^2}{4\alpha^2} - \frac{\gamma}{\alpha}} = -\frac{\beta}{2\alpha} \pm \frac{1}{2\alpha}\sqrt{\beta^2 - 4\alpha\gamma}\,.$$

Die beiden Lösungen fallen somit genau dann zusammen, wenn $\beta^2 - 4\alpha\gamma$ verschwindet. In unserem Fall ist $\alpha = (2u - p)$, $\beta = -q$ und $\gamma = u^2 - r$; wir erhalten also die Bedingung

$$q^2 - 4(2u - p)(y^2 - r) = -8u^3 + 4pu^2 + 8ru + q^2 - 4pr = 0\,.$$

Dies ist eine kubische Gleichung für u, die wir mit der Methode aus dem vorigen Abschnitt (vielleicht) lösen können. Ist u_0 eine der Lösungen, so steht in der Gleichung

$$(y^2 + u_0)^2 = (2u_0 - p)y^2 - qy + u_0^2 - r$$

rechts das Quadrat eines linearen Polynoms $sy + t$, das wir – da wir alle Koeffizienten kennen – problemlos hinschreiben können. Dies führt dann nach Wurzelziehen zu den beiden quadratischen Gleichungen

$$y^2 + u_0 = \pm(sy + t)$$

für y, deren Wurzeln die Nullstellen von $y^4 + py^2 + qy + r = 0$ sind.

Es wäre nicht schwer, mit Hilfe der Lösungsformel für kubische Gleichungen, eine explizite Formel für die vier Lösungen hinzuschreiben; sie ist allerdings erstens deutlich länger und zweitens für die praktische Berechnung reeller Nullstellen mindestens genauso problematisch wie die für kubische Gleichungen. Auf Beispiele zur Lösung biquadratischer Gleichungen verzichte ist, denn schon in einfachen Fällen ist die kubische Gleichung für u selbst dann sehr kompliziert, wenn es eine reelle Lösung gibt, die in der Lösungsformel in rein reeller Form auftaucht.

Für numerische Berechnungen sind übrigen sowohl die Lösungsformel für kubische Gleichungen als auch die für biquadratische eher nicht geeignet, da es beim Einsetzen in die Formeln oft vorkommen kann, daß zwei ungefähr gleich große Zahlen voneinander subtrahiert werden, so daß die Anzahl der geltenden Ziffern dramatisch kleiner wird. Die klassischen numerischen Verfahren zur Nullstellenbestimmung liefern genauere Ergebnisse. Sie sind auch einfacher anzuwenden und oft schneller.

§3: Gleichungen höheren Grades

Nach der (mehr oder weniger) erfolgreichen Auflösung der kubischen und biquadratischen Gleichungen in der ersten Hälfte des sechzehnten Jahrhunderts beschäftigten sich natürlich viele Mathematiker mit dem nächsten Fall, der Gleichung fünften Grades. Hier gab es jedoch über 250 Jahre lang keinerlei Fortschritt, bis zu Beginn des neunzehnten Jahrhunderts ABEL glaubte, eine Lösung gefunden zu haben. Er entdeckte dann aber recht schnell seinen Fehler und bewies stattdessen 1824, daß es keine allgemeinen Lösungsformel für Gleichungen fünften (oder höheren) Grades geben kann, die nur mit Grundrechenarten und Wurzeln auskommt.

Die Grundidee seines Beweises liegt in der Betrachtung von Symmetrien innerhalb der Lösungsmenge: Man betrachtet die Menge aller Permutationen der Nullstellenmenge, die durch Abbildungen $\varphi\colon \mathbb{C} \to \mathbb{C}$ erreicht werden können, wobei φ sowohl mit der Addition als auch der Multiplikation verträglich sein muß. ABEL zeigt, daß diese Permutationen für allgemeine Gleichungen vom Grad größer vier eine (in heutiger Terminologie) *nichtauflösbare* Gruppe bilden und daß es aus diesem Grund keine Lösungsformel geben kann, in der nur Grundrechenarten und Wurzeln vorkommen. Für Einzelheiten sei auf die Vorlesung *Algebra* verwiesen, in der dieser Satz ein wesentlicher Bestandteil ist.

Der norwegische Mathematiker NILS HENRIK ABEL (1802–1829) ist trotz seinen frühen Todes (an Tuberkulose) Initiator vieler Entwicklungen der Mathematik des neunzehnten Jahrhunderts; Begriffe wie abelsche Gruppen, abelsche Integrale, abelsche Funktionen, abelsche Varietäten, die auch in der heutigen Mathematik noch allgegenwärtig sind, verdeutlichen seinen Einfluß. Zu seinem 200. Geburtstag stiftete die norwegische Regierung einen ABEL-Preises für Mathematik mit gleicher Ausstattung und Vergabebedingungen wie die Nobelpreise; erster Preisträger war 2003 JEAN-PIERRE SERRE (*1926) vom Collège de France für seine Arbeiten über algebraische Geometrie, Topologie und Zahlentheorie.

Der ABELsche Satz besagt selbstverständlich nicht, daß Gleichungen höheren als vierten Grades *unlösbar* seien; er sagt nur, daß es *im all-*

gemeinen nicht möglich ist, die Lösungen durch Wurzelausdrücke in den Koeffizienten darzustellen: Für eine allgemeine Lösungsformel muß man also außer Wurzeln und Grundrechenarten noch weitere Funktionen zulassen. Beispielsweise fanden sowohl HERMITE als auch KRONECKER 1858 Lösungsformeln für Gleichungen fünften Grades mit sogenannten elliptischen Modulfunktionen; 1870 löste JORDAN damit Gleichungen beliebigen Grades.

§4: Der Wurzelsatz von Viète

Auch wenn es nach ABELs Satz keine allgemeine algebraische Lösungsformel für Gleichungen höheren Grades als vier geben kann, lassen sich doch in speziellen Fällen oft auch Lösungen von Gleichungen sehr viel höheren Grades leicht finden. In diesem Paragraphen wollen wir eine Methode kennenlernen, um ganzzahlige Lösungen zu finden.

Ausgangspunkt dazu ist eine Umkehrung des Problems: Wir fragen uns nicht, wie wir aus den Koeffizienten der Gleichung die Lösungen bestimmen können, sondern wie wir aus den Lösungen die Koeffizienten erhalten.

Dazu erinnern wir uns zunächst an die den meisten wohl aus der Schule oder aus Anfängervorlesungen bekannte Polynomdivision mit Rest: Ein Polynom ist bekanntlich eine formale Summe

$$f = a_d X^d + a_{d-1} X^{d-1} + \cdots + a_1 X + a_0$$

mit Koeffizienten $a_i \in k$ und einem „Symbol“ X. Falls alle a_i verschwinden, reden wir vom *Nullpolynom;* ansonsten nehmen wir, wie bei den Gleichungen, an, daß a_d nicht verschwindet und bezeichnen $d = \deg f$ als den *Grad* und a_d als den führenden Koeffizienten von f. Das Nullpolynom hat keinen Grad.

Oft ist es üblich, den Buchstaben x sowohl als Bezeichnung für eine Variable als auch für eine konkrete Lösung einer Gleichung zu verwenden. Die folgenden Überlegungen werden aber wohl klarer, wenn wir zwischen den beiden Bedeutungen unterscheiden: Große Buchstaben

stehen für Variablen und kleine für Zahlen. Der Wert des obigen Polynoms f an der Stelle x ist somit die Zahl

$$f(x) = x^d + a_{d-1}x^{d-1} + a_{d-2}x^{d-2} + \cdots + a_2x^2 + a_1x + a_0 \,.$$

Nun seien f und g Polynome mit Koeffizienten aus einem Körper k mit $\deg f = d$ und $\deg g = e$. Der Divisionsalgorithmus konstruiert dazu Polynome q und r mit Koeffizienten aus k für die $f = qg + r$ gilt, wobei r entweder das Nullpolynom ist oder einen kleineren Grad als g hat. Wir bezeichnen q als den *Quotienten* und r als den *Divisionsrest.* Sie werden wie folgt bestimmt:

0. Schritt: Setze $r = f$ und $q = 0$. b_e sei der führende Koeffizient von g.

i, $i \geq 1$***. Schritt:*** Falls $r = 0$ ist oder $\deg r < \deg g$, endet der Algorithmus. Andernfalls sei a der führende Koeffizient von r. Wir eliminieren den führenden Term von r, indem wir r ersetzen durch $r - \frac{a}{b_e}X^{\deg r - e}g$. Gleichzeitig ersetzen wir q durch $q + \frac{a}{b_e}X^{\deg r - e}$.

Dieser Algorithmus endet nach endlich vielen Schritten, denn in jedem Schritt ab dem ersten wird der Summand von r mit der höchsten X-Potenz eliminiert, so daß der Grad von r um mindestens eins kleiner wird oder r sogar zum Nullpolynom wird. Nach endlich vielen Schritten ist daher entweder $r = 0$ oder $\deg r < \deg g$, so daß der Algorithmus endet.

Nach Schritt 0 ist $qg + r = 0 \cdot g + f = f$, und wenn diese Gleichung $f = qg + r$ vor Beginn des i-ten Schritts gilt, gilt sie auch danach, denn q wird ersetzt durch $q + \frac{a}{b_e}X^{\deg r - e}$ und r durch $r - \frac{a}{b_e}X^{\deg r - e}g$, und

$$\left(q + \frac{a}{b_e}X^{\deg r - e}\right)q + \left(r - \frac{a}{b_e}X^{\deg r - e}g\right) = qg + r = f \,.$$

Nach Beendigung des Algorithmus ist außerdem noch $r = 0$ oder $\deg r < \deg e$, so daß der Algorithmus das gewünschte Ergebnis liefert.

Da wir wiederholt durch b_e dividieren, wobei die Werte von a von Schritt zu Schritt variieren können, mußten wir annehmen, daß die Koeffizienten in einem Körper liegen. Es gibt allerdings eine Ausnahme: Falls der führende Koeffizient von g gleich eins ist, sind keine Divisionen

notwendig, und der Algorithmus funktioniert auch bei Koeffizienten aus einem (kommutativen) Ring wie beispielsweise den ganzen Zahlen.

Das wollen wir anwenden auf die Nullstellen eines Polynoms, die im betrachteten Koeffizientenbereich liegen. x sei also eine Nullstelle des Polynoms

$$f = a_d X^d + a_{d-1} X^{d-1} + \cdots + a_1 X + a_0 \,,$$

das heißt

$$f(x) = a_d x^d + a_{d-1} x^{d-1} + \cdots + a_1 x + a_0 = 0 \,.$$

Wir wenden den Divisionsalgorithmus an auf f und $g = X - x$. Er liefert Polynome q, r derart, daß $f = qg + r$ ist mit $r = 0$ oder $\deg r < \deg g = 1$. Der Divisionsrest r ist also in jedem Fall eine Konstante. Wenn wir x einsetzen, ergibt sich $0 = f(x) = q(x)g(x) + r = r$, da $g(x) = x - x = 0$ ist. Somit gibt es ein Polynom q derart, daß $f = q \cdot (X - x)$ ist. Falls dabei auch $q(x) = 0$ ist, gibt es ein weiteres Polynom q_2, so daß $q = q_2 \cdot (X - x)$ ist und damit $f = q_2 \cdot (X - x)^2$. Wenn auch $q_2(x)$ verschwindet, können wir weitermachen, bis wir schließlich eine Darstellung $f = q_n \cdot (X - x)^n$ erhalten mit $q_n(x) \neq 0$. Wir sagen dann, x sei eine n-fache Nullstelle oder die Vielfachheit der Nullstelle x sei n.

Falls wir eine weitere Nullstelle $x' \neq x$ von f kennen, muß $q_n(x')$ verschwinden, denn $X - x$ verschwindet natürlich nicht an der Stelle x'. Wenn wir ℓ verschiedene Nullstellen $x_1, \ldots, x_\ell$ kennen mit Vielfachheiten $n_1, \ldots, n_\ell$, erhalten wir somit eine Darstellung

$$f = \widetilde{q} \cdot (X - x_1)^{n_1} \cdots (X - x_\ell)^{n_\ell} \,.$$

Ist d der Grad von f, so ist $d = \deg \widetilde{q} + n_1 + \cdots n_\ell$; damit folgt

Lemma: Die Anzahl der Nullstellen eines Polynoms vom Grad d ist, auch mit Vielfachheiten gezählt, höchstens gleich dem Grad. ∎

In der Vorlesung *Algebra* wird gezeigt, daß es stets einen Körper gibt, in dem die Anzahl der Nullstellen des Polynoms mit Vielfachheiten gezählt gleich dem Grad ist. Für Polynome mit reellen Koeffizienten ist das beispielsweise der Körper $\mathbb{C}$ der komplexen Zahlen, aber es gibt stets auch noch deutlich kleinere Körper mit dieser Eigenschaft.

Die Tatsache, daß $X - x$ für eine Nullstelle x eines Polynoms f ein Teiler von f ist, läßt sich für Polynome mit ganzzahligen Koeffizienten zum Erraten zumindest der ganzzahligen Nullstellen verwenden:

Lemma: $f = a_d X^d + a_{d-1} X^{d-1} + \cdots + a_1 X + a_0$ sei ein Polynom mit ganzzahligen Koeffizienten. Falls $f(x)$ für eine ganze Zahl $x \neq 0$ verschwindet, ist x ein Teiler von a_0. ($f(0) = 0$ ist natürlich äquivalent zu $a_0 = 0$.)

Beweis: Wie wir oben gesehen haben, gibt es ein Polynom q derart, daß $f = q \cdot (X - x)$ ist. Da $X - x$ den höchsten Koeffizienten eins hat und $x \in \mathbb{Z}$, entstehen in jedem Schritt des Divisionsalgorithmus wieder Polynome mit ganzzahligen Koeffizienten; daher hat auch der Quotient q ganzzahlige Koeffizienten. Ist b_0 der konstante Koeffizient von q, so ist a_0 das Produkt von b_0 mit dem konstanten Koeffizienten $-x$ von $X - x$, d.h. $a_0 = -b_0 x$, d.h. a_0 ist ein Vielfaches von x und x damit ein Teiler von a_0.

■

Als Beispiel betrachten wir die kubische Gleichung $x^3 - 7x + 6 = 0$. Das Polynom $f = X^3 - 7X + 6$ hat den konstanten Koeffizienten 6; falls es ganzzahlige Nullstellen gibt, müssen diese also unter den Zahlen $\pm 1, \pm 2, \pm 3$ und ± 6 sein. Einsetzen zeigt, daß $f(1) = f(2) = 0$, $f(-1) = f(-2) = f(3) = 12$ und $f(-3) = 0$ ist. Da es nicht mehr als drei Nullstellen geben kann, hat die kubische Gleichung daher die drei Lösungen $x_1 = 1$, $x_2 = 2$ und $x_3 = -3$.

Bei einer Gleichung ohne ganzzahlige Lösungen führt dieser Ansatz natürlich nicht zum Ziel, aber da wir im Voraus nicht wissen, ob es ganzzahlige Lösungen gibt, ist er doch oft einen Versuch wert. Selbst wenn wir damit nur einen Teil der Nullstellen finden, können wir die zugehörigen Linearfaktoren abdividieren und erhalten für die restlichen Nullstellen eine Gleichung kleineren Grades.

Tatsächlich läßt sich die Methode noch ausbauen. Wir nehmen dazu an, wir hätten ein Polynom f, das über einem hinreichend großen Körper

in Linearfaktoren zerfällt, d.h.

$$\begin{aligned} f &= a_d X^d + a_{d-1} X^{d-1} + \cdots + a_1 X + a_0 \\ &= a_d (X - x_1)(X - x_2) \cdots (X - x_d)\,, \end{aligned}$$

wobei die x_i natürlich nicht alle verschieden sein müssen. Ausmultiplizieren und Koeffizientenvergleich führt auf die Gleichungen

$$\begin{aligned} a_{d-1} &= -a_d \sigma_1(x_1, \ldots, x_d) \quad \text{mit} \quad \sigma_1(x_1, \ldots, x_d) = x_1 + \cdots + x_n \\ a_{d-2} &= a_d \sigma_2(x_1, \ldots, x_d) \quad \text{mit} \quad \sigma_2(x_1, \ldots, x_d) = \sum_{i<j} x_i x_j \\ a_{d-3} &= -a_d \sigma_3(x_1, \ldots, x_d) \quad \text{mit} \quad \sigma_3(x_1, \ldots, x_d) = \sum_{i<j<k} x_i x_j x_k \\ &\vdots \qquad\qquad \vdots \\ a_0 &= (-1)^d a_d \sigma_d(x_1, \ldots, x_d) \quad \text{mit} \quad \sigma_d(x_1, \ldots, x_d) = x_1 \cdots x_d\,. \end{aligned}$$

Allgemein ist a_{d-r} bis aufs Vorzeichen gleich der Summe alle Produkte aus r Werten x_i mit verschiedenem Index. Diese Summen bezeichnet man als die *elementarsymmetrischen Funktionen* $\sigma_r(x_1, \ldots, x_d)$ und die obigen Gleichungen als den Wurzelsatz von VIÈTE.

FRANÇOIS VIÈTE (1540–1603) studierte Jura an der Universität Poitiers, danach arbeitete er als Hauslehrer. 1573, ein Jahr nach dem Massaker an den Hugenotten, berief ihn CHARLES IX (obwohl VIÈTE Hugenotte war) in die Regierung der Bretagne; unter HENRI III wurde er geheimer Staatsrat. 1584 wurde er auf Druck der katholischen Liga vom Hofe verbannt und beschäftigte sich fünf Jahre lang nur mit Mathematik. Unter HENRI IV arbeitete er wieder am Hof und knackte u.a. verschlüsselte Botschaften an den spanischen König PHILIP II. In seinem Buch *In artem analyticam isagoge* rechnete er als erster systematisch mit symbolischen Größen, führte also die „Buchstabenrechnung“ ein. Auch die mathematische Formelschreibweise geht auf ihn zurück, insbesondere auch die Zeichen „+“ und „-“ für Addition und Subtraktion.

Für eine quadratische Gleichung $x^2 + px + q = 0$ besagt der Satz von VIÈTE einfach, daß die Summe der Lösungen gleich $-p$ und das Produkt gleich q ist, was die Babylonier bekanntlich schon vor rund vier Jahrtausenden wußten.

Diese elementarsymmetrischen Funktionen sind für r-Werte im mittleren Bereich recht umfangreiche Summen, die beiden Fälle $r = 0$ und $r = d - 1$ können aber gelegentlich sehr nützlich sein, um Lösungen zu erraten, vor allem wenn $a_d = 1$ ist:

Bei der oben betrachteten Gleichung $f(x) = x^3 - 7x + 6 = 0$ etwa ist das Produkt aller Nullstellen gleich -6 und ihre Summe verschwindet. Aus den Zahlen $\pm 1, \pm 2, \pm 3$ und ± 6 müssen wir also drei (nicht notwendigerweise verschiedene) finden mit Summe Null und Produkt -6. Das geht offensichtlich nur mit $1, 2$ und -3; Einsetzen zeigt, daß dies auch tatsächlich Nullstellen sind. Damit mußten wir nur drei Zahlen einsetzen, statt wie oben sechs, um alle Lösungen zu finden.

Man beachte, daß dieses Einsetzen unbedingt notwendig ist: Bei der Gleichung $g(x) = x^3 - 6x + 6 = 0$ hätten wir genauso vorgehen können und wären auf dieselben drei Kandidaten gekommen, aber $g(1) = 1, g(2) = 2$ und $g(-3) = -3$. (Daß die Lösungsmenge nicht$\{1, 2, -3\}$ sein kann, erkennt man auch daran, daß

$$\sigma_2(1, 2, -3) = 1 \cdot 2 + 1 \cdot (-3) + 2 \cdot (-3) = -7$$

nicht gleich dem Koeffizienten Null von x^2 ist.)

Auch die Gleichung $x^3 - 21x - 20 = 0$ läßt sich leicht nach VIÈTE lösen: Hier ist das Produkt aller Nullstellen gleich 20; *falls* sie alle ganzzahlig sind, kommen also nur $\pm 1, \pm 2, \pm 4, \pm 5, \pm 10$ und ± 20 in Frage. Aus diesen zwölf Zahlen müssen wir drei (nicht notwendigerweise verschiedene) auswählen mit Produkt 20 und Summe null. Das geht offensichtlich nur mit $-1, -4$ und 5, und wieder zeigt Einsetzen, daß dies auch tatsächlich Nullstellen sind.

Betrachten wir als nächstes Beispiel das Polynom

$$f = X^4 + 14X^3 - 52X^2 - 14X + 51$$

mit $a_0 = 51 = 3 \cdot 17$. Da das Produkt aller Nullstellen diesen Wert haben muß, kommen – *falls* alle Nullstellen ganzzahlig sind – für diese nur die Werte $\pm 1, \pm 3, \pm 17$ und ± 51 in Frage. Wäre eine der Nullstellen ± 51, müßten alle anderen den Betrag eins haben und die Summe könnte nicht gleich -14 sein. Daher muß eine Nullstelle Betrag drei und

eine Betrag 17 haben, die beiden anderen Betrag eins. Produkt 51 und Summe -14 erzwingt dabei offensichtlich, daß sowohl +1 als auch -1 Nullstellen sind, außerdem -17 und +3. Einsetzen zeigt, daß alle vier auch tatsächlich Nullstellen sind.

Beim Polynom $X^3 - 3X - 2$ ist das Produkt aller Nullstellen -2 und die Summe verschwindet. Die Teiler von -2 sind ± 1 und ± 2; Einsetzen zeigt, daß nur -1 und 2 Nullstellen sind. Sowohl aus dem Verschwinden der Summe als auch aus dem Produkt -2 folgt, daß -2 eine doppelte Nullstelle sein muß, d.h. $X^3 - 3X - 2 = (X + 1)^2(X - 2)$.

Beim Polynom

$$f = X^6 + 27X^5 - 318X^4 - 5400X^3 - 10176X^2 + 27648X + 32768$$

ist $a_0 = 32768 = 2^{15}$; hier wissen wir also nur, daß – sofern alle Nullstellen ganzzahlig sind – jede Nullstelle die Form $\pm 2^i$ haben muß, wobei die Summe aller Exponenten gleich 15 sein muß und die Anzahl der negativen Vorzeichen gerade. Einsetzen zeigt, daß

$$-1, \quad 2, \quad -4, \quad -8, \quad 16, \quad -32$$

die Nullstellen sind.

Gelegentlich lassen sich auch nicht ganzzahlige Nullstellen mit Hilfe des Satzes von VIÈTE erraten: Bei Polynom $X^4+X^3-7X^2-5X+10$ ist das Produkt der Nullstellen zehn. Wie Einsetzen zeigt, sind 1 und -2 Nullstellen. Ihre Summe ist -1 und ihr Produkt -2. Die beiden restlichen Nullstellen haben somit die Summe Null und das Produkt -5. Wir suchen also Zahlen x_3 und x_4 mit $x_4 = -x_3$ und $x_3x_4 = -x_3^2 = -5$. Das geht nur, wenn x_3 und x_4 gleich $\pm\sqrt{5}$ sind.

Man beachte, daß die Anwendung des Satzes von VIÈTE nur deshalb so gut funktionierte, weil die betrachteten Polynome höchsten Koeffizient eins hatten. Ist das nicht der Fall, ist das Produkt der Nullstellen gleich dem Quotienten aus konstantem Koeffizienten und führendem Koeffizienten mal $(-1)^{\text{Grad}}$, und wenn das keine ganze Zahl ist, können wir nicht mehr mit Teilbarkeit argumentieren, sondern müssen uns auf der Suche nach rationalen Lösungen auch mit den möglichen Nennern beschäftigen. Dabei genügt es leider nicht, alle möglichen Faktoren von Zähler

und Nenner miteinander zu kombinieren, da ein Produkt rationaler Zahlen einen kleineren Nenner haben kann als die Faktoren. Im Falle der quadratischen Gleichung $12x^2 - x - 6 = 0$ etwa bekommen wir nach Division durch den führenden Koeffizienten einen konstanten Koeffizienten von $-\frac{1}{2}$, der auch tatsächlich das Produkt der beiden Nullstellen $\frac{3}{4}$ und $-\frac{2}{3}$ ist, aber beide haben einen Nenner größer zwei. Nur wenn für für den Zähler alle Teiler von sechs und für den Nenner alle Teiler von zwölf ausprobiert hätten, hätten wir (nach recht vielen Versuchen) die Lösungen gefunden, denn offensichtlich gilt

Lemma: Ist der gekürzte Bruch $x = \frac{p}{q}$ eine Nullstelle des Polynoms $a_n X^n + \cdots + a_1 X + a_0$ mit ganzzahligen Koeffizienten, so ist p ein Teiler von a_0 und q ein Teiler von a_n.

Beweis: Setzen wir $x = \frac{p}{q}$ in die Gleichung ein und multiplizieren sie mit q^n erhalten wir die rein ganzzahlige Gleichung

$$a_n p^n + a_{n-1} p^{n-1} q + \cdots + a_1 p q^{n-1} + a_0 q^n = 0\,.$$

Auf der linken Seite enthalten alle Summanden außer dem ersten mindestens einen Faktor q. Da die Summe alle Terme verschwindet, muß daher auch der erste Summand durch q teilbar sein. Damit muß q ein Teiler von a_n sein, denn q und p^n haben keinen gemeinsamen Teiler. Genauso folgt, daß a_0 durch p teilbar sein muß, denn alle Summanden außer dem letzten enthalten mindestens einen Faktor p. ∎

Korollar: Bei einem Polynom mit ganzzahligen Koeffizienten und höchstem Koeffizienten eins sind alle rationalen Nullstellen ganz. ∎

§5: Andere Ansätze für Gleichungen höheren Grades

Der Satz von ABEL besagt, daß es Polynome vom Grad größer vier gibt, deren Nullstellen sich nicht mit Grundrechenarten und Wurzeln als Ausdrücke in den Koeffizienten schreiben lassen. Selbstverständlich lassen sie sich trotzdem numerisch mit beliebiger vorgegebener Genauigkeit approximieren, aber in der Computeralgebra interessieren wir uns nicht

für Approximationen, sondern für exakte Lösungen. Wie kann die exakte Lösung einer solchen Polynomgleichung aussehen?

Wenn wir uns auf reelle Nullstellen beschränken, können wir diese jedenfalls eindeutig charakterisieren durch ein Paar bestehend aus einem Polynom f und aus einem Intervall $[a, b]$ mit der Eigenschaft, daß f dort genau eine Nullstelle hat. Dabei kann f das Ausgangspolynom sein; falls sich dieses aber über $\mathbb{Z}$, $\mathbb{Q}$ oder einem sonstigen für effektives Rechnen geeigneten Zahlbereich in ein Produkt von Polynomen niedrigeren Grades zerlegen läßt, könnte man stattdessen auch den Faktor nehmen, zu dem die Nullstelle gehört, was wegen des kleineren Grades effizienter sein sollte.

Nun wollen wir Nullstellen aber nicht nur charakterisieren, wir wollen auch mit ihnen rechnen. Eine solche Darstellung von Nullstellen ist daher nur dann nützlich, wenn wir mit den so charakterisierten Zahlen auch rechnen können. Außerdem brauchen wir natürlich Algorithmen, die in der Lage sind, zu einem gegebenen Polynom Intervalle zu finden, in denen jeweils genau eine Nullstelle liegt. Im nächsten Kapitel werden wir sehen, daß beides möglich ist.

Wenn wir auch an komplexen Nullstellen interessiert sind, können wir diese charakterisieren, indem wir sowohl den Realteil als auch den Imaginärteil als Paare aus einem Polynom und einem Intervall darstellen. Das ist zwar etwas umständlich, aber kein größeres Problem.

Kapitel 2
Im Umkreis des Euklidischen Algorithmus

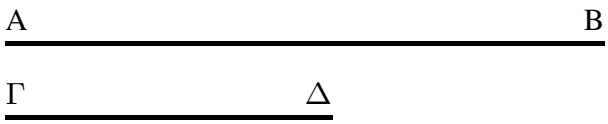
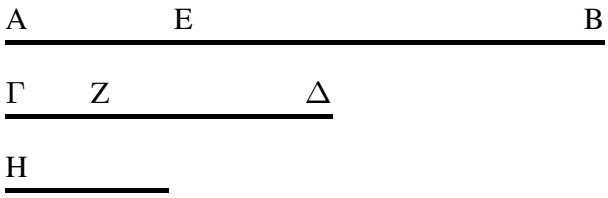
Der EUKLIDische Algorithmus sollte allen Hörern aus der Linearen Algebra bekannt sein. Er geht mit ziemlicher Sicherheit nicht auf EUKLID zurück, sondern war wohl bereits rund zweihundert Jahre vorher den Pythagoräern bekannt, die ihn möglicherweise aus noch älteren Quellen kannten. Die im zugrunde liegende Technik ist die *Wechselwegnahme* oder wechselseitige Subtraktion, griechisch Antanairesis ($\grave{\alpha}\nu\tau\alpha\nu\alpha\iota\rho\epsilon\sigma\iota\varsigma$) oder auch Anthyphairesis ($\grave{\alpha}\nu\theta\upsilon\phi\alpha\iota\rho\epsilon\sigma\iota\varsigma$), und zumindest diese wurde von den Pythagoräern definitiv verwendet. Unter anderem gelangten sie dadurch zu der für sie schockierenden Erkenntnis, daß nicht alle Größen in einem rationalen Verhältnis zueinander stehen, sondern daß das ausgerechnet bei ihrem Wahrzeichen, dem Pentagramm, die Länge von dessen Seiten in keinem rationalen Verhältnis zur Seite des umgebenden regelmäßigen Fünfecks stehen kann. Das Prinzip der Wechselwegnahme fand nicht nur damals sondern auch später noch zahlreiche weitere Anwendungen, von denen wir in diesem Kapitel einige für die Computeralgebra relevanten betrachten wollen.

§1: Geschichtliche Vorbemerkung

So wie wir den EUKLIDischen Algorithmus heute kennen, hat er nichts mit wechselseitiger Subtraktion zu tun. Bei EUKLID, in Proposition 2 des siebten Buchs seiner *Elemente* aber sieht es anders aus. In der Übersetzung

EUKLID: Die Elemente, aus dem Griechischen übersetzt und herausgegeben von CLEMENS THAER, *Ostwalds Klassiker der exakten Wissenschaften* **235**, *Verlag Harri Deutsch,* 1997

lesen wir:

> *Zu zwei gegebenen Zahlen, die nicht prim gegeneinander sind, ihr größtes gemeinsames Maß zu finden.*
>
> Die zwei gegebenen Zahlen, die nicht prim, gegeneinander sind, seien AB, ΓΔ. Man soll das größte gemeinsame Maß von AB, ΓΔ finden.
>
> 
>
>
> Wenn ΓΔ hier AB mißt – sich selbst mißt es auch – dann ist ΓΔ gemeinsames Maß von ΓΔ, AB. Und es ist klar, daß es auch das größte ist, denn keine Zahl größer ΓΔ kann ΓΔ messen.
>
> Wenn ΓΔ aber AB nicht mißt, und man nimmt bei AB, ΓΔ abwechselnd immer das kleinere vom größeren weg, dann muß (schließlich) eine Zahl übrig bleiben, die die vorangehende mißt. Die Einheit kann nämlich nicht übrig bleiben; sonst müßten AB, ΓΔ gegeneinander prim sein, gegen die Voraussetzung. Also muß eine Zahl übrig bleiben, die die vorangehende mißt. ΓΔ lasse, indem es BE mißt, EA, kleiner als sich selbst übrig; und EA lasse, indem es ΔZ mißt, ZΓ, kleiner als sich selbst übrig; und ΓZ messe AE.
>
> 
>
>
> Da ΓZ AE mißt und AE ΔZ, muß ΓZ auch ΔZ messen; es mißt aber auch sich selbst, muß also auch das Ganze ΓΔ messen. ΓΔ mißt aber BE; also mißt ΓZ auch BE; es mißt aber auch EA, muß also auch das Ganze BA messen. Und es mißt auch ΓΔ; ΓZ mißt also AB und ΓΔ; also ist ΓZ gemeinsames Maß von AB, ΓΔ. Ich behaupte, daß es auch das größte ist. Wäre nämlich ΓZ nicht das größte gemeinsame Maß von AB, ΓΔ, so müßte irgendeine Zahl größer ΓZ die Zahlen AB und ΓΔ messen. Dies geschehe; die Zahl sei H. Da H dann ΓΔ mäße und ΓΔ BE mißt, mäße H auch BE; es soll aber auch das Ganze BA messen, müßte also auch den Rest AE messen. AE mißt aber ΔZ; also müßte H auch ΔZ messen; es soll aber auch das Ganze ΔΓ messen, müßte also auch den Rest ΓZ messen, als größere Zahl die kleinere; dies ist unmöglich. Also kann keine Zahl größer ΓZ die Zahlen AB und ΓΔ messen; ΓZ ist also das größte gemeinsame Maß von AB, ΓΔ; dies hatte man beweisen sollen.

Aus heutiger Sicht erscheint die Voraussetzung, daß die betrachteten Größen nicht teilerfremd sein dürfen, seltsam; sie erklärt sich daraus,

daß in der griechischen Philosophie wie auch der Mathematik die Einheit eine Sonderrolle einnahm und nicht als Zahl angesehen wurde: Die Zahlen begannen erst mit der Zwei. Dementsprechend führt er in Proposition 1 des siebten Buchs fast wörtlich dieselbe Konstruktion durch für den Fall von teilerfremden Größen. Schon wenig später wurde die Eins auch in Griechenland als Zahl anerkannt, und für uns heute ist die Unterscheidung natürlich ohnehin bedeutungslos; wir können die Bedingung und ihre Anwendung auf den Beweis, daß der ggT ungleich eins ist, also einfach ignorieren.

Formal unterscheidet sich EUKLIDs Algorithmus zur Bestimmung des größten gemeinsamen Teilers vor allem in einem Punkt von der heute gebräuchlichen Formulierung: Während EUKLID von einer längeren Strecke eine kürzere sooft es geht wegnimmt, führen wir heute eine Division mit Rest durch. Am Ergebnis ändert sich dadurch natürlich nichts.

Nichts spricht dafür, daß das, was wir heute als erweiterten EUKLIDischen Algorithmus bezeichnen, zu Zeiten EUKLIDs bekannt war. Die älteste überlieferte Beschreibung des Verfahrens zur Darstellung des größten gemeinsamen Teilers als ganzzahlige Linearkombination erschien 1624 in der zweiten Auflage des Buchs *Problèmes plaisants et délectables qui se fonts par les nombres* von BACHET DE MÉZIRIAC. Dort geht es allerdings nicht um eine abstrakte Darstellung des Algorithmus, sondern, wie schon der Titel des Buches sagt, um amüsante und köstliche Probleme im Zusammenhang mit Zahlen. Im zehnten Problem des zweiten Teils geht es um ein Bankett von 41 Personen, Männer, Frauen und Kinder, wobei sowohl die Gesamtkosten bekannt sind als auch das, was ein Mann, eine Frau und ein Kind jeweils ausgibt. Daraus soll die Anzahl der Männer, Frauen und Kinder berechnet werden. 1993 brachte der Verlag Blanchard eine vereinfachte Version der Auflage von 1874 neu heraus, die unter `cnum.cnam.fr/DET/8PY45.html` auch online verfügbar ist, aber leider ist dieser Vereinfachung unter anderem auch die mathematische Beschreibung von MÉZIRIACs Vorgehensweise zum Opfer gefallen. Offensichtlich glaubte der Herausgeber, daß man Lesern des späten zwanzigsten Jahrhundert keine so komplizierte Mathematik mehr zumuten kann wie denen des siebzehnten.

Der Name MÉZIRIAC wird allerdings nur selten im Zusammenhang mit diesem Algorithmus erwähnt. Wenn man nicht einfach vom erweiterten EUKLIDischen Algorithmus spricht, bezeichnet man die entsprechende Gleichung als Identität von BÉZOUT nach dem französischen Mathematiker ETIENNE BÉZOUT, der sie 1766 in einem Lehrbuch beschrieb und auf Polynome verallgemeinerte. Auch dabei spielte die (Polynom-)Division mit Rest eine entscheidende Rolle.

Wir wollen daher im nächsten Paragraphen eine abstrakte algebraische Struktur betrachten, in der eine Division mit Rest möglich ist.

CLAUDE GASPAR BACHET SIEUR DE MÉZIRIAC (1581-1638) verbrachte den größten Teil seines Lebens in seinem Geburtsort Bourg-en-Bresse. Er studierte zwar bei den Jesuiten in Lyon und Milano und trat 1601 in den Orden ein, trat aber bereits 1602 wegen Krankheit wieder aus und kehrte nach Bourg zurück. Sein Buch erschien erstmals 1612, Am bekanntesten ist BACHET für seine lateinische Übersetzung der *Arithmetika* von DIOPHANTOS. In einem Exemplar davon schrieb FERMAT seine Vermutung an den Rand. Auch Gedichte von BACHET sind erhalten. 1635 wurde er Mitglied der französischen Akademie der Wissenschaften.

ETIENNE BÉZOUT (1730-1783) wurde in Nemours in der Ile-de-France geboren, wo seine Vorfahren Magistrate waren. Er ging stattdessen an die Akademie der Wissenschaften; seine Hauptbeschäftigung war die Zusammenstellung von Lehrbüchern für die Militärausbildung. Im 1766 erschienenen dritten Band (von vier) seines *Cours de Mathématiques à l'usage des Gardes du Pavillon et de la Marine* ist die Identität von BÉZOUT dargestellt. Seine Bücher waren so erfolgreich, daß sie ins Englische übersetzt und z.B. in Harvard als Lehrbücher benutzt wurden. Heute ist er vor allem bekannt durch seinen Beweis, daß sich zwei Kurven der Grade n und m in höchstens nm Punkten schneiden können.

§2: Euklidische Ringe

Erinnern wir uns zunächst an die Definition eines Rings:

Definition: *a)* Ein Ring ist eine Menge R zusammen mit zwei Rechenoperationen „+“ und „$\cdot$“ von $R \times R$ nach R, für die gilt:

1.) R bildet bezüglich „+“ eine abelsche Gruppe, d.h. für die Addition gilt das Kommutativgesetz $f + g = g + f$ sowie das Assoziativgesetz $(f + g) + h = f + (g + h)$ für alle $f, g, h \in \mathbb{R}$, es gibt ein Element $0 \in R$, so daß $0 + f = f + 0 = f$ für alle $f \in R$, und zu jedem $f \in R$ gibt es ein Element $-f \in R$, so daß $f + (-f) = 0$ ist.

2.) Die Verknüpfung „$\cdot$“: $R \times R \to R$ erfüllt das Assoziativgesetz $f(gh) = (fg)h$, und es gibt ein Element $1 \in R$, so daß $1f = f1 = f$.

3.) „+“ und „$\cdot$“ erfüllen die Distributivgesetze $f(g + h) = fg + fh$ und $(f + g)h = fh + gh$.

b) Ein Ring heißt *kommutativ,* falls zusätzlich noch das Kommutativgesetz $fg = gf$ der Multiplikation gilt.
c) Ein Ring heißt *nullteilerfrei* wenn gilt: Falls ein Produkt $fg = 0$ verschwindet, muß mindestens einer der beiden Faktoren f, g gleich Null sein. Ein nullteilerfreier kommutativer Ring heißt *Integritätsbereich.*

Natürlich ist jeder Körper ein kommutativer Ring; für einen Körper werden schließlich genau dieselben Eigenschaften gefordert und zusätzlich noch die Existenz multiplikativer Inverser. Aus letzterer folgt sofort, daß Körper Integritätsbereiche sind.

Das bekannteste Beispiel eines Rings, der kein Körper ist, sind die ganzen Zahlen; auch sie bilden einen Integritätsbereich.

Auch die Menge

$$k[X] = \left\{ \sum_{i=0}^{n} a_i X^i \;\middle|\; n \in \mathbb{N}_0, a_i \in k \right\}$$

aller Polynome mit Koeffizienten aus einem Körper k ist ein Integritätsbereich; ersetzt man den Körper k durch einen beliebigen kommutativen Ring R, ist $R[X]$ immerhin noch ein Ring. Man überlegt sich leicht, daß $R[X]$ genau dann ein Integritätsbereich ist, wenn auch R einer ist.

Als Beispiel eines nichtkommutativen Rings können wir die Menge aller $n \times n$-Matrizen über einem Körper betrachten; dieser Ring hat

auch Nullteiler, denn beispielsweise ist

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} ,$$

obwohl keiner der beiden Faktoren die Nullmatrix ist.

Was uns nun noch fehlt, ist eine Division mit Rest. Für Zahlen a, b, q, r aus $\mathbb{N}_0$ ist die Aussage

$$a : b = q \text{ Rest } r$$

äquivalent zu den beiden Bedingungen

$$a = bq + r \quad \text{und} \quad 0 \leq r < b\,.$$

Die erste dieser Bedingungen können wir in einem beliebigen Ring hinschreiben, eine Kleinerrelation haben wir dort allerdings nicht. Andererseits brauchen wir aber etwas nach Art der zweiten Bedingung: Falls der Divisionsrest nicht in irgendeiner Weise kleiner als der Divisor sein muß, könnten wir einfach *immer* sagen $a : b = 0$ Rest a, was nicht sonderlich viel nützt.

Wir fordern deshalb die Existenz einer Funktion $\nu\colon R \setminus \{0\} \to \mathbb{N}_0$, die im Falle eines von Null verschiedenen Divisionsrests für den Rest einen kleineren Wert annimmt als für den Divisor:

Definition: Ein EUKLIDischer Ring ist ein Integritätsbereich R zusammen mit einer Abbildung $\nu\colon R \setminus \{0\} \to \mathbb{N}_0$, so daß gilt: Ist $f = gh$, so ist $\nu(f) \geq \max\big(\nu(g), \nu(h)\big)$, und zu je zwei Elementen $f, g \in R$ gibt es Elemente $q, r \in R$ mit

$$f = qq + r \quad \text{und} \quad r = 0 \text{ oder } \nu(r) < \nu(g)\,.$$

Wir schreiben auch $f : g = q$ Rest r und bezeichnen r als Divisionsrest bei der Division von f durch g.

Standardbeispiel sind auch hier wieder die ganzen Zahlen, wo wir als ν einfach die Betragsfunktion nehmen können. Quotient und Divisionsrest sind durch die Forderung $\nu(r) < \nu(y)$ allerdings nicht eindeutig festgelegt, beispielsweise ist im Sinne dieser Definition

$$11 : 3 = 3 \text{ Rest } 2 \quad \text{und} \quad 11 : 3 = 4 \text{ Rest } -1\,.$$

Die Definition des EUKLIDischen Rings verlangt nur, daß es *mindestens* eine Darstellung gibt; Eindeutigkeit ist nicht gefordert.

Das für uns im Augenblick wichtigste Beispiel ist der Polynomring $k[X]$ über einem Körper k; hier zeigt die bekannte Polynomdivision mit Rest, daß die Bedingungen erfüllt sind bezüglich der Abbildung

$$\nu\colon \begin{cases} k[X] \setminus \{0\} \to \mathbb{N}_0 \\ f \mapsto \text{Grad } f \end{cases}.$$

Hier ist es allerdings wichtig, daß k ein Körper ist: Bei der Polynomdivision mit Rest müssen wir schließlich die führenden Koeffizienten durcheinander dividieren, und das ist etwa im Polynomring $\mathbb{Z}[X]$ nicht immer möglich.

Dies beweist freilich nicht, daß $\mathbb{Z}[X]$ *kein* EUKLIDischer Ring wäre, denn in der Definition war ja nur gefordert, daß es für *irgendeine* Funktion ν *irgendein* Divisionsverfahren gibt. Die Nichtexistenz eines solchen Verfahrens ist sehr schwer zu zeigen – es sei denn, eine der im folgenden hergeleiteten Eigenschaften eines EUKLIDischen Rings ist nicht erfüllt. Bei $\mathbb{Z}[X]$ ist dies, wie wir bald sehen werden, bei der linearen Kombinierbarkeit des ggT in der Tat der Fall, so daß $\mathbb{Z}[X]$ kein EUKLIDischer Ring sein kann.

Ein weiteres bekanntes Beispiel eines EUKLIDischen Rings ist der Ring der GAUSSschen Zahlen, d.h. die Menge aller komplexer Zahlen mit ganzzahligem Real- und Imaginärteil; hier können wir $\nu(x+iy) = x^2+y^2$ setzen. Da dieser Ring hier keine Rolle spielen wird, sei auf einen Beweis verzichtet.

§3: Der größte gemeinsame Teiler

Bevor wir uns mit der Berechnung des größten gemeinsamen Teilers zweier Elemente eines EUKLIDischen Rings beschäftigen, müssen wir zunächst definieren, was das sein soll. Da es bei der Division durch einen Nullteiler keinen eindeutigen Quotienten geben kann, beschränken wir uns auf Integritätsbereiche.

Definition: R sei ein Integritätsbereich.
a) Ein Element $h \in R$ heißt Teiler von $f \in R$, in Zeichen $h|f$, wenn es ein $q \in R$ gibt, so daß $f = qh$ ist.
b) $h \in R$ heißt *größter gemeinsamer Teiler* (kurz ggT) der beiden Elemente f und g aus R, wenn h Teiler von f und von g ist, und wenn für jeden anderen gemeinsamen Teiler r von f und g gilt: $r|h$.
c) Zwei Elemente $f, g \in R$ heißen *assoziiert,* wenn f Teiler von g und g Teiler von f ist.
d) Ein Element $u \in R$ heißt *Einheit,* falls es ein $v \in R$ gibt mit $uv = 1$. Die Menge aller Einheiten von R bezeichnen wir mit $R^\times$.

In einem Körper ist natürlich jedes von null verschiedene Element Teiler eines jeden anderen Elements und damit auch eine Einheit; in $\mathbb{Z}$ dagegen sind ± 1 die beiden einzigen Einheiten, und zwei ganze Zahlen sind genau dann assoziiert, wenn sie sich höchstens im Vorzeichen unterscheiden.

Man beachte, daß wir beim größten gemeinsamen Teiler die „Größe“ über Teilbarkeit definieren; von daher ist in $\mathbb{Z}$ außer 2 auch -2 ein größter gemeinsamer Teiler von 8 und 10. Insbesondere ist „der“ größte gemeinsame Teiler also im allgemeinen nicht eindeutig bestimmt, was uns bei seiner Berechnung in Polynomringen noch einiges an Problemen schaffen wird.

In einem Polynomring über einem Integritätsbereich ist der Grad des Produkts zweier Polynome gleich der Summe der Grade der Faktoren. Da das konstante Polynom eins Grad null hat, muß daher jede Einheit Grad null haben. Die Einheiten von $R[x]$ sind somit genau die Einheiten von R. Speziell für Polynomringe über Körpern sind dies genau die von null verschiedenen Konstanten.

Einheiten hängen auch eng zusammen mit Assoziiertheit:

Lemma: Zwei von null verschiedene Elemente f, g eines Integritätsbereichs sind genau dann assoziiert, wenn es eine Einheit u gibt, so daß $f = ug$ ist.

Beweis: Eine Einheit $u \in R$ hat nach Definition ein Inverses $u^{-1} \in R$, und aus $f = ug$ folgt $g = u^{-1}f$. Somit ist f Teiler von g und g Teiler

von f; die beiden Elemente sind also assoziiert.

Sind umgekehrt $f, g \in R \setminus \{0\}$ assoziiert, so gibt es Elemente $u, v \in R$ derart, daß $g = uf$ und $f = vg$ ist. Damit ist $g = uf = uvg$ und $f = vg = vuf$, also $(1 - uv)g = 0$ und $(1 - vu)f = 0$. Da wir in einem Integritätsbereich sind und f, g nicht verschwinden, muß somit $uv = vu = 1$ sein, d.h. u und v sind Einheiten. ∎

Damit sind also zwei Polynome über einem Körper genau dann assoziiert, wenn sie sich nur durch eine von null verschiedene multiplikative Konstante unterscheiden. Nur bis auf eine solche Konstante können wir auch den größten gemeinsamen Teiler zweier Polynome bestimmen, denn allgemein gilt:

Lemma: Der größte gemeinsame Teiler zweier Polynome ist bis auf Assoziiertheit eindeutig. Sind also h und $\tilde{h}$ zwei größte gemeinsame Teiler der beiden Elemente f und g, so sind h und $\tilde{h}$ assoziiert; ist umgekehrt h ein größter gemeinsamer Teiler von f und g und ist $\tilde{h}$ assoziiert zu h, so ist auch $\tilde{h}$ ein größter gemeinsamer Teiler von f und g.

Beweis: Sind h und $\tilde{h}$ größte gemeinsame Teiler, so sind sie insbesondere gemeinsame Teiler und damit Teiler eines jeden größten gemeinsamen Teilers. Somit müssen h und $\tilde{h}$ einander teilen, sind also assoziiert. Ist h ein größter gemeinsamer Teiler und $\tilde{h}$ assoziiert zu h, so teilt $\tilde{h}$ jedes Vielfache von h, ist also auch ein gemeinsamer Teiler, und da h jeden gemeinsamen Teiler teilt, gilt dasselbe auch für $\tilde{h}$. Somit ist auch $\tilde{h}$ ein größter gemeinsamer Teiler. ∎

In heutiger Sprache ausgedrückt beruht der EUKLIDische Algorithmus auf den folgenden beiden Tatsachen:

1. Wenn wir zwei Elemente f, g eines EUKLIDischen Rings mit Rest durcheinander dividieren, so ist $f : g = q$ Rest r äquivalent zu jeder der beiden Gleichungen

$$f = qg + r \quad \text{und} \quad r = f - qg\,.$$

   Diese zeigen, daß jeder gemeinsame Teiler von f und g auch ein gemeinsamer Teiler von g und r ist und umgekehrt. Die beiden

Paare (f, g) und (g, r) haben also dieselben gemeinsamen Teiler und damit auch denselben größten gemeinsamen Teiler:

$$\text{ggT}(f, g) = \text{ggT}(g, r)\,.$$

2. $\text{ggT}(f, 0) = f$, denn jedes Element eines Integritätsbereichs teilt die Null.

Aus diesen beiden Beobachtungen folgt nun leicht

Satz: In einem EUKLIDischen Ring gibt es zu je zwei Elementen $f, g \in R$ stets einen größten gemeinsamen Teiler. Dieser kann nach folgendem Algorithmus berechnet werden:
Schritt 0: Setze $r_0 = f$ und $r_1 = g$
Schritt i, $i \geq 1$**:** Falls $r_i = 0$ ist, endet der Algorithmus mit dem Ergebnis $\text{ggT}(f, g) = r_{i-1}$. Andernfalls wird r_{i-1} mit Rest durch r_i dividiert, wobei r_{i+1} der Divisionsrest sei.
Der Algorithmus endet nach endlich vielen Schritten und liefert den größten gemeinsamen Teiler.

Beweis: Wir überlegen uns als erstes, daß im i-ten Schritt für $i \geq 1$ stets $\text{ggT}(f, g) = \text{ggT}(r_{i-1}, r_i)$ ist. Für $i = 1$ gilt dies nach der Konstruktion im nullten Schritt. Falls es im i-ten Schritt für ein $i \geq 1$ gilt und der Algorithmus nicht mit dem i-ten Schritt abbricht, wird dort r_{i+1} als Rest bei der Division von r_{i-1} durch r_i berechnet; wie wir oben gesehen haben, ist somit $\text{ggT}(r_i, r_{i+1}) = \text{ggT}(r_{i-1}, r_i)$, und das ist nach Induktionsvoraussetzung gleich dem ggT von f und g.

Falls der Algorithmus im i-ten Schritt abbricht, ist dort $r_i = 0$. Außerdem ist dort wie in jedem anderen Schritt auch $\text{ggT}(f, g) = \text{ggT}(r_{i-1}, r_i)$. Somit ist r_{i-1} der ggT von f und g.

Schließlich muß noch gezeigt werden, daß der Algorithmus nach endlich vielen Schritten abbricht. Dazu dient die Funktion ν: Nach Definition eines EUKLIDischen Rings ist im i-ten Schritt entweder $\nu(r_i) < \nu(r_{i-1})$ oder $r_i = 0$. Da ν nur natürliche Zahlen und die Null als Werte annimmt und es keine unendliche absteigende Folge natürlicher Zahlen gibt, muß nach endlich vielen Schritten $r_i = 0$ sein, womit der Algorithmus abbricht.

■

§4: Der erweiterte Euklidische Algorithmus

Zur Bestimmung des ggT zweier Elemente eines EUKLIDischen Rings R berechnen wir eine Reihe von Elementen r_i, wobei r_0 und r_1 die Ausgangsdaten sind und alle weiteren r_i durch Division mit Rest ermittelt werden:

$$r_{i-1} : r_i = q_i \text{ Rest } r_{i+1}$$

Damit ist $r_{i+1} = r_{i-1} - q_i r_i$ als Linearkombination seiner beiden Vorgänger r_i und r_{i-1} mit Koeffizienten aus R darstellbar, die wiederum R-Linearkombinationen *ihrer* Vorgänger sind, *usw.* Wenn wir alle diese Darstellungen ineinander einsetzen, erhalten wir r_i schließlich als Linearkombination der Ausgangselemente. Dies gilt insbesondere für das letzte nichtverschwindende r_i, den ggT. Der ggT zweier Elemente f, g eines EUKLIDischen Rings ist somit darstellbar als R-Linearkombination von f und g.

Algorithmisch sieht dies folgendermaßen aus:

Schritt 0: Setze $r_0 = f$, $r_1 = g$, $\alpha_0 = \beta_1 = 1$ und $\alpha_1 = \beta_0 = 0$. Mit $i = 1$ ist dann

$$r_{i-1} = \alpha_{i-1} a + \beta_{i-1} b \quad \text{und} \quad r_i = \alpha_i a + \beta_i b \,.$$

Diese Relationen werden in jedem der folgenden Schritte erhalten:

Schritt i, $i \geq 1$**:** Falls $r_i = 0$ ist, endet der Algorithmus mit

$$\text{ggT}(f, g) = r_{i-1} = \alpha_{i-1} f + \beta_{i-1} g \,.$$

Andernfalls dividiere man r_{i-1} mit Rest durch r_i mit dem Ergebnis

$$r_{i-1} = q_i r_i + r_{i+1} \,.$$

Dann ist

$$\begin{aligned} r_{i+1} = -q_i r_i + r_{i-1} &= -q_i(\alpha_i f + \beta_i g) + (\alpha_{i-1} f + \beta_{i-1} g) \\ &= (\alpha_{i-1} - q_i \alpha_i) f + (\beta_{i-1} - q_i \beta_i) g \,; \end{aligned}$$

man setze also

$$\alpha_{i+1} = \alpha_{i-1} - q_i \alpha_i \quad \text{und} \quad \beta_{i+1} = \beta_{i-1} - q_i \beta_i \,.$$

Da die Schritte hier einfach Erweiterungen der entsprechenden Schritte des klassischen EUKLIDischen Algorithmus sind, ist klar, daß auch dieser

Algorithmus nach endlich vielen Schritten abbricht und als Ergebnis den ggT liefert. Da die beiden Relationen aus Schritt 0 in allen weiteren Schritten erhalten bleiben, ist auch klar, daß dieser ggT am Ende als Linearkombination dargestellt ist.

Lemma: Der Ring $\mathbb{Z}[X]$ aller Polynome mit ganzzahligen Koeffizienten ist nicht EUKLIDisch.

Beweis: Wir wissen zwar noch nicht, daß zwei beliebige Elemente von $\mathbb{Z}[X]$ auch in $\mathbb{Z}[X]$ einen größten gemeinsamen Teiler haben, es ist aber klar, daß der größte gemeinsame Teiler der beiden Polynome X und 2 existiert und eins ist: Die einzigen Teiler von 2 sind ± 1 und ± 2, und ± 2 sind keine Teiler von X. Wäre $\mathbb{Z}[X]$ ein EUKLIDischer Ring, gäbe es also Polynome $\alpha, \beta \in \mathbb{Z}[x]$, so daß $\alpha X + 2\beta = 1$ wäre. Der konstante Koeffizient von $\alpha X + 2\beta$ ist aber das Doppelte des konstanten Koeffizienten von β, also eine gerade Zahl. Somit kann es keine solche Darstellung geben. ∎

(In $\mathbb{Q}[X]$ gibt es selbstverständlich so eine Darstellung: $1 = 0 \cdot X + \frac{1}{2} \cdot 2$. Im übrigen ist zwei dort ein Teiler von X.)

§5: Der Satz von Sturm

JACQUES CHARLES-FRANÇOIS STURM wurde 1803 in Genf als Sohn eines Mathematiklehrers geboren. Ab 1821 studierte er an der dortigen Akademie Mathematik. 1823, nach Ende seines Studiums, wurde er Tutor des Sohns von Mme DE STAËL, was ihm genügend Zeit für mathematische Arbeiten ließ. Als die Familie für sechs Monate nach Paris zog, traf er dort im Haus von ARAGO unter anderem LAPLACE, POISSON, FOURIER, GAY-LUSSAC und AMPÈRE. 1825 kehrte er nach Paris zurück, wo er zwar als Tutor für ARAGOs Sohn arbeitete, vor allem aber Vorlesungen besuchte. Zeitweise arbeitete er auch als Assistent von FOURIER. Nach der Revolution von 1830 wurde es auch für einen Protestanten möglich, eine Professor in Frankreich zu bekommen; so kam er 1830 ans *Collège Rollin* und 1838 an die *Ecole normale supérieure;* 1833 wurde er französischer Staatsbürger. In seinen späten Arbeiten beschäftige er sich vor allem, zusammen mit LIOUVILLE, mit Differentialgleichungen. Er starb 1855 in Paris.

STURMs Ansatz beruht auf einer leichten Modifikation des EUKLIDischen Algorithmus: Die Folge der Divisionsreste bei der Berechnung des ggT

zweier Polynome f und g nach EUKLID können wir beschreiben durch die Rekursionsvorschrift

$$r_0 = f, \quad r_1 = g, \quad r_{i+2} = \text{Rest bei der Division von } r_i \text{ durch } r_{i+1},$$

wobei abgebrochen wird, sobald ein r_j verschwindet. Für $i \geq 2$ ist also $r_i = q_i r_{i+1} + r_{i+2}$. STURM betrachtet zu einem Polynom f und dessen Ableitung f' stattdessen die Folge der *negativen* Divisionsreste:

Definition: Die STURMsche Kette zum Polynom $f \in \mathbb{R}[X]$ wird berechnet nach der Rekursionsvorschrift

$$f_0 = f, \quad f_1 = f', \quad f_{i+2} = -\text{Rest bei der Division von } f_i \text{ durch } f_{i+1},$$

wobei abgebrochen wird, sobald ein f_j verschwindet. Für $i \geq 2$ ist also $f_i = q_i f_{i+1} - f_{i+2}$, wobei q_i der Quotient bei der Polynomdivision mit Rest von f_i durch f_{i+1} ist.

Sind r_i die Divisionsreste bei der Anwendung des EUKLIDischen Algorithmus auf f und f', ist also für $i \geq 2$ jeweils $f_i = -r_i$. Damit ist klar, daß die Kette nach endlich vielen Schritten abbricht, und mit r_s ist auch $f_s = -r_s$ ein größter gemeinsamer Teiler von f und f'. Insbesondere ist f_s konstant, falls f keine mehrfachen Nullstellen hat.

Lemma: Die STURMsche Kette $(f_0, \ldots, f_s)$ eines Polynoms f ohne mehrfache Nullstellen hat folgende Eigenschaften:

a) $f_0 = f$

b) f_s hat keine Nullstellen

c) Ist für ein i mit $0 < i < s$ der Punkt x_0 eine Nullstelle von f_i, so ist $f_{i-1}(x_0) f_{i+1}(x_0) < 0$, d.h. $f_{i-1}(x_0)$ und $f_{i+1}(x)$ haben verschiedene Vorzeichen

d) Ist x_0 eine Nullstelle von $f = f_0$, so sind in einer Umgebung von x_0 die Funktionswerte von $f_0(x) f_1(x)$ links von x_0 negativ und rechts davon positiv.

Beweis: a) ist Teil der Definition der STURMschen Kette von f, und wie wir uns gerade überlegt haben, ist f_s im Falle eines Polynoms ohne mehrfache Nullstellen konstant, hat also keine Nullstellen. Für *c)* beachten wir, daß

$$f_{i+1}(x_0) = q_i(x_0) f_i(x_0) - f_{i-1}(x_0) = -f_{i-1}(x_0)$$

ist, da $f_i(x_0)$ verschwindet. *d)* schließlich folgt daraus, daß x_0 eine einfache Nullstelle ist, d.h. $f'(x_0) \neq 0$. Somit hat

$$\frac{d}{dx} f_0(x) f_1(x) = f'(x)^2 - f(x) f''(x)$$

bei x_0 den positiven Wert $f'(x_0)^2$. Damit ist diese Ableitung auch in einer Umgebung von x_0 positiv. Da $f_0(x) f_1(x)$ an der Stelle x_0 verschwindet, muß es also links davon negativ und rechts positiv sein.

■

Definition: *a)* Eine Folge $(f_0, \ldots, f_s)$ von Polynomen aus $\mathbb{R}[X]$ heißt STURMsche Folge zum Polynom $f \in \mathbb{R}[X]$, wenn sie die vier Eigenschaften *a)* bis *d)* aus obigem Lemma hat.
b) Eine Folge $(a_0, \ldots, a_s)$ von reellen Zahlen hat einen *Vorzeichenwechsel* an der Stelle a_i, wenn für den kleinsten Index $j > i$ mit $a_j \neq 0$ das Produkt $a_i a_j$ negativ ist.
c) Die Variation $v(a)$ einer Folge $(f_0, \ldots, f_s)$ von Polynomen aus $\mathbb{R}[X]$ an der Stelle a ist die Anzahl der Vorzeichenwechsel in der Folge reeller Zahlen $\big(f_0(a), f_1(a), \ldots, f_s(a)\big)$.

Nach obigem Lemma ist also die STURMsche Kette eines Polynoms ohne mehrfache Nullstellen eine STURMsche Folge zu diesem Polynom.

Für jede STURMsche Folge zu einem Polynom f gilt:

Satz: Die Anzahl der Nullstellen des Polynoms f mit $a < x \leq b$ ist $v(a) - v(b)$.

Beweis: Wir überlegen uns zunächst, in der Umgebung welcher Punkte sich in der Folge $\big(f_0(x), f_1(x), \ldots, f_s(x)\big)$ etwas an den Vorzeichen ändern kann.

Sind alle $f_i(x) \neq 0$, so bleiben auch in einer Umgebung von x alle Vorzeichen gleich, also ist $v(x)$ konstant in der Umgebung von x.

Ist $f(x_0) \neq 0$, aber (mindestens) ein $f_i(x_0) = 0$, so folgt aus *a),* daß $i > 0$ ist, und nach *b)* ist $i < s$. Damit gibt es Funktionen f_{i-1} und f_{i+1}. Nach *c)* ist $f_{i-1}(x_0) f_{i+1}(x_0) < 0$. Somit haben $f_{i-1}(x_0)$ und $f_{i+1}(x_0)$ verschiedene Vorzeichen, sind also insbesondere ungleich Null. Die Vorzeichen von $f_{i-1}(x)$ und $f_{i+1}(x)$ sind daher in einer Umgebung von x_0

konstant und verschieden. Zwischen $f_{i-1}(x)$ und $f_{i+1}(x)$ gibt es somit genau einen Vorzeichenwechsel, egal welchen Wert $f_i(x)$ annimmt. Dies zeigt, daß $v(x)$ in einer Umgebung von x_0 konstant ist.

Bleibt noch der Fall, daß f selbst im Punkt x_0 verschwindet. Nach *d)* ist dann in einer Umgebung von x_0 das Produkt $f_0(x)f_1(x)$ links von x_0 negativ und rechts positiv, d.h. für $x < x_0$ gibt es dort einen Vorzeichenwechsel zwischen $f_0(x)$ und $f_1(x)$, für $x > x_0$ nicht mehr. Für $x > x_0$ aus dieser Umgebung ist daher $v(x)$ um eins kleiner als für $x < x_0$.

Damit ist gezeigt, daß die Funktion v genau in den Punkten um eins kleiner wird, in denen f eine Nullstelle hat, und daraus folgt der Satz. ∎

Satz von Sturm: Ist $[a, b]$ ein Intervall, an dessen Endpunkten a, b das Polynom f nicht verschwindet, und bezeichnet $v(x)$ die Variation der STURMschen Kette zu f, so hat f in $[a, b]$ genau $v(a) - v(b)$ verschiedene Nullstellen.

Beweis: Falls f keine mehrfachen Nullstellen hat, ist die STURMsche Kette eine STURMsche Folge, also folgt die Behauptung aus dem gerade bewiesenen Satz.

Andernfalls sei $(f_0, \ldots, f_s)$ die STURMsche Kette von f. Wegen deren Konstruktion über den EUKLIDischen Algorithmus ist dann $g = f_s$ ein größter gemeinsamer Teiler von f und f', und alle f_i sind durch g teilbar. Somit besteht auch die Folge $(f_0/g, \ldots, f_s/g)$ nur aus Polynomen, und in jedem Punkt, in dem g keine Nullstelle hat, ist ihre Variation gleich der der STURMschen Kette. Da die Intervallenden a und b nach Voraussetzung keine Nullstellen sind, hat sie also insbesondere an den Stellen a und b dieselbe Variation wie die STURMsche Kette von f.

Die Funktion f/g hat dieselben Nullstellen wie f, aber jeweils nur einfach. Falls wir also zeigen können, daß $(f_0/g, \ldots, f_s/g)$ eine STURMsche Folge zu f/g ist, folgt der Satz auch in diesem Fall aus dem gerade bewiesenen.

Die Eigenschaften *a)* und *b)* einer STURMsche Folge sind trivial.

Für *c)* müssen wir eine Nullstelle x von f_i/g für ein i zwischen Null und $s - 1$ betrachten. f_{i+1} ist der negative Rest bei der Division von

f_{i-1} durch f_i. Ist q_i der Quotient, so ist also $f_{i-1} = q_i f_i - f_{i+1}$. Damit ist auch $f_{i-1}/g = q_i \cdot f_i/g - f_{i+1}/g$, also haben f_{i-1}/g und f_{i+1}/g in jeder Nullstelle von f_i/g verschiedene Vorzeichen. (Sie können nicht Null sein, denn wenn zwei aufeinanderfolgende f_i/g in einem Punkt verschwinden, müßte $f_s/g = 1$ dort wegen der gerade gezeigten Rekursionsbeziehung ebenfalls verschwinden.)

Bleibt noch *d):* Wir betrachten eine k-fache Nullstelle x_0 von f und schreiben $f(x) = (x - x_0)^k h(x)$, wobei h ein Polynom ist, das in x_0 nicht verschwindet. Dann ist

$$f'(x) = (x - x_0)^k h'(x) + k(x - x_0)^{k-1} h(x)\,;$$

also gibt es ein Polynom q mit $q(x_0) \neq 0$ und $g(x) = (x - x_0)^{k-1} q(x)$. Damit ist $\dfrac{f_1(x)}{g(x)} = (x - x_0)\dfrac{h'(x)}{q(x)} + k\dfrac{h(x)}{q(x)}$ und

$$\frac{f_0(x)}{g(x)} \cdot \frac{f_1(x)}{g(x)} = (x - x_0)^2 \frac{h(x)h'(x)}{q(x)^2} + k(x - x_0)\frac{h(x)^2}{q(x)^2}\,.$$

An der Stelle x_0 verschwindet $f_0(x)f_1(x)$, und überall dort, wo $f_0(x)f_1(x)$ nicht verschwindet, hat sein Wert das gleiche Vorzeichen wie

$$\frac{f_0(x)f_1(x)}{g(x)^2} = \frac{f_0(x)}{g(x)} \cdot \frac{f_1(x)}{g(x)} = k\frac{h(x)^2}{q(x)^2}(x - x_0) + \frac{h(x)h'(x)}{q(x)^2}(x - x_0)^2\,.$$

In einer hinreichend kleinen Umgebung von x_0 kann $(x-x_0)^2$ gegenüber $(x - x_0)$ vernachlässigt werden; dort ist das Vorzeichen also gleich dem des ersten Summanden. Dieser ist negativ für $x < x_0$ und positiv für $x > x_0$, womit auch *d)* gezeigt ist.

■

Als erstes Beispiel betrachten wir das Polynom

$$f = X^4 + 3X^3 + 2X^2 + X + \tfrac{1}{2}\,.$$

Seine STURMsche Kette ist

$$\left(f,\ 4X^3 + 9X^2 + 4X + 1,\ \tfrac{11}{16}X^2 - \tfrac{5}{16},\ -\tfrac{64}{11}X - \tfrac{56}{11},\ -\tfrac{219}{1024}\right).$$

Für eine Zahl x mit hinreichend großem Betrag wird das Vorzeichen des Werts einer Polynomfunktion durch den höchsten Term bestimmt;

wir haben also für stark negative Werte von x die Vorzeichenverteilung $(+,-,+,+,-)$ mit drei Vorzeichenwechseln; für große positive x erhalten wir $(+,+,+,-,-)$ mit nur einem Vorzeichenwechsel. Somit gibt es insgesamt zwei reelle Nullstellen.

Um deren Vorzeichen zu bestimmen, werten wir die STURMsche Kette an der Stelle Null aus: $(\frac{1}{2}, 1, -\frac{5}{16}, -\frac{56}{11}, -\frac{219}{1024})$ hat einen Vorzeichenwechsel, also sind alle Nullstellen negativ. Wenn wir $x = -1$ in die STURMsche Kette einsetzen, erhalten wir die Folge $(-\frac{1}{2}, 2, \frac{3}{8}, \frac{8}{11}, -\frac{219}{1024})$ mit zwei Vorzeichenwechsel. Somit gibt es eine Nullstelle z_1 mit $-1 < z_1 < 0$ und eine Nullstelle $z_2 < -1$.

Für $x = -2$ erhalten wir die Folge $(-\frac{3}{2}, -3, \frac{39}{16}, \frac{72}{11}, -\frac{219}{1024})$ mit ebenfalls zwei Vorzeichenwechseln, so daß $z_2 < -2$ sein muß, und für $x = -3$ haben wir in $(\frac{31}{2}, -38, \frac{47}{8}, \frac{136}{11}, -\frac{219}{1024})$ drei Vorzeichenwechsel, also ist $-3 < z_2 \le -2$. Damit kennen wir immerhin schon die ganzzahligen Anteile der beiden Nullstellen.

Als nächstes „Beispiel“ wollen wir untersuchen, wie viele reelle Nullstellen das quadratische Polynom $f = aX^2 + bX + c$ mit $a \neq 0$ hat. Seine Ableitung ist $f_1 = 2aX + b$, und

$$(aX^2 + bX + c) : (2aX + b) = \frac{X}{2} + \frac{b}{4a} \qquad \text{Rest } \frac{b^2 - 4ac}{4a}\,.$$

Also ist f_2 die Konstante $\Delta/4a$ mit $\Delta = b^2 - 4ac$, und die STURMsche Kette von f ist

$$\Big(aX^2 + bX + c,\ \ 2aX + b, \frac{\Delta}{4a}\Big)\,.$$

Ist $a > 0$, so haben wir für große x die Vorzeichenfolge $\big(+,+,\mathrm{sgn}(\Delta)\big)$, für sehr negative x erhalten wir $\big(+,-,\mathrm{sgn}(\Delta)\big)$.

Für $\Delta > 0$ haben wir daher für $x \to \infty$ die Variation $v(x) = 0$, für $x \to -\infty$ dagegen $v(x) = 2$. Somit gibt es zwei reelle Nullstellen. Für $\Delta = 0$ folgt entsprechend, daß es nur eine gibt. Für $\Delta < 0$ haben wir die beiden Vorzeichenverteilungen $(+,+,-)$ und $(+,-,-)$, die beide Variation eins haben, also gibt es für $\Delta < 0$ keine reelle Nullstelle. Für $a < 0$ drehen sich alle Vorzeichen um; an den Variationen und somit

am Ergebnis ändert sich nichts. Beruhigenderweise stimmen alle diese Ergebnisse überein mit dem, was wir auch direkt aus der Lösungsformel für quadratische Gleichungen ablesen können.

Bei kubischen Polynomen können wir uns bekanntlich auf solche der Form $f_0 = f = X^3+pX+q$ beschränken. Die Ableitung ist $f_1 = 3X^2+p$, und

$$(X^3 + pX + q) : (3X^2 + p) = \frac{X}{3} \quad \text{Rest } \frac{3p}{2}X + q\,,$$

so daß $f_2 = -\frac{3p}{2}X - q$ ist. Weiter ist für $p \neq 0$

$$(3X^2 + p) : \left(-\frac{3p}{2}X - q\right) = -\frac{9X}{2p} + \frac{27q}{4p^2} \quad \text{Rest } \left(p^3 + \frac{27}{4}q^2\right) \Big/ p^2\,,$$

die STURMsche Kette endet also mit $f_3 = -\left(p^3 + \frac{27}{4}q^2\right) \Big/ p^2$. Im Falle $p = 0$ dividieren wir durch die Konstante $-q$, sofern diese nicht auch verschwindet, und damit ist der Divisionsrest gleich null.

Für die Anzahl reeller Lösungen ist das asymptotische Verhalten relevant: Da $f(x) = x^3 + px + q$ durch den führenden Term x^3 dominiert wird, ist hier das Vorzeichen unabhängig von p und q für große negative x stets negativ und für große positive x stets positiv. Entsprechend haben wir für $f_1(x) = 3x^2 + p$ in beiden Fällen positive Vorzeichen. Auch bei der linearen Funktion $f_2(x) = -\frac{3}{2}px - q$ ist unabhängig von q das Vorzeichen für stark negative x stets gleich dem von p, für positive dagegen gleich dem von $-p$. (Der ziemlich triviale Fall $p = 0$ sei dem Leser überlassen.) Das Vorzeichen von $f_3(x)$ schließlich ist das von $-\Delta$ mit $\Delta = p^3 + \frac{27}{4}q^2$, denn $p^2 \geq 0$.

Die Vorzeichenfolge wird dann für große negative Werte von x zu $(-, +, \operatorname{sgn} p, -\operatorname{sgn}\Delta)$; für große positive zu $(+, +, -\operatorname{sgn} p, -\operatorname{sgn}\Delta)$. Für $\Delta > 0$ und haben wir also die Folgen $(-, +, \operatorname{sgn} p, -)$ und $(+, +, -\operatorname{sgn} p, -)$. Da $\pm\operatorname{sgn} p$ zwischen einem + und einem $-$ steht, haben wir im ersten Quadrupel immer zwei Vorzeichenwechsel und im zweiten immer nur einen; es gibt daher für $\Delta < 0$ nur eine reelle Nullstelle (und zwei komplexe).

Im Fall $\Delta = 0$ haben wir die Folgen $(-, +, \operatorname{sgn} p, 0)$ und $(+, +, -\operatorname{sgn} p, 0)$. Da $q^2 \geq 0$ aber $\Delta = 0$ ist, muß hier entweder $p = q = 0$ sein oder $p < 0$. Im letzteren Fall hat $(-, +, \operatorname{sgn} p, 0)$ zwei Vorzeichenwechsel und $(+, +, -\operatorname{sgn} p, 0)$ keinen; es gibt also zwei reelle Nullstellen (von denen eine die Vielfachheit zwei hat). Im ersten Fall hat $(+, +, -\operatorname{sgn} p, 0)$ nur einen Vorzeichenwechsel und $(+, +, -\operatorname{sgn} p, 0)$ wieder keinen; es gibt also nur eine reelle Nullstelle. Da wir für $p = q = 0$ die Gleichung $x^3 = 0$ bekommen, ist das die dreifache Nullstelle $x = 0$.

Für $\Delta < 0$ schließlich ist notwendigerweise $p < 0$, denn q^2 kann nicht negativ werden. Wir bekommen daher die Folgen $(-, +, -, +)$ mit drei Vorzeichenwechseln und $(+, +, +, +)$ ohne Vorzeichenwechsel. Hier gibt es also drei reelle Nullstellen.

Wenn wir mit der Lösungsformel

$$y = u - \frac{p}{3u} \quad \text{mit} \quad u = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{\Delta}{27}}}$$

vergleichen, sehen wir, warum wir in Kapitel 1, §2 bei den Beispielen mit drei verschiedenen reellen Nullstellen Schwierigkeiten hatten: Wie wir gerade gesehen haben, muß Δ dann negativ sein; die Quadratwurzel in der Lösungsformel liefert also einen imaginären Wert. Obwohl es drei reelle Nullstellen gibt, müssen wir also zu deren Berechnung die Kubikwurzel einer nichtreellen komplexen Zahl finden.

Der Satz von STURM sagt uns für jedes Intervall $[a,\ b]$, wie viele Nullstellen des Polynoms f dort liegen. Wenn wir uns für die Nullstellen eines Polynoms interessieren, geht es aber eher darum, eine Liste möglichst kleiner Intervalle zu finden die jeweils genau eine Nullstelle von f enthalten. STURM hat auch gezeigt, wie das möglich ist: Sobald wir ein Intervall kennen, in dem alle reellen Nullstellen liegen, können wir durch fortgesetzte Intervallhalbierungen zu einer Liste von Intervallen kommen, die jeweils genau eine Nullstelle enthalten. Durch weitere Halbierungen können wir gegebenenfalls auch die Länge dieser Intervalle beliebig kurz machen.

Für das Ausgangsintervall brauchen wir eine Abschätzung für die Größe der reellen Nullstellen eines Polynoms

$$f = a_n X^n + a_{n-1} X^{n-1} + \cdots + a_1 X + a_0 \,.$$

An den Nullstellen dieses Polynoms ändert sich nichts, wenn wir es mit einer von Null verschiedenen Konstanten multiplizieren; eine gute Schranke sollte daher nur von den Quotienten a_i/a_n abhängen. Um nicht ständig mit diesen Quotienten hantieren zu müssen, beschränken wir uns in der folgenden Diskussion auf normierte Polynome, d.h. auf Polynome mit führendem Koeffizienten eins.

Es gibt eine ganze Reihe von Schranken für die Nullstellen eines Polynoms. Wohl am einfachsten und für uns völlig ausreichend ist die folgende 1829 von CAUCHY veröffentlichte:

Lemma: z sei eine reelle Nullstelle des nichtkonstanten Polynoms

$$f = X^n + a_{n-1} X^{n-1} + \cdots + a_1 X + a_0 \,,$$

und J sei die Menge aller Indizes i mit $a_i < 0$; die Elementanzahl von J sei m. Für $m = 0$ ist $z \leq 0$, ansonsten ist z kleiner als das Maximum aus eins und den Zahlen $\sqrt[n-k]{|ma_k|}$ für $k \in J$.

Beweis: Ist $J = \emptyset$, so nimmt $f(x)$ für jedes positive x einen positiven Wert an. Es kann daher keine positiven Nullstellen geben, also ist $z \leq 0$. Andernfalls sei S das Maximum aus eins und den Zahlen $\sqrt[n-k]{|ma_k|}$ mit $k \in J$. Für jedes solche k ist dann $|ma_k| \leq S^{n-k}$. Damit ist auch für jedes $x > S$ zunächst $x^{n-k} > |ma_k|$ und somit

$$x^n = x^{n-k} \cdot x^k > |ma_k|\, x^k = -ma_k x^k \,.$$

Addieren wir diese Ungleichungen über alle $k \in J$, folgt

$$mx^n > -m \sum_{k \in J} a_k x^k \quad \text{und} \quad x^n + \sum_{k \in J} a_k x^k > 0 \,.$$

Da $a_k x^k$ für alle $k \notin J$ größer oder gleich null ist, ist damit auch $f(x) > 0$; ein $x > S$ kann also keine Nullstelle sein.

■

Baron AUGUSTIN LOUIS CAUCHY (1789–1857) stellte als erster durch die exakte Definition von Begriffen wie *Konvergenz* und *Stetigkeit* die Analysis auf ein sicheres Fundament. In insgesamt 789 Arbeiten beschäftigte er sich u.a. auch mit komplexer Analysis, Variationsrechnung, Differentialgleichungen, FOURIER-Analysis, Permutationsgruppen, der Diagonalisierung von Matrizen und der theoretischen Mechanik. Als überzeugter Royalist hatte er häufig Schwierigkeiten mit den damaligen Regierungen; er lebte daher mehrere Jahre im Exil in Turin und später in Prag, wo er (mit sehr mäßigem Erfolg) den französischen Thronfolger unterrichtete.

Als Beispiel betrachten wir das Polynoms $f = X^5-2X^4-3X^3+2X^2-1$. Hier ist $J = \{0, 3, 4\}$, also $m = 3$. Unter den Zahlen 6, $\sqrt{9} = 3$ und $\sqrt[5]{3}$ ist 6 die größte, also ist jede reelle Nullstelle kleiner oder gleich sechs.

Um auch eine untere Schranke für die reellen Nullstellen zu erhalten, betrachten wir das Polynom $f(-X)$ oder besser, da $f(-X)$ den höchsten Term $-X^5$ hat, das Polynom $-f(-X) = X^5 + 2X^4 - 3X^3 - 2X^2 + 1$. Hier ist $J = \{2, 3\}$, also $m = 2$, und unter den Zahlen $\sqrt{6}$ und $\sqrt[3]{4}$ ist $\sqrt{6}$ die größere, denn $\sqrt[3]{4} < \sqrt[3]{8} = 2$, aber $\sqrt{6} > \sqrt{4} = 2$. Damit wissen wir, daß alle reellen Nullstellen z von f die Ungleichung $-\sqrt{6} \leq z \leq 6$ erfüllen.

Sobald wir ein Intervall $[a, b]$ kennen, in dem alle reellen Nullstellen eines Polynoms f liegen, ist eigentlich klar, wie das STURMsche Intervallhalbierungsverfahren funktioniert: Wir suchen eine Liste $\mathcal{M}$ von Intervallen $[a_i, b_i]$ mit der Eigenschaft, daß jedes dieser Intervalle genau eine Nullstelle von f enthält. Eventuell können wir auch noch fordern, daß die Länge jedes dieser Intervalle unter einer gewissen Schranke s liegt.

Wir arbeiten mit einer Liste $\mathcal{L}$ bestehend aus noch zu untersuchenden Intervallens $[c, d]$ zusammen mit den Variationen $v(c)$ und $v(d)$ der STURMschen Kette von f. Zu Beginn enthält $\mathcal{L}$ nur das Intervall $[a, b]$, in dem alle reellen Nullstellen liegen, zusammen mit den Variationen $v(a)$ und $v(b)$.

So lange die Liste $\mathcal{L}$ nicht leer ist, wählen wir eines der dort befindlichen Intervalle $[c, d]$ aus und berechnen nach STURM die Anzahl der

dort befindlichen Nullstellen. Wenn es keine gibt, eliminieren wir das Intervall, falls es nur eine ist (und gegebenenfalls die Intervallänge unter der Schranke s liegt), kommt das Intervall in die Ergebnisliste $\mathcal{M}$. Andernfalls wählen wir einen Punkt $t \in (c,\, d)$i mit $f(t) \neq 0$, z.B. den Mittelpunkt $t = \frac{1}{2}(c + d)$, und berechnen die Variation $v(t)$. Ist $v(c) > v(t)$, kommt das Intervall $[c,\, t]$ zusammen mit den beiden Variationen in die Liste, und entsprechend auch $[t,\, d]$, falls $v(t) > v(d)$ ist. Der Eintrag zu $[c,\, d]$ wird dann aus $\mathcal{L}$ entfernt.

Um diesen Algorithmus auf das obige Beispiel anwenden zu können, müssen wir zunächst die STURMsche Kette von f berechnen. Da es uns nur um Vorzeichen geht, können wir die einzelnen Polynome problemlos ersetzen durch ihre Produkte mit geeigneten positiven Zahlen und so Nenner vermeiden.

$$f_0 = f = X^5 - 2X^4 - 3X^3 + 2X^2 - 1$$

$$f_1 = f' = 5X^4 - 8X^3 - 9X^2 + 4X$$

$$f_2 = \frac{46}{25}X^3 - \frac{12}{25}X^2 + 1 - \frac{8}{25}X$$

Wir nehmen stattdessen

$$f_2 = 46X^3 - 12X^2 + 25 - 8X$$

$$f_3 = \frac{5225}{529}X^2 - \frac{125}{1058}X - \frac{1925}{529}$$

Wir multiplizieren mit $\frac{1058}{25}$ und betrachten

$$f_3 = 418X^2 - 5X - 154$$

$$f_4 = -\frac{769695}{87362}X - \frac{82524}{3971}$$

$$f_5 = -\frac{513601198}{235225}$$

Wir wissen, daß alle reellen Nullstellen zwischen $-\sqrt{6}$ und 6 liegen; um ganze Zahlen zu haben, starten wir mit dem Intervall $[-3,\, 6]$ und werten die STURMsche Kette an dessen Endpunkten aus. Dazu müssen wir in der Folge $(f_0, f_1, f_2, f_3, f_4, f_5)$ den jeweils betrachteten Punkt x

einsetzen und die Vorzeichenwechsel zählen. Im Maxima können wir das Einsetzen einfach dadurch verlangen, daß wir hinter die Liste der Polynome $X = x$ schreiben, also beispielsweise

```
[f0, f1, f2, f3, f4, f5], X=-3;
```

mit Ergebnis

$$\left[-307, \quad 528, \quad -1301, \quad 3623, \quad \frac{493557}{87362}, \quad -\frac{513601198}{235225}\right].$$

Bequemer wird es, wenn wir noch die Signum-Funktion `sign` darauf anwenden und das ganze als eine Funktion schreiben:

```
VZW(x):= map(sign, ev([f0, f1, f2, f3, f4, f5], X=x));
```

Damit bekommen wir auf das Kommando `VZW(-3);` als Ergebnis die Folge `[neg, pos, neg, pos, pos, neg]`, und sehen, daß wir hier vier Vorzeichenwechsel haben. Entsprechend erhalten wir für $x = 6$ zunächst vier positive, dann zwei negative Werte, haben also einen Vorzeichenwechsel. Damit wissen wir, daß das Polynom f drei reelle Nullstellen hat.

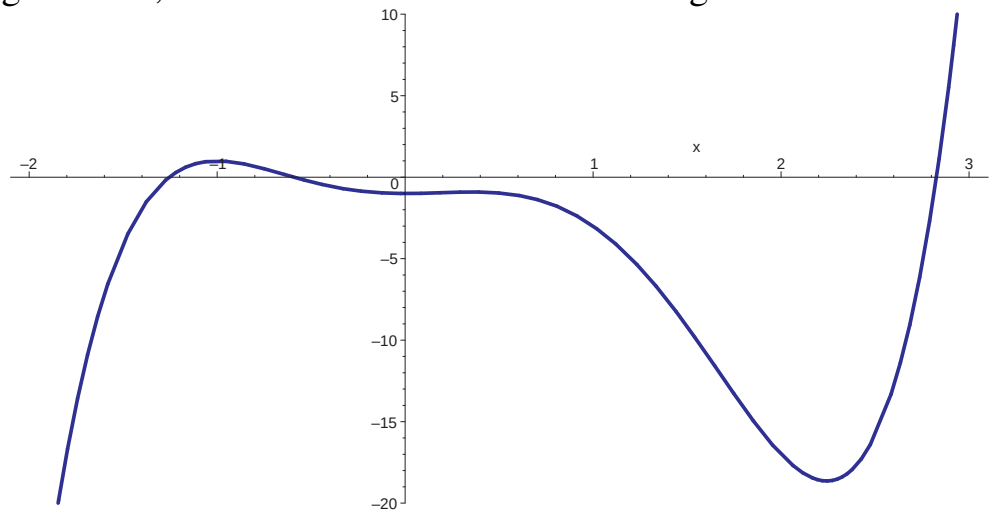
Wir unterteilen das Intervall an der Stelle $x = 2$ und berechnen die Variation dort: Einsetzen liefert uns die Vorzeichenfolge $(-, -, +, +, -, -)$ mit zwei Vorzeichenwechseln. Somit gibt es zwei Nullstellen zwischen -3 und 2 und eine zwischen 2 und 6. Das Intervall $[2, 6]$ enthält also bereits nur eine einzige Nullstelle und kommt somit, falls wir keine Ansprüche an die Intervallängen stellen, in die Ergebnisliste.

Das Intervall $[-3, 2]$ muß weiter zerlegt werden, z.B. an der Stelle $x = 0$; hier erhalten wir die Vorzeichenfolge $(-, +, +, -, -, -)$. Wieder zwei Vorzeichenwechsel, also gibt es keine Nullstelle in $[0, 2]$, aber zwei in $[-3, 0]$. Wir zerlegen weiter an der Stelle $x = -1$ mit $(+, +, -, +, -, -)$, also drei Vorzeichenwechsel. Damit haben wir eine Nullstelle in $[-3, -1]$ und eine in $[-1, 0]$.

Wenn uns die Intervallängen nicht interessieren, sind wir damit fertig; wenn wir allerdings Intervalle der Länge eins wollen, müssen wir $[-3, -1]$ und $[2, 6]$ noch weiter unterteilen: An der Stelle -2 haben wir vier Vorzeichenwechsel in $(-, +, -, +, -, -)$, die Nullstelle liegt also in

$[-2, -1]$. Für $x = 4$ erhalten wir die Folge $(+, +, +, +, -, -)$ mit einem Vorzeichenwechsel; die Nullstelle liegt also in $[2, 4]$. Für $x = 3$ ergibt sich dieselbe Folge, sie liegt also in $[2, 3]$.

Wir haben also drei reelle Nullstellen, und sie liegen in den Intervallen $[-2, -1]$, $[-1, 0]$ und $[2, 3]$. Durch eine Zeichnung können wir uns vergewissern, daß die Nullstellen wirklich so liegen:



Natürlich hätten wir die Information über die Lage der Nullstellen auch direkt aus so einer Zeichnung ablesen können, aber es kann problematisch werden, aus Graphik *exakte* Schlüsse zu ziehen. Hätten wir f statt zwischen -2 und 3 zwischen -3 und 4 gezeichnet, hätte uns der Graph so gut wie nichts über die negativen Nullstellen sagen können, und eine Dartellung zwischen -5 und 5 wäre völlig nutzlos gewesen. Auch ist zu bedenken, daß Graphigsysteme praktisch immer mit numerischen Methoden arbeiten und das Ergebnis anschließend noch durch die Diskretisierung zur Pixelgraphik weiter vergröbert wird. Die Anwendung des Satzes von STURM dagegen führt zu beweisbar korrekten Ergebnissen.

Wie gut (und schnell) man die Nullstellen im konkreten Fall isolieren kann, hängt natürlich ab von deren Abstand. Erfahrungsgemäß funktioniert der Algorithmus recht gut; er ist auch in vielen Computeralgebrasystemen standardmäßig vorhanden.

Da der Algorithmus in der Praxis gut funktioniert, könnte man es dabei bewenden lassen und an eine Bemerkung von ZASSENHAUS denken,

der in einem Kolloquiumsvortrag an der Universität Karlsruhe einmal sagte: „In der experimentellen Mathematik haben wir es nicht nötig, die Sätze zu beweisen, die wir für wahr halten.“ Tatsächlich aber ist von ZASSENHAUS so gut wie keine unbewiesene Behauptung überliefert, und auch für unser Problem gibt es bewiesene Resultate:

KURT MAHLER (1903–1988) gab 1964 in seiner (inzwischen auch frei im Netz zugänglichen) Arbeit

K. MAHLER: An inequality for the discriminant of a polynomial, *Michigan Math. J.* **11** (1964), 257–262

eine untere Schranke für den Abstand zwischen zwei Nullstellen eines Polynoms mit komplexen Koeffizienten. Man findet sein Resultat auch in §7.2.4 des Buchs

ALKIVAIDIS G. AKRITAS: Elements of Computer Algebra with Applications, *Wiley,* 1989

Spätestens nach dem Ende dieser Vorlesung sollte beides gut verständlich sein.

§6: Resultanten

Um Resultanten mit dem EUKLIDischen Algorithmus in Verbindung zu bringen, könnten wir uns fragen, wann zwei Polynome in einer Veränderlichen einen nichtkonstanten gemeinsamen Teiler haben. Das ist natürlich genau dann der Fall, wenn der Grad ihres größten gemeinsamen Teilers positiv ist, aber man kann es auch dadurch charakterisieren, daß ein gewisses Polynom in ihren Koeffizienten, die *Resultante* verschwindet.

Resultanten erschienen erstmalig 1840 in einer Arbeit von JAMES JOSEPH SYLVESTER. Dabei ging es allerdings nicht um gemeinsame Teiler, sondern um gemeinsame Nullstellen, was – wenn man Nullstellen in hinreichend großen Erweiterungskörpern zuläßt – natürlich dasselbe ist. Wie wir sehen werden, können sie auch verwendet werden, um die Lösungsmenge eines nichtlinearen Gleichungssystems besser zu verstehen und eventuell sogar explizit zu bestimmen.

JAMES JOSEPH SYLVESTER (1814–1897) wurde geboren als JAMES JOSEPH; erst als sein Bruder nach USA auswanderte und dazu einen dreiteiligen Namen brauchte, erweiterte er aus Solidarität auch seinem Namen. 1837 bestand er das berüchtigte Tripos-Examen der Universität Cambridge als Zweitbester, bekam aber keinen akademischen Abschluß, da er als Jude den dazu vorgeschriebenen Eid auf die 39 Glaubensartikel der Church of England nicht leisten konnte. Trotzdem wurde er Professor am University College in London. Seine akademischen Grade bekam er erst 1841 aus Dublin, wo die Vorschriften gerade mit Rücksicht auf die Katholiken geändert worden waren. Während seiner weiteren Tätigkeit an sowohl amerikanischen als auch englischen Universitäten beschäftigte er sich mit Matrizen, fand die Diskriminante kubischer Gleichungen und entwickelte auch die allgemeine Theorie der Diskriminanten. In seiner Zeit an der Johns Hopkins University in Baltimore gründete er das American Journal of Mathematics, das noch heute zu den wichtigsten mathematischen Fachzeitschriften Amerikas zählt. Von 1883 bis zu seinem Tod hatte er einen Lehrstuhl für Geometrie an der Universität Oxford.

Wir beginnen mit zwei Polynomen

$$f = a_d X^d + a_{d-1} X^{d-1} + \cdots + a_1 X + a_0 \quad \text{mit} \quad a_d \neq 0$$

und

$$g = b_e X^e + b_{e-1} X^{e-1} + \cdots + b_1 X + b_0 \quad \text{mit} \quad b_e \neq 0$$

in einer Veränderlichen, lassen aber für die Koeffizienten nicht nur Elemente aus einem Körper zu, sondern aus einem beliebigen Integritätsbereich R.

Die Resultante der beiden Polynome f und g soll uns ein Kriterium dafür geben, daß f und g einen gemeinsamen Teiler h mit positivem Grad haben. In diesem Fall ist

$$\frac{fg}{h} = \frac{f}{h} \cdot g = \frac{g}{h} \cdot f$$

ein gemeinsames Vielfaches von f und g, dessen Grad

$$\deg f + \deg g - \deg h = d + e - \deg h$$

höchstens gleich $d + e - 1$ ist. Setzen wir $u = g/h$ und $v = f/h$ so ist also $uf = vg$, wobei $\deg u < \deg g$ und $\deg v < \deg f$ ist.

Diese Bedingung schreiben wir um in ein lineares Gleichungssystem für die Koeffizienten von u und v: Da $\deg u \leq \deg g - 1 = e - 1$ ist und $\deg v \leq \deg f - 1 = d - 1$, lassen sich die beiden Polynome schreiben als

$$u = u_{e-1}X^{e-1} + u_{e-2}X^{e-2} + \cdots + u_1X + u_0$$

und

$$v = v_{d-1}X^{d-1} + v_{d-2}X^{d-2} + \cdots + v_1X + v_0 \,.$$

Die Koeffizienten von X^r in uf und vg sind

$$\sum_{i,j \text{ mit } i+j=r} a_i u_j \quad \text{und} \quad \sum_{i,j \text{ mit } i+j=r} b_i v_j \,,$$

f und g haben daher genau dann einen gemeinsamen Teiler vom Grad mindestens r, wenn es nicht allesamt verschwindende Körperelemente $u_0, \ldots, u_{e-1}$ und $v_0, \ldots, v_{d-1}$ gibt, so daß

$$\sum_{i,j \text{ mit } i+j=r} a_i u_j - \sum_{i,j \text{ mit } i+j=r} b_i v_j = 0 \quad \text{für } r = 0, \ldots, d+e-1$$

ist. Dies ist ein homogenes lineares Gleichungssystem aus $d + e$ Gleichungen für die $d + e$ Unbekannten $u_0, \ldots, u_{e-1}$ und $v_0, \ldots, v_{d-1}$; es hat genau dann eine nichttriviale Lösung, wenn seine Matrix kleineren Rang als $d + e$ hat, wenn also deren Determinante verschwindet.

Ausgeschrieben wird das Gleichungssystem, wenn wir mit dem Koeffizienten von x^{d+e-1} anfangen, zu

$$\begin{aligned}
a_d u_{e-1} - b_e v_{d-1} &= 0 \\
a_{d-1}u_{e-1} + a_d u_{e-2} - b_{e-1}v_{d-1} - b_e v_{d-2} &= 0 \\
a_{d-2}u_{e-1} + a_{d-1}u_{e-2} + a_d u_{e-3} & \\
\qquad - b_{e-2}v_{d^{-1}} - b_{e-1}v_{d-2}b_e v_{d-3} &= 0 \\
\cdots & \\
a_0u_3 + a_1u_2 + a_2u_1 + a_3u_0 - b_0v_3 - b_1v_2 - b_2v_1 - b_3v_0 &= 0 \\
a_0u_2 + a_1u_1 + a_2u_0 - b_0v_2 - b_1v_1 - b_2v_0 &= 0 \\
a_0u_1 + a_1u_0 - b_0v_1 - b_1v_0 &= 0 \\
a_0u_0 - b_0v_0 &= 0
\end{aligned}$$

Natürlich ändert sich nichts an der nichttrivialen Lösbarkeit oder Unlösbarkeit dieses Gleichungssystems, wenn wir anstelle der Variablen v_j die Variablen $-v_j$ betrachten, womit alle Minuszeichen im obigen Gleichungssystem zu Pluszeichen werden; außerdem hat es sich – der größeren Übersichtlichkeit wegen – eingebürgert, die Transponierte der Matrix des Gleichungssystems zu betrachten. Dies führt auf die $(d+e) \times (d+e)$-Matrix

$$\begin{pmatrix}
a_d & a_{d-1} & a_{d-2} & \dots & a_1 & a_0 & 0 & 0 & \dots & 0 \\
0 & a_d & a_{d-1} & \dots & a_2 & a_1 & a_0 & 0 & \dots & 0 \\
0 & 0 & a_d & \dots & a_3 & a_2 & a_1 & a_0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \dots & a_d & a_{d-1} & a_{d-2} & a_{d-3} & \dots & a_0 \\
b_e & b_{e-1} & b_{e-2} & \dots & b_2 & b_1 & b_0 & 0 & \dots & 0 \\
0 & b_e & b_{e-1} & \dots & b_3 & b_2 & b_1 & b_0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \dots & 0 & b_e & b_{e-1} & b_{e-2} & \dots & b_0
\end{pmatrix},$$

in der e Zeilen aus Koeffizienten von f stehen und d Zeilen aus Koeffizienten von g.

Definition: Die obige Matrix heißt SYLVESTER-Matrix. Ihre Determinante ist die *Resultante* $\mathrm{Res}(f,g) = \mathrm{Res}_X(f,g)$ der beiden Polynome f und g bezüglich der Variablen X.

Der Index X ist dann notwendig, wenn schon der Ring R ein Polynomring ist, so daß f und g Polynome in mehreren Veränderlichen sind. Dann kann man bezüglich jeder der Variablen eine Resultante bilden, die dann ein Polynom in den übrigen Variablen ist, und natürlich sind diese Resultanten im Allgemeinen voneinander verschieden. Ist R ein Zahlbereich, können wir auf den Index X verzichten.

Damit haben wir gezeigt

Satz: Wenn zwei Polynome $f, g \in R[X]$ über dem Intgegritätsbereich R einen gemeinsamen Faktor positiven Grades haben, verschwindet ihre Resultante $\mathrm{Res}_X(f,g)$.

■

In den meisten Fällen gilt auch die Umkehrung dieses Resultats, allerdings reicht es dazu wohl nicht aus, R nur als einen Integritätsbereich vorauszusetzen: Wir brauchen so etwas wie die eindeutige Primzerlegung in den natürlichen Zahlen.

Definition: *a)* Ein Element f eines Integritätsbereichs R heißt *irreduzibel,* falls bei jeder Produktdarstellung $f = gh$ genau einer der beiden Faktoren eine Einheit sein muß.
b) R heißt *faktoriell,* wenn sich jede Nichteinheit außer der Null bis auf Reihenfolge und Assoziertheit eindeutig als Produkt irreduzibler Elemente darstellen läßt.

Im Ring der ganzen Zahlen sind alle Primzahlen irreduzibel, aber auch ihre Negativen. Wir können also die Zehn darstellen als $10 = 2 \cdot 5$, aber auch als $10 = (-1) \cdot (-5)$, und natürlich auch als $5 \cdot 2$ oder $(-5) \cdot (-2)$. Wenn wir die eindeutige Primzerlegung für natürliche Zahlen als bekannt voraussetzen, folgt also, daß $\mathbb{Z}$ ein faktorieller Ring ist.

Genauso folgt aus dem Fundamentalsatz der Algebra, daß der Polynomring $\mathbb{C}[X]$ faktoriell ist, denn jedes Polynom vom Grad $n \geq 1$ läßt sich darstellen in der Form

$$f = a_n(X - x_1) \cdots (X - x_n) .$$

Die die Einheiten die von Null verschiedenen komplexen Zahlen sind, folgt, daß die irreduziblen Polynome gerade die linearen sind. Indem wir den Faktor a_n irgendwie auf die Polynome $X - x_j$ verteilen, können wir jedes nichtkonstante Polynom als Produkt von irreduziblen darstellen, und da die Nullstellenmenge eines Polynoms eindeutig bestimmt ist, folgt die Eindeutigkeit bis auf Reihenfolge und Assoziiertheit: Die einzigen linearen Polynome mit Nullstelle x_j sind die Polynome $cX - cx_j$ mit $c \in \mathbb{C}^\times$, und die sind alle assoziiert zueinander, da c eine Einheit ist.

In $\mathbb{R}[X]$ sind natürlich nicht alle irreduziblen Polynome linear; auch quadratische Polynome ohne reelle Nullstellen sind irreduzibel. In $\mathbb{Z}[X]$ schließlich ist nicht jedes lineare Polynom irreduzibel: Beispielsweise läßt sich $2X-4$ als Produkt der Nichteinheiten zwei und $X-2$ darstellen.

Als Beispiel eines nichtfaktoriellen Rings können wir $\mathbb{Z} \oplus \mathbb{Z}\sqrt{-5}$ betrachten: Dort ist $6 = 2 \cdot 3 = \left(1 + \sqrt{-5}\right)\left(1 - \sqrt{-5}\right)$, und alle vier Zahlen $2, 3$ und $1 \pm \sqrt{-5}$ sind irreduzibel: Das Betragsquadrat eines Ringelements $a + b\sqrt{-5}$ ist $a^2 + 5b^2$, und ist $a + b\sqrt{-5}$ ein Teiler von $2, 3$ oder $1 \pm \sqrt{-5}$, so muß $a^2 + 5b^2$ deren Betragsquadrat $4, 9$ *bzw.* 6 teilen. Betragsquadrat eins haben nur die beiden Einheiten ± 1; bei einer Produktdarstellung ohne Einheiten brauchen wir also Zahlen, deren Betragsquadrate echte Teiler von $4, 9$ oder 6 sind. Offensichtlich gibt es aber keine ganzen Zahlen a, b mit $a^2 + 5b^2 = 2$ oder 3, so daß die vier Zahlen in der Tat irreduzibel sind. Natürlich ist $1 \pm \sqrt{-5}$ auch nicht assoziiert zu 2 oder 3, so daß wir hier zwei wesentlich verschiedene Zerlegungen der Sechs in ein Produkt zweier irreduzibler Elemente haben.

Der Beweis des Hauptsatzes der elementaren Zahlentheorie, wonach sich jede natürliche Zahl größer eins bis auf Reihenfolge eindeutig in ein Produkt von Primzahlen zerlegen läßt, führt praktisch sofort zu einem Beweis dafür, daß jeder EUKLIDische Ring faktoriell ist: Abgesehen von einigen recht einfachen Aussagen über die Funktion ν braucht man dafür vor allem die Aussage, daß ein irreduzibles p ein Produkt ab nur dann teilen kann, wenn es mindestens einen der beiden Faktoren teilt. Falls ein Teiler p von ab kein Teiler von a ist, sind a und p teilerfremd, denn jeder Teiler von p ist entweder assoziiert zu p oder eine Einheit. Somit gibt es Ringelemente u, v, für die $1 = ua + vp$ ist. Multiplikation mit b macht daraus die Gleichung $b = u(ab) + p(vb)$, deren rechte Seite offensichtlich durch p teilbar ist.

Umgekehrt ist aber nicht jeder faktorielle Ring EUJLIDisch: Beispielsweise zeigte bereits GAUSS, daß der Polynomring $\mathbb{Z}[X]$ (in heutiger Terminologie) faktoriell ist. Mit seiner Beweisidee läßt sich allgemeiner zeigen, daß für jeden faktoriellen Ring R auch der Polynomring $R[X]$ faktoriell ist. Induktiv folgt daraus sofort, daß auch die Polynomringe $R[X_1, \ldots, X_n]$ in mehreren Variablen faktoriell sind.

Für Beweise der obigen Aussagen sei auf Vorlesungen, Skripten oder Lehrbücher der Algebra verwiesen. Sie sind auch in einigen Versionen dieses Skriptums für vergangene Semester zu finden.

Für faktorielle Ringe gilt nun der gewünschte Satz:

Satz: Zwei Polynome $f, g \in R[X]$ über einem faktoriellen Ring R haben genau dann einen gemeinsamen Faktor positiven Grades, wenn die Resultante $\mathrm{Res}_X(f, g)$ verschwindet.

Beweis: Wie wir beim Beweis des obigen Satzes gesehen haben, ist das Verschwinden der Resultante äquivalent dazu, daß ein homogenes lineares Gleichungssystem eine nichttriviale Lösung hat, und eine solche Lösung führt auf Polynome $u, v \in R[X]$ mit $\deg u < \deg g$ und $\deg v < \deg f$, für die $uf = vg$ ist. Mit R ist nach GAUSS auch der Polynomring $R[X]$ faktoriell; wir können also die Zerlegungen von f und g in irreduzible Faktoren betrachten. Falls f und g keinen gemeinsamen Teiler positiven Grades haben, kann in der Primzerlegung von f kein Polynom vorkommen, das assoziiert ist zu einem Polynom aus der Zerlegung von g. In der Zerlegung eines gemeinsamen Vielfachen h müssen daher alle irreduziblen Polynome vorkommen, die in der Zerlegung von f oder g vorkommen, und sie müssen auch mit mindestens der gleichen Potenz vorkommen wie dort. Damit ist $\deg h \geq \deg f + \deg g$, im Widerspruch zu der Tatsache, daß $uf = vg$ ein gemeinsames Vielfaches mit kleinerem Grad ist. Daher müssen f und g einen gemeinsamen Teiler positiven Grades haben.

■

Resultanten wurden ursprünglich eingeführt zum Lösen nichtlinearer Gleichungssysteme, und das ist auch heute noch eine wichtige Anwendung. Betrachten wir zunächst den einfachsten Fall von zwei Polynomen in zwei Veränderlichen.

Angenommen, wir haben zwei Polynome $f, g \in k[X, Y]$ über einem Körper k und suchen deren gemeinsame Nullstellenmenge. Wie wir schon bei Polynomen in einer Veränderlichen gesehen haben, sollten wir uns dabei nicht auf Nullstellen aus k beschränken, sondern auch in Erweiterungskörpern K von k nach Lösungen suchen.

Wir können f und g auffassen als Polynome in X über $k[Y]$ und ihre Resultante $\mathrm{Res}_X(f, g) \in k[Y]$ betrachten. Für einen speziellen Wert $y \in K$ können wir auch die Resultante der beiden Polynome $f(X, y)$

und $g(X, y)$ aus $K[X]$ betrachten. Falls die Koeffizienten der jeweils höchsten X-Potenz von f und g an der Stelle y nicht verschwinden, hat die SYLVESTER-Matrix von $f(X, y)$ und $g(X, y)$ die gleiche Gestalt wie die von f und g über $k[Y]$, und diese geht durch Einsetzen von y über in die von $f(X, y)$ und $g(X, y)$. Somit ist die Resultante von $f(X, y)$ und $g(X, y)$ gleich dem Wert des Polynoms $\mathrm{Res}_X(f, g)$ an der Stelle y und verschwindet genau dann, wenn y eine Nullstelle von $\mathrm{Res}_X(f, g)$ ist. Die beiden Polynome $f(X, y)$ und $g(X, y)$ haben in diesem Fall also genau dann einen gemeinsamen Faktor positiven Grades, wenn $\mathrm{Res}_X(f, g)$ an der Stelle y verschwindet. Zumindest in einem geeigneten Erweiterungskörper von K haben sie dann auch eine gemeinsame Nullstelle x, und (x, y) ist eine Lösung unseres Gleichungssystems.

Falls einer der führenden X-Koeffizienten von f oder g an der Stelle y verschwindet, können wir nicht so argumentieren: In diesem Fall ist die SYLVESTER-Matrix von $f(X, y)$ und $g(X, y)$ kleiner als die von f und g, und wir müssen die Resultante von $f(X, y)$ und $g(X, y)$ direkt berechnen. Da die führenden X-Koeffizienten von f und g Polynome in Y sind, gibt es aber höchstens endlich viele Werte, für die das nötig wird.

Falls $\mathrm{Res}_X(f, g)$ nicht verschwindet, ist das ein Polynom in Y, das ebenfalls höchstens endlich viele Nullstellen hat. Insgesamt hat man damit, sofern man die Nullstellen der auftretenden Polynome in einer Veränderlichen bestimmen kann, eine endliche Menge von Werten y mit der Eigenschaft, daß $f(X, y)$ und $g(X, y)$ nur für diese Werte gemeinsame Nullstellen haben können. Wenn die Nullstellen eines der beiden Polynome einfach zu erkennen sind, genügt es, diese in das andere einzusetzen. Ansonsten empfiehlt es sich, den ggT der beiden Polynome zu bestimmen und dessen Nullstellen zu berechnen, was wegen des kleineren Grads meist einfacher sein dürfte.

Natürlich können wir die Rolle von X und Y auch vertauschen und $\mathrm{Res}_Y(f, g) \in k[X]$ bestimmen. Gelegentlich kann sich dadurch der Aufwand dramatisch ändern, da die Nullstellen einer der beiden Resultanten einfacher zu finden sind als die der anderen.

Als Beispiel für die Lösung eines nichtlinearen Gleichungssystems mit

Resultanten betrachten wir die beiden Gleichungen

$f(x,y) = x^2+2y^2+8x+8y-40$ und $g(x,y) = 3x^2+y^2+18x+4y-50$.

Ihre Resultante bezüglich X ist

$$\mathrm{Res}_X(f,g) = 25Y^4 + 200Y^3 - 468Y^2 - 3472Y + 6820\,;$$

Maxima gibt ihre Nullstellen an als

$$y = -2 \pm \frac{1}{5}\sqrt{534 \pm 24\sqrt{31}}\,.$$

Diese können wir beispielsweise in g einsetzen, die entstehende quadratische Gleichung für x lösen, um dann zu testen, ob das Lösungspaar (x,y) auch eine Nullstelle von g ist. Zumindest mit einem Computeralgebrasystem ist dass durchaus machbar.

Einfacher wird es aber, wenn wir Y an Stelle von X eliminieren:

$$\mathrm{Res}_Y(f,g) = (5X^2 + 28X - 60)^2$$

ist das Quadrat eines quadratischen Polynoms; dessen Nullstellen

$$x = -\frac{14}{5} \pm \frac{4}{5}\sqrt{31}$$

uns die wohlbekannte Lösungsformel liefert. Diese Werte können wir nun in f oder g einsetzen, die entstehende Gleichung lösen und das Ergebnis ins andere Polynom einsetzen.

Alternativ können wir auch mit *beiden* Resultanten arbeiten: Ist (x,y) eine gemeinsame Nullstelle von f und g, so muß x eine Nullstelle von $\mathrm{Res}_y(f,g)$ sein und y eine von $\mathrm{Res}_X(f,g)$. Da es nur $4 \times 2 = 8$ Kombinationen gibt, können wir diese hier einfach durch Einsetzen testen. Wie sich zeigt, hat das System die vier Lösungen

$$\left(-\frac{14}{5} + \frac{4}{5}\sqrt{31},\ -2 - \frac{1}{5}\sqrt{534 - 24\sqrt{31}}\right)$$
$$\left(-\frac{14}{5} + \frac{4}{5}\sqrt{31},\ -2 + \frac{1}{5}\sqrt{534 - 24\sqrt{31}}\right)$$
$$\left(-\frac{14}{5} - \frac{4}{5}\sqrt{31},\ -2 - \frac{1}{5}\sqrt{534 + 24\sqrt{31}}\right)$$
$$\left(-\frac{14}{5} - \frac{4}{5}\sqrt{31},\ -2 + \frac{1}{5}\sqrt{534 + 24\sqrt{31}}\right).$$

Für ein allgemeines Gleichungssystem

$$f_1(x_1, \ldots, x_n) = \cdots = f_m(x_1, \ldots, x_n) = 0$$

betrachten wir die $f_i \in k[X_1, \ldots, X_n]$ als Polynome in X_n mit Koeffizienten aus $k[X_1, \ldots, X_{n-1}]$. Falls die Resultante $\mathrm{Res}_{X_n}(f_i, f_j)$ für zwei Polynome f_i, f_j das Nullpolynom ist, haben f_i und f_j einen gemeinsamen Faktor, aber das wird wohl nur selten der Fall sein. Falls wir die Polynome vorher faktorisieren und dann das Gleichungssystem ersetzen durch mehrere Systeme aus irreduziblen Polynomen, können wir das sogar ausschließen.

Häufiger und interessanter ist der Fall, daß die Resultante nur für gewisse $(n-1)$-tupel $(x_1, \ldots, x_{n-1}) \in k^{n-1}$ verschwindet. Dann wissen wir, daß die Polynome

$$f_i(x_1, \ldots, x_{n-1}, X_n) \quad \text{und} \quad f_j(x_1, \ldots, x_{n-1}, X_n)$$

aus $k[X_n]$ zumindest in einem Erweiterungskörper von k eine gemeinsame Nullstelle haben. Falls wir $x_1, \ldots, x_{n-1}$ kennen, können wir diese Nullstelle(n) bestimmen, indem wir die Nullstellen zweier Polynome in einer Veränderlichen berechnen und miteinander vergleichen oder, meist einfacher, die Nullstellen des größten gemeinsamen Teilers berechnen.

Um das obige Gleichungssystem zu lösen, führen wir es also zurück auf das Gleichungssystem

$$\mathrm{Res}_{x_n}(f_i, f_{i+1})(x_1, \ldots, x_{n-1}) = 0 \quad \text{für } i = 1, \ldots, m-1\,,$$

lösen dieses und betrachten für jedes Lösungstupel jenes Gleichungssystem in x_n, das entsteht, wenn wir im Ausgangssystem für die ersten $n-1$ Variablen die Werte aus dem Tupel einsetzen. Die Lösungen dieses Gleichungssystems sind gerade die Nullstellen des größten gemeinsamen Teilers aller Gleichungen.

Man beachte, daß dieser ggT durchaus gleich eins sein kann, daß es also nicht notwendigerweise eine Erweiterung des Tupels $(x_1, \ldots, x_{n-1})$ zu einer Lösung des gegebenen Gleichungssystems gibt: Wenn alle Resultanten verschwinden, haben nach Einsetzen zwar f_1 und f_2 eine

gemeinsame Nullstelle und genauso auch f_2 und f_3, aber diese beiden Nullstellen können verschieden sein. Es muß also keine gemeinsame Nullstelle von f_1, f_2 und f_3 geben. Als Beispiel betrachten wir die drei Polynome

$$f_1 = (X-1)^2(X-2)^2 + (Y-2)^2$$
$$f_2 = (X-1)^2(X-3)^2 + (Y-2)^2$$
$$f_3 = (X-2)^2(X-3)^2 + (Y-2)^2 .$$

Da im Reellen eine Summe von Quadraten nur verschwinden kann, wenn alle Summanden verschwinden, hat f_1 in $\mathbb{R}^2$ nur die beiden Nullstellen $(1,2)$ und $(2,2)$. Bei f_2 sind es entsprechend $(1,2)$ und $(3,2)$, und bei f_3 schließlich $(2,2)$ und $(3,2)$. Somit haben f_1 und f_2 die gemeinsame Nullstelle $(1,2)$, bei f_1 und f_3 ist es $(3,2)$ und bei f_2 und f_3 haben wir die Lösung $(2,2)$. Das System aus allen drei Gleichungen aber hat zumindest im Reellen keine Lösung.

Läßt man ein Computeralgebrasystem die Resultanten ausrechnen und faktorisieren, erhält man

$$\mathrm{Res}_X(f_1, f_2) = \mathrm{Res}_X(f_2, f_3) = (Y-2)^4(16Y^2 - 64Y + 73)$$
$$\text{und} \qquad \mathrm{Res}_X(f_1, f_3) = 256(Y-2)^6 .$$

Der quadratische Faktor hat die beiden konjugiert komplexen Nullstellen $2 \pm \frac{3}{4}i$; setzt man das für Y ein, haben f_1 und f_3 die gemeinsame Nullstelle $x = \frac{5}{2}$, bei f_2 und f_3 ist es $x = \frac{3}{2}$. Auch im Komplexen ist das System aus allen drei Gleichungen somit unlösbar.

Für die Lösung eines nichtlinearen Gleichungssystems mit Resultanten ist es natürlich unerläßlich, die Nullstellen der Resultanten *exakt* zu berechnen. Verwendet man nur eine numerische Approximation, wird man beim Einsetzen praktisch immer zwei Gleichungen ohne gemeinsame Nullstellen bekommen. Wie wir im vorigen Kapitel gesehen haben, lassen sich aber die Nullstellen eines Polynoms vom Grad mindestens fünf in den meisten Fällen nicht mehr durch Wurzelausdrücke darstellen. Als Ausweg hatten wir gesehen, daß man über den Satz von STURM zumindest für die reellen Nullstellen Intervalle finden kann, in denen

jeweils genau eine der Nullstellen liegt, womit diese eindeutig charakterisiert ist. Wenn wir eine so dargestellte Nullstelle der Resultante in ein Polynom einsetzen wollen, müssen wir allerdings mit solchen Ausdrücken rechnen können, und auch dabei helfen Resultanten.

Definition: Eine Zahl $z \in \mathbb{C}$ heißt algebraisch, wenn es ein Polynom f aus $\mathbb{Q}[X]$ gibt, das an der Stelle z verschwindet. Andernfalls heißt z transzendent.

Da sich an den Nullstellen eines Poynoms nichts ändert, wenn wir es mit dem Hauptnenner seiner Koeffizienten multiplizieren, können wir auch sagen, die algebraischen Zahlen seien die Nullstellen der Polynome mit ganzzahligen Koeffizienten.

So ist beispielsweise $\sqrt{2}$ als Nullstelle des Polynoms X^2-2 algebraisch, π dagegen ist nach einem 1882 bewiesenen Resultat von FERDINAND VON LINDEMANN (1852–1939) transzendent.

Man kann zeigen, daß die algebraischen Zahlen einen abzählbaren Teilkörper von $\mathbb{C}$ bilden und daß dieser Körper auch algebraisch abgeschlossen ist, d.h. jedes Polynom mit algebraischen Zahlen als Koeffizienten hat auch algebraische Nullstellen. Hier beschränken wir uns auf reelle algebraische Zahlen. Diese bilden zwar auch einen Körper, aber der ist natürlich nicht algebraisch abgeschlossen, da beispielsweise das Polynom $X^2 + 1$ keine reelle Nullstelle hat.

Eine reelle algebraische Zahl ist eindeutig festgelegt durch die Angabe eines Polynoms f aus $\mathbb{Q}[X]$ und eines Intervalls $[a,\, b]$, in dem f nur eine Nullstelle hat. Die Intervallenden wählen wir – um exakt rechnen zu können – als rationale Zahlen. Wir wollen uns überlegen, daß wir mit solchen Paaren $\big(f, [a,\, b]\big)$ alle Grundrechenarten durchführen können, was gleichzeitig zeigen wird, daß die reellen algebraischen Zahlen einen Körper bilden. Wir müssen uns auch überlegen, wie wir Gleichheit sowie Größerbeziehungen entscheiden können.

Dazu seien u und v zwei reelle algebraische Zahlen, gegeben durch die Paare $\big(f, [a,\, b]\big)$ und $\big(g, [c,\, d]\big)$.

Beginnen wir mit der Addition: Für $w = u+v$ ist $g(w-u) = g(v) = 0$. Wir führen eine neue Variable Z ein und schreiben $g(Z - X)$ als Polynom

in X mit Koeffizienten aus $\mathbb{Q}[Z]$. Natürlich können wir auch $f \in \mathbb{Q}[X]$ als ein solches Polynom auffassen: Die Koeffizienten von f sind rationale Zahlen und damit auch (konstante) Polynome aus $\mathbb{Q}[Z]$. Wir haben damit zwei Polynome in X mit Koeffizienten aus $\mathbb{Q}[Z]$, die für $z = u+v$ die gemeinsame Nullstelle $X = u$ haben.

Zwei Polynome in X über einem faktoriellen Ring R haben genau dann eine gemeinsame Nullstelle, wenn ihre Resultante verschwindet. Diese ist hier ein Polynom $h \in \mathbb{Q}[Z]$, das somit für $Z = u + v$ verschwinden muß. Damit haben wir ein Polynom h mit rationalen Koeffizienten gefunden, das $w = u + v$ als Nullstelle hat.

Da $a \le u \le b$ und $c \le v \le d$, liegt w natürlich im Intervall $[a+c,\ b+d]$. Es ist allerdings nicht klar, ob w dort die einzige Nullstelle von h ist. Das läßt sich aber mit dem Satz von STURM überprüfen; gegebenenfalls müssen die Intervalle $[a,\ b]$ und $[c,\ d]$ so lange verkleinert werden, bis das Intervall $[a + c,\ b + d]$ nur noch eine Nullstelle enthält

Damit wissen wir, wie man Summen berechnet; um auch Differenzen berechnen zu können, müssen wir uns nur überlegen, wie man für eine durch $\big(f, [a,\ b]\big)$ dargestellte reelle algebraische Zahl u ihr Negatives darstellt. Mit

$$f = \sum_{i=0}^{n} a_i X^i \quad \text{ist} \quad h = \sum_{i=0}^{n} a_i (-X)^i = \sum_{i=0}^{n} (-1)^i a_i X^i$$

offensichtlich ein Polynom, das $-u$ als Nullstelle hat, und $-u$ liegt im Intervall $[-b,\ -a]$. Es ist dort auch die einzige Nullstelle, denn jede Nullstelle von h ist das Negative einer Nullstelle von f.

Für Produkte können wir fast genauso vorgehen wir für Summen: Ist $w = uv$ und $u \neq 0$, so ist $g(w/u) = 0$. Um aus $g(Z/X)$ ein Polynom zu machen, müssen wir mit X^m multiplizieren, wobei m den Grad von g bezeichnet: Für $g = b_m X^m + \cdots + b_0$ betrachten wir also

$$X^m g(Z/X) = b_m Z^m + b_{m-1} Z^{m-1} X + \cdots + b_1 Z X^{m-1} + b_0 X^m\,,$$

ein Polynom vom Grad m in X mit Koeffizienten aus $\mathbb{Q}[Z]$. Auch f können wir als so ein Polynom auffassen und erhalten wie oben, daß die Resultante dieser beiden Polynome ein rationales Polynom ist, das in w verschwindet. Das Intervall, in dem w liegt, läßt sich leicht aus

a, b, c und d berechnen, allerdings sind Fallunterscheidungen bezüglich der Vorzeichen nötig. Auch hier kann es wieder notwendig werden, die Ausgangsintervalle zu verkleinern um sicherzustellen, daß w die einzige Nullstelle im Intervall ist.

Fehlen schließlich noch die Quotienten, und dazu reicht es, wenn wir zu einer gegebenen reellen algebraischen Zahl u ein Polynom und ein Intervall für ihren Kehrwert finden. Ist u eine Nullstelle von f, so ist $1/u$ offensichtlich Nullstelle von

$$X^n f(1/X) = a_0 X^n + a_1 X^{n-1} + \cdots + a_{n-1} X + a_n \,.$$

Beim Intervall $[a,\ b]$ für u müssen wir zunächst sicherstellen, daß es die Null nicht enthält, daß also a und b dasselbe Vorzeichen haben; dann liegt $1/u$ im Intervall $[1/b,\ 1/a]$.

Damit lassen sich alle vier Grundrechenarten algorithmisch ausführen, und wir haben auch gezeigt, daß die reellen algebraischen Zahlen einen Körper bilden.

Als Beispiel wollen wir Summe und Produkt von $\sqrt{2}$ und $\sqrt{3}$ auf diese Weise behandeln. $\sqrt{2}$ ist Nullstelle des irreduziblen Polynoms $f = X^2 - 2$ und liegt im Intervall $[0,\ 2]$, entsprechend ist $\sqrt{3}$ Nullstelle des Polynoms $g = X^2 - 3$ und liegt im Intervall $[0,\ 3]$.

$$g(Z - X) = (Z - X)^2 - 3 = X^2 - 2ZX + Z^2 - 3\,;$$

wir müssen also die Resultante

$$\begin{vmatrix} 1 & 0 & -2 & 0 \\ 0 & 1 & 0 & -2 \\ 1 & -2Z & Z^2 - 3 & 0 \\ 0 & 1 & -2Z & Z^2 - 3 \end{vmatrix} = Z^4 - 10Z^2 + 1$$

von f und diesem Polynom berechnen.

Da f auch die Nullstelle $-\sqrt{2}$ hat und g entsprechend die Nullstelle $-\sqrt{3}$, sind alle vier Zahlen $\pm\sqrt{2} \pm \sqrt{3}$ Nullstellen dieser Resultante; da sie Grad vier hat kennen wir somit ihre sämtlichen Nullstellen.

Die Summe einer Zahl aus $[0,\ 2]$ und einer aus $[0,\ 3]$ liegt in $[0,\ 5]$; somit ist $\sqrt{2} + \sqrt{3}$ eine Nullstelle des Polynoms $X^4 - 10X^2 + 1$ aus diesem Intervall. Wir müssen überprüfen, ob es die einzige ist.

Da wir alle vier Nullstellen kennen, brauchen wir dazu keinen Satz von STURM, sondern sehen auch so, daß die Nullstelle $\sqrt{3}-\sqrt{2}$ ebenfalls in diesem Intervall liegt. Wir müssen wir die Ausgangsintervalle also verkleinern.

Wegen $1^2 < 2 < 3 < 2^2$ liegen sowohl $\sqrt{2}$ als auch $\sqrt{3}$ im Intervall $[1,\,2]$; ihre Summe liegt daher in $[2,\,4]$. In diesem Intervall liegt keine weitere Nullstelle, denn $\sqrt{3}-\sqrt{2} \in [0,\,1]$. Somit können wir $\sqrt{2}+\sqrt{3}$ charakterisieren als *die* Nullstelle von X^4-10X^2+1 im Intervall $[2,\,4]$.

Für das Produkt $\sqrt{2}\cdot\sqrt{3}=\sqrt{6}$ müssen wir, wenn wir strikt nach Schema vorgehen, zunächst das Polynom

$$X^2 g(Z/X) = X^2\left(\frac{Z^2}{X^2}-3\right) = Z^2-3X^2$$

bestimmen und seine Resultante mit f berechnen:

$$\begin{vmatrix} 1 & 0 & -2 & 0 \\ 0 & 1 & 0 & -2 \\ -3 & 0 & Z^2 & 0 \\ 0 & -3 & 0 & Z^2 \end{vmatrix} = \left(Z^2-6\right)^2 .$$

Dieses Polynom hat nur die Nullstellen $\pm\sqrt{6}$, wir können das Produkt von $\sqrt{2}$ und $\sqrt{3}$ also charakterisieren als die Nullstelle von $(Z^2-6)^2$ in $[0,\,6]$. Natürlich reicht auch Z^2-6; die Resultante liefert offensichtlich nicht immer das kleinstmögliche Ergebnis. Glücklicherweise erlaubt uns die Computeralgebra, jedes Polynom über $\mathbb{Q}$ zu faktorisieren, und wir können dann nachprüfen, welcher irreduzible Faktor für die betrachtete Zahl verschwindet.

Wir müssen uns noch überlegen, daß wir auch entscheiden können, wann zwei Darstellungen $\big(f,[a,\,b]\big)$ und $\big(g,[c,\,d]\big)$ die gleiche Zahl darstellen. Wenn der Durchschnitt der beiden Intervalle leer ist, kann das unmöglich der Fall sein, ebenso wenig, wenn f und g teilerfremd sind, denn dann gibt es keine gemeinsame Nullstelle.

Falls $\mathrm{ggT}(f,g)$ positiven Grad hat und $[a,\,b]\cap[c,\,d]$ nicht leer ist. können wir zunächst (z.B. nach STURM) überprüfen, ob der ggT im

Durchschnitt der beiden Intervalle eine Nullstelle hat. Falls nein, können die Ausgangszahlen nicht gleich sein.

Andernfalls ist diese Nullstelle auch eine Nullstelle von f, und sie liegt insbesondere im Intervall $[a,\ b]$. Da f dort nur die eine Nullstelle x hat, muß sie also gleich x sein. Genauso folgt, daß sie gleich y sein muß, und das zeigt die Gleichheit von x und y.

Die Relationen $<$ und $>$ lassen sich ebenfalls leicht entscheiden, z.B. durch Verkleinerung der Intervalle oder durch Berechnung der Differenz und Untersuchung auf positive oder negative Nullstellen von deren definierendem Polynom.

Für die praktische Anwendung von Resultanten ist es unerläßlich, daß wir diese effizient berechnen können. Die Resultante zweier Polynome der Grade 30 und 40 ist eine 70×70-Determinante. Der Entwicklungssatz von LAPLACE stellt diese dar als eine Summe mit $70! \approx 1,2 \cdot 10^{100}$ Summanden. Wegen der speziellen Form der SYLVESTER-Matrix werden zwar viele davon verschwinden zwar viele Summanden, aber trotzdem ist es unrealistisch, die Resultante so berechnen zu wollen: In der Kryptographie geht man davon aus, daß ein Verschlüsselungsverfahren schon dann praktisch sicher ist, wenn ein Gegner zum Knacken etwa $2^{128} \approx 3{,}4 \cdot 10^{38}$ Rechenschritte benötigt. Tatsächlich verwendet aber natürlich ohnehin niemand den Entwicklungssatz von LAPLACE um eine große Determinante zu berechnen; dessen Nützlichkeit beschränkt sich definitiv auf theoretische Fragen und kleinere Spielzeugdeterminanten, wie sie vor allem in Mathematikklausuren vorkommen. In realistischen Anwendungen wird man die Matrix durch Zeilen- und/oder Spaltenoperationen auf Dreiecksform bringen und dann die Determinante einfach als Produkt der Diagonaleinträge berechnen, oder man führt eine LR- oder QR-Zerlegung durch. Das dauert für die SYLVESTER-Matrix zweier Polynome der Grade dreißig und vierzig auf heutigen Computern weniger als eine halbe Minute.

Stellt man allerdings keine Matrix auf, sondern verlangt von einem Computeralgebrasystem einfach, daß es die Resultante der beiden Polynome berechnen soll, hat man das Ergebnis nach weniger als einem Zehntel

der Zeit. Einer der Schlüssel dazu ist wieder einmal der EUKLIDische Algorithmus.

Angenommen, wir haben zwei Polynome f, g in einer Variablen X über einem faktoriellen Ring R:

$$f = a_d X^d + a_{d-1} X^{d-1} + \cdots + a_1 X + a_0 \quad \text{und}$$
$$g = b_e X^e + b_{e-1} X^{e-1} + \cdots + b_1 X + b_0 \quad \text{mit} \quad d \leq e\,.$$

Falls $f = a_0$ konstant ist, also $d = 0$, gibt es in der SYLVESTER-Matrix null Zeilen aus Koeffizienten von g und e Zeilen aus Koeffizienten von f; die Matrix ist also einfach a_0 mal der $e \times e$-Einheitsmatrix, und die Resultante als ihre Determinante ist a_0^e.

Andernfalls dividieren wir g durch f und erhalten einen Rest h:

$$g : f = q \text{ Rest } h \quad \text{oder} \quad h = g - qf\,.$$

Das ist freilich nur dann möglich, wenn $R[X]$ ein EUKLIDischer Ring ist, also im wesentlichen nur dann, wenn R ein Körper ist. Das ist aber keine so große Einschränkung wie es scheint, denn zu jedem Integritätsbereich R können wir den sogenannten *Quotientenkörper* $K = \operatorname{Quot} R$ konstruieren als die Menge aller Paare (f, g) mit $f, g \in R$ und $g \neq 0$. Zwei Paare (f, g) und (f^*, g^*) sollen dabei als das gleiche Element von K gelten, wenn $fg^* = f^*g$ ist. Die Summe zweier Paare (f, g) und (f^*, g^*) ist $(fg^* + f^*g, gg^*)$ und ihr Produkt (ff^*, gg^*), wie es den Regeln der Bruchrechnung für die Brüche f/g und f^*/g^* entspricht.

Im Polynomring $K[X]$ haben wir eine Division mit Rest, und wenn wir dort die Resultante zweier Polynome aus $R[X]$ berechnen, bekommen wir bei der hier betrachteten Methode zwar Zwischenergebnisse aus $K[X]$, aber die Resultante ist unabhhängig von der Art der Berechnung ein eindeutig bestimmtes Element von R. Die Zwischenergebnisse können freilich recht große Nenner bekommen – ein wohlbekanntes Problem der Computeralgebra, das uns bereits beim EUKLIDischen Algorithmus für Polynome begegnet ist.

Der zentrale Punkt beim EUKLIDischen Algorithmus ist, daß die gemeinsamen Teiler von f und g genau dieselben sind wie die von f und h. Insbesondere haben also f und g genau dann einen gemeinsamen Teiler von

positivem Grad, wenn f und h einen haben, d.h. $\mathrm{Res}_X(f,g)$ verschwindet genau dann, wenn $\mathrm{Res}_X(f,h)$ verschwindet. Damit sollte es also einen Zusammenhang zwischen den beiden Resultanten geben, und den können wir zur Berechnung von $\mathrm{Res}_X(f,g)$ ausnützen, denn natürlich ist die Determinante $\mathrm{Res}_X(f,h)$ kleiner und einfacher als $\mathrm{Res}_X(f,g)$.

Bei der Polynomdivision von g durch f berechnen wir eine Folge von Polynomen $g_0 = g, g_1, \ldots, g_r = h$, wobei g_i aus seinem Vorgänger dadurch entsteht, daß wir ein Vielfaches von $X^j f$ subtrahieren, wobei $j = \deg g_i - \deg f$ ist. Der maximale Wert, den j annehmen kann, ist offenbar

$$\deg g - \deg f = e - d\,.$$

Wir wollen uns überlegen, wie sich die SYLVESTER-Matrix ändert, wenn wir dort die Koeffizienten von $g_0 = g$ nacheinander durch die der nachfolgenden g_i ersetzen. Um die Gestalt der Matrix nicht zu verändern, behandeln wir dabei auch die g_i wie Polynome vom Grad e, indem wir die Koeffizienten aller X-Potenzen mit einem Exponent oberhalb $\deg g_i$ auf Null setzen.

Die Zeilen der SYLVESTER-Matrix sind Vektoren in R^{d+e}; die ersten e sind die Koeffizientenvektoren von $X^{e-1}f, \ldots, Xf, f$, danach folgen die von $X^{d-1}g, \ldots, Xg, g$.

Im ersten Divisionschritt subtrahieren wir von g ein Vielfaches $\lambda X^j f$ mit $j = e - d$; damit subtrahieren wir auch von jeder Potenz $X^i g$ das Polynom $\lambda X^{i+j} f$. Für $0 \le i < d$ und $0 \le j \le e - d$ ist $0 \le i + j < e$. Was wir subtrahieren entspricht auf dem Niveau der Koeffizientenvektoren also stets einem Vielfachen einer Zeile der SYLVESTER-Matrix. Damit ändert sich nichts am Wert der Determinanten, wenn wir den Koeffizientenvektor von g nacheinander durch den von $g_1, \ldots, g_r = h$ ersetzen.

Die Resultante ändert sich also nicht, wenn wir in der SYLVESTER-Matrix jede Zeile mit Koeffizienten von g ersetzen durch die entsprechende Zeile mit Koeffizienten von h, wobei h wie ein Polynom vom Grad e behandelt wird, dessen führende Koeffizienten verschwinden.

Ist $h = c_s X^s + \cdots + c_1 X + c_0$, so ist also $\mathrm{Res}_X(f,g)$ gleich

$$\begin{vmatrix}
a_d & a_{d-1} & a_{d-2} & \dots & a_1 & a_0 & 0 & 0 & \dots & 0 \\
0 & a_d & a_{d-1} & \dots & a_2 & a_1 & a_0 & 0 & \dots & 0 \\
0 & 0 & a_d & \dots & a_3 & a_2 & a_1 & a_0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \dots & a_d & a_{d-1} & a_{d-2} & a_{d-3} & \dots & a_0 \\
c_e & c_{e-1} & c_{e-2} & \dots & c_2 & c_1 & c_0 & 0 & \dots & 0 \\
0 & c_e & c_{e-1} & \dots & c_3 & c_2 & c_1 & c_0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \dots & 0 & c_e & c_{e-1} & c_{e-2} & \dots & c_0
\end{vmatrix},$$

wobei die Koeffizienten $c_e, \dots, c_{s+1}$ alle verschwinden.

Somit beginnt im unteren Teil der Matrix jede Zeile mit mindestens $e-s$ Nullen.

In den ersten $e-s$ Spalten der Matrix stehen daher nur noch Koeffizienten von f: In der ersten ist dies ausschließlich der führende Koeffizient a_d von f in der ersten Zeile. Entwickeln wir nach der ersten Zeile, können wir also einfach die erste Zeile und die erste Spalte streichen; die Determinante ist dann a_d mal der Determinante der übrigbleibenden Matrix. Diese hat (falls $e > s+1$) wieder dieselbe Gestalt, wir können also wieder einen Faktor a_d ausklammern und bekommen eine Determinante mit einer Zeile und einer Spalte weniger, *usw.* Das Ganze funktioniert $e-s$ mal; dann ist der führende Koeffizient von h in die erste Spalte gerutscht, und die übriggebliebene Matrix ist die SYLVESTER-Matrix von f und h – falls etwas übrigbleibt. Offensichtlich bleibt genau dann nichts übrig, wenn h das Nullpolynom ist: Dann sind die unteren m Zeilen Null, d.h. die Resultante verschwindet.

Andernfalls ist $\mathrm{Res}_X(f,g) = a_d^{e-s}\,\mathrm{Res}_X(f,h)$, und da diese Formel auch für $h = 0$ gilt, haben wir gezeigt

Lemma: Hat f keinen größeren Grad als g und ist h der Divisionsrest von g durch f, der den Grad s habe, so ist $\mathrm{Res}_X(f,g) = a_d^{e-s}\,\mathrm{Res}_X(f,h)$. ∎

Dies läßt sich nun nach Art des EUKLIDischen Algorithmus iterieren: Berechnen wir wie dort die Folge der Reste $r_1 = h$ der Division von g

durch f und dann (mit $r_0 = g$) weiter r_{i+1} gleich dem Rest bei der Division von r_i durch r_{i-1}, so können wie die Berechnung von $\mathrm{Res}_X(f,g)$ durch Multiplikation mit Potenzen der führenden Koeffizienten der Divisoren zurückführen auf die viel kleineren Resultanten $\mathrm{Res}_X(r_i, r_{i+1})$. Sobald r_{i+1} eine Konstante ist, egal ob Null oder nicht, haben wir eine explizite Formel und der Algorithmus endet. Für den Fall, daß f größeren Grad als g hat brauchen wir noch

Lemma: Für ein Polynom, f vom Grad d und ein Polynom g vom Grad e ist $\mathrm{Res}_X(f,g) = (-1)^{de}\,\mathrm{Res}_X(g,f)$.

Beweis: Wir müssen in der SYLVESTER-Matrix e Zeilen zu f mit den d Zeilen zu g vertauschen. Dies kann beispielsweise so realisiert werden, daß wir die unterste f-Zeile nacheinander mit jeder der g-Zeilen vertauschen, bis sie nach d Vertauschungen schließlich unten steht. Dies müssen wir wiederholen, bis alle f-Zeilen unten stehen, wir haben also insgesamt de Zeilenvertauschungen. Somit ändert sich das Vorzeichen der Determinante um den Faktor $(-1)^{de}$. ∎

Zum Abschluß dieses Paragraphen wollen wir uns noch überlegen, daß die Resultante zweier Polynome noch aus einem anderen Grund für jede gemeinsame Nullstelle verschwinden muß: Genau wie der größte gemeinsame Teiler von f und g läßt sich auch deren Resultante als Linearkombination von f und g darstellen:

Lemma: R sei ein Ring und $f, g \in R[X]$ seien Polynome über R. Dann gibt es Polynome $p, q \in R[X]$, so daß $\mathrm{Res}_X(f,g) = pf + qg$ ist.

Man beachte, daß p, q, f und g zwar Polynome sind, die Resultante aber nur ein Element von R.

Beweis: Wir schreiben

$$f = a_d X^d + \cdots + a_1 X + a_0 \quad \text{und} \quad g = b_e X^e + \cdots + b_1 X + b_0\,,$$

wobei wir annehmen können, daß a_d und b_e beide nicht verschwinden. Die Gleichungen

$$X^i f = a_d X^{d+i} + \cdots + a_1 X^{1+i} + a_0 X^i \quad \text{für } i = 0, \ldots, e-1$$

und

$$X^j g = b_e X^{e+j} + \cdots + b_1 X^{1+j} + b_0 X^j \quad \text{für } j = 0, \ldots, d-1$$

können wir in Vektorschreibweise so zusammenfassen, daß wir den $(d+e)$-dimensionalen Vektor

$$F = \left(X^{e-1} f, \ldots, Xf, f, X^{d-1} g, \ldots, Xg, g\right)^T \in R[X]^{d+e}$$

darstellen in der Form

$$F = X^{d+e-1} r_1 + \cdots + X^1 r_{d+e-1} + X^0 r_{d+e}$$

mit Vektoren $r_k \in R^{d+e}$, deren Einträge Koeffizienten von f und g sind. Die Resultante ist nach Definition gleich der Determinanten der $(d+e) \times (d+e)$-Matrix mit den r_k als Spaltenvektoren.

Nun gehen wir vor, wie bei der Herleitung der CRAMERschen Regel: Wir betrachten obige Vektorgleichung als ein lineares Gleichungssystem mit rechter Seite F in den „Unbekannten“ X^k und tun so, als wollten wir den Wert von $X^0 = 1$ aus diesem Gleichungssystem bestimmen. Dazu ersetzen wir nach CRAMER in der Determinante des Gleichungssystems die letzte Spalte durch die rechte Seite, berechnen also die Determinante

$$\begin{aligned}
\det(r_1, \ldots, r_{d+e-1}, F) &= \det\Big(r_1, \ldots, r_{d+e-1}, \sum_{k=1}^{d+e} X^{d+e-k} r_k\Big) \\
&= \sum_{k=1}^{d+e} X^{d+e-k} \det(r_1, \ldots, r_{d+e-1}, r_k) \\
&= \det(r_1, \ldots, r_{d+e-1}, r_{d+e}) \,,
\end{aligned}$$

denn für $k \neq d+e$ steht die Spalte r_k zweimal in der Matrix, so daß die Determinante verschwindet.

Wenn wir bei der Berechnung von $\det(f_1, \ldots, r_{d+e-1})$ nach dem LAPLACEschen Entwicklungssatz die Polynome f und g in F stehen lassen, erhalten wir die Determinante als Ausdruck der Form $pf + qg$ mit Polynomen p und q aus $R[X]$: Da f und g beide nur in der letzten Spalte vorkommen, dort aber in jedem Eintrag genau eines der beiden,

enthält jedes der $(d+e)!$ Produkte, die nach LAPLACE aufsummiert werden, genau eines der beiden Polynome. Nach der obigen Rechnung ist $pf + qg$ gleich der Determinante der r_k, also der Resultante.

■

Kapitel 3
Modulare Berechnung des ggT

§1: Rechnen mit homomorphen Bildern

Beim EUKLIDischen Algorithmus für Polynome haben wir gesehen, daß selbst teilerfremde Polynome mit relativ kleinen, ganzzahligen Koeffizienten zu Zwischenergebnisse mit recht großen Nennern führen können. Diese Explosion der Zwischenergebnisse ist ein wohlbekanntes Problem der Computeralgebra, das beileibe nicht nur beim EUKLIDischen Algorithmus auftritt. Es ist auch der Grund dafür, daß in der Computeralgebra die Speichergröße oft wichtiger ist als die Rechengeschwindigkeit.

Natürlich bemühen sich Mathematiker seit sie Computer für symbolisches Rechnen benutzen, mit diesem Problem fertig zu werden. Mittlerweile sind eine ganze Reihe von Algorithmen bekannt, mit denen sich dieses Problem zumindest abmildern läßt. Beim EUKLIDischen Algorithmus für Polynome mit ganzzahligen Koeffizienten könnte man Nenner vermeiden, indem man vor jeder Polynomdivision den Dividenden mit einer hinreichend hohen Potenz des führenden Koeffizienten des Divisors multipliziert. Das führt dann aber leider zu einem exponentiellen Anstieg der Koeffizienten, so daß dies keine wirkliche Lösung ist. Mit der sogenannten Subresultantenmethode kann man, mit einem etwas höheren mathematischen Aufwand, einen Kompromiss finden, wie man die Koeffizienten ganzzahlig halten und trotzdem nicht zu groß werden lassen kann.

Hier in diesem Kapitel soll es aber um zwei andere Methoden gehen: Wenn wir nicht mit ganzen oder rationalen Zahlen rechnen, sondern mit Zahlen modulo einer festen natürlichen Zahl N, sind automatisch auch

alle Zwischenergebnisse Elemente von $\mathbb{Z}/N$. Wenn wir dann eine obere Schranke M für die Beträge der Koeffizienten des ggT kennen und wenn $N \geq 2M + 1$ ist, ist der ggT eindeutig bestimmt durch seine Reduktion modulo N. Es ist freilich nicht klar, daß für zwei Polynome $f, g \in \mathbb{Z}[X]$ die Reduktion des ggT modulo N gleich dem ggT von f mod N und g mod N ist: Die beiden Polynome

$$f = X^2 - 3X + 2 = (X - 1)(X - 2) \quad \text{und}$$
$$g = X^2 - 12X + 32 = (X - 4)(X - 8)$$

aus $\mathbb{Z}[X]$ sind offensichtlich teilerfremd, haben also den ggT eins. f mod $2 = X^2 - X$ und g mod $2 = X^2$ haben aber den ggT X, modulo drei sind die beiden Polynome sogar gleich, und modulo sieben ist der ggT gleich $X - 1$. Ein weiteres Problem besteht darin, daß wir modulo N manche Rechenoperationen nicht ausführen können, beispielsweise die Division durch ein Zahl, die nicht teilerfremd zu N ist. Auch sind Probleme zu erwarten, wenn eine solche Zahl der führender Koeffizient eines der beiden Polynome ist.

Es gibt also eine ganze Reihe von Gründen, warum eine modulare Berechnung schiefgehen kann, und unsere Algorithmen müssen so aufgebaut sein, daß wir Probleme möglichst früh erkennen oder besser noch vermeiden können. Auf jeden Fall muß am Ende überprüft werden, ob das Endergebnis wirklich ein gemeinsamer Teiler von f und g ist.

Falls die Schranke M für die Beträge der Koeffizienten groß ist und wir modulo einer Zahl $N \geq 2M + 1$ rechnen wollen, brauchen wir Langzahlarithmetik. Darüber verfügt natürlich jedes Computeralgebrasystem, aber im Vergleich zur Hardwarearithmetik des Rechners ist sie doch recht langsam.

Üblicherweise rechnet man daher modulo einer Primzahl p, für die man alle Rechenoperationen modulo p mit dem Datentyp *integer* des Rechners ausführen kann. Bei einem 64-Bit-Rechner kann p dann höchstens 32 Bit haben, d.h. $p < 2^{32} = 4\,294\,967\,296$. Das wird natürlich oft nicht ausreichen.

Es gibt zwei Strategien, wie man trotzdem über die jeweilige Schranke kommen kann. Die eine besteht darin, modulo mehrerer Primzahlen

zu rechnen und dann die Ergebnisse über den chinesischen Restesatz zusammenzusetzen. Falls die modularen ggTs verschiedene Grade haben, sieht man sofort, daß eine oder mehrere der verwendeten Primzahlen hier nicht zur modularen Berechnung des ggT geeignet sind. Da das Produkt aller Primzahlen kleiner 2^{32} eine Zahl mit mehr als 1,8 Milliarden Dezimalstellen ist, kommen wir mit solchen Primzahlen über jede realistische Schranke für praktisch lösbare Probleme.

Alternativ kann man eine sogenannte p-adische Strategie anwenden. Hier hebt man das modulo p berechnete Ergebnis schrittweise hoch zu einem Ergebnis modulo p^2, p^3, *usw.,* bis die Primzahlpotenz über der jeweiligen Schranke liegt. Beim Hochheben macht man bei bekanntem ggT h_{n-1} von f mod p^{n-1} und g mod p^{n-1} für den ggT h_n der Polynome modulo p^n den Ansatz $h_n = h_{n-1} + p^{n-1}h^*$, wobei die Koeffizienten von h^* wieder zwischen 0 und $p-1$ liegen, so daß man mit gewöhnlichen Maschinenzahlen arbeiten kann.

Mit modularen Methoden lassen sich nicht nur Probleme mit Polynomen aus $\mathbb{Z}[X]$ zurückführen auf solche mit Polynomen über endlichen Körpern, sondern auch Probleme mit Polynomen in mehreren Veränderlichen auf solche mit Polynomen in nur einer Veränderlichen. Das Analogon zum chinesischen Restesatz ist hier die Interpolation: Wenn wir den Wert eines Polynoms vom Grad d aus einem Polynomring $R[X]$ für $d+1$ Argumente aus R kennen, können wir das Polynom eindeutig rekonstruieren.

Falls wir beispielsweise den ggT zweier Polynome f, g aus $\mathbb{Z}[X, Y]$ bestimmen wollen, betrachten wir sie als Polynome in X über $\mathbb{Z}[Y]$. Auch der ggT liegt in $\mathbb{Z}[Y][X]$, und seine Koeffizienten können offensichtlich keinen höheren Grad haben als das Minimum m der Y-Grade von f und g. Dann wählen wir $m+1$ ganze Zahlen $y_0, \ldots, y_m$, für die keiner der höchsten X-Koeffizienten von f und g verschwindet, und berechnen die $m+1$ größten gemeinsamen Teiler der Polynome $f(X, y_i)$ und $g(X, y_i)$ in $\mathbb{Z}[X]$. Falls wir Glück hatten und alle haben den gleichen X-Grad wie ggT(f, g), können wir dessen X-Koeffizienten jeweils durch Interpolation der entsprechenden Koeffizienten dieser ggTs bestimmen. Bei Polynomen in mehr als zwei Veränderlichen können wir schritt-

weise durch Einsetzen für eine der Variablen reduzieren auf den Fall von Polynomen mit einer Veränderlichen weniger.

§2: Modulare Berechnung der Resultante

Zur Illustration der modularen Vorgehensweise möchte ich mit der Berechnung der Resultante zweier Polynome aus $\mathbb{Z}[X]$ beginnen. Die beiden Polynome seien

$$f = a_d X^d + \cdots + a_1 X + a_0 \quad \text{und} \quad g = b_e X^e + \cdots + b_1 X + b_0 \,.$$

Ihre Resultante

$$\mathrm{Res}_Y(f,g) = \begin{vmatrix} a_d & & \cdots & & a_0 & & \\ & \ddots & & & & \ddots & \\ & & a_d & \cdots & & & a_0 \\ b_e & & \cdots & & b_0 & & \\ & \ddots & & & & \ddots & \\ & & & b_e & \cdots & & b_0 \end{vmatrix}$$

ist als Determinante einer ganzzahligen Matrix selbst ganzzahlig, läßt sich aber schon für moderat große Grade d und e nicht mehr mit einem realistischen Aufwand über den LAPLACEschen Entwicklungssatz berechnen. Die Berechnung nach Art des EUKLIDischen Algorithmus in $\mathbb{Q}[X]$ erfordert deutlich weniger Rechenoperationen, führt allerdings meist relativ schnell zu rationalen Zahlen mit riesigen Nennern. Wenn wir modulo einer Primzahl p rechnen, wenden wir diesen Algorithmus in $\mathbb{F}_p[X]$ an, und dort haben natürlich auch alle Zwischenergebnisse Koeffizienten aus $\mathbb{F}_p$.

Sei also p prim. Wir betrachten

$$f^{(p)} = (a_d \bmod p) X^d + \cdots + (a_1 \bmod p) X + (a_0 \bmod p) \in \mathbb{F}_p[X]$$

und

$$g^{(p)} = (b_e \bmod p) X^e + \cdots + (b_1 \bmod p) X + (b_0 \bmod p) \in \mathbb{F}_p[X] \,.$$

Wenn a_d durch p teilbar ist, hat $f^{(p)}$ einen kleineren Grad als f, und wenn b_e durch p teilbar ist, hat $g^{(p)}$ einen kleineren Grad als g. In beiden

Fällen hat die SYLVESTER-Matrix von $f^{(p)}$ und $g^{(p)}$ eine andere Gestalt als die von f und g; wir können daher nicht erwarten, daß $\mathrm{Res}_X(f,g)$ und $\mathrm{Res}_X(f^{(p)},g^{(p)})$ viel miteinander zu tun haben. In diesen Fällen sagen wir, das Problem habe schlechte Reduktion bei p.

Wenn p keinen der beiden führenden Koeffizienten teilt, ist auch $\deg f^{(p)} = d$ und $\deg g^{(p)} = e$; die Resultante der beiden Polynome sieht also genauso aus wie die obige Determinante, nur daß wir alle Einträge modulo p betrachten. Da die Determinante ein Polynom in ihren Einträgen ist, erhalten wir genau das gleiche Ergebnis, wenn wir sie zunächst in $\mathbb{Z}$ berechnen und dann zur Restklasse modulo p übergehen. Somit ist

$$\mathrm{Res}_X(f^{(p)},g^{(p)}) = \mathrm{Res}_X(f,g) \bmod p$$
$$\text{falls } p \text{ kein Teiler von } a_d \text{ oder } b_e \text{ ist.}$$

In diesen Fällen sagen wir, das Problem habe gute Reduktion modulo p.

Hier können wir den beiden Polynomen sofort ansehen, ob das Problem modulo einer gegebenen Primzahl schlechte Reduktion hat oder nicht; bei der Bestimmung des ggT wird das leider nicht mehr so einfach sein.

Um die Resultante modular berechnen zu können, brauchen wir noch eine obere Schranke für ihren Betrag.

Betrachten wir zunächst die Determinante einer beliebigen $r \times r$-Matrix $A = (a_{ij})$. Für jeden Index i sei eine Schranke b_i gegeben derart, daß $\left|a_{ij}\right| \le b_i$ für alle j. Dann ist

$$\begin{aligned}|\det A| &= \left|\sum_{\pi\in S_r} \operatorname{sgn}\pi \cdot a_{1\pi(1)}\cdots a_{r\pi(r)}\right| \le \sum_{\pi\in S_r} \left|a_{1\pi(1)}\right|\cdots\left|a_{r\pi(r)}\right| \\ &\le \sum_{\pi\in S_r} b_1\cdots b_r = r!\cdot b_1\cdots b_r\end{aligned}$$

Um dies auf den Fall einer Resultanten anzuwenden, brauchen wir Schranken für die Koeffizienten eines Polynoms.

Definition: Die *Höhe* $H(P)$ eines Polynoms $P = c_m X^m + \cdots c_1 X + c_0$ aus $\mathbb{C}[X]$ ist das Maximum der Beträge der Koeffizienten $c_0,\ldots,c_m$.

Die ersten e Spalten der Resultante von f und g haben Koeffizienten von f als Einträge; ihre Beträge sind kleiner oder gleich der Höhe $H(f)$ von f. Die restlichen d Zeilen enthalten Koeffizienten von g, deren Beträge durch $H(g)$ beschränkt sind.

$$|\mathrm{Res}_X(f,g)| \leq (d+e)! \cdot H(f)^e \cdot H(g)^d .$$

(Diese Schranke ist natürlich alles andere als optimal; wegen der speziellen Struktur der SYLVESTER-Matrix könnte man sie leicht verbessern.)

Damit ist klar, wie wir die Resultante modular berechnen können:

1. Ansatz: Wir wählen eine Primzahl $p > 2(d+e)! \cdot H(f)^e \cdot H(g)^d$. Da $H(f) \geq |a_d|$ und $H(g) \geq |b_e|$, kann p für positive d und e kein Teiler von a_d oder b_e sein, es gibt also keine schlechte Reduktion. Wir berechnen daher $\mathrm{Res}_X(f^{(p)}, g^{(p)})$, und $\mathrm{Res}_X(f,g)$ ist die einzige ganze Zahl mit Betrag höchstens $(d+e)! \cdot H(f)^e \cdot H(g)^d$, die modulo p diese Restklasse hat.

2. Ansatz: Wir wählen verschiedene Primzahlen $p_1, \ldots, p_r$, modulo derer das Problem gute Reduktion hat und deren Produkt N größer ist als $2(d+e)! \cdot H(f)^e \cdot H(g)^d$. Für jede der Primzahlen p_i berechnen wir $\mathrm{Res}_X(f^{(p_i)}, g^{(p_i)})$ und setzen die Ergebnisse nach dem chinesischen Restesatz zusammen zu einer Restklasse modulo N. Wieder ist die Resultante die einzige Zahl mit Betrag höchstens $(d+e)! \cdot H(f)^e \cdot H(g)^d$, die modulo N diese Restklasse hat.

Auch Resultanten von Polynomen in mehreren Veränderlichen lassen sich modular berechnen: Seien etwa $f, g \in \mathbb{Z}[X,Y]$ zwei Polynome zweier Veränderlichen, deren Resultante bezüglich Y wir berechnen wollen. Dazu fassen wir f und g auf als Polynome in X mit Koeffizienten aus $\mathbb{Z}[Y]$, also

$$f = f(X,Y) = a_d(Y)X^d + \cdots + a_1(Y)X + a_0(Y)$$

und

$$g = g(X,Y) = b_e(Y)X^e + \cdots + b_1(Y)X + b_0(Y)\,;$$

die Resultante ist die Determinante der SYLVESTER-Matrix mit diesen Koeffizienten als Einträgen.

Setzen wir für Y eine ganze Zahl c ein, werden

$$f(X,c) = a_d(c)X^d + \cdots + a_1(c)X + a_0(c) \in \mathbb{Z}[X]$$

und

$$g(X,c) = b_e(c)X^e + \cdots + b_1(c)X + b_0(c) \in \mathbb{Z}[X]$$

zu Polynomen in einer Veränderlichen, deren Resultante

$$\begin{vmatrix}
a_d(c) & & \cdots & & a_0(c) & & \\
& \ddots & & & & \ddots & \\
& & a_d(c) & \cdots & & & a_0(c) \\
b_e(c) & & \cdots & & b_0(c) & & \\
& \ddots & & & & \ddots & \\
& & & b_e(c) & \cdots & & b_0(c)
\end{vmatrix}$$

wir nach obigem Algorithmus modular berechnen können.

Auch hier müssen wir darauf achten, daß wir keine schlechte Reduktion haben, daß der Grad von $f(X,c)$ also nicht kleiner wird als der X-Grad von f und entsprechend für g. Der eingesetzte Wert c darf also weder eine Nullstelle von a_d noch eine von b_e sein.

Alsdann hat die SYLVESTER-Matrix von $f(X,c)$ und $g(X,c)$ wie die von f und g bezüglich X, und wieder gilt: Zumindest vom Ergebnis her ist es gleichgültig, ob wir zunächst die Resultante von f und g als Polynom aus $\mathbb{Z}[Y]$ bestimmen und dann den Wert c einsetzen, oder ob wir c gleich in alle Einträge einsetzen und dann die Resultante als ganze Zahl berechnen:

$$\begin{aligned}\mathrm{Res}_X(f,g)(c) &= \mathrm{Res}_X\big(f(X,c), g(X,c)\big) \\ &\text{falls } a_d(c) \neq 0 \quad \text{und} \quad b_e(c) \neq 0\,.\end{aligned}$$

Falls $\mathrm{Res}_X(f,g) \in \mathbb{Z}[Y]$ den Grad n hat, benötigen wir $n+1$ Werte c, um das Polynom durch Interpolation zu bestimmen. Für einen Algorithmus zur Berechnung von $\mathrm{Res}_X(f,g)$ fehlt also noch eine obere Schranke für den Grad der Resultante.

Beginnen wir wieder mit der Determinante einer beliebigen $r \times r$-Matrix mit Polynomen in X als Einträgen. Falls alle Einträge der i-ten Zeile höchstens den Grad n_i haben, hat jedes der Produkte aus

dem LAPLACEschen Entwicklungssatz höchstens die Summe der n_i als Grad, und da sich der Grad durch Addition nicht erhöhen kann, ist diese Summe auch eine obere Schranke für den Grad der Determinante.

Hat also jeder der Koeffizienten $a_i = a_i(X) \in \mathbb{Z}[X]$ höchstens den Grad n und jeder der Koeffizienten b_j höchstens den Grad m, so hat die Resultante höchstens den Grad $ne + md$. Wir können die Resultante daher als Interpolationspolynom vom Grad höchstens $ne + md$ bestimmen, wenn wir die Resultante von $f(c, Y)$ und $g(c, Y)$ für $ne + md + 1$ verschiedene Werte c berechnen, wobei diese allesamt weder Nullstellen von a_d noch von b_e sein dürfen.

Resultanten von Polynomen in drei Veränderlichen kann man mit der gleichen Strategie zurückführen auf solche in zwei Veränderlichen, solche in vier auf solche in drei, und so weiter. Wenn man sich überlegt, wie viele Berechnungen von Resultanten ganzzahliger Polynome in einer Veränderlichen hinter dieser Vorgehensweise stehen, mag dies als eine sehr ineffiziente Methode erscheinen, aber GEORGE E. COLLINS hat 1971 die verschiedenen bekannten Methoden verglichen und kam zum Schluß, daß dies die effizienteste Vorgehensweise ist; siehe

> GEORGE E. COLLINS: The Calculation of Multivariate Polynomial Resultants, *J. ACM.* **18** (1971), S. 515–532

(Das *Journal of the ACM* (Association for Computing Machinery) ist im Netz der Universität Mannheim online verfügbar und steht auch als Papierausgabe in der Bibliothek.)

Seit 1971 gab es natürlich viele Fortschritte in der Computeralgebra und Verbesserungen an allen bekannten Verfahren, aber soweit ich weiß, gab es nichts, was eine andere Vorgehensweise besser als die modulare Berechnung werden ließ.

§3: Die Landau-Mignotte-Schranke

Hauptthema dieses Kapitels ist die Anwendung modularer Methoden auf die Berechnung des größten gemeinsamen Teilers zweier Polynome. Hier wird sich herausstellen, daß die Identifikation der Stellen schlechter

Reduktion sehr viel schwieriger ist als im Falle der Resultantenberechnung. Daher soll die modulare Berechnung des ggT auf mehrere Abschnitte verteilt werden. Hier im ersten davon geht es um eine Schranke für die Koeffizienten des ggT zweier ganzzahliger Polynome in einer Veränderlichen. Da wir später auch Schranken für die Koeffizienten der irreduziblen Faktoren eines Polynoms benötigen werden, beginnen wir mit der allgemeineren Frage nach Schranken für beliebige Teiler.

$f \in \mathbb{Z}[X]$ sei also ein bekanntes Polynom mit ganzzahligen Koeffizienten, und $g \in \mathbb{Z}[X]$ sei ein (im allgemeinen noch unbekannter) Teiler von f. Wir wollen eine obere Schranke für die Koeffizienten von g finden.

Dazu ordnen wir jedem Polynom

$$f = \sum_{k=0}^{d} a_k X^k \in \mathbb{C}[X]$$

mit komplexen Koeffizienten a_k eine Reihe von Maßzahlen für die Größe der Koeffizienten zu: Wir kennen bereits die Höhe

$$H(f) = \max_{k=0}^{d} |a_k| \ ,$$

und unser Ziel ist es, für ein gegebenes Polynom $f \in \mathbb{Z}[X]$ die Höhe seiner Teiler abzuschätzen. Auf dem Weg zu dieser Abschätzung werden uns noch eine Reihe anderer Größen nützlich sein, darunter die L^1- und die L^2-Norm

$$\|f\|_1 = \sum_{k=0}^{d} |a_k| \quad \text{und} \quad \|f\|_2 = \sqrt{\sum_{k=0}^{d} a_k \, \overline{a_k}} = \sqrt{\sum_{k=0}^{d} |a_k|^2} \ .$$

Für die drei bislang definierten Größen gilt

Lemma 1: $H(f) \le \|f\|_2 \le \|f\|_1 \le \sqrt{d+1}\ \|f\|_2 \le (d+1)H(f)$

Beweis: Ist a_ν der betragsgrößte Koeffizient von f, so ist

$$H(f) = |a_\nu| = \sqrt{|a_\nu|^2}$$

offensichtlich kleiner oder gleich $\|f\|_2$. Dies wiederum ist nach der Dreiecksungleichung kleiner oder gleich $\|f\|_1$, denn schreiben wir in $\mathbb{C}^{d+1}$

den Koeffizientenvektor von f als Summe von Vielfachen der Basisvektoren, d.h.

$$\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} a_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ a_1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ a_d \end{pmatrix},$$

so steht links ein Vektor der Länge $\|f\|_2$, und rechts stehen Vektoren, deren Längen sich zu $\|f\|_1$ summieren.

Das nächste Ungleichheitszeichen ist die CAUCHY-SCHWARZsche Ungleichung, angewandt auf die Vektoren

$$\begin{pmatrix} |a_0| \\ |a_1| \\ \vdots \\ |a_d| \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Das Skalarprodukt dieser beiden Vektoren ist die Summe der $|a_i|$, also die L^1-Norm $\|f\|_1$; die Länge des ersten Vektors ist gleich $\|f\|_2$, und die des zweiten ist $\sqrt{d+1}$.

Schließlich ist noch

$$\|f\|_2 = \sqrt{\sum_{j=0}^{d} |a_j|^2} \leq \sqrt{\sum_{j=0}^{d} |a_\nu|^2} = \sqrt{d+1}\,|a_\nu| = \sqrt{d+1}\,H(f);$$

■

multipliziert man diese Ungleichung mit $\sqrt{d+1}$ folgt die letzte Ungleichung aus der Behauptung.

■

Es ist alles andere als offensichtlich, wie sich die drei bislang definierten Maßzahlen für einen Teiler eines Polynoms durch die entsprechende Größen für das Polynom selbst abschätzen lassen, denn über die Koeffizienten eines Teilers können wir leider nur sehr wenig sagen. Über seine Nullstellen allerdings schon: Die Nullstellen eines Teilers bilden natürlich eine Teilmenge der Nullstellen des Polynoms. Also sollten wir versuchen, auch die Nullstellen ins Spiel zu bringen.

Der Wurzelsatz von VIÈTE. liefert uns bekanntlich einen Zusammenhang zwischen den Nullstellen und den Koeffizienten: Für ein Polynom

$$f = X^d + a_{d-1}d^{n-1} + a_{d-2}X^{d-2} + \cdots + a_2X^2 + a_1X + a_0$$

mit höchstem Koeffizienten eins und mit (nicht notwendigerweise verschiedenen) Nullstellen $z_1, \ldots, z_d$ ist auch

$$f = (X - z_1)(X - z_2)\cdots(X - z_d)\,,$$

und

$$\begin{aligned} a_{d-1} &= -(z_1 + \cdots + z_d) \\ a_{d-2} &= \sum_{i<j} z_i z_j \\ a_{d-3} &= -\sum_{i<j<k} z_i z_j z_k \\ &\vdots \qquad\qquad \vdots \\ a_0 &= (-1)^d z_1 \cdots z_d\,. \end{aligned}$$

Um die Koeffizienten eines Polynoms durch die Nullstellen abschätzen zu können, brauchen wir also obere Schranken für die Beträge der Produkte aus k Nullstellen. Natürlich ist jedes solche Produkt ein Teilprodukt des Produkts $z_1 \cdots z_d$ aller Nullstellen, aber das führt zu keiner Abschätzung, da unter den fehlenden Nullstellen auch welche sein können, deren Betrag kleiner als eins ist. Um eine obere Schranke für den Betrag zu kommen, müssen wir diese Nullstellen im Produkt $z_1 \cdots z_d$ durch Einsen ersetzen; dann können wir sicher sein, daß kein Produkt von k Nullstellen einen größeren Betrag hat als das so modifizierte Produkt. Diese Überlegungen führen auf die

Definition: Das *Maß* $\mu(f)$ eines nichtkonstanten Polynoms

$$f = a_d \prod_{j=1}^{d} (X - z_j)$$

ist das Produkt der Beträge aller Nullstellen von Betrag größer eins mal dem Betrag des führenden Koeffizienten a_d von f:

$$\mu(f) = |a_d| \prod_{j=1}^{d} \max\bigl(1, |z_j|\bigr) .$$

Dieses Maß ist im allgemeinen nur schwer explizit berechenbar, da man dazu die sämtlichen Nullstellen des Polynoms explizit kennen muß. Es hat aber den großen Vorteil, daß für zwei Polynome f und g trivialerweise gilt

$$\mu(f \cdot g) = \mu(f) \cdot \mu(g) .$$

Auch können wir es nach dem Wurzelsatz von VIÈTE leicht für eine Abschätzung der Koeffizienten verwenden: a_k/a_d ist bis aufs Vorzeichen die Summe aller Produkte von $d-k$ Nullstellen, und jedes einzelne solche Produkt mal $|a_d|$ hat höchstens den Betrag $\mu(f)$. Die Anzahl der Summanden ist die Anzahl von Möglichkeiten, aus d Indizes eine k-elementige Teilmenge auszuwählen, also $\binom{d}{k}$. Damit folgt

Lemma 2: Für ein nichtkonstantes Polynom $f = \sum_{k=0}^{d} a_k X^k \in \mathbb{C}[X]$ ist

$$|a_k| \le \binom{d}{k} \mu(f) .$$

■

Der größte unter den Binomialkoeffizienten $\binom{d}{k}$ ist bekanntlich der mittlere *bzw.* sind die beiden mittleren, und die Summe aller Binomialkoeffizienten $\binom{d}{k}$ ist, wie die binomische Formel für $(1+1)^d$ zeigt, gleich 2^d. Damit folgt

Korollar: Für ein nichtkonstantes Polynom $f \in \mathbb{C}[X]$ ist

$$H(f) \le \binom{d}{[d/2]} \mu(f) \quad \text{und} \quad H(f) \le \|f\|_1 \le 2^d \mu(f) .$$

■

Zur Abschätzung des Maßes durch eine Norm zeigen wir zunächst

Lemma 3: Für jedes Polynom $f \in \mathbb{C}[X]$ und jede komplexe Zahl z ist

$$\|(X - z)f\|_2 = \|(\overline{z}X - 1)f\|_2 .$$

Beweis durch explizite Berechnung der beiden Seiten: Sei $f = \sum_{k=0}^{d} a_k X^d$.

Das Quadrat von $\|(X-z)f\|_2 = \left\| a_d X^{d+1} + \sum_{k=1}^{d} (za_k - a_{k-1}) X^k - a_0 z \right\|_2$ ist die Summe aller Koeffizientenquadrate, also

$$\begin{aligned} & a_d\,\overline{a_d} + \sum_{k=1}^{d} (za_k - a_{k-1})\,\overline{(za_k - a_{k-1})} + a_0 z \overline{a_0 z} \\ = {} & |a_d|^2 + \sum_{k=1}^{d} \left(|a_k|^2 |z|^2 - 2\,\mathfrak{Re}(za_k \overline{a_{k-1}}) + |a_{k-1}|^2 \right) + |a_0|^2 |z|^2 \\ = {} & (1+|z|^2) \sum_{k=0}^{d} |a_k|^2 - 2 \sum_{k=1}^{d} \mathfrak{Re}(za_k \overline{a}_{k-1})\,. \end{aligned}$$

Entsprechend ist $(\overline{z}X - 1)f = a_d \overline{z} X^{d+1} + \sum_{k=1}^{d} (\overline{z}a_{k-1} - a_k) X^k - a_0$ und auch $\|(\overline{z}X-1)f\|_2^2$ wird zu

$$\begin{aligned} & a_d \overline{z} \cdot \overline{a_d} z + \sum_{k=1}^{d} (\overline{z}a_{k-1} - a_k)(z\overline{a_{k-1}} - \overline{a_k}) + a_0 \overline{a_0} \\ = {} & |za_d|^2 + \sum_{k=1}^{d} \left(|za_{k-1}|^2 - 2\,\mathfrak{Re}(za_k \overline{a_{k-1}}) + |a_k|^2 \right) + |a_0|^2 \\ = {} & (1+|z|^2) \sum_{k=0}^{d} |a_k|^2 - 2 \sum_{k=1}^{d} \mathfrak{Re}(za_k \overline{a}_{k-1})\,. \end{aligned}$$

■

Für das Polynom $f = a_d \prod_{j=1}^{d} (X - z_j)$ bedeutet dies, daß wir den Faktor $(X - z_j)$ durch $(\overline{z}_j X - 1)$ ersetzen können, ohne daß sich die L^2-Norm ändert. Wenden wir dies an auf alle Faktoren $(X - z_j)$, für die $|z_j| > 1$ ist, erhalten wir ein Polynom, dessen sämtliche Nullstellen Betrag kleiner oder gleich eins haben, denn $\overline{z}_j X - 1$ verschwindet für $X = 1/\overline{z}_j$, was für $|z|_j > 1$ einen Betrag kleiner Eins hat. Das Maß des modifizierten Polynoms ist also gleich dem Betrag des führenden Koeffizienten, und dieser wiederum ist natürlich kleiner oder gleich der L^2-Norm. Andererseits ist das Maß des modifizierten Polynoms gleich dem des ursprünglichen, denn für jeden Faktor $(X - z_j)$ wird der führende Koeffizient bei der Modifikation mit $\overline{z}_j$ multipliziert, was denselben Betrag hat wie z_j. Damit folgt:

Lemma 4: Für ein nichtkonstantes Polynom $f \in \mathbb{C}[X]$ ist

$$\mu(f) \leq \|f\|_2 \ .$$

■

Nach diesen Vorbereitungen können wir uns an die Abschätzung der Koeffizienten eines Teilers machen. Sei dazu

$$g = \sum_{j=0}^{e} b_j X^j \quad \text{Teiler von} \quad f = \sum_{i=0}^{d} a_i X^i \ .$$

Da jede Nullstelle von g auch Nullstelle von f ist, lassen sich die Maße der beiden Polynome leicht vergleichen:

$$\mu(g) \leq \left|\frac{b_e}{a_d}\right| \cdot \mu(f) \, .$$

Kombinieren wir dies mit dem Korollar zu Lemma 2 und mit Lemma 4, erhalten wir die LANDAU-MIGNOTTE-Schranke:

$$H(g) \leq \binom{e}{[e/2]} \left|\frac{b_e}{a_d}\right| \|f\|_2 \quad \text{und} \quad \|g\|_1 \leq 2^e \left|\frac{b_e}{a_d}\right| \|f\|_2 \ .$$

Der ggT zweier Polynome f und g muß diese Abschätzung für beide Polynome erfüllen, allerdings kennen wir *a priori* weder den Grad noch den führenden Koeffizienten des ggT. Falls wir Polynome mit ganzzahligen Koeffizienten betrachten und einen ggT in $\mathbb{Z}[X]$ suchen, wissen wir nur, daß sein führender Koeffizient die führenden Koeffizienten sowohl von f als auch von g teilen muß, und daß sein Grad natürlich weder den von f noch den von g übersteigen kann. Damit erhalten wir die LANDAU-MIGNOTTE-Schranke für den ggT zweier Polynome: Schreiben wir f und g wie oben, so ist für $f, g \in \mathbb{Z}[X]$

$$\begin{aligned} H\big(\mathrm{ggT}(f,g)\big) &\leq \|\mathrm{ggT}(f,g)\|_1 \\ \leq \mathrm{LM}(f,g) &\underset{\text{def}}{=} 2^{\min(d,e)} \, \mathrm{ggT}(a_d, b_e) \min\left(\frac{\|f\|_2}{|a_d|}, \frac{\|g\|_2}{|b_e|}\right) \ . \end{aligned}$$

EDMUND GEORG HERMANN LANDAU (1877–1938) wurde in Berlin geboren und studierte an der dortigen Universität, wo er auch von 1899 bis 1909 lehrte. Dann bekam er einen Ruf an die damals führende deutsche Mathematikfakultät in Göttingen. 1933 verlor er seinen dortigen Lehrstuhl, denn die Studenten boykottierten seine Vorlesungen, da sie meinten, sie könnten Mathematik unmöglich bei einem jüdischen Professor lernen. LANDAUS zahlreiche Publikationen beschäftigen sich vor allem mit der Zahlentheorie, über die er auch ein bedeutendes Lehrbuch schrieb. Sehr bekannt sind insbesondere seine Arbeiten über Primzahlverteilung.

MAURICE MIGNOTTE promovierte 1974 in Paris und arbeitete dann bis zu seiner Emeritierung am Institut de Recherche Mathématique Avancée der Universität Straßburg. Sein Hauptforschungsgebiet sind diophantische Gleichungen. Er ist Autor mehrerer Lehrbücher, unter anderem aus dem Gebiet der Computeralgebra.

Als Beispiel betrachten wir noch einmal die beiden Polynome

$$f = X^8 + X^6 - 3X^4 - 3X^3 + 8X^2 + 2X - 5 \quad \text{und}$$
$$g = 3X^6 + 5X^4 - 4X^2 - 9X + 21\,.$$

f hat die L^2-Norm

$$\|f\|_2 = \sqrt{1^2 + 1^2 + 3^2 + 3^2 + 8^2 + 2^2 + 5^2} = \sqrt{113}$$

und den führenden Koeffizienten eins; während

$$\|g\|_2 = \sqrt{3^2 + 5^2 + 4^2 + 9^2 + 21^2} = \sqrt{572} = 2\sqrt{143}$$

isti, und hier ist der führende Koeffizienten gleich drei.

Da $3^2 \cdot 113 > 900$ größer ist als $2^2 \cdot 143 < 600$, ist die LANDAU-MIGNOTTE-Schranke für diese beiden Polynome

$$\mathrm{LM}(f, g) = 2^6 \cdot \frac{2}{3}\sqrt{143} \approx 510{,}2191249\,.$$

Die Koeffizienten des ggT sind ganze Zahlen; daher kann der Betrag eines jeden Koeffizienten höchstens gleich 510 sein. Wie wir wissen, ist der tatsächliche ggT gleich eins; die Schranke ist also zumindest in diesem Fall recht pessimistisch. Sie ist allerdings bedeutend kleiner als Zahlen, die bei direkter Anwendung des EUKLIDischen Algorithmus als Koeffizienten auftreten.

§4: Gute und schlechte Reduktion beim ggT

Nachdem wir eine obere Schranke für die Koeffizienten des größten gemeinsamen Teilers zweier Polynome aus $\mathbb{Z}[X]$ gefunden haben, stellt sich als nächstes die Frage, bei welchen Primzahlen wir gute Reduktion haben.

Angenommen, f und g aus $\mathbb{Z}[X]$ sind zwei Polynome mit ganzzahligen Koeffizienten. Ihr ggT $h \in \mathbb{Z}[X]$ ist bis auf eine Einheit eindeutig bestimmt, also bis aufs Vorzeichen. Sein Grad sei d.

Nun sei p eine Primzahl und $f^{(p)}, g^{(p)} \in \mathbb{F}_p[X]$ seien die Polynome, die aus f und g entstehen, wenn wir alle Koeffizienten modulo p reduzieren. Wann wissen wir, daß auch deren ggT in $\mathbb{F}_p[X]$ den Grad d hat?

Ist $f = hf_1$, $g = hg_1$, und sind $h^{(p)}, f_1^{(p)}, g_1^{(p)}$ die Reduktionen von h, f_1 und g_1 modulo p, so ist offensichtlich $f^{(p)} = h^{(p)} f_1^{(p)}$ und $g^{(p)} = h^{(p)} g_1^{(p)}$. Somit ist $h^{(p)}$ auf jeden Fall ein gemeinsamer Teiler von $f^{(p)}$ und $g^{(p)}$, muß also deren größten gemeinsamen Teiler teilen. Daraus folgt nun aber nicht, daß dessen Grad mindestens gleich d sein muß, denn wenn der führende Koeffizient von h durch p teilbar ist, hat $h^{(p)}$ kleineren Grad als h.

Ein Beispiel dafür können wir uns leicht konstruieren: Wie der EUKLIDische Algorithmus zeigt, sind die Polynome $f_1 = X^3 - X^2 + 2$ und $g_1 = X^2 + X + 1$ teilerfremd. Mit $h = 3X + 1$ haben

$$f = hf_1 = 3X^4 - 2X^3 - X^2 + 6X + 2 \quad \text{und} \quad g = hg_1 = 3X^3 + 4X^2 + 4X + 1$$

daher den ggT h. Wenden wir den EUKLIDischen Algorithmus aber an auf $f^{(3)} = X^3 - X^2 + 2$ und $g^{(3)} = X^2 + X + 1$, erhalten wir den ggT eins, der zumindest in diesem Fall auch gleich $h^{(3)}$ ist.

Die Tatsache, daß der Grad von $h^{(p)}$ genau dann kleiner als der von h, wenn der führende Koeffizient von h durch p teilbar ist, hat für die Wahl geeigneter Primzahlen keinen großen Nutzen, denn zu Beginn der Rechnung kennen wir h ja noch nicht. Nun ist aber $f = hf_1$ und $g = hg_1$; der führende Koeffizient von f *bzw.* g ist also das Produkt der führenden

Koeffizienten von h und von f_1 *bzw.* g_1. Wenn daher der führende Koeffizient von h durch p teilbar ist, so gilt dasselbe auch für die führenden Koeffizienten von f und von g. Die Umkehrung dieser Aussage gilt natürlich nicht, aber da wir eine große Auswahl an Primzahlen haben, stört uns das nicht weiter. Wir können also festhalten:

Lemma: Falls für die beiden Polynome $f, g \in \mathbb{Z}[X]$ die Primzahl p nicht beide führende Koeffizienten teilt, hat der ggT von $f^{(p)}$ und $g^{(p)}$ in $\mathbb{F}_p[X]$ mindestens denselben Grad wie $h = \mathrm{ggT}(f, g) \in \mathbb{Z}[X]$ und ist ein Vielfaches von $h^{(p)}$.

■

Falls er unter diesen Bedingungen größeren Grad als $h^{(p)}$ hat, müssen $f^{(p)}/h^{(p)} = (f/h)^{(p)}$ und $g^{(p)}/h^{(p)} = (g/h)^{(p)}$ einen gemeinsamen Faktor positiven Grades haben, d.h.

$$\mathrm{Res}_X(f^{(p)}/h^{(p)}, g^{(p)}/h^{(p)}) = \mathrm{Res}_X(f/h, g/h) \bmod p$$

muß verschwinden, Falls p keinen der führenden Koeffizienten von f und g teilt, teilt es auch keinen der führenden Koeffizienten von f/h und g/h, so daß diese Resultante gleich $\mathrm{Res}_X(f/h, g/h) \bmod p$ ist, d.h. p muß Teiler dieser Resultanten sein. Da wir h nicht kennen, können wir sie nicht ausrechnen; aber wir wissen nun, daß höchstens für endlich viele Primzahlen p der ggT von $f^{(p)}$ und $g^{(p)}$ etwas anderes als $\mathrm{ggT}(f, g) \bmod p$ sein kann.

Damit können wir das bisherige Ergebnis dieses Paragraphen für Zwecke der ggT-Berechnung in $\mathbb{Z}[X]$ folgendermaßen zusammenfassen:

Satz: Für zwei Polynome $f, g \in \mathbb{Z}[X]$ mit $\mathrm{ggT}(f, g) = h$ und ihre Reduktionen $f^{(p)}, g^{(p)} \in \mathbb{F}_p[X]$ mit $\mathrm{ggT}(f^{(p)}, g^{(p)}) = h^*$ gilt:
a) Falls p nicht die führenden Koeffizienten von sowohl f als auch g teilt, ist die Reduktion $h^{(p)}$ von h ein Teiler von h^* und $\deg h^* \geq \deg h$.
b) Es gibt höchstens endlich viele Primzahlen p, für die $h^{(p)}$ nicht gleich dem ggT von $f^{(p)}$ und $g^{(p)}$ ist.

■

Nun haben wir nur noch eine Schwierigkeit: Da $\mathbb{F}_p$ ein Körper ist, können wir den modulo p berechneten ggT stets so normieren, daß er führenden

Koeffizienten eins hat. Für Polynome mit ganzzahligen Koeffizienten ist das nicht möglich: Der ggT von

$$f = (2X+1)^2 = 4X^2+4X+1 \quad \text{und} \quad g = (2X+1)(2X-1) = 4X^2-1$$

ist $h = 2X+1$, was wir in $\mathbb{Z}[X]$ nicht zu $X+\frac{1}{2}$ kürzen können. Berechnen wir dagegen in $\mathbb{F}_5[X]$ den ggT der beiden Reduktionen modulo fünf, so ist $X+3$ ein genauso akzeptables Ergebnis wie $2(X+3) = 2X+1$ oder $3(X+3) = 3X+4$ oder $4(X+3) = 4X+2$. Welches dieser Polynome sollen wir nach $\mathbb{Z}[X]$ hochheben?

Wir wissen, daß der führende Koeffizient des ggT in $\mathbb{Z}[X]$ die führenden Koeffizienten beider Polynome teilen muß; er muß daher ein Teiler des ggT c dieser beiden führenden Koeffizienten sein. Wie wir am obigen Beispiel sehen, ist er freilich im Allgemeinen nicht gleich diesem ggT.

Betrachten wir zunächst den Fall, daß f und g zwei primitive Polynome sind im Sinne der folgenden

Definition: R sei ein faktorieller Ring.
a) Ein Polynom $f \in R[X]$ heißt *primitiv,* wenn der größte gemeinsame Teiler seiner Koeffizienten eins ist.
b) Der Inhalt $I(f)$ eines Polynoms $f \in R[X]$ ist der größte gemeinsame Teiler seiner Koeffizienten.

Offensichtlich läßt sich jedes Polynom aus $\mathbb{Z}[X]$ zerlegen in das Produkt aus einem primitiven Polynom und einer ganzen Zahl, seinem Inhalt.

Wenn der ggT zweier Polynome nicht primitiv ist, müssen beide Polynome durch seinen Inhalt teilbar sein. Ist also nur eines der beiden Polynome primitiv, so ist es auch der ggT.

Wenn wir von zwei primitiven Polynomen f, g mit führenden Koeffizienten a_d und b_e ausgehen, können wir den modulo p berechneten ggT zunächst so liften, daß er ggT(a_d, b_e) als führenden Koeffizienten hat. Im obigen Beispiel bekämen wir dann das Polynom $4X+2$.

Da die Polynomen f und g primitiv sind, muß auch ihr ggT primitiv sein. In $4X+2$ sind alle Koeffizienten durch zwei teilbar; der primitive Anteil ist der ggT $2X+1$.

Allgemein gilt: Wenn h den echten Teiler c_0 von c als führenden Koeffizienten hat, so ist $\widetilde{h} = \frac{c}{c_0} h$ ein Polynom mit führendem Koeffizienten c, das modulo p ein ggT von f mod p und g mod p ist. Sein primitiver Anteil ist der korrekte ggT h. im obigen Beispiel ist das $2X + 1$.

§5: Der Algorithmus

Nach vielen Vorbereitungen sind wir nun endlich in der Lage, einen Algorithmus zur modularen Berechnung des ggT in $\mathbb{Z}[X]$ oder $\mathbb{Q}[X]$ zu formulieren. Wesentlich ist für beide Fälle nur die Berechnung des ggT zweier primitiver Polynome aus $\mathbb{Z}[X]$: Zwei Polynome aus $\mathbb{Q}[X]$ lassen sich stets schreiben als λf und μg mit $\lambda, \mu \in \mathbb{Q}^\times$ und primitiven Polynomen $f, g \in \mathbb{Z}[X]$, und sie haben denselben ggT wie f und g. Für Polynome aus $\mathbb{Z}[X]$ sind $\lambda, \mu \in \mathbb{Z}$ die Inhalte, und der ggT in $\mathbb{Z}[X]$ ist $\mathrm{ggT}(\lambda, \mu) \cdot \mathrm{ggT}(f, g)$. Für Polynome aus $\mathbb{Q}[X]$ kommt es auf nicht auf Konstanten an, und wir können $\mathrm{ggT}(f, g)$ als ggT nehmen.

Seien nun also $f, g \in \mathbb{Z}[X]$ primitive Polynome.

a) Wir arbeiten nur mit Primzahlen, die nicht die führenden Koeffizienten sowohl von f als auch von g teilen. Wie wir im vorigen Paragraphen gesehen haben, hat dann der ggT h_p von $f^{(p)}$ und $g^{(p)}$ mindestens denselben Grad wie $\mathrm{ggT}(f, g)$, und es gibt höchstens endlich viele Primzahlen, für die sich die beiden Grade unterscheiden. Für alle anderen p ist $h_p = \mathrm{ggT}(f, g)^{(p)}$.

b) Diese endlich vielen Primzahlen, für die das Problem h_p größeren Grad hat, lassen sich nicht schon *a priori* ausschließen. Wir können sie aber anhand zweier Kriterien nachträglich erkennen: Falls wir eine Primzahl q (die nicht beide führende Koeffizienten teilt) finden, für die $\deg h_q < \deg h_p$ ist, muß das Problem bei p schlechte Reduktion haben. Wenn wir mehrere Primzahlen haben, die uns modulare ggTs desselben Grads liefern, so können wir diese nach dem chinesischen Restesatz zusammensetzen. Falls wir hier keine Lösung finden, bei der sämtliche Koeffizienten einen Betrag unterhalb der LANDAU-MIGNOTTE-Schranke liegen, oder wenn wir eine solche Lösung finden, diese aber

kein gemeinsamer Teiler von f und g ist, dann waren alle betrachteten Primzahlen schlecht.

Um die Übersicht zu behalten fassen wir bei der Rechnung alle bereits betrachteten Primzahlen zusammen zu einer Menge $\mathcal{P}$ und wir berechnen auch in jedem Schritt das Produkt N aller Elemente von $\mathcal{P}$, die wir noch nicht als schlecht erkannt haben. Falls sie wirklich nicht schlecht sind, kennen wir den ggT modulo N.

Diese Ideen führen zu folgendem Rechengang zur modularen Berechnung des ggT zweier primitiver Polynome aus $\mathbb{Z}[X]$:

1. Schritt (Initialisierung): Berechne den ggT c der führenden Koeffizienten von f und g sowie die LANDAU-MIGNOTTE-Schranke $\mathrm{LM}(f,g)$ und setze $M = 2c\big[\mathrm{LM}(f,g)\big] + 1$. Setze außerdem $\mathcal{P} = \emptyset$ und $N = 1$.

Da der Betrag eines jeden Koeffizienten des ggT höchstens gleich $\big[\mathrm{LM}(f,g)\big]$ ist und wir höchstens das c-fache dieses ggT berechnen, kennen wie die Koeffizienten in $\mathbb{Z}$, sobald wir sie modulo M kennen.

2. Schritt: Wähle eine zufällige Primzahl $p \notin \mathcal{P}$, die nicht die führenden Koeffizienten von sowohl f als auch g teilt, ersetze $\mathcal{P}$ durch $\mathcal{P} \cup \{p\}$ und berechne in $\mathbb{F}_p[X]$ den ggT h_p von $f^{(p)}$ und $g^{(p)}$. Dieser sei so normiert, daß sein höchster Koeffizient gleich eins ist. Falls $h_p = 1$ ist, endet der Algorithmus und $\mathrm{ggT}(f,g) = 1$. Andernfalls wird $N = p$ gesetzt und ein Polynom $h \in \mathbb{Z}[X]$ berechnet, dessen Reduktion modulo p gleich ch_p ist.

3. Schritt: Falls $N \geq M$ ist, ändere man die Koeffizienten von h modulo N nötigenfalls so ab, daß ihre Beträge höchstens gleich $c\,\mathrm{LM}(f,g)$ sind. Falls das nicht möglich ist, haben wir bislang modulo lauter schlechter Primzahlen gerechnet, können also alle bisherigen Ergebnisse vergessen und gehen zurück zum zweiten Schritt.

Andernfalls wird h durch seinen primitiven Anteil ersetzt und wir überprüfen, ob h sowohl f als auch g teilt. Falls ja, ist h der gesuchte ggT, und der Algorithmus endet; andernfalls müssen wir ebenfalls zurück zum zweiten Schritt und dort von Neuem anfangen.

4. Schritt: Im Fall $N < M$ wählen wir eine zufällige Primzahl $p \notin \mathcal{P}$, die nicht die führenden Koeffizienten von sowohl f als auch g teilt,

ersetzen $\mathcal{P}$ durch $\mathcal{P} \cup \{p\}$ und berechnen in $\mathbb{F}_p[X]$ den ggT h_p von $f^{(p)}$ und $g^{(p)}$. Falls dieser gleich eins ist, endet der Algorithmus und $\mathrm{ggT}(f, g) = 1$. Falls sein Grad größer als der von h ist, war p eine schlechte Primzahl; wir vergessen h_p und gehen zurück an den Anfang des vierten Schritts, d.h. wir wiederholen die Rechnung mit einer neuen Primzahl,

Falls der Grad von h_p kleiner ist als der von h, waren alle bisher betrachteten Primzahlen mit der eventuellen Ausnahme von p schlecht; wir setzen N deshalb zurück auf p und konstruieren ein Polynom $h \in \mathbb{Z}[X]$, dessen Reduktion modulo p gleich ch_p ist.

Ist schließlich $\deg h = \deg h_p$, so konstruieren wir nach dem chinesischen Restesatz ein neues Polynom h, das modulo N gleich dem alten h und modulo p gleich ch_p ist. Danach geht es weiter mit dem dritten Schritt.

Der Algorithmus muß enden, da es nur endlich viele Primzahlen p gibt, für die der in $\mathbb{F}_p[X]$ berechnete ggT nicht einfach die Reduktion von $\mathrm{ggT}(f, g)$ modulo p ist, und nach endlich vielen Durchläufen sind genügend viele gute Primzahlen zusammengekommen, daß ihr Produkt die Zahl M übersteigt. Da der ggT in $\mathbb{F}_p[X]$ für Primzahlen, die nicht beide führende Koeffizienten teilen, höchstens höheren Grad als $\mathrm{ggT}(f, g)$ haben kann, ist auch klar, daß der Algorithmus mit einem korrekten Ergebnis abbricht.

Betrachten wir dazu ein Beispiel:

$$f = X^6 - 124X^5 - 125X^4 - 2X^3 + 248X^2 + 249X + 125$$
$$g = X^5 + 127X^4 + 124X^3 - 255X^2 - 381X - 378$$

Eine einfache, aber langweilige Rechnung zeigt, daß die LANDAU-MIGNOTTE-Schranke von f und g ungefähr den Wert 13199,21452 hat. Wegen möglicher Rundungsfehler sollten wir zur Sicherheit vielleicht besser von 13200 ausgehen. Die Zahl, modulo derer wir die Koeffizienten mindestens kennen müssen, ist somit $M = 26401$.

Als erste Primzahl wählen wir zum Beispiel $p = 107$ und berechnen in

$\mathbb{F}_{107}$ den ggT von $f^{(107)}$ und $g^{(107)}$. Das Ergebnis ist

$$X^3 + 90X^2 + 90X + 89\,.$$

Damit ist $\mathcal{P} = \{107\}$ und $N = 107 < M$. Also wählen wir eine weitere Primzahl, etwa $p = 271$.

$$\mathrm{ggT}(f^{(271)}, g^{(271)}) = X^3 + 127X^2 + 127X + 126\,.$$

Auch dieser modulare ggT hat Grad drei, wir können die beiden also zusammensetzen, indem wir den chinesischen Restesatz auf die Koeffizienten anwenden:

$$x \equiv 90 \bmod 107 \wedge x \equiv 127 \bmod 271 \Longrightarrow x \equiv 5547 \bmod 28997$$

$$x \equiv 89 \bmod 107 \wedge x \equiv 126 \bmod 271 \Longrightarrow x \equiv 5546 \bmod 28997$$

Damit ist also $h = X^3 + 5547X^2 + 5547X + 5546$, $\mathcal{P} = \{107, 271\}$ und $N = 107 \times 271 = 28997$.

Dies ist größer als M, und alle Koeffizienten von h liegen unterhalb der LANDAU-MIGNOTTE-Schranke, also müssen wir untersuchen, ob h Teiler von f und von g ist. Bei der Division von f durch h erhalten wir den Rest

$$967384732340761X^2 + 967384732340761X + 967384732340761\,,$$

d.h. h ist kein Teiler von f. Somit sind 107 und 271 schlechte Primzahlen für diese ggT-Berechnung. Versuchen wir unser Glück als nächstes mit $p = 367$:

$$\mathrm{ggT}(f^{(371)}, g^{(371)}) = X^2 + X + 1\,.$$

Damit wird $\mathcal{P} = \{107, 271, 367\}$ und $N = 367$. Wir erwarten, daß der gesuchte ggT modulo 367 gleich $X^2 + X + 1$ ist. Um von 367 aus über die Schranke M zu kommen reicht eine relativ kleine Primzahl, z.B. $p = 73$.

$$\mathrm{ggT}(f^{(73)}, g^{(73)}) = X^3 + 22X^2 + 22X + 21\,.$$

Dieser ggT hat zu großen Grad, also ist auch 73 schlecht für uns. Wir lassen daher $N = 367$ und haben nun $\mathcal{P} = \{73, 107, 271, 367\}$.

Die nächste Primzahl nach 73 ist 79 und

$$\mathrm{ggT}(f^{(79)}, g^{(79)}) = X^2 + X + 1\,.$$

Wieder erhalten wir ein quadratisches Polynom, also setzen wir

$$N = 367 \times 79 = 44503, \qquad \mathcal{P} = \{73, 79, 107, 271, 367\}$$

und natürlich $h = X^2+X+1$. Da $N > M$ ist und alle Koeffizienten von h unter der LANDAU-MIGNOTTE-Schranke liegen, müssen wir nun testen, ob h Teiler von f und von g ist. Da wir in beiden Fällen Divisionsrest Null erhalten, ist $\mathrm{ggT}(f, g) = X^2 + X + 1$.

Bei diesem Beispiel habe ich natürlich absichtlich möglichst viele schlechte Primzahlen verwendet; wählt man seine Primzahlen wirklich zufällig, wird man nur selten eine erwischen.

Bei den beiden Polynomen

$$f = X^8 + X^6 - 3X^4 - 3X^3 + 8X^2 + 2X - 5$$

und

$$g = 3X^6 + 5X^4 - 4X^2 - 9X + 21$$

hatten wir gesehen, daß die Anwendung des EUKLIDischen Algorithmus zu Koeffizienten mit riesigen Nennern führt. Wenden wir auch hier die modulare Methode an! In §3 hatten wir bereits die LANDAU-MIGNOTTE-Schranke

$$\mathrm{LM}(f, g) = 2^6 \cdot \frac{2}{3}\sqrt{143} \approx 510{,}2191249$$

berechnet; da der ggT der führenden Koeffizienten gleich eins ist, reicht es also, den ggT modulo 1021 zu kennen. Dies ist eine Primzahl und paßt gut in ein Maschinenwort, also sollten wir als erstes die modulare Berechnung mit $p = 1021$ durchführen:

Division von $f^{(1021)}$ durch $g^{(1021)}$ führt auf den Divisionsrest

$$r_2 = 907X^4 + 227X^2 + 340\,,$$

Division von $g^{(1021)}$ durch r_2 auf

$$r_3 = 77X^2 + 1012X + 181\,,$$

die von r_2 durch r_3 auf

$$r_4 = 405X + 581\,,$$

und die von r_3 durch r_4 schließlich auf $r_5 = 956$.

Somit ist 956 ein ggT in $\mathbb{F}_{1021}[X]$, und damit natürlich auch die Eins. Nach dem, was wir in diesem Paragraphen gesehen haben, folgt daraus, daß auch der ggT von f und g in $\mathbb{Z}[X]$ gleich eins ist.

Vergleicht man mit dem entsprechenden Rechengang in $\mathbb{Q}[X]$, hat sich abgesehen von den modularen Polynomdivisionen nichts wesentliches geändert, jedoch sind die Zwischenergebnisse erheblich angenehmer geworden.

Die Resultante von f und g ist in diesem Fall $260708 = 2^2 \cdot 7 \cdot 9311$. Für deren Primteiler erhalten wir

$$\mathrm{ggT}(f^{(2)}, g^{(2)}) = X^2 + X + 1, \quad \mathrm{ggT}(f^{(7)}, g^{(7)}) = X + 3$$
$$\text{und} \quad \mathrm{ggT}(f^{(9311)}, g^{(9311)}) = X - 820\,;$$

für alle anderen Primzahlen p ist $\mathrm{ggT}(f^{(p)}, g^{(p)}) = 1$.

§6: Polynome in mehreren Veränderlichen

Wie im vorigen Kapitel erwähnt, ist auch der Polynomring in mehreren Veränderlichen über den ganzen Zahlen oder über einem Körper faktoriell; somit existieren auch dort größte gemeinsame Teiler. Wir haben gerade gesehen, wie sich diese im Falle einer Veränderlichen berechnen lassen. Hier soll nun im wesentlichen die gleiche Technik angewendet werden, um die ggT-Bestimmung für Polynome in n Veränderlichen zurückzuführen auf die in $n-1$ Veränderlichen.

Wir betrachten also zwei Polynome f, g in $n \geq 2$ Veränderlichen $X_1, \ldots, X_n$ über einem Körper oder einem faktoriellen Ring k; wichtig sind vor allem die Fälle $k = \mathbb{Z}$, $k = \mathbb{Q}$ und $k = \mathbb{F}_p$. Wir betrachten die Polynome aus $R_n = k[X_1, \ldots, X_n]$ als Polynome in der einer Veränderlichen X_n über dem Polynomring $R_{n-1} = k[X_1, \ldots, X_{n-1}]$, schreiben also $R_n = R_{n-1}[X_n]$. Durch ggT-Berechnungen in R_{n-1} können wir diese Polynome zerlegen in ihre Inhalte und primitiven Anteile; der ggT der Inhalte läßt sich wieder in R_{n-1} berechnen.

Bleibt noch der ggT der primitiven Anteile; diese seien f und g, jeweils aufgefaßt als Polynome in X_n mit Koeffizienten aus R_{n-1}. Um deren

ggT zu berechnen, könnten wir den EUKLIDischen Algorithmus über dem Quotientenkörper von R_{n-1} anwenden, allerdings steigen hier die Grade von Zähler und Nenner der Koeffizienten sowie *deren* Koeffizienten im allgemeinen so stark an, daß dies nur bei wenigen Variablen und sehr kleinen Graden praktisch durchführbar ist. Daher müssen wir auch hier wieder nach Alternativen suchen.

Im vorigen Paragraphen hatten wir, um die Explosion der Koeffizienten beim EUKLIDischen Algorithmus in $\mathbb{Q}[X]$ zu vermeiden, den Umweg über die ganzen Zahlen modulo einer Primzahl p genommen, also zunächst einen ggT (oder mehrere) in Körpern $\mathbb{F}_p[X]$ berechnet. Wie wir uns schon in den ersten beiden Paragraphen überlegt haben, können wir genauso auch die Variablenanzahl reduzieren, indem wir für eine der Variablen Werte einsetzen, z.B. für X_{n-1}. Wir betrachten also anstelle eines Polynoms $f \in R_n = R_{n-1}[X_n]$ das Polynom

$$f^{(c)} = f(X_1, \ldots, X_{n-2}, c, X_n) \in k[X_1, \ldots, X_{n-2}, X_n] = R_{n-2}[X_n]\,,$$

in dem für X_{n-1} der Wert c eingesetzt wird. Jeder Koeffizient $a_j \in R_{n-1}$ wird also ersetzt durch $a_j(X_1, \ldots, X_{n-2}, c) \in R_{n-2}$. Auch hier stellt sich die Frage, was der ggT von $f^{(c)}$ und $g^{(c)}$ mit dem von f und g zu tun hat.

Ist $h \in R_{n-1}[X]$ ein Teiler von f, etwa $f = qh$, so ist $f^{(c)} = q^{(c)}h^{(c)}$, d.h. auch $h^{(c)}$ ist ein Teiler von $f^{(c)}$. Dieser Teiler könnte aber einen kleineren Grad haben als h; dies passiert offensichtlich genau dann, wenn der führende Koeffizient von h durch Einsetzen von $X_{n-1} = c$ zum Nullpolynom aus R_{n-2} wird. Da der führende Koeffizient von f das Produkt der führenden Koeffizienten von h und q ist, gilt dann dasselbe auch für den führenden Koeffizienten von f; wir können dieses Problem also vermeiden, indem wir c so wählen, daß der führende Koeffizient von f durch Einsetzen von $X_{n-1} = c$ nicht zum Nullpolynom wird. Wenn wir das für f *oder* g sicherstellen, wissen wir daher, daß $\mathrm{ggT}(f,g)^{(c)}$ ein Teiler von $f^{(c)}$ und $g^{(c)}$, also auch von $\mathrm{ggT}(f^{(c)}, g^{(c)})$ ist, und daß beide größte gemeinsame Teiler denselben Grad in X_n haben. Da die führenden Koeffizienten von f und g als Polynome in X_{n-1} geschrieben werden können, gibt es nur endlich viele Werte von c, die wir vermeiden müssen, und diese lassen sich einfach identifizieren.

Auch dann wissen wir allerdings nur, daß $h^{(c)} = \text{ggT}(f,g)^{(c)}$ ein Teiler von $\text{ggT}(f^{(c)}, g^{(c)})$ ist. $h^{(c)}$ ist genau dann ein echter Teiler, wenn $f^{(c)}/h^{(c)}$ und $g^{(c)}/h^{(c)}$ einen gemeinsamen Faktor haben, der keine Einheit ist, wenn also die Resultante von $f^{(c)}/h^{(c)}$ und $g^{(c)}/h^{(c)}$ bezüglich X_n verschwindet. Bezeichnet h den ggT von f und g, so entsteht diese Resultante im Falle, daß *keiner* der führenden Koeffizienten von f oder g an der Stelle $X_{n-1} = c$ verschwindet, aus $\text{Res}_{X_n}(f/h, g/h) \in R_{n-1}$ durch Einsetzen von $X_{n-1} = c$. Da diese Resultante als Polynom in X_{n-1} geschrieben werden kann, gibt es daher wieder höchstens endlich viele Werte von c, für die dies der Fall ist. Da wir h nicht kennen, können wir diese Werte allerdings nicht im voraus identifizieren – ganz analog zur Situation bei der modularen Berechnung des ggT in $\mathbb{Z}[X]$.

Als nächstes stellt sich das Problem, was wir aus der Kenntnis von $\text{ggT}(f^{(c)}, g^{(c)})$ für $\text{ggT}(f,g)$ folgern können. Offensichtlich nicht sonderlich viel, denn wenn wir ein Polynom nur an einer Stelle $X_{n-1} = c$ kennen, gibt uns das noch kaum Information. Wenn wir allerdings ein Polynom vom Grad d in X_{n-1} an $d+1$ verschiedenen Punkten kennen, dann kennen wir es vollständig.

Die theoretisch einfachste Konstruktion des Polynoms aus seinen Funktionswerten an $d+1$ verschiedenen Stellen geht auf JOSEPH-LOUIS COMTE DE LAGRANGE zurück und benutzt dieselbe Strategie, die wir vom chinesischen Restesatz her kennen: Ist R ein Integritätsbereich und suchen wir ein Polynom $h \in R[X]$ vom Grad d, das an den Stellen $c_i \in R$ für $i = 0, \ldots, d$ die Werte $h_i \in R$ annimmt, so konstruieren wir zunächst Polynome α_i mit $\alpha_i(c_i) = 1$ und $\alpha_i(c_j) = 0$ für $j \neq i$. Das Verschwinden an den Stellen c_j können wir erreichen, indem wir die Linearfaktoren $(X - c_j)$ für $j \neq i$ miteinander multiplizieren. Um an der Stelle c_i den Wert eins zu erhalten, müssen wir allerdings noch durch das Produkt der $(c_i - c_j)$ dividieren, und damit kommen wir eventuell aus R hinaus und müssen im Quotientenkörper rechnen. Mit den so definierten Polynomen

$$\alpha_i(X) = \frac{\prod_{j \neq i}(X - c_j)}{\prod_{j \neq i}(c_i - c_j)}$$

ist das Interpolationspolynom dann $f(X) = \sum_{i=1}^{d} \alpha_i(X) h_i$.

(Das Interpolationsverfahren von LAGRANGE ist zwar einfach zu verstehen und führt auf eine elegante Formel, es gibt jedoch effizientere Verfahren, die auch hier anwendbar sind, z.B. das von ISAAC NEWTON. Für Einzelheiten sei auf die Numerik-Vorlesung verwiesen.)

Die Nenner in der LAGRANGEschen (oder auch NEWTONschen) Interpolationsformel stören uns nicht besonders, da wir ja spezialisieren, indem wir für X_{n-1} jeweils Konstanten einsetzen, die c_i liegen also alle im Ring k der Konstanten. Falls es sich dabei um einen Körper handelt, haben wir überhaupt keine Probleme mit den Divisionen; im wohl wichtigsten Fall, daß wir über den ganzen Zahlen arbeiten, erhalten wir zwar Interpolationspolynome mit rationalen Koeffizienten, können diese aber zerlegen in einen konstanten Faktor mal einem ganzzahligen Polynom mit teilerfremden Koeffizienten, das für die Berechnung des ggT zweier primitiver ganzzahliger Polynome an Stelle des Interpolationspolynoms verwendet werden kann.

JOSEPH-LOUIS LAGRANGE (1736–1813) wurde als GIUSEPPE LODOVICO LAGRANGIA in Turin geboren und studierte dort zunächst Latein. Erst eine alte Arbeit von HALLEY über algebraische Methoden in der Optik weckte sein Interesse an der Mathematik, woraus ein ausgedehnter Briefwechsel mit EULER entstand. In einem Brief vom 12. August 1755 berichtete er diesem unter anderem über seine Methode zur Berechnung von Maxima und Minima. 1756 wurde er, auf EULERs Vorschlag, Mitglied der Berliner Akademie; zehn Jahre später zog er nach Berlin und wurde dort EULERs Nachfolger als deren mathematischer Direktor. 1787 wechselte er an die Pariser Académie des Sciences, wo er bis zu seinem Tod blieb und unter anderem an der Einführung des metrischen Systems beteiligt war. Seine Arbeiten umspannen weite Teile der Analysis, Algebra und Geometrie.

Damit ergibt sich folgender Algorithmus zur Zurückführung des ggT zweier Polynome in n Veränderlichen auf die Berechnung von ggTs von Polynomen in $n-1$ Veränderlichen:

Wir gehen aus von zwei Polynomen $F, G \in R_n = k[X_1, \ldots, X_n]$, mit

$k = \mathbb{Z}, \mathbb{Q}$ oder $\mathbb{F}_p$ (oder sonst einem faktoriellen Ring, über dem wir den ggT zweier Polynome in einer Veränderlichen berechnen können).

1. Schritt (Initialisierung): Schreibe

$$F = \sum_{i=0}^{d} a_i(X_1, \ldots, X_{n-1}) X_n^i \quad \text{und}$$

$$G = \sum_{j=0}^{e} b_j(X_1, \ldots, X_{n-1}) X_n^j ,$$

wobei die führenden Koeffizienten a_d und b_e nicht identisch verschwinden sollen. Weiter sei $\mathcal{C} = \emptyset$ die Menge aller bislang betrachteten Spezialisierungen und $\mathcal{M} = \emptyset$ die Teilmenge der nach unserem jeweiligen Erkenntnisstand „guten“ Spezialisierungen. Als „Ersatz“ für die LANDAU-MIGNOTTE-Schranke haben wir das um eins vergrörte Maximum der X_{n-1}-Grade der a_i und b_j.

Als nächstes werden die Inhalte $I(F)$ und $I(G)$ von F und G bezüglich obiger Darstellung berechnet, d.h. $I(F)$ ist der ggT der $a_i(X_1, \ldots, X_{n-1})$ und $I(G)$ der von $b_0(X_1, \ldots, X_{n-1})$ bis $b_e(X_1, \ldots, X_{n-1})$. Beides kann bestimmt werden durch eine Folge von ggT-Berechnungen in $n-1$ Veränderlichen, ebenso auch der ggT I_0 dieser beiden Inhalte. Weiter seien $f = F/I(F)$ und $g = G/I(G)$ die primitiven Anteile von F und G. Der ggT von F und G ist I_0 mal dem in den folgenden Schritten berechneten ggT von f und g.

2. Schritt: Wähle so lange ein neues zufälliges Element $c \in k \setminus \mathcal{C}$ und ersetze $\mathcal{C}$ durch $\mathcal{C} \cup \{c\}$, bis $a_d(X_1, \ldots, X_{n-2}, c)$ und $b_e(X_1, \ldots, X_{n-2}, c)$ nicht beide gleich dem Nullpolynom sind. (Meist wird dies bereits beim ersten Versuch der Fall sein.) Berechne dann den ggT h_c von

$$f^{(c)} = \sum_{i=0}^{d} a_i(X_1, \ldots, X_{n-2}, c) X_n^i \quad \text{und}$$

$$g^{(c)} = \sum_{j=0}^{e} b_j(X_1, \ldots, X_{n-2}, c) X_n^j .$$

Falls $h_c = 1$, endet der Algorithmus mit dem Ergebnis $\mathrm{ggT}(f, g) = 1$. Andernfalls wird $\mathcal{M} = \{c\}$ und $N = \deg_{X_n} h_c$.

3. Schritt: Falls die Elementanzahl $\#\mathcal{M}$ von $\mathcal{M}$ gleich m ist, wird das Interpolationspolynom $h \in k[X_1, \ldots, X_n]$ berechnet, das für jedes $c \in \mathcal{M}$ die Gleichung

$$h(X_1, \ldots, X_{n-1}, c, X_n) = h_c(X_1, \ldots, X_{n-2}, X_n)$$

erfüllt. *Falls* h sowohl f als auch g teilt, ist $h = \mathrm{ggT}(f, g)$ und der Algorithmus endet mit diesem Ergebnis. Andernfalls waren alle bisherigen Spezialisierungen schlecht, und wir müssen von Neuem mit Schritt 2 beginnen.

4. Schritt: Falls $\#\mathcal{M} < m$, wählen wir ein zufälliges $c \in k \setminus \mathcal{C}$ solange, bis $a_d(X_1, \ldots, X_{n-2}, c)$ und $b_e(X_1, \ldots, X_{n-2}, c)$ nicht beide gleich dem Nullpolynom sind. Wir berechnen wieder den ggT h_c von

$$f^{(c)} = \sum_{i=0}^{d} a_i(X_1, \ldots, X_{n-2}, c) X_n^i \quad \text{und}$$

$$g^{(c)} = \sum_{j=0}^{e} b_j(X_1, \ldots, X_{n-2}, c) X_n^j \,.$$

Falls $h_c = 1$, endet der Algorithmus mit dem Ergebnis $\mathrm{ggT}(f, g) = 1$.

Falls $\deg_{X_n} h_c > N$ ist, haben wir ein schlechtes c gewählt und gehen zurück zum Anfang des vierten Schritts.

Falls $\deg_{X_n} h_c < N$ ist, waren alle zuvor betrachteten Werte von c schlecht; wir setzen $\mathcal{M} = \{c\}$ und $N = \deg_{X_n} h_c$.

Falls schließlich $\deg_{X_n} h_c = N$ ist, ersetzen wir $\mathcal{M}$ durch $\mathcal{M} \cup \{c\}$, und es geht weiter mit Schritt 3.

Da es nur endlich viele schlechte Werte für c gibt, muß der Algorithmus nach endlich vielen Schritten enden.

Als Beispiel wollen wir den ggT der beiden Polynome

$$f = X^3 + X^2Y + X^2Z + XYZ + Y^2Z + YZ^2$$

und

$$g = X^3 + X^2Y + X^2Z + XY^2 + XZ^2 + Y^3 + Y^2Z + YZ^2 + Z^3$$

aus $\mathbb{Z}[X, Y, Z]$ berechnen. Wir fassen Sie zunächst auf als Polynome in Z mit Koeffizienten aus $\mathbb{Z}[X, Y]$:

$$f = YZ^2 + (X^2 + XY + Y^2)Z + X^3 + X^2Y$$

und

$$g = Z^3 + (X + Y)Z^2 + (X^2 + Y^2)Z + X^3 + X^2Y + XY^2 + Y^3$$

Der führende Koeffizient von f ist Y, der von g ist eins. Wie man leicht sieht, sind beide Polynome bereits primitiv.

Der höchste Y-Grad eines Koeffizienten ist drei; wir brauchen daher vier zufällig gewählte Spezialisierungen. Der Einfachheit und vor allem der Übersichtlichkeit halber seien hierfür die (nicht gerade „zufälligen“) Werte $c = 1, 2, 3$ und 4 gewählt.

Für $c = 1$ ist

$$f(X, 1, Z) = Z^2 + (X^2 + X + 1)Z + X^3 + X^2$$

und

$$g(X, 1, Z) = Z^3 + (X + 1)Z^2 + (X^2 + 1)Z + X^3 + X^2 + X + 1\,;$$

wir müssen den ggT dieser beiden Polynome berechnen.

Dies leistet der entsprechende Algorithmus für Polynome in zwei Veränderlichen; da die Polynome wieder primitiv sind und der höchste X-Grad eines Koeffizienten gleich drei ist, müssen wir vier Spezialisierungen für X betrachten. Auch diese seien zufälligerweise gerade $1, 2, 3$ und 4. Wir erhalten folgende Ergebnisse:

d	$f(d, 1, Z)$	$g(d, 1, Z)$	ggT
1	$Z^2 + 3Z + 2$	$Z^3 + 2Z^2 + 2Z + 4$	$Z + 2$
2	$Z^2 + 7Z + 12$	$Z^3 + 3Z^2 + 5Z + 15$	$Z + 3$
3	$Z^2 + 13Z + 36$	$Z^3 + 4Z^2 + 10Z + 40$	$Z + 4$
4	$Z^2 + 21Z + 80$	$Z^3 + 5Z^2 + 17Z + 85$	$Z + 5$

Auch ohne Interpolationsformel sehen wir, daß

$$h_1(X, Z) = X + 1 + Z$$

das Interpolationspolynom ist. Division zeigt, daß

$$\frac{f(X,1,Z)}{h_1(X,Z)} = X^2 + Z \quad \text{und} \quad \frac{g(X,1,Z)}{h_1(X,Z)} = X^2 + Z^2 + 1$$

beides Polynome sind; somit ist

$$\mathrm{ggT}\big(f(X,1,Z), g(X,1,Z)\big) = X + 1 + Z\,.$$

Als nächstes setzen wir $c = 2$ für Y ein; wir erhalten

$$f(X,2,Z) = 2Z^2 + (X^2 + 2X + 4)Z + X^3 + 2X^2$$

und

$$g(X,2,Z) = Z^3 + (X+2)Z^2 + (X^2+4)Z + X^3 + 2X^2 + 4X + 8$$

und spezialisieren darin wieder X zu $1, 2, 3, 4$:

d	$f(d,2,Z)$	$g(d,2,Z)$	ggT
1	$2Z^2 + 7Z + 3$	$Z^3 + 3Z^2 + 5Z + 15$	$Z + 3$
2	$2Z^2 + 12Z + 16$	$Z^3 + 4Z^2 + 8Z + 32$	$Z + 4$
3	$2Z^2 + 19Z + 45$	$Z^3 + 5Z^2 + 13Z + 65$	$Z + 5$
4	$2Z^2 + 28Z + 96$	$Z^3 + 6Z^2 + 20Z + 120$	$Z + 6$

Hier ist unser ggT-Kandidat somit $h_2(X,Z) = X + 2 + Z$, und wieder zeigt Division, daß dies tatsächlich ein Teiler beider Polynome und somit deren ggT ist.

Für $c = 3$ ist

$$f(X,3,Z) = 3Z^2 + (X^2 + 3X + 9)Z + X^3 + 3X^2$$

und

$$g(X,3,Z) = Z^3 + 4Z^2 + 10Z + 40\,.$$

Die Spezialisierungen in X und ihre größten gemeinsamen Teiler sind

d	$f(d,3,Z)$	$g(d,3,Z)$	ggT
1	$3Z^2 + 13Z + 4$	$Z^3 + 4Z^2 + 10Z + 40$	$Z + 4$
2	$3Z^2 + 19Z + 20$	$Z^3 + 5Z^2 + 13Z + 65$	$Z + 5$
3	$3Z^2 + 27Z + 54$	$Z^3 + 6Z^2 + 18Z + 108$	$Z + 6$
4	$3Z^2 + 37Z + 112$	$Z^3 + 7Z^2 + 25Z + 175$	$Z + 7$

Hier ist entsprechend $h_3(X, Z) = X + 3 + Z$.

Für $c = 4$ schließlich erhalten wir

$$f(X, 4, Z) = 4Z^2 + (X^2 + 4X + 16)Z + X^3 + 4X^2$$

und

$$g(X, 4, Z) = Z^3 + (X + 4)Z^2 + (X^2 + 16)Z + X^3 + 4X^2 + 16X + 64\,.$$

Die Spezialisierungen in X und ihre größten gemeinsamen Teiler sind

d	$f(d, 4, Z)$	$g(d, 4, Z)$	ggT
1	$4Z^2 + 21Z + 5$	$Z^3 + 5Z^2 + 17Z + 85$	$Z + 5$
2	$4Z^2 + 28Z + 24$	$Z^3 + 6Z^2 + 20Z + 120$	$Z + 6$
3	$4Z^2 + 37Z + 63$	$Z^3 + 7Z^2 + 25Z + 175$	$Z + 7$
4	$4Z^2 + 48Z + 128$	$Z^3 + 8Z^2 + 32Z + 256$	$Z + 8$

Dies führt auf $h_4(X, Z) = X + 4 + Z$.

Auch das Polynom $h(X, Y, Z)$ mit $h(X, c, Z) = h_c(X, Z)$ für die Werte $c = 1, 2, 3, 4$ läßt sich ohne Interpolationsformel leicht erraten: Offensichtlich ist

$$h(X, Y, Z) = X + Y + Z\,.$$

Division zeigt, daß

$$\frac{f}{h} = X^2 + YZ \quad \text{und} \quad \frac{g}{h} = X^2 + Y^2 + Z^2$$

ist; somit ist

$$\mathrm{ggT}(f, g) = h = X + Y + Z\,.$$

Dieses Ergebnis hätten wir natürlich schon sehr viel früher erraten können, und in der Tat wird der Algorithmus oft so implementiert, daß man bereits nach eigentlich zu wenigen Spezialisierungen interpoliert und nachprüft, ob man einen gemeinsamen Teiler gefunden hat; wenn ja, ist dies der ggT. Falls nein, läßt sich aber noch nicht schließen, daß alle bisherigen Spezialisierungen schlecht waren; vielleicht waren auch nur die Grade einiger Koeffizienten zu klein, was sich nur durch weitere Spezialisierungen und Interpolationen feststellen läßt.

§7: Das Henselsche Lemma

Wir kennen bislang zwei Verfahren für die modulare Berechnung des ggT zweier Polynome aus $\mathbb{Z}[X]$: Entweder wir rechnen modulo einer hinreichend großen Primzahl, oder wir rechnen modulo mehrerer kleiner Primzahlen und versuchen anschließend, die Ergebnisse über den chinesischen Restesatz zusammenzusetzen.

Eine dritte Möglichkeit wäre, modulo einer kleinen Primzahl p zu rechnen und das Ergebnis dann schrittweise hochzuheben zu einem Ergebnis modulo p^2, p^3, und so weiter, bis zu einer p-Potenz, die über der notwendigen Schranke liegt. Diesen Ansatz stellte HANS JULIUS ZASSENHAUS erstmals 1969 vor mit einem Algorithmus zur Faktorisierung ganzzahliger Polynome; vier Jahre später machten JOEL MOSES und DAVID Y.Y. YUN daraus ein Verfahren zur Berechnung des größten gemeinsamen Teilers, das sie EZ GCD Algorithm nannten. Die Abkürzung steht für *Extended Zassenhaus Greatest Common Divisor.*

Grundlage sowohl des Faktorisierungsalgorithmus von ZASSENHAUS als auch von EZ GCD ist das HENSELsche Lemma, das es gestattet, Faktorisierungen modulo p hochzuheben zu Faktorisierungen modulo beliebiger p-Potenzen. Sie läßt sich allerdings im allgemeinen *nicht* hochheben zu einer Faktorisierung über den ganzen Zahlen. HENSEL definierte deshalb eine Erweiterung des Rings der ganzen Zahlen, die sogenannten p-adischen Zahlen, und sein eigentliches Resultat war, daß sich die Faktorisierung hochheben läßt in den Polynomring über den p-adischen Zahlen. Da uns Faktorisierungen modulo p^n genügen, möchte ich darauf nicht näher eingehen.

Lemma: f, g, h seien Polynome aus $\mathbb{Z}[X]$ derart, daß $f \equiv gh \bmod p$. Außerdem seinen seien $g \bmod p$ und $h \bmod p$ teilerfremd über $\mathbb{F}_p[X]$. Dann gibt es für jede natürliche Zahl n Polynome g_n, h_n derart, daß

$$g_n \equiv g \bmod p, \quad h_n \equiv h \bmod p \quad \text{und} \quad f \equiv g_n h_n \bmod p^n .$$

Beweis durch vollständige Induktion: Der Fall $n = 1$ ist die Voraussetzung des Lemmas. Ist das Lemma für ein n bewiesen, machen wir den Ansatz

$$g_{n+1} = g_n + p^n g^* \quad \text{und} \quad h_{n+1} = h_n + p^n h^* .$$

Nach Induktionsvoraussetzung ist $f \equiv g_n h_n \bmod p^n$, die Differenz $f - g_n h_n$ ist also durch p^n teilbar und es gibt ein Polynom $f^* \in \mathbb{Z}[X]$, so daß $f = g_n h_n + p^n f^*$ ist. Wir möchten, daß

$$f \equiv (g_n + p^n g^*)(h_n + p^n h^*) = g_n h_n + p^n (g_n h^* + h_n g^*) + p^{2n} \bmod p^{n+1}$$

wird. Da $2n \geq n+1$ ist, können wir den letzten Summanden vergessen; zu lösen ist also die Kongruenz

$$f \equiv g_n h_n + p^n f^* = g_n h_n + p^n (g_n h^* + h_n g^*) \bmod p^{n+1}$$

oder

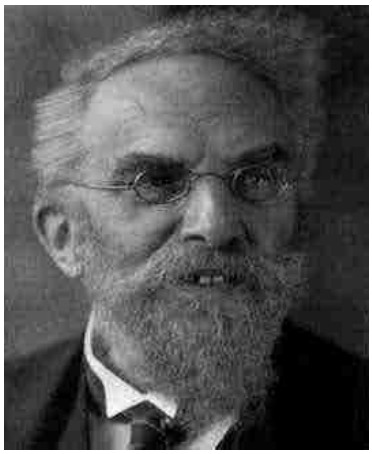
$$p^n f^* \equiv p^n (g_n h^* + h_n g^*) \bmod p^{n+1} .$$

Division durch p^n macht daraus

$$f^* \equiv g_n h^* + h_n g^* \bmod p \quad \text{oder} \quad f^* \equiv g h^* + h g^* \bmod p ,$$

denn $g_n \equiv g \bmod p$ und $h_n \equiv h \bmod p$. Die letztere Kongruenz können wir als Gleichung in $\mathbb{F}_p[X]$ auffassen und dort lösen, indem wir den erweiterten EUKLIDischen Algorithmus auf die Polynome $g \bmod p$ und $h \bmod p$ aus $\mathbb{F}_p[X]$ anwenden: Da diese nach Voraussetzung teilerfremd sind, können wir ihren ggT Eins und damit auch jedes andere Polynom über $\mathbb{F}_p$ als Linearkombination der beiden darstellen. Da der Grad von f die Summe der Grade von g und h ist und f^* höchstens denselben Grad wie f hat, können wir dann auch eine Darstellung $f^* = g h^* + h g^*$ in $\mathbb{F}_p[X]$ finden mit $\deg g^* \leq \deg g$ und $\deg h^* \leq \deg h$. Ersetzen wir g^* und h^* durch irgendwelche Repräsentantanten gleichen Grades aus $\mathbb{Z}[X]$, erfüllen $g_{n+1} = g_n + p^n g^*$ und $h_{n+1} = h_n + p^n h^*$ die Kongruenz $f \equiv g_n h_n$ modulo p^{n+1}.

■



KURT HENSEL wurde 1861 im damaligen Königsberg geboren; als er neun Jahre alt war, zog die Familie nach Berlin. Er studierte dort und in Bonn. 1884 promovierte er in Berlin bei KRONECKER, 1886 folgte die Habilitation. Er blieb bis 1901 als Privatdozent in Berlin. 1901 bekam er einen Lehrstuhl in Marburg, den er bis zu seiner Emeritierung 1930 innehatte. Er starb 1941 in Marburg. Seine Arbeiten drehen sich hauptsächlich um die Zahlentheorie und die eng damit verwandte Arithmetik von Funktionenkörpern. Bekannt wurde er vor allem durch die Einführung der p-adischen Zahlen. Er ist Autor dreier Lehrbücher.

§8: Der EZ GCD Algorithmus

Die Grundidee zur Anwendung des HENSELschen Lemmas auf die ggT-Berechnung ist einfach: Für zwei primitive Polynome $f, g \in \mathbb{Z}[X]$ wählt man eine zufällige Primzahl p, die nicht beide führende Koeffizienten a_d und b_e teilt, und berechnet nach dem EUKLIDischen Algorithmus „den“ ggT von $f^{(p)}$ und $g^{(p)}$. Das Polynom $h \in \mathbb{Z}[X]$ mit führendem Koeffizienten $\mathrm{ggT}(a_d, b_e)$ sei modulo p ein skalares Vielfaches dieses ggT, und $f^*, g^* \in \mathbb{Z}[X]$ seien Polynome, die modulo p gleich den Quotienten $f^{(p)}/h$ und $g^{(p)}/h$ sind. In $\mathbb{Z}[X]$ ist dann

$$f \equiv h \cdot f^* \bmod p \quad \text{und} \quad g \equiv h \cdot g^* \bmod p\,.$$

Falls h und f^* oder h und g^* zueinander teilerfremd sind, kann man die entsprechende Faktorisierung nach dem HENSELschen Lemma hochheben zu einer Faktorisierung von f oder g modulo einer beliebig großen Potenz von p. Sobald diese Potenz p^n größer ist als das Doppelte der LANDAU-MIGNOTTE-Schrankei mal dem ggT der führenden Koeffizienten von f und g, ersetzt man h gegebenenfalls durch ein modulo p^n dazu kongruentes Polynom, dessen Koeffizienten einen Betrag höchstens gleich dieser Schranke haben. Falls das nicht möglich ist, wissen wir wieder, daß der ggT der Polynome modulo p einen größeren Grad hat als der ggT der Ausgangspolynome, so daß wir den Algorithmus mit einer neuen Primzahl wiederholen müssen. Andernfalls müssen wir noch testen, ob das so konstruierte Polynom sowohl f als auch g teilt. Wenn ja, haben wir den ggT gefunden, wenn nein, müssen wir mit einer neuen Primzahl noch einmal von vorne anfangen.

Man beachte, daß es durchaus vorkommen kann, daß weder h und f^* noch h und g^* teilerfremd sind: Die beiden Polynome

$$f = (X-1)^2(X-2)(X-3) \quad \text{und} \quad g = (X-1)(X-2)^2(X-4)$$

etwa haben den ggT $h = (X-1)(X-2)$. Somit ist $f = h \cdot (X-1)(X-3)$ und $g = h \cdot (X-2)(X-4)$, und in beiden Fällen hat der Kofaktor einen gemeinsamen Teiler mit h.

In solchen Fällen hilft ein Trick von DAVID SPEAR: Für jedes Paar (a, b) ganzer Zahlen mit $b \neq 0$ ist $\mathrm{ggT}(f, g) = \mathrm{ggT}(f, af + bg)$. Ist nun

$h^{(p)} = \mathrm{ggT}(f^{(p)}, g^{(p)})$ und $f \equiv h \cdot f^* \bmod p$ sowie $g \equiv h \cdot g^* \bmod p$, so ist $af + bg \equiv h \cdot (af^* + bg^*) \bmod p$, und da $f^* \bmod p$ und $g^* \bmod p$ teilerfremd sind, gibt es Paare (a, b), für die $af^* + bg^* \bmod p$ teilerfremd ist zu $h^{(p)}$. Man wählt also so lange zufällig ein Paar (a, b) bis dies der Fall ist. Danach kann man nach dem HENSELschen Lemma die Faktorisierung $af + bg \equiv h \cdot (af^* + bg^*) \bmod p$ hochheben und weiter vorgehen wie oben. Im obigen Beispiel etwa ist $f^* = (X-1)(X-3)$ und $g^* = (X-2)(X-4)$; das Polynom $f^* + g^*$ hat für $X = 1$ den Wert drei und für $X = 2$ den Wert -1, ist also teilerfremd zu $h = (X-1)(X-2)$, so daß wir hier $a = b = 1$ wählen könnten.

Ohne Kenntnis der Faktorisierungen von f und g müßten wir ausgehen von

$$f = X^4 - 7X^3 + 17X^2 - 17X + 6 \quad \text{und} \quad g = X^4 - 9X^3 + 28X^2 - 36X + 16\,;$$

die LANDAU-MIGNOTTE-Schranke liegt knapp unter 412,3. Wir wählen die Primzahl $p = 101$ und erhalten $\mathrm{ggT}(f^{(101)}, g^{(101)}) = X^2 + 98X + 2$. Dividieren wir f und g modulo 101 durch dieses Polynom, erhalten wir

$$f \equiv (X^2 + 98X + 2)(X^2 + 97X + 3) \bmod 101$$

und

$$g \equiv (X^2 + 98X + 2)(X^2 + 95X + 8) \bmod 101\,;$$

Anwendung des EUKLIDischen Algorithmus in $\mathbb{F}_{101}[X]$ zeigt, daß

$$\mathrm{ggT}(X^2 + 98X + 2, X^2 + 97X + 3) = X + 100$$

und

$$\mathrm{ggT}(X^2 + 98X + 2, X^2 + 95X + 8) = 3X + 95$$

ist. Somit haben wir in keinem der beiden Fälle einen zu h teilerfremden Kofaktor und probieren es mit einer Linearkombination:

$$f + g = 2X^4 - 16X^3 + 45X^2 - 53X + 22 \equiv h \cdot (2X^2 + 91X + 11)\,,$$

und $q = 2X^2 + 91X + 11$ ist teilerfremd zu h. Genauer zeigt eine Anwendung des erweiterten EUKLIDischen Algorithmus in $\mathbb{F}_{101}[X]$, daß

$$(70X + 25)h + (66X + 69)q \equiv 1 \bmod 101$$

ist.

Zur Anwendung des HENSELschen Lemmas machen wir den Ansatz

$$f + g \equiv (h + 101h^*)(q + 101q^*) \equiv hq + 101(hq^* + h^*q) \bmod 101^2 .$$

$hq = 2X^4 + 287X^3 + 8933X^2 + 1260X + 22$ ist kongruent zu $f + g$ modulo 101, und

$$\frac{f + g - hq}{101} = -3X^3 - 88X^2 - 13X .$$

Wir müssen also Polynome q^* und h^* finden, so daß $hq^* + h^*q$ modulo 101 gleich diesem Polynom ist. Die obige Darstellung der Eins als Linearkombination von h und q führt auf

$$(70X+25)(-3X^3-88X^2-13X)h+(66X+69)(-3X^3-88X^2-13X)q \equiv -3X^3 - 88X^2 - 13X) \bmod 101 .$$

Ausmultipliziert modulo 101 wird das zu

$$(93X^4 + 27X^3 + 21X^2 + 19X)h + (4X^4 + 25X^3 + 39X^2 + 12X)q \equiv -3X^3 - 88X^2 - 13X \bmod 101 .$$

Die Vorfaktoren von h und q haben hier natürlich noch einen viel zu hohen Grad. Da wir beliebige Vielfache der Gleichung $qh - hq = 0$ subtrahieren dürfen, können wir den Faktor vor h ersetzen durch seinen Rest bei der Division durch q, und den vor q durch seinen Rest bei der Division durch h, wobei wir natürlich alle Divisionen in $\mathbb{F}_{101}[X]$ ausführen müssen. Beide Reste sind gleich $100X$; wir erhalten also die erheblich einfachere Kongruenz

$$100Xh + 100Xq \equiv -3X^3 - 88X^2 - 13X \bmod 101 .$$

Somit können wir $q^* = h^* = 100X$ setzen und erhalten die Kongruenz

$$\begin{aligned} f + q &\equiv (h + 101h^*)(q + 101q^*) \\ &= (X^2 + 10198X + 2)(X^2 + 10191X + 11) \bmod 101^2 . \end{aligned}$$

101^2 liegt über dem Doppelten der LANDAU-MIGNOTTE-Schranke; da $10198 \equiv -3 \bmod 101^2$, ist $X^2 - 3X + 2$ ein Polynom aus $\mathbb{Z}[X]$, das modulo 101^2 kongruent zum ersten Faktor ist und Koeffizienten vom Betrag unterhalb der LANDAU-MIGNOTTE-Schranke hat und somit ein

Kandidat für den ggT ist. Tatsächlich zeigt Polynomdivision in $\mathbb{Z}[X]$, daß dieses Polynom sowohl f als auch g teilt und damit gleich dem größten gemeinsamen Teiler dieser beiden Polynome ist.

Wenn wir die Restklassen modulo 101 nicht durch die Zahlen von Null bis hundert repräsentiert hätten, sondern durch die zwischen -50 und 50, hätten wir schon modulo 101 den ggT $X^2 - 3X + 2$ erhalten und hätten durch eine Division in $\mathbb{Z}[X]$ sehen können, daß dieses Polynom ein Teiler sowohl von f als auch von g ist. Da 101 keinen führenden Koeffizienten teilt und der modulare ggT somit mindestens den Grad des ggT in $\mathbb{Z}[X]$ hat, hätten wir bereits an dieser Stelle erkannt, daß dies der ggT sein muß. Es kann sich also durchaus lohnen, schon modulo „zu kleiner“ Primzahlpotenzen zu überprüfen, ob ein gemeinsamer Teiler gefunden ist. Die LANDAU-MIGNOTTE-Schranke ist schließlich in den meisten Fällen recht pessimistisch.

Kapitel 4
Faktorisierung von Polynomen

Nach einem Satz von GAUSS läßt sich jedes Polynom in endlich vielen Veränderlichen über einem faktoriellen Ring bis auf Reihenfolge und Einheiten eindeutig als Produkt irreduzibler Faktoren schreiben. Der Beweis, den GAUSS in Artkel 42 seiner 1801 erschienenen *Disquisitiones Arithmeticae* gab, ist allerdings nicht konstruktiv. Ein zumindest im Prinzip konstruktives Verfahren zur Bestimmung der Faktoren entwickelte erst Jahrzehnte später LEOPOLD KRONECKER, das erste wirklich brauchbare Verfahren sogar erst 1969 HANS JULIUS ZASSENHAUS.

§1: Der Algorithmus von Kronecker

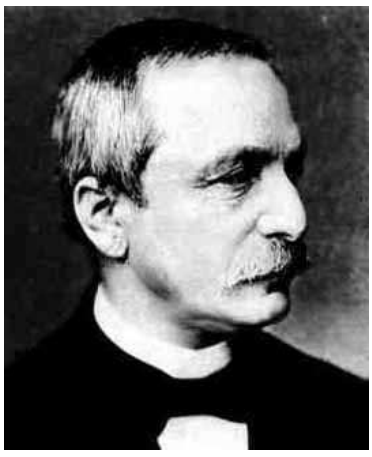
Der erste zumindest im Prinzip konstruktive Beweis für die Faktorisierbarkeit von Polynomen geht zurück auf LEOPOLD KRONECKER. Er führt das Problem der Faktorisierung eines Polynoms über $\mathbb{Z}$ zurück auf das (zumindest für große Zahlen alles andere als einfache) Problem der Faktorisierung ganzer Zahlen. Ausgangspunkt ist die folgende triviale Beobachtung: Angenommen, wir haben in $\mathbb{Z}[X]$ eine Zerlegung $f = gh$. Für jede ganze Zahl a ist dann $f(a) = g(a)h(a)$. Somit ist $g(a)$ für jedes $a \in \mathbb{Z}$ ein Teiler von $f(a)$.

Ein Polynom vom Grad d ist eindeutig bestimmt durch seine Werte an $d+1$ verschiedenen Stellen $a_0, \ldots, a_n$ und kann mit Hilfe wohlbekannter Interpolationsformeln leicht aus den $d + 1$ Paaren $\big(a_i, g(a_i)\big)$ bestimmt werden; die möglichen Werte $g(a_i)$ wiederum sind Teiler der $f(a_i)$.

Daher berechnet KRONECKER auf der Suche nach einem Teiler vom Grad d für $d + 1$ beliebig gewählte ganzzahlige Werte $a_0, \ldots, a_d$ die

Funktionswerte $f(a_0), \ldots, f(a_d)$, und konstruiert für jedes $(d+1)$-tupel $(b_0, \ldots, b_d)$ ganzer Zahlen mit $b_i | f(a_i)$ ein Interpolationspolynom g mit $g(a_i) = b_i$. Falls keines der Polynome Teiler von f ist, hat f keinen Teiler vom Grad d, andernfalls werden diese Teiler gefunden.

Über den Grad d eines potentiellen Teilers ist natürlich *a priori* nichts bekannt; wir wissen nur eines: Wenn es einen nichttrivialen Teiler gibt, dann gibt es auch einen, dessen Grad höchstens gleich der Hälfte des Grads von f ist. Im Extremfall müssen wir daher alle diese Grade ausprobieren, bis sich dann möglicherweise herausstellt, daß f irreduzibel ist. Falls dann noch einige der Zahlen $f(a_i)$ viele Teiler haben, läßt sich leicht vorstellen, daß der Aufwand schon für sehr moderate Grade von f astronomisch wird. Hinzu kommt, daß auch die Primzerlegung großer Zahlen sehr aufwendig sein kann, auch wenn die algorithmische Zahlentheorie inzwischen Verfahren kennt, die für große Zahlen deutlich effizienter sind als Probedivisionen. Zum Glück kennt die Computeralgebra auch Faktorisierungsverfahren für Polynome, die deutlich effizienter sind als das von KRONECKER, so daß dessen Algorithmus heute keine Rolle mehr spielt.



LEOPOLD KRONECKER (1823–1891) ist heute zwar Vielen nur im Zusammenhang mit dem KRONECKER-δ bekannt, er war aber einer der bedeutendsten deutschen Mathematiker seiner Zeit. Seine Arbeiten befaßten sich mit Algebra, Zahlentheorie und Analysis, wobei er insbesondere die Verbindungen zwischen der Analysis und den beiden anderen Gebieten erforschte. Bekannt ist auch seine Ablehnung jeglicher mathematischer Methoden, die, wie die Mengenlehre oder Teile der Analysis, unendliche Konstruktionen verwenden. Er war deshalb mit vielen anderen bedeutenden Mathematikern seiner Zeit verfeindet, z.B. mit CANTOR und mit WEIERSTRASS.

Als Beispiel wollen wir versuchen, das Polynom

$$f = 8X^7 - 16X^6 - 20X^5 + 15X^4 + 13X^3 + 9X^2 + 10X + 2$$

vom Grad sieben nach KRONECKER in $\mathbb{Z}[X]$ zu faktorisieren. Falls f nicht irreduzibel ist, muß es mindestens einen Faktor vom Grad höchstens drei haben.

Linearfaktoren haben die Form $(bX + c)$. Um sie mit KRONECKERs Methode zu finden, müssen wir die Funktion an zwei Stellen mit möglichst einfachen Funktionswerten betrachten; dazu bieten sich $x_0 = -1$ mit $f(x_0) = -1$ und $x_1 = 0$ mit $f(x_1) = 2$ an. Für einen Teiler $g \in \mathbb{Z}[X]$ von f muß dann $g(x_0) = \pm 1$ und $g(x_1) = \pm 1$ oder ± 2 sein.

Tatsächlich können wir uns auf Polynome mit $g(x_0) = 1$ beschränken, denn g ist genau dann ein Teiler, wenn auch $-g$ einer ist, und wenn das eine Polynom an der Stelle -1 den Wert 1 hat, ist das andere dort gleich -1. Wir müssen also die Interpolationspolynome zu den vier Wertepaaren $\big((-1, 1), (0, y_0)\big)$ mit $y_0 \in \{-2, -1, 1, 2\}$ konstruieren und testen, ob sie f teilen. Interpolation nach LAGRANGE oder NEWTON führt auf das konstante Polynom 1, das natürlich keinen Faktor liefert. Somit gibt es keine linearen Faktoren.

Auf der Suche nach quadratischen Faktoren brauchen wir einen weiteren Interpolationspunkt; da $f(1) = 21$ nur zwei Primteiler hat, bietet sich $x = 1$ an, wo ein Teiler von f einen der acht Werte $\pm 1, \pm 3, \pm 7$ oder ± 21 annehmen muß. Nun müssen also schon $4 \times 8 = 32$ Interpolationspolynome konstruiert und durchprobiert werden. Nachprüfen per Computer zeigt, daß keines davon auf einen Faktor führt; es gibt also auch keine quadratischen Teiler.

Für kubische Faktoren brauchen wir einen weiteren Interpolationspunkt. $f(2) = -238$ und $f(-2) = -1254$ sehen nicht gerade vielversprechend aus, aber Versuche mit anderen betragskleinen x-Werten liefern auch nichts Besseres. $238 = 2 \cdot 7 \cdot 17$ hat drei Primteiler, $1254 = 2 \cdot 3 \cdot 11 \cdot 19$ aber vier, also versuchen wir unser Glück mit dem Punkt $(2, -238)$, wobei wir nun für $g(2)$ schon 16 Werte betrachten müssen. Insgesamt müssen wir somit $4 \times 8 \times 16 = 512$ Interpolationspolynome berechnen. Eines davon führt auf den Faktor $-2X^3 + 3X^2 + 5X + 1$. Dieser ist irreduzibel, denn jeder Teiler wäre ja erst recht ein Teiler von f. Aus dem gleichen Grund ist in der Zerlegung

$$\begin{aligned} f &= (-2X^3 + 3X^2 + 5X + 1)(-4X^4 + 2X^3 + 3X^2 + 2) \\ &= (2X^3 - 3X^2 - 5X - 1)(4X^4 - 2X^3 - 3X^2 - 2)\,. \end{aligned}$$

auch der zweite Faktor irreduzibel, denn andernfalls müßte er einen höchstens quadratischen Teiler haben.

§2: Die quadratfreie Zerlegung eines Polynoms

Die vollständige Faktorisierung eines Polynoms ist recht aufweendig. Wir betrachten daher zunächst eine Vorstufe, die mit Hilfe des EUKLIDischen Algorithmus schnell und effizient gefunden werden kann und den weiteren Verlauf der Faktorisierung vereinfachen wird.

Wir betrachten ein Polynom f über einem Körper oder einem faktoriellen Ring k. Da der Polynomring $k[X]$ faktoriell ist, zerfällt f dort in ein Produkt aus einer Einheit $u \in k^\times$ und Potenzen irreduzibler Polynome aus $k[X]$:

$$f = u \prod_{i=1}^{N} q_i^{e_i} .$$

Falls alle $e_i = 1$ und kein zwei q_i zueinander assoziiert sind, bezeichnen wir f als quadratfrei. In diesem Fall gibt es kein Polynom positiven Grades, dessen Quadrat ein Teiler von f ist.

Ziel der quadratfreien Zerlegung ist es, ein beliebiges Polynom f in der Form

$$f = \prod_{j=1}^{M} g_j^j$$

zu schreiben, wobei die g_j paarweise teilerfremde quadratfreie Polynome sind. Vergleichen wir mit der obigen Darstellung und vernachlässigen wir für den Augenblick die Einheit u, so folgt, daß g_j das Produkt aller q_i mit $e_i = j$ ist.

a) Quadratfreie Zerlegung über den reellen Zahlen

Wenn ein Polynom $f \in \mathbb{R}[X]$ eine mehrfache Nullstelle hat, verschwindet dort auch die Ableitung f'. Allgemeiner gilt, daß für ein Polynom $h \in \mathbb{R}[X]$, dessen e-te Potenz f teilt, zumindest h^{e-1} auch die Ableitung f' teilen muß, denn ist $f = h^e g$, so ist

$$f' = eh^{e-1}h'g + h^e g' = h^{e-1}(eh'g + hg') .$$

Falls f *genau* durch h^e teilbar ist, ist auch f' genau durch h^{e-1} teilbar, denn wäre es sogar durch h^e teilbar, so wäre auch $eh^{e-1}h'g$ durch h^e teilbar, so daß h ein Teiler von g wäre.

Damit ist $\mathrm{ggT}(f, f') = \prod_{i=1}^{r} f_i^{e_i-1}$ und

$$h_1 = \frac{f}{\mathrm{ggT}(f, f')} = \prod_{i=1}^{r} q_i$$

ist das Produkt aller irreduzibler Faktoren von f. Alle irreduziblen Faktoren von f, die dort mindestens in der zweiten Potenz vorkommen, sind auch Teiler von f', also ist

$$g_1 = \frac{h_1}{\mathrm{ggT}(h_1, f')}$$

das Produkt aller irreduzibler Faktoren von f, die dort genau in der ersten Potenz vorkommen.

In $f_1 = f/h_1$ kommen alle irreduziblen Faktoren von f mit einem um eins verminderten Exponenten vor; insbesondere sind also die mit $e_i = 1$ verschwunden. Wenden wir darauf dieselbe Konstruktion an, erhalten wir die Zerlegung $\mathrm{ggT}(f_1, f_1') = \prod_{i=1}^{r} f_i^{\max(e_i-2,0)}$, und

$$h_2 = \frac{f_1}{\mathrm{ggT}(f_1, f_1')} = \prod_{i=1}^{r} q_i$$

ist das Produkt aller irreduzibler Faktoren von f_1, also das Produkt aller Faktoren von f, die mit einem Exponenten von mindestens zwei vorkommen. Damit ist

$$g_2 = \frac{h_2}{\mathrm{ggT}(h_2, f_1')}$$

das Produkt aller Faktoren, die in f mit Multiplizität genau zwei vorkommen.

Nach dem gleichen Schema können wir, falls $f_2(x)$ nicht konstant ist, weitermachen und rekursiv für $i \geq 3$ definieren

$$h_i = \frac{f_{i-1}}{\mathrm{ggT}(f_{i-1}, f_{i-1}')}, \quad g_i(x) = \frac{h_i}{\mathrm{ggT}(h_i, f_{i-1}')} \quad \text{und} \quad f_i(x) = \frac{f_{i-1}}{h_i},$$

bis wir für ein konstantes f_i erhalten. Dann hat jedes Polynom g_i nur einfache Nullstellen, und diese Nullstellen sind genau die i-fachen Nullstellen des Ausgangspolynoms f.

Bis auf eine eventuell notwendige Konstante c ist damit f das Produkt der Polynome g_j^j, und falls wir alle Nullstellen der Polynome g_j bestimmen können, kennen wir alle Nullstellen von f.

Als Beispiel betrachten wir das Polynom

$$f(x) = X^4 - 5X^2 + 6X - 2 \quad \text{mit} \quad f'(X) = 4X^3 - 10X + 6\,.$$

Wir berechnen zunächst den ggT von f und f':

$$(X^4 - 5X^2 + 6X - 2) : (4X^3 - 10X + 6) = \frac{X}{4} \text{ Rest } -\frac{5}{2}X^2 + \frac{9}{2}X - 2$$

$$(4X^3 - 10X + 6) : \left(-\frac{5}{2}X^2 + \frac{9}{2}X - 2\right) = -\frac{8}{5}X - \frac{72}{25} \text{ Rest } -\frac{6}{25}X + \frac{6}{25}$$

$$\left(-\frac{5}{2}X^2 + \frac{9}{2}X - 2\right) : \left(-\frac{6}{25}X + \frac{6}{25}\right) = \frac{125}{12}X - \frac{25}{3}$$

Somit ist der ggT gleich $-\frac{6}{25}(X - 1)$; da es auf Konstanten nicht ankommt, rechnen wir besser mit $(X - 1)$.

Eigentlich sind wir damit schon fertig: Der ggT hat nur die einfache Nullstelle $x = 1$, also hat $f(x)$ an der Stelle eins eine doppelte Nullstelle, und alles andere sind einfache Nullstellen. Da

$$(X^4 - 5X^2 + 6X - 2) : (X - 1)^2 = X^2 + 2X - 2$$

ist, haben wir die quadratfreie Zerlegung

$$f(X) = (X^2 + 2X - 2) \cdot (X - 1)^2\,.$$

Zur Illustration können wir aber auch strikt nach Schema weiterrechnen. Dann brauchen wir als nächstes

$$h_1 = \frac{f}{\text{ggT}(f, f')} = \frac{X^4 - 5X^2 + 6X - 2}{X - 1} = X^3 + X^2 - 4X + 2\,,$$

das Polynom das an jeder Nullstelle von $f(X)$ eine einfache Nullstelle hat. Sodann brauchen wir den ggT von $h_1(X)$ und $f'(X)$; da wir schon wissen, daß f und f' außer der Eins keine gemeinsame Nullstelle haben, muß das $(X - 1)$ sein. Somit ist

$$g_1 = \frac{X^3 + X^2 - 4X + 2}{X - 1} = X^2 + 2X - 2 = (X + 1)^2 - 3$$

das Polynom, das genau bei den einfachen Nullstellen von f verschwindet, also bei $-1 \pm \sqrt{3}$. Als nächstes muß

$$g_1(X) = \frac{f(X)}{h_1(X)} = \frac{X^4 - 5X^2 + 6X - 2}{X^3 + X^2 - 4X + 2} = X - 1$$

untersucht werden; da es nur für $X = 1$ verschwindet, ist die Eins eine doppelte Nullstelle von f. Damit sind alle Nullstellen von $f(X)$ sowie auch deren Vielfachheiten gefunden.

b) Ableitungen über einem beliebigen Körper

Auch wenn Ableitungen ursprünglich über Grenzwerte definiert sind, ist doch die Ableitung eines Polynoms rechnerisch gesehen eine rein algebraische Operation, die sich im Prinzip über jedem beliebigen Körper oder sogar Ring k erklären läßt: Wir *definieren* die Ableitung eines Polynoms

$$f = a_d X^d + a_{d-1} X^{d-1} + \cdots + a_1 X + a_0 \in k[X]$$

als das Polynom

$$f' = d a_d X^{d-1} + (d-1) a_{d-1} X^{d-2} + \cdots + a_1 \in k[X]\,.$$

Es ist klar, daß die so definierte Abbildung $k[X] \to k[X]$, die jedem Polynom $f \in k[X]$ seine Ableitung f' zuordnet, k-linear ist. Auch die LEIBNIZsche Produktregel $(fg)' = fg' + fg'$ ist erfüllt: Wegen der Linearität der Ableitung und der Linearität beider Seiten der Formel sowohl in f als auch in g genügt es, dies für X-Potenzen nachzurechnen, und für $f = X^n, g = X^m$ ist $(fg)' = (n+m)X^{n+m-1}$ gleich

$$fg' + f'g = X^n m X^{m-1} + n X^{n-1} X^m = (m+n) X^{n+m-1}\,.$$

Damit gelten die üblichen Ableitungsregeln auch für die formale Ableitung von Polynomen aus $k[X]$.

c) Die Charakteristik eines Körpers

Es gibt allerdings einen wesentlichen Unterschied: In der Analysis folgt durch eine einfache Anwendung des Mittelwertsatzes, daß die Ableitung einer differenzierbaren Funktion genau dann identisch verschwindet,

wenn die Funktion konstant ist. Bei einer rein algebraischen Behandlung des Themas können wir natürlich nicht auf den Mittelwertsatz der Differentialrechnung zurückgreifen, sondern müssen direkt nachrechnen, wann die Ableitung eines Polynoms gleich dem Nullpolynom ist.

Mit den obigen Bezeichnungen ist dies genau dann der Fall, wenn alle Koeffizienten ia_i der Ableitung verschwinden. Bei den Faktoren dieses Produkts handelt es sich um die Zahl $i \in \mathbb{N}_0$ und das Körperelement a_i.

Falls der Grundkörper k die rationalen Zahlen enthält, können wir auch i als Element von k auffassen und haben somit ein Produkt zweier Körperelemente. Dieses verschwindet genau dann, wenn mindestens einer der beiden Faktoren verschwindet. Die Ableitung ist somit genau dann das Nullpolynom, wenn $a_i = 0$ für alle $i \neq 0$, wenn das Polynom also konstant ist.

Auch wenn $\mathbb{N}_0$ keine Teilmenge von k ist, muß k als Körper zumindest die Eins enthalten. Wir können daher rekursiv eine Abbildung φ von $\mathbb{N}_0$ nach k definieren durch die Vorgaben $\varphi(0) = 0$ und $\varphi(n+1) = \varphi(n) + 1$ für alle $n \in \mathbb{N}_0$. Diese Abbildung läßt sich auf $\mathbb{Z}$ fortsetzen durch die weitere Forderung $\varphi(-n) = -\varphi(n)$.

Da die Addition in $\mathbb{N}$ rekursiv über Summen von Einsen definiert wird, überlegt man sich schnell, daß für zwei ganze Zahlen $a, b \in \mathbb{Z}$ gilt: $\varphi(a+b) = \varphi(a) + \varphi(b)$. Da die Multiplikation in $\mathbb{Z}$ rekursiv definiert ist über die Addition, folgt daraus wiederum, daß auch $\varphi(ab) = \varphi(a)\varphi(b)$ ist; φ ist also mit Addition und Multiplikation verträglich. (In der Algebra sagt man, φ sei ein Ringhomomorphismus.)

Falls $\varphi(n)$ nur für $n = 0$ verschwindet, kann $\mathbb{Z}$ und damit auch $\mathbb{Q}$ als Teilmenge von k aufgefaßt werden; andernfalls gibt es eine kleinste natürliche Zahl p, so daß $\varphi(p) = 0$ ist. Da $\varphi(1) = 1 \neq 0$, ist $p \geq 2$.

Ist a eine weitere ganze Zahl mit $\varphi(a) = 0$, so können wir a mit Rest durch p dividieren: $a = pb + r$ mit $0 \leq r < p$. Dabei ist

$$\varphi(r) = \varphi(a) - \varphi(pb) = \varphi(a) - \varphi(p)\varphi(b) = 0\,,$$

also ist auch $r = 0$, denn p ist die kleinste positive Zahl mit $\varphi(p) = 0$. Somit verschwindet $\varphi(p)$ genau für die Vielfachen von p.

Schreiben wir $p = ab$ als Produkt zweier natürlicher Zahlen a, b, so ist $0 = \varphi(p) = \varphi(a)\varphi(b)$. Da $\varphi(a)$ und $\varphi(b)$ Körperelemente sind, muß daher mindestens eines der beiden verschwinden; da beides natürliche Zahlen und höchstens gleich p sind, geht das nur, wenn eines gleich eins und das andere gleich p ist. Somit ist p eine Primzahl.

Definition: Wir sagen, ein Körper k habe die Charakteristik null, wenn die Abbildung $\varphi\colon \mathbb{Z} \to k$ injektiv ist. Andernfalls sagen wir, die Charakteristik von k sei gleich p, wobei p die kleinste natürliche Zahl ist mit $\varphi(p) = 0$.

Die Charakteristik eines Körpers ist somit entweder null oder eine Primzahl. Wir schreiben char $k = 0$ *bzw.* char $k = p$.

Offensichtlich bilden die rationalen, reellen und auch komplexen Zahlen Körper der Charakteristik null, und char $\mathbb{F}_p = p$.

Gehen wir zurück zur Ableitung eines Polynoms! Das Produkt ia_i ist gleich dem in k berechneten Produkt $\varphi(i)a_i$, verschwindet also genau dann, wenn mindestens einer der beiden Faktoren verschwindet. Für einen Körper der Charakteristik null verschwindet $\varphi(i)$ nur für $i = 0$; hier müssen also alle a_i mit $i \geq 1$ verschwindet, d.h. das Polynom ist konstant.

Für einen Körper der Charakteristik $p > 0$ verschwindet $\varphi(i)$ allerdings auch für alle Vielfachen von p, so daß die entsprechenden Koeffizienten nicht verschwinden müssen. Ein Polynom f über einem Körper der Charakteristik $p > 0$ hat daher genau dann das Nullpolynom als Ableitung, wenn es sich als Polynom in X^p schreiben läßt.

Wir wollen uns überlegen, daß dies genau dann der Fall ist, wenn das Polynom die p-te Potenz eines anderen Polynoms ist, dessen Koeffizienten allerdings möglicherweise in einem größeren Körper liegen. Ausgangspunkt dafür ist das folgende

Lemma: Ist k ein Körper der Charakteristik $p > 0$, so gilt für zwei Polynome $f, g \in k[X]$ und zwei Elemente a, b des Körpers die Gleichung $(af + bg)^p = a^p f^p + b^p g^p$. Insbesondere ist

$$(a_d X^d + a_{d-1} X^{d-1} + \cdots + a_0)^p = a_d^p X^{dp} + a_{d-1}^p X^{(d-1)p} + \cdots + a_0^p \,.$$

Beweis: Nach dem binomischen Lehrsatz ist

$$(af + bg)^p = \sum_{i=0}^{p} \binom{p}{i} (af)^i (bg)^{p-i} \quad \text{mit} \quad \binom{p}{i} = \frac{p \cdots (p-i+1)}{i!} .$$

Für $i = 0$ und $i = p$ ist $\binom{p}{i} = 1$, für alle anderen i steht p zwar im Zähler, nicht aber im Nenner des obigen Bruchs. Daher ist $\binom{p}{i}$ durch p teilbar, die Multiplikation mit $\binom{p}{i}$ ist also die Nullabbildung. Somit ist

$$(af + bg)^p = (af)^p + (bg)^p = a^p f^p + b^p g^p .$$

Durch Anwendung auf die Summanden des Polynoms folgt daraus induktiv auch die zweite Formel. ∎

Wir können dieses Lemma auch speziell auf eine Summe von lauter Einsen anwenden; dann folgt

$$\underbrace{(1 + \cdots + 1)}_{n \text{ mal}}^p = \underbrace{1^p + \cdots + 1^p}_{n \text{ mal}} = \underbrace{1 + \cdots + 1}_{n \text{ mal}} ;$$

solche Summen sind also gleich ihrer p-ten Potenz. Somit gilt

Kleiner Satz von Fermat: Für jedes Element $a \in \mathbb{F}_p$ ist $a^p = a$. ∎

Der französische Mathematiker PIERRE DE FERMAT (1601–1665) wurde in Beaumont-de-Lomagne geboren. Bekannt ist er heutzutage vor allem für seine 1994 von ANDREW WILES bewiesene Vermutung, wonach die Gleichung $x^n + y^n = z^n$ für $n \geq 3$ keine ganzzahlige Lösung mit $xyz \neq 0$ hat. Dieser „große" Satzes von FERMAT, von dem FERMAT lediglich in einer Randnotitz behauptete, daß er ihn beweisen könne, erklärt den Namen der obigen Aussage. Obwohl FERMAT sich sein Leben lang sehr mit Mathematik beschäftigte und wesentliche Beiträge zur Zahlentheorie, Wahrscheinlichkeitstheorie und Analysis lieferte, war er hauptberuflich Jurist und Chef der Börse von Toulouse.

Speziell für den Körper $\mathbb{F}_p$ vereinfacht sich das obige Lemma nach dem kleinen Satz von FERMAT zum folgenden

Korollar: Für ein Polynom mit Koeffizienten aus $\mathbb{F}_p$ ist

$$(a_d X^d + a_{d-1} X^{d-1} + \cdots + a_0)^p = a_d X^{dp} + a_{d-1} X^{(d-1)p} + \cdots + a_0 .$$ ∎

d) Quadratfreie Zerlegung über beliebigen Körpern

Falls ein Polynom f durch das Quadrat q^2 eines anderen teilbar ist, gibt es ein Polynom $g \in k[X]$ mit $f = q^2 g$, und nach der Produktregel ist $f' = 2qg + q^2 g' = q(2g + qg')$, d.h. q teilt auch f' und damit den ggT von f und f'.

Ist umgekehrt ein irreduzibles Polynom $q \in k[X]$ Teiler von f, etwa $f = qh$, so ist $f' = q'h + qh'$ genau dann durch q teilbar, wenn $q'h$ durch q teilbar ist. Da q irreduzibel ist, muß dann entweder q' oder h durch q teilbar sein. Wäre q ein Teiler von q', so müßte q' aus Gradgründen das Nullpolynom sein, q selbst also entweder konstant oder – über einem Körper positiver Charakteristik – eine p-te Potenz. Beides ist durch die Definition eines irreduziblen Polynoms ausgeschlossen. Somit muß dann h durch q teilbar sein und $f = qh$ durch q^2. Damit haben wir bewiesen:

Lemma: Ein irreduzibles Polynom q ist genau dann ein mindestens quadratischer Faktor von f, wenn es den ggT von f und f' teilt. ∎

Genauer: Wenn q in der Primfaktorzerlegung von f in der Potenz q^e auftritt, d.h. $f = q^e g$ mit $q \nmid g$, so ist $f' = eq^{e-1} g + q^e g'$.

Über $\mathbb{R}$ würde daraus folgen, daß q^{e-1} die höchste q-Potenz ist, die f' teilt. Da wir aber über einem beliebigen Körper arbeiten, könnte es sein, daß der erste Faktor verschwindet: Dies passiert genau dann, wenn der Exponent e durch die Charakteristik p des Grundkörpers teilbar ist. In diesem Fall ist $f' = q^e g$ mindestens durch q^e teilbar. Da f genau durch q^e teilbar ist, folgt

Lemma: Ist $f = u \prod q_i^{e_i}$ mit $u \in k^\times$ die Zerlegung eines Polynoms $f \in k[X]$ in irreduzible Faktoren, so ist der ggT von f und f' gleich $\prod q_i^{d_i}$ mit $d_i = \begin{cases} e_i - 1 & \text{falls } p \nmid e_i \\ e_i & \text{falls } p \mid e_i \end{cases}$. ∎

Nach dem Lemma ist zumindest klar, daß $h_1 = f/\operatorname{ggT}(f, f')$ ein quadratfreies Polynom ist, nämlich das Produkt aller jener Primfaktoren von f, deren Exponent nicht durch p teilbar ist. In Charakteristik null ist also $f/\operatorname{ggT}(f, f')$ einfach das Produkt der sämtlichen irreduziblen Faktoren von f. Diejenigen Faktoren, die mindestens quadratisch vorkommen, sind gleichzeitig Teiler des ggT; das Produkt g_1 der Faktoren, die genau in der ersten Potenz vorkommen, ist also $h_1/\operatorname{ggT}\big(h_1, \operatorname{ggT}(f, f')\big)$.

Falls $\operatorname{ggT}(f, f')$ kleineren Grad als f hat, können wir rekursiv weitermachen und nach derselben Methode das Produkt aller Faktoren bilden, die in $f_1 = \operatorname{ggT}(f, f')$ genau mit Exponent eins vorkommen; in f selbst sind das quadratische Faktoren. Weiter geht es mit $f_2 = \operatorname{ggT}(f_2, f_2')$, dessen Faktoren mit Exponent eins kubisch in f auftreten, *usw.*

Über einem Körper der Charakteristik null liefert diese Vorgehensweise die gesamte quadratfreie Zerlegung; in Charakteristik $p > 0$ kann es allerdings vorkommen, daß $\operatorname{ggT}(f, f') = f$ ist, nämlich wenn sich f als Polynom in X^p schreiben läßt. Für $f \in \mathbb{F}_p[X]$ ist dann, wie wir oben gesehen haben,

$$\begin{aligned} f &= a_{dp}X^{dp} + a_{(d-1)p}X^{(d-1)p} + \cdots + a_pX^p + a_0 \\ &= \left(a_{dp}X^p + a_{(d-1)p}X^{(d-1)} + \cdots + a_pX + a_0\right)^p , \end{aligned}$$

f ist dann also die p-te Potenz eines anderen Polynoms, und wir können den Algorithmus auf dieses anwenden. Im Endergebnis müssen dann natürlich alle hier gefundenen Faktoren in die p-te Potenz gehoben werden.

In anderen Körpern der Charakteristik p ist die Situation etwas komplizierter: Dort müssen wir zunächst Elemente b_i finden mit $b_i^p = a_{ip}$; dann ist

$$f = \left(b_dX^d + b_{d-1}X^{d-1} + \cdots + b_1X + b_0\right)^p .$$

Solche Elemente müssen nicht existieren, es gibt aber eine große Klasse von Körpern, in denen sie stets existieren:

Definition: Ein Körper k der Charakteristik $p > 0$ heißt vollkommen, wenn die Abbildung $k \to k$; $x \mapsto x^p$ surjektiv ist.

Beispielsweise sind die Körper $\mathbb{F}_p$ offenswichtlich vollkommen, denn dort ist $x^p = x$ für alle x. Auch alle anderen endlichen Körper sind vollkommen, denn ein endlicher Köper muß ein endlichdimensionaler Vektorraum über einem der Körper $\mathbb{F}_p$ sein, so daß seine Elementanzahl ein Potenz p^n ist. Die multiplikative Gruppe dieses Körpers hat $p^n - 1$ Elemente; daher ist $x^{p^n-1} = 1$ für alle $x \neq 0$ und $x^{p^n} = x$ für alle x. Somit ist x die p-te Potenz von $y = x^{p^{n-1}}$. Ein Beispiel eines nicht vollkommenen Körpers wäre der Körper $\mathbb{F}_p(X)$ aller rationaler Funktionen mit Koeffizienten aus $\mathbb{F}_p$, wo X offensichtlich nicht als p-te Potenz eines anderen Körperelements geschrieben werden kann.

Über einem vollkommenen Körper der Charakteristik $p > 0$ kann man also jedes Polynom, dessen Ableitung das Nullpolynom ist, als p-te Potenz eines anderen Polynoms schreiben und so, falls man die p-ten Wurzeln auch effektiv berechnen kann, den Algorithmus zur quadratfreien Zerlegung durchführen. Insbesondere gibt es keinerlei Probleme mit den Körpern $\mathbb{F}_p$, denn dort ist jedes Element seine eigene p-te Wurzel.

§3: Der Berlekamp-Algorithmus

Wir gehen aus von einem *quadratfreien* Polynom über dem Körper $\mathbb{F}_p$ mit p Elementen, d.h. $f \in \mathbb{F}_p[X]$ ist ein Produkt von *verschiedenen* irreduziblen Polynomen $f_1, \ldots, f_N$, von denen auch keine zwei assoziert sind. Durch quadratfreie Zerlegung läßt sich jedes Faktorisierungsproblem in $\mathbb{F}_p[X]$ auf diesen Fall zurückführen.

Um zu sehen, wie wir die f_i bestimmen können, nehmen wir zunächst an, sie seien bereits bekannt. Wir wählen uns dann irgendwelche Zahlen $s_1, \ldots, s_N \in \mathbb{F}_p$ und suchen ein Polynom $g \in \mathbb{F}_p[X]$ mit

$$g \equiv s_i \bmod f_i \quad \text{für alle } i = 1, \ldots, N\,.$$

Falls die s_i paarweise verschieden sind, können wir den Faktor f_i bestimmen als

$$f_i = \mathrm{ggT}(g - s_i, f)\,.$$

Nun können wir freilich nicht immer erreichen, daß die s_i alle paarweise verschieden sind: Wenn N größer als p ist, gibt es dazu einfach nicht

genügend Elemente in $\mathbb{F}_p$. In diesem Fall ist ggT$(g - s_i, f)$ das Produkt aller f_j mit $s_j = s_i$. Sofern nicht alle s_i gleich sind, führt das immerhin zu einer partiellen Faktorisierung von f, die wir dann mit einem neuen Polynom $\widetilde{g}$ zu neuen Elementen $\widetilde{s}_i$ weiter zerlegen müssen, und so weiter.

Nach dem chinesischen Restesatz ist klar, daß es zu jeder Wahl von N Elementen $s_1, \ldots, s_N$ ein Polynom g gibt mit $g \equiv s_i \bmod f_i$, denn wegen der Quadratfreiheit von f sind die f_i ja paarweise teilerfremd. Das Problem ist nur, daß wir die f_i erst berechnen wollen und g daher nicht wie im Beweis des chinesischen Restesatzes konstruieren können. Wir müssen g also auch noch anders charakterisieren.

Nach dem kleinen Satz von FERMAT ist jedes s_i gleich seiner p-ten Potenz, also ist

$$g^p \equiv s_i^p = s_i \equiv g \bmod f_i \quad \text{für } i = 1, \ldots, N\,.$$

Da f das Produkt der paarweise teilerfremden f_i ist, gilt daher auch

$$g^p \equiv g \bmod f\,.$$

Falls umgekehrt ein Polynom $g \in \mathbb{F}_p[X]$ diese Kongruenz erfüllt, so ist f ein Teiler von $g^p - g$. Letzteres Polynom können wir weiter zerlegen:

Lemma: *a)* Über einem Körper k der Charakteristik $p > 0$ ist

$$X^p - X = \prod_{j=0}^{p-1}(X - j) \quad \text{und} \quad X^{p-1} - 1 = \prod_{j=1}^{p-1}(X - j)\,.$$

b) Für jedes Polynom $g \in k[X]$ ist

$$g^p - g = \prod_{j=0}^{p-1}(g - j) \quad \text{und} \quad g^{p-1} - 1 = \prod_{j=1}^{p-1}(g - j)\,.$$

Beweis: a) Nach dem kleinen Satz von FERMAT sind alle $i \in \mathbb{F}_p$ Nullstellen des Polynoms $X^p - X$, und da ein von null verschiedenes Polynom vom Grad p nicht mehr als p Nullstellen haben kann, gibt es keine weiteren. Die Gleichheit beider Seiten folgt somit daraus, daß die führenden Koeffizienten beider Polynome eins sind.

b) Im Polynomring $k[X]$ ist, wie wir gerade gesehen haben, $X^p - X$ gleich dem Produkt aller Polynome $(X - j)$. Diese Identität, genau wie die für $X^{p-1} - 1$, bleibt natürlich erhalten, wenn man auf beiden Seiten für X irgendein Polynom aus $k[X]$ einsetzt.

■

Für ein Polynom $g \in \mathbb{F}_p[X]$ mit $g^p \equiv g \bmod f$ ist f somit ein Teiler von

$$\prod_{j=0}^{p}(g - j)\,.$$

Jeder irreduzible Faktor f_i von f muß daher genau eines der Polynome $g - j$ teilen. Somit gibt es zu jedem Faktor f_i ein Element $s_i \in \mathbb{F}_p$, so daß $g \equiv s_i \bmod p$.

Wenn wir uns auf Polynome g beschränken, deren Grad kleiner ist als der von f, so ist g durch die Zahlen s_i eindeutig bestimmt, denn nach dem chinesischen Restesatz unterscheiden sich zwei Lösungen des Systems

$$g \equiv s_i \bmod f_i \quad \text{für } i = 1, \ldots, N$$

um ein Vielfaches des Produkts der f_i, also ein Vielfaches von f.

Die Menge V aller Polynome g mit kleinerem Grad als f, für die es Elemente $s_1, \ldots, s_N \in \mathbb{F}_p$ gibt, so daß die obigen Kongruenzen erfüllt sind, ist offensichtlich ein $\mathbb{F}_p$-Vektorraum: Für eine Linearkombination zweier Polynome aus V sind solche Kongruenzen erfüllt für die entsprechenden Linearkombinationen der s_i. Die Abbildung

$$\begin{cases} V \to \mathbb{F}_p^N \\ g \mapsto (g \bmod f_1, \ldots, g \bmod f_N) \end{cases}$$

ist nach dem chinesischen Restesatz ein Isomorphismus; somit ist die Dimension von V gleich der Anzahl N irreduzibler Faktoren von f.

Wie die obige Diskussion zeigt, ist V auch der Vektorraum aller Polynome g mit kleinerem Grad als f, für die $g^p \equiv g \bmod f$ ist. In dieser Form läßt sich V berechnen: Ist $\deg f = d$, so können wir jedes $g \in V$ schreiben als

$$g = g_{d-1}X^{d-1} + a_{d-2}X^{d-2} + \cdots + g_1X + g_0$$

mit geeigneten Koeffizienten $g_i \in \mathbb{F}_p$, und

$$g^p = g_{d-1}X^{(d-1)p} + g_{d-2}X^{(d-2)p} + \cdots + g_1X^p + g_0 \,.$$

Modulo f müssen g und g^p übereinstimmen. Um dies in eine Bedingung an die Koeffizienten g_i zu übersetzen, dividieren wir die Potenzen X^{ip} für $0 \le i < d$ mit Rest durch f und erhalten Kongruenzen

$$X^{ip} \equiv \sum_{j=0}^{d-1} b_{ij}X^j \bmod f \,.$$

Dann ist

$$g^p = \sum_{i=0}^{d-1} g_iX^{ip} \equiv \sum_{i=0}^{d-1}\sum_{j=0}^{d-1} ig_ib_{ij}X^j = \sum_{j=0}^{d-1}\left(\sum_{i=0}^{d-1} b_{ij}g_i\right)X^j \bmod f \,.$$

g liegt genau dann in V, wenn $g^p \equiv g \bmod f$ ist, und wie ein Koeffizientenvergleich zeigt, ist das genau dann der Fall, wenn die Koeffizienten g_j Lösungen des folgenden homogenen linearen Gleichungssystems sind:

$$\sum_{i=0}^{d-1} b_{ij}g_i = g_j \quad \text{für } j = 0, \ldots, d-1 \,.$$

V ist also auch beschreibbar als der Lösungsraum eines Gleichungssystems. Dieser läßt sich allein auf Grund der Kenntnis von f explizit berechnen, und seine Dimension ist gleich der Anzahl N der irreduziblen Faktoren von f. Insbesondere ist V genau dann eindimensional, wenn f irreduzibel ist.

Andernfalls wählen wir irgendein Element $g \in V$ und berechnen die Polynome $\mathrm{ggT}(g - j, f)$ für alle $j \in \mathbb{F}_p$. Falls wir dabei N mal ein nichtkonstantes Polynom bekommen, haben wir f faktorisiert. Wenn wir weniger Faktoren bekommen, waren für das betrachtete Polynom g einige der Werte s_i gleich. Wir bilden eine Liste der gefundenen (und zumindest noch nicht in allen Fällen irreduziblen) Faktoren, wählen ein von g linear unabhängiges neues Polynom $h \in V$ und verfahren damit genauso. Indem wir für jedes nichtkonstante Polynom $\mathrm{ggT}(h - j, f)$ den ggT mit den in der Liste stehenden Faktoren bilden, können wir die Listenelemente weiter zerlegen. Bei jeder gefundenen Zerlegung

ersetzen wir das zerlegte Element durch seine Faktoren. Sobald die Liste N Faktoren enthält, sind wir fertig.

Falls die sämtlichen ggT$(h - j, f)$ immer noch nicht ausreichen, um N Faktoren zu produzieren, müssen wir ein neues, von g und h linear unabhängiges Element von V wählen und damit weitermachen, *usw.*

Das Verfahren muß spätestens mit dem N-ten Polynom enden, denn dann haben wir eine Basis $g_1, \ldots, g_N$ von V durchprobiert. Hätten wir dann noch nicht alle N Faktoren isoliert, müßte es (mindestens) zwei Faktoren f_i und f_j geben, so daß g mod f_i für alle Polynome g einer Basis von V gleich g mod f_j ist und damit für alle $g \in V$. Das ist aber nicht möglich, denn nach dem chinesischen Restesatz enthält V beispielsweise auch ein Element g mit g mod $f_i = 0$ und g mod $f_j = 1$.

Damit liefert uns dieser Algorithmus von BERLEKAMP zusammen mit der quadratfreien Zerlegung für jedes Polynom über $\mathbb{F}_p$ eine Zerlegung in irreduzible Faktoren. Mit einigen offensichtlichen Modifikationen schafft er dasselbe auch für Polynome über jedem der in dieser Vorlesung nicht behandelten anderen endlichen Körpern, allerdings ist für solche Körper gelegentlich ein alternativer Algorithmus von HARALD NIEDERREITER effizienter.

ELWYN BERLEKAMP wurde 1940 in Dover, Ohio geboren. Er studierte Elektrotechnik am MIT, wo er 1964 mit einer Arbeit aus dem Gebiet der Kodierungstheorie promovierte. Seine anschließenden Arbeiten und auch Positionen sowohl in der Wirtschaft als auch an Universitäten bewegen sich im Grenzgebiet zwischen Mathematik, Elektrotechnik und Informatik; einen gewissen Schwerpunkt bilden Bücher und Zeitschriftenartikel über die Mathematik von Spielen sowie Arbeiten zur Informationstheorie. 2006 emeritierte er als Mathematikprofessor in Berkeley und starb 2019. Seine home page `math.berkeley.edu/~berlek/` existiert noch.

Als Beispiel wollen wir $f = X^6 + 2X^5 + 4X^4 + X^3 - X^2 - X - 1$ aus $\mathbb{F}_7[X]$ faktorisieren. Wir setzen ein Polynom vom Grad fünf mit unbestimmten Koeffizienten an:

$$g = g_5X^5 + g_4X^4 + g_3X^3 + g_2X^2 + g_1X + g_0$$

Die siebte Potenz davon ist

$$g^7 = g_5X^{35} + g_4X^{28} + g_3X^{21} + g_2X^{14} + g_1X^7 + g_0\,.$$

Um sie modulo f auszudrücken, müssen wir die Divisionsreste h_i bei der Division von X^{7i} durch f berechnen. Wir erhalten

$$\begin{aligned}
X^7 &\equiv h_1 = 3X^3 + 6X^2 + 6X + 5 \bmod f\\
X^{14} &\equiv h_2 = 4X^5 + X^4 + 2X^3 + 6X + 6 \bmod f\\
X^{21} &\equiv h_3 = X^5 + 4X^4 + 6X^3 + 6X^2 + 2X + 5 \bmod f\\
X^{28} &\equiv h_4 = X^5 + 6X^4 + 5X^3 + X^2 + X + 2 \bmod f\\
X^{35} &\equiv h_5 = 2X^5 + 4X^4 + 6X^3 + 4X^2 + 2X + 5 \bmod f
\end{aligned}$$

Damit ist

$$g^7 \equiv g_5h_5 + g_4h_4 + g_3h_3 + g_2h_2 + g_1h_1 + g_0 \bmod f\,.$$

Sortiert nach X-Potenzen ist die rechte Seite gleich

$$\begin{aligned}
&(4g_2 + g_3 + g_4 + 2g_5)X^5 + (g_2 + 4g_3 + 6g_4 + 4g_5)X^4\\
&+ (3g_1 + 2g_2 + 6g_3 + 5g_4 + 6g_5)X^3 + (6g_1 + 6g_3 + g_4 + 4g_5)X^2\\
&+ (6g_1 + 6g_2 + 2g_3 + g_4 + 2g_5)X + (g_0 + 5g_1 + 6g_2 + 5g_3 + 2g_4 + 5g_5)\,.
\end{aligned}$$

Wenn $g^7 \equiv g \bmod f$ ist, muß das gleich g sein, was auf ein homogenes lineares Gleichungssystem für die sechs Variablen $g_i \in \mathbb{F}_7$ führt:

$$\begin{aligned}
4g_2 + g_3 + g_4 + 2g_5 &= g_5\\
g_2 + 4g_3 + 6g_4 + 4g_5 &= g_4\\
3g_1 + 2g_2 + 6g_3 + 5g_4 + 6g_5 &= g_3\\
6g_1 + 6g_3 + g_4 + 4g_5 &= g_2\\
6g_1 + 6g_2 + 2g_3 + g_4 + 2g_5 &= g_1\\
g_0 + 5g_1 + 6g_2 + 5g_3 + 2g_4 + 5g_5 &= g_0
\end{aligned}$$

Wie der GAUSS-Algorithmus zeigt, ist die Lösungsmenge zweidimensional. Setzen wir $g_0 = \lambda$ und $g_5 = \mu$, erhalten wir die Lösungen

$$g_0 = \lambda, \quad g_1 = 6\mu, \quad g_2 = 3\mu, \quad g_3 = 5\mu, \quad g_4 = 3\mu, \quad g_5 = \mu\,.$$

Da der Lösungsraum zweidimensional ist, hat f zwei irreduzible Faktoren. Wir können uns eine Basis des Lösungsraums verschaffen, indem wir für das erste Basispolynom $\lambda = 1$ und $\mu = 0$ setzen und für das zweite $\lambda = 0$ und $\mu = 1$. Die erste Substitution führt zum nutzlosen konstanten Polynom 1, die zweite zu

$$g = X^5 + 3X^4 + 5X^3 + 3X^2 + 6X\ .$$

Auf der Suche nach Faktoren von f müssen wir für alle $j \in \mathbb{F}_7$ die größten gemeinsamen Teiler von $g - j$ und f berechnen. Für $j = 0$ erhalten wir den Faktor $X^3 + 5X + 2$ und für $j = 2$ das Polynom $X^3 + 2X^2 + 6X + 3$. In allen anderen Fällen sind f und $g - j$ teilerfremd. Somit ist $f = (X^3 + 5X + 2)(X^3 + 2X^2 + 6X + 3)$ die Zerlegung von f in irreduzible Faktoren in $\mathbb{F}_7[X]$.

§4: Faktorisierung über den ganzen Zahlen und über endlichen Körpern

Wie bei der Berechnung des ggT zweier Polynome wollen wir auch bei der Faktorisierung den Umweg über endliche Körper gehen, um das Problem für Polynome über $\mathbb{Z}$ zu lösen. Allerdings kann es hier häufiger passieren, daß sich Ergebnisse über $\mathbb{F}_p$ deutlich unterscheiden von denen über $\mathbb{Z}$:

Zunächst einmal muß ein quadratfreies Polynom aus $\mathbb{Z}[X]$ modulo p nicht quadratfrei bleiben: $f = (X + 10)(X - 20)$ etwa ist modulo zwei oder fünf gleich X^2 und modulo drei $(X + 1)^2$. Dieses Problem tritt allerdings nur bei endlich vielen Primzahlen auf und kann vermieden werden: Ist $f \in \mathbb{Z}[X]$ quadratfrei, seine Reduktion $f^{(p)} \in \mathbb{F}_p[X]$ aber nicht, so haben $f^{(p)}$ und seine Ableitung einen gemeinsamen Faktor, ihre Resultante verschwindet also. Sofern der Grad von f nicht durch p teilbar ist, ist diese Resultante die Reduktion modulo p der Resultante von f und f', so daß p ein Teiler der über $\mathbb{Z}$ berechneten Resultante sein muß, und das läßt sich leicht nachprüfen. Dazu müssen wir zwar eine Resultante berechnen, was wir im vorigen Kapitel aus Effizienzgründen vermieden hatten, aber wie wir bald sehen werden, ist das Problem der Faktorisierung deutlich komplexer als der EUKLIDische Algorithmus, so

daß hier der Aufwand für die Resultantenberechnung nicht weiter ins Gewicht fällt.

Tatsächlich betrachtet man in der Algebra meist nicht die Resultante von f und f', sondern die sogenannte *Diskriminante*

$$D(f) = \frac{(-1)^{\frac{1}{2}d(d-1)}}{a_n} \operatorname{Res}_X(f, f') \quad \text{mit} \quad d = \deg f\,,$$

deren algebraische Eigenschaften etwas besser sind. Für praktische Rechnungen ist der Unterschied hier aber unbedeutend, denn wie man der SYLVESTER-Matrix von f und f' leicht ansieht, ist die Resultante eines nichtkonstanten Polynoms durch eine höhere Potenz des führenden Koeffizienten a_n teilbar als nur die erste; die Primteiler von Resultante und Diskriminante sind also dieselben.

Vermeidet man diese und die Primteiler des Grades von f, bleibt f somit auch modulo p quadratfrei. Die Zerlegungen von f und $f^{(p)}$ in irreduzible Faktoren in $\mathbb{Z}[X]$ und $\mathbb{F}_p[X]$ können sich jedoch deutlich unterscheiden:

Betrachten wir dazu als erstes Beispiel das Polynom $X^2 + 1$ aus $\mathbb{Z}[X]$. Es ist irreduzibel, da eine Zerlegung die Form $(X - a)(X + a)$ haben müßte mit $a \in \mathbb{Z}$, und in $\mathbb{Z}$ gibt es kein Element a mit $a^2 = -1$.

Auch über dem Körper $\mathbb{F}_p$ muß eine eventuelle Faktorisierung die Form $(X - a)(X + a)$ haben mit $a^2 = -1$; wir müssen uns also überlegen, wann das der Fall ist. Die elementare Zahlentheorie sagt uns:

Lemma: Genau dann gibt es im endlichen Körper $\mathbb{F}_p$ ein Element a mit $a^2 = -1$, wenn $p = 2$ oder $p \equiv 1 \bmod 4$ ist.

Beweis: Für $p = 2$ ist natürlich $1^2 = 1 = -1$ die Lösung. Für ungerade $p \equiv 1 \bmod 4$ schreiben wir $p = 4k + 1$. Nach dem kleinen Satz von FERMAT und der dritten binomischen Formel ist für alle $x \in \mathbb{F}_p^\times$

$$x^{p-1} - 1 = x^{4k} - 1 = (x^{2k} + 1)(x^{2k} - 1) = 0\,,$$

das Polynom $X^{p-1} - 1$ hat somit $p - 1 = 4k$ Nullstellen und zerfällt daher über $\mathbb{F}_p$ in Linearfaktoren. Damit gilt dasselbe für die beiden Faktoren

$X^{2k} \pm 1$. Insbesondere gibt es daher ein $x \in \mathbb{F}_p$ mit $x^{2k} + 1 = 0$. Für $a = x^k$ ist dann $a^2 = x^{2k} = -1$.

Ist $p \equiv 3 \bmod 4$ und $a^2 = -1$ für ein $a \in \mathbb{F}_p$, so ist $a^4 = 1$. Außerdem ist nach dem kleinen Satz von FERMAT $a^{p-1} = 1$. Wegen $p \equiv 3 \bmod 4$ gibt es ein $k \in \mathbb{N}_0$, so daß $p - 1 = 4k + 2$ ist, also ist $a^{p-1} = a^2$ und damit $1 = -1$. Dieser Widerspruch zeigt, daß es für $p \equiv 3 \bmod 4$ keine Elemente mit Quadrat -1 in $\mathbb{F}_p$ geben kann.

■

$X^2 + 1$ ist also genau dann irreduzibel über $\mathbb{F}_p$, wenn $p \equiv 3 \bmod 4$ ist. In allen anderen Fällen zerfällt das Polynom in zwei Linearfaktoren. Nach einem berühmten Satz von DIRICHLET über Primzahlen in arithmetischen Progressionen bleibt $X^2 + 1$ damit nur modulo der Hälfte aller Primzahlen irreduzibel; insbesondere gibt es also unendlich viele Primzahlen, modulo derer das Problem schlechte Reduktion hat.

Noch schlimmer ist es bei $X^4 + 1$: Auch dieses Polynom ist irreduzibel über $\mathbb{Z}$: Da seine Nullstellen $\frac{1}{2}\sqrt{2}(\pm 1 \pm i)$ nicht in $\mathbb{Z}$ liegen, gibt es keinen linearen Faktor, und wäre

$$\begin{aligned} X^4 + 1 &= (X^2 + aX + b)(X^2 + cX + d) \\ &= X^4 + (a + c)X^3 + (b + d + ac)X^2 + (ad + bc)X + bd \end{aligned}$$

eine Zerlegung in quadratische Faktoren, so zeigen die Koeffizienten von X^3 und der konstante Term, daß $c = -a$ und $b = d = \pm 1$ sein müßte. Die Produkte

$$(X^2 + aX + 1)(X^2 - aX + 1) = X^4 + (2 - a^2)X^2 + 1$$

und

$$(X^2 + aX - 1)(X^2 - aX - 1) = X^4 - (2 + a^2)X^2 + 1$$

zeigen aber, daß beides nur für $a^2 = \pm 2$ zu einer Faktorisierung führen könnte, was in $\mathbb{Z}$ nicht erfüllbar ist.

In den Körpern $\mathbb{F}_p$ dagegen kann es sehr wohl Elemente geben, deren Quadrat ± 2 ist, und dann zeigen die obigen Formeln, daß $X^4 + 1$ dort in ein Produkt zweier quadratischer Polynome zerlegt werden kann. Auch

wenn es ein Element $a \in \mathbb{F}_p$ gibt mit $a^2 = -1$, können wir $X^4 + 1$ als Produkt schreiben, nämlich als $X^4 + 1 = (X^2 + a)(X^2 - a)$.

Somit ist $X^4 + 1$ über dem Körper $\mathbb{F}_p$ zumindest dann reduzibel, wenn dort wenigstens eines der drei Elemente -1 und ± 2 ein Quadrat ist. Um zu sehen, daß $X^4 + 1$ über jedem dieser Körper zerfällt, müssen wir uns daher überlegen, daß in keinem der Körper $\mathbb{F}_p$ alle drei Elemente *keine* Quadrate sind. Da $-2 = -1 \cdot 2$ ist, folgt dies aus

Lemma: Sind im Körper $\mathbb{F}_p$ die zwei Elemente a, b beide nicht als Quadrate darstellbar, so ist ihr Produkt ab ein Quadrat.

Beweis: Für $p = 2$ ist jedes Element ein Quadrat und nichts zu beweisen.

Ansonsten betrachten wir die Abbildung $\varphi\colon \mathbb{F}_p^\times \to \mathbb{F}_p^\times$, die jedes von Null verschiedene Element von $\mathbb{F}_p$ auf sein Quadrat abbildet. Für zwei Elemente $x, y \in \mathbb{F}_p^\times$ ist offensichtlich $\varphi(x) = \varphi(y)$ genau dann, wenn $x = \pm y$ ist. Daher besteht das Bild von φ aus $\frac{1}{2}(p-1)$ Elementen, und genau die Hälfte der Elemente von $\mathbb{F}_p^\times$ sind Quadrate. Ist a keines, so ist auch ax^2 für kein $x \in \mathbb{F}_p^\times$ ein Quadrat, denn wäre $ax^2 = y^2$, sonst wäre auch $a = y^2 x^{-2} = (y/x)^2$ ein Quadrat.

Da es $\frac{1}{2}(p-1)$ Quadrate und genauso viele Nichtquadrate gibt, läßt sich somit jedes Nichtquadrat b als $b = ax^2$ schreiben mit einem geeigneten Element $x \in \mathbb{F}_p$. Damit ist $ab = a \cdot ax^2 = (ax)^2$ ein Quadrat. ∎

Die Situation ist also deutlich schlechter als im Fall des EUKLIDischen Algorithmus, wo wir sicher sein konnten, daß es höchstens endlich viele Primzahlen gibt, modulo derer das Problem schlechte Reduktion hat: Das Problem der Faktorisierung des Polynoms $X^4 + 1$ hat, wie wir gerade gesehen haben, modulo *jeder* Primzahl schlechte Reduktion, und auch für $X^2 + 1$ gibt es unendlich viele schlechte Primzahlen.

Aus diesem Grund empfiehlt sich für die Faktorisierung definitiv kein Ansatz mit dem chinesischen Restesatz: Wenn wir die Faktorisierung modulo verschiedener Primzahlen durchführen, können wir praktisch sicher sein, daß es darunter auch schlechte gibt, und meist werden auch

die Ergebnisse modulo verschiedener Primzahlen entweder nicht zusammenpassen, oder aber wir haben mehrere Faktoren gleichen Grades, von denen wir nicht wissen, welche wir via chinesischen Restesatz miteinander kombinieren sollen. Es hat daher keinen Zweck, zufällig Primzahlen zu wählen und dann eine Rückfallstrategie für schlechte Primzahlen zu entwickeln.

Der modulare Ansatz verfolgt deshalb im Falle der Faktorisierung eine andere Strategie als die mit dem chinesischen Restesatz: Wir beschränken uns auf eine einzige Primzahl p und heben die Faktorisierung modulo p nach dem HENSELschen Lemma hoch zu einer Faktorisierung modulo einer hinreichend hohen p-Potenz.

§6: Der Algorithmus von Zassenhaus

HANS JULIUS ZASSENHAUS wurde 1912 in Koblenz geboren, ging aber in Hamburg zur Schule und zur Universität. Er promovierte 1934 über Permutationsgruppen; seine Habilitation 1940 handelte von LIE-Ringen in positiver Charakteristik. Da er nicht der NSdAP beitreten wollte, arbeitete er während des Krieges als Meteorologe bei der Marine. Nach dem Krieg war er von 1949 bis 1959 Professor in Montréal, dann fünf Jahre lang in Notre Dame und schließlich bis zu seiner Emeritierung an der Ohio State University in Columbus. Dort starb er 1991. Bekannt ist er vor allem für seine Arbeiten zur Gruppentheorie und zur algorithmischen Zahlentheorie.

Aus Kapitel 3, §3, wissen wir, daß für einen Teiler $g \in \mathbb{C}[X]$ vom Grad e eines Polynoms $f \in \mathbb{C}[X]$ vom Grad d gilt:

$$H(g) \leq \binom{e}{[e/2]} \left| \frac{b_e}{a_d} \right| \|f\|_2 \ ,$$

wobei a_d und b_e die führenden Koeffizienten von f und g sind. Für $g, f \in \mathbb{Z}[X]$ muß b_e ein Teiler von a_d sein, der Quotient b_e/a_d hat also höchstens den Betrag eins. Der Grad e eines Teilers kann höchstens gleich dem Grad d von f sein, also ist für jeden Teiler $g \in \mathbb{Z}[X]$ von $f \in \mathbb{Z}[X]$

$$H(g) \leq \binom{d}{[d/2]} \|f\|_2 \ .$$

Nach ZASSENHAUS gehen wir zur Faktorisierung eines Polynoms f aus $\mathbb{Z}[X]$ oder $\mathbb{Q}[X]$ folgendermaßen vor:

Erster Schritt: Berechne die quadratfreie Zerlegung von f und ersetze die quadratfreien Faktoren durch ihre primitiven Anteile g_i. Dann gibt es eine Konstante c, so daß $f = c\prod_{i=1}^{r} g_i^i$ ist. Falls f in $\mathbb{Z}[X]$ liegt, ist c eine ganze Zahl. Für eine Faktorisierung in $\mathbb{Z}[X]$ muß auch c in seine Primfaktoren zerlegt werden; für eine Faktorisierung in $\mathbb{Q}[X]$ kann c als Einheit aus $\mathbb{Q}^\times$ stehen bleiben. Die folgenden Schritte werden einzeln auf jedes der g_i angewandt, danach werden die Ergebnisse zusammengesetzt zur Faktorisierung von f. Für das Folgende sei g eines der g_i.

Zweiter Schritt: Wir setzen $L = \binom{\deg g}{[\frac{1}{2}\deg g]} \|g\|_2$, $M = 2L+1$ und wählen eine Primzahl p, die weder den führenden Koeffizienten noch die Diskriminante noch den Grad von g teilt. Dann ist auch das Polynom $g^{(p)} = g \bmod p \in \mathbb{F}_p[X]$ quadratfrei.

Dritter Schritt: Wir faktorisieren $g^{(p)}$ nach BERLEKAMP in $\mathbb{F}_p[X]$.

Vierter Schritt: Die Faktorisierung wird nach dem HENSELschen Lemma hochgehoben zu einer Faktorisierung modulo p^n für eine natürliche Zahl n mit $p^n \geq M$.

Fünfter Schritt: Setze $m = 1$ und teste für jeden der gefundenen Faktoren, ob er ein Teiler von g ist. Falls ja, kommt er in die Liste $\mathcal{L}_1$ der Faktoren von g, andernfalls in eine Liste $\mathcal{L}_2$.

Sechster Schritt: Falls die Liste $\mathcal{L}_2$ keine Einträge hat, endet der Algorithmus und g ist das Produkt der Faktoren aus $\mathcal{L}_1$. Andernfalls setzen wir $m = m + 1$ und testen für jedes Produkt aus m verschiedenen Polynomen aus $\mathcal{L}_2$, ob ihr Produkt modulo p^n (mit Koeffizienten vom Betrag höchstens L) ein Teiler von g ist. Falls ja, entfernen wir die m Faktoren aus $\mathcal{L}_2$ und fügen ihr Produkt in die Liste $\mathcal{L}_1$ ein. Wiederhole diesen Schritt.

Auch wenn der sechste Schritt wie eine Endlosschleife aussieht, endet der Algorithmus natürlich nach endlich vielen Schritten, denn $\mathcal{L}_2$ ist eine endliche Liste, und spätestens das Produkt aller Elemente aus $\mathcal{L}_2$ muß Teiler von g sein, da sein Produkt mit dem Produkt aller Elemente

von $\mathcal{L}_1$ gleich g ist. Tatsächlich kann man schon abbrechen, wenn die betrachteten Faktoren einen größeren Grad als $\frac{1}{2}\deg g$ haben, denn falls f reduzibel ist, gibt es einen Faktor, der höchstens diesen Grad hat.

§7: Swinnerton-Dyer Polynome

Der potentiell aufwendigste Schritt des obigen Algorithmus ist der letzte: Vor allem, wenn wir auch Produkte von mehr als zwei Faktoren betrachten müssen, kann dieser sehr teuer werden. Ein Beispiel dafür bieten die sogenannten SWINNERTON-DYER-Polynome: Zu n paarweise verschiedenen Primzahlen $p_1, \ldots, p_n$ gibt es genau ein Polynom f vom Grad 2^n mit führendem Koeffizienten eins, dessen Nullstellen genau die 2^n Zahlen

$$\pm\sqrt{p_1} \pm \sqrt{p_2} \pm \cdots \pm \sqrt{p_n}$$

sind. Dieses Polynom hat folgende Eigenschaften:

1. Es hat ganzzahlige Koeffizienten.
2. Es ist irreduzibel über $\mathbb{Z}$.
3. Modulo jeder Primzahl p zerfällt es in Faktoren vom Grad höchstens zwei.

Beweisen läßt sich das am besten mit Methoden der abstrakten Algebra, wie sie in jeder Vorlesung *Algebra I* präsentiert werden. Da hier keine Algebra I vorausgesetzt wird, sei nur kurz die Idee angedeutet: Um den kleinsten Teilkörper von $\mathbb{C}$ zu konstruieren, in dem alle Nullstellen von f liegen, können wir folgendermaßen vorgehen: Wir konstruieren als erstes einen Körper K_1, der $\sqrt{p_1}$ enthält. Das ist einfach: Der Vektorraum $K_1 = \mathbb{Q} \oplus \mathbb{Q}\sqrt{p_1}$ ist offensichtlich so ein Körper. Als nächstes konstruieren wir einen Körper K_2, der sowohl K_1 als auch $\sqrt{p_2}$ enthält. Dazu können wir genauso vorgehen: Wir betrachten einfach den zweidimensionalen K_1-Vektorraum $K_2 = K_1 \oplus K_1\sqrt{p_2}$. Als Vektorraum über $\mathbb{Q}$ ist K_2 natürlich vierdimensional. Weiter geht es mit $K_3 = K_2 \oplus K_2\sqrt{p_3}$ *usw.* bis $K_n = K_{n-1} \oplus K_{n-1}\sqrt{p_n}$. Als $\mathbb{Q}$-Vektorraum hat dieser Körper die Dimension 2^n.

Offensichtlich ist K_n der kleinste Erweiterungskörper von $\mathbb{Q}$, der alle Quadratwurzeln der p_i enthält. Er enthält natürlich auch alle der oben

postulierten Nullstellen, und umgekehrt muß ein Körper, der diese enthält, auch alle $\sqrt{p_i}$ enthalten, denn $2\sqrt{p_i}$ läßt sich als Summe zweier solcher Nullstellen schreiben. K_n ist also auch der kleinste Körper, der alle diese Nullstellen enthält, der sogenannte *Zerfällungskörper* des Polynoms.

In der Algebra ordnet man einem solchen Zerfällungskörper die Gruppe seiner Automorphismen zu, die sogenannte GALOIS-Gruppe. Nach einen Satz aus der GALOIS-Theorie ist ihre Ordnung gleich der Vektorraumdimension des Körpers über $\mathbb{Q}$, hier also 2^n. Da sie offensichtlich die Abbildungen $\sqrt{p_i} \mapsto -\sqrt{p_i}$ enthält, ist sie die von diesen Automorphismen erzeugte Gruppe. Sie läßt die Nullstellenmenge von f (als Menge) fest, also nach dem Wurzelsatz von VIÈTE auch die Koeffizienten. Somit hat f rationale Koeffizienten, und da alle Nullstellen ganz sind (im Sinne der algebraischen Zahlentheorie), liegen diese Koeffizienten sogar in $\mathbb{Z}$. Außerdem operiert die GALOIS-Gruppe transitiv auf der Nullstellenmenge von f; also ist f irreduzibel in $\mathbb{Q}[X]$ und damit auch $\mathbb{Z}[X]$.

Betrachten wir f modulo einer Primzahl p, so können wir die analoge Konstruktion durchführen ausgehend vom Körper $\mathbb{F}_p$ anstelle von $\mathbb{Q}$. Während wir aber im Falle der rationalen Zahlen sicher sein konnten, daß $\sqrt{p_i}$ nicht bereits im Körper K_{i-1} liegt, ist dies hier nicht mehr der Fall: Für ungerades p gibt es $\frac{1}{2}(p+1)$ Quadrate in $\mathbb{F}_p$; dazu könnte auch p_i gehören. Falls nicht, ist $K = \mathbb{F}_p \oplus \mathbb{F}_p\sqrt{p_i}$ ein Körper mit p^2 Elementen. Wie man in der Algebra lernt, gibt es aber bis auf Isomorphie nur einen solchen Körper; K enthält daher die Quadratwurzeln *aller* Elemente von $\mathbb{F}_p$ und somit *alle* Nullstellen von f mod p. Spätestens über K zerfällt f mod p also in Linearfaktoren, und da alle Koeffizienten in $\mathbb{F}_p$ liegen, lassen sich je zwei Linearfaktoren, die nicht in $\mathbb{F}_p[X]$ liegen, zu einem quadratischen Faktor aus $\mathbb{F}_p[X]$ zusammenfassen. Somit hat f mod p höchstens quadratische Faktoren.

(Für eine ausführlichere und etwas elementarere Darstellung siehe etwa §6.3.2 in MICHAEL KAPLAN: Computeralgebra, *Springer,* 2005.)

Falls wir f nach dem oben angegebenen Algorithmus faktorisieren, erhalten wir daher modulo *jeder* Primzahl p mindestens 2^{n-1} Faktoren. Diese lassen sich über das HENSELsche Lemma liften zu Fak-

toren über $\mathbb{Z}$, und wir müssen alle Kombinationen aus mindestens 2^{n-2} Faktoren ausprobieren bis wir erkennen, daß f irreduzibel ist, also mindestens $2^{2^{n-2}}$ Möglichkeiten. Für $n = 10$ etwa ist f ein Polynom vom Grad 1024, dessen Manipulation durchaus im Rahmen der Möglichkeiten eines heutigen Computeralgebrasystems liegt. Das Ausprobieren von $2^{256} \approx 10^{77}$ Möglichkeiten überfordert aber selbst heutige Supercomputer oder parallel arbeitende Cluster aus Millionen von Computern ganz gewaltig: Der heutige Sicherheitsstandard der Kryptographie geht davon aus, daß niemand in der Lage ist, 2^{128} Rechenoperationen in realistischer Zeit (d.h. in wenigen Jahren) auszuführen.

SIR HENRY PETER FRANCIS SWINNERTON-DYER, 16th Baronet, wurde 1927 geboren. Er studierte und lehrte an der Universität Cambridge, wo er unter anderem Dean des Trinity College, Master von St. Catherine College und Vizekanzler der Universität war. Er starb am 26. Dezember 2019. Obwohl er hauptsächlich für seine Beiträge zur Zahlentheorie bekannt ist, beschäftigte er sich zunächst mit nichtlinearen Differentialgleichungen. Am bekanntesten ist er durch die Vermutung von BIRCH und SWINNERTON-DYER über den Zusammenhang zwischen der Arithmetik einer elliptischen Kurve und analytischen Eigenschaften ihrer ζ-Funktion.

§8: Faktoren und Gittervektoren

In diesem Paragraphen soll eine Methode vorgestellt werden, die für Polynome einer Veränderlichen über $\mathbb{Z}$ (oder $\mathbb{Q}$), aber leider auch nur für diese, eine Alternative zum stumpfsinnigen Ausprobieren im sechsten Schritt des Algorithmus von ZASSENHAUS bietet.

Sie wurde 1982 vorgestellt in

A.K. LENSTRA, H.W. LENSTRA, L. LOVÁSZ: Factoring Polynomials with Rational Coefficients, *Math. Ann.* **261** (1982), 515–534

und wird nach den Initialen der drei Autoren kurz als LLL bezeichnet.

ARJEN K. LENSTRA wurde 1956 in Groningen geboren und studierte Mathematik an der Universität Amsterdam, wo er 1984 über das Thema Faktorisierung von Polynomen promovierte. Danach arbeitete er zunächst als Gastprofessor an der Informatikfakultät der Universität von Chicago, dann ab 1989 in einem Forschungszentrum von Bellcore. 1996 wurde er Vizepräsident am Corporate Technology Office der City Bank in New York, von 2000 bis 2006 auch Teilzeitprofessor an der Technischen Universität Eindhoven. 2004 wechselte er von der City Bank zu Lucent Technology, den ehemaligen Bell Labs. Seit

2006 ist er (inzwischen emeritierter) Professor für Kryptologie an ders Eidgenössischen Technischen Hochschule in Lausanne. Seine Arbeiten befassen sich mit der Faktorisierung von Zahlen und Polynomen sowie mit kryptographischen Verfahren und Attacken.

Sein Bruder HENDRIK W. LENSTRA wurde 1949 geboren. Auch er studierte an der Universität Amsterdam, wo er 1977 bei dem Algebraischen Geometer und Zahlentheoretiker FRANS OORT über Zahlkörper mit EUKLIDischem Algorithmus promovierte und 1978 Professor wurde. 1987 wechselte er nach Berkeley; von 1998 bis zu seiner Emeritierung in Berkeley lehrte er sowohl in Berkeley als auch an der Universität Leiden, danach bis zu seiner dortigen Emeritierung nur noch in Leiden. Seine Arbeiten beschäftigen sich hauptsächlich mit der algorithmischen Seite der Zahlentheories. Außer für den LLL-Algorithmus ist er vor allem bekannt für seinen Algorithmus zur Faktorisierung ganzer Zahlen mit elliptischen Kurven.

LÁSZLÓ LOVÁSZ wurde 1948 in Budapest geboren und promovierte 1971 an der dortigen Universität. Nach kürzeren Aufenthalten an verschiedenen ungarischen und ausländischen Universitäten (darunter 1984/85 in Bonn) ging er 1993 nach Yale, wo er bis 2000 eine Professur hatte. Von 1999 bis 2006 war er Senior Researcher bei Microsoft Research. Seit 2006 ist er Direktor des mathematischen Instituts der Eötvös Loránd Universität in Budapest. Für die Wahlperiode 2007–2010 war er auch Präsident der Internationalen Mathematikervereinigung IMU. 2021 erhielt er den ABEL-Preis für Mathematik, den die norwegische Regierung 2002 zum 200. Geburtstag von NIELS HENRIK ABEL stiftete und der seit 2003 in ähnlicher Form und Dotierung wie die Nobelpreise im Beisein des norwegischen Königs vergeben wird.

Der LLL-Algorithmus berechnet zu jedem irreduzibhlen Faktor h von $f^{(p)}$ in $\mathbb{F}_p[X]$ direkt einen Faktor von f in $\mathbb{Z}[X]$, der modulo p durch h teilbar ist, ohne Berücksichtigung der übrigen Faktoren von $f^{(p)}$. Ausgangspunkt ist das folgende

Lemma: p sei eine Primzahl, k eine beliebige natürliche Zahl. Außerdem sei $f \in \mathbb{Z}[X]$ ein Polynom vom Grad d und $h \in \mathbb{Z}[X]$ eines vom Grad e mit folgenden Eigenschaften:

1.) h hat führenden Koeffizienten eins
2.) $h \bmod p^k$ ist in $(\mathbb{Z}/p^k)[X]$ ein Teiler von $f \bmod p^k$
3.) $h \bmod p$ ist irreduzibel in $\mathbb{F}_p[X]$
4.) $(h \bmod p)^2$ ist kein Teiler von $f \bmod p$ in $\mathbb{F}_p[X]$.

Dann gilt:

a) f hat in $\mathbb{Z}[X]$ einen irreduziblen Faktor h_0, der modulo p ein Vielfaches von $h \bmod p$ ist.

b) h_0 ist bis aufs Vorzeichen eindeutig bestimmt.

c) Für einen beliebigen Teiler g von f in $\mathbb{Z}[X]$ sind folgende Aussagen äquivalent:

(i) $h \bmod p$ teilt $g \bmod p$ in $\mathbb{F}_p[X]$

(ii) $h \bmod p^k$ teilt $g \bmod p^k$ in $(\mathbb{Z}/p^k)[X]$

(iii) h_0 teilt g in $\mathbb{Z}[X]$.

Beweis: Die irreduziblen Faktoren h_i von f in $\mathbb{Z}[X]$ können modulo p in $\mathbb{F}_p[X]$ eventuell weiter zerlegt werden. Da $(h \bmod p)^2$ kein Teiler von $f \bmod p$ ist, teilt $h \bmod p$ genau eines der $h_i \bmod p$; wir setzen $h_0 = h_i$. Da irreduzible Faktoren in $\mathbb{Z}[X]$ bis aufs Vorzeichen eindeutig bestimmt sind, ist auch *b)* klar. In *c)* folgt *(i)* sofort aus *(ii)* wie auch aus *(iii)*; zu zeigen sind die Umkehrungen *(i)* $\Rightarrow$ *(iii)* und *(i)* $\Rightarrow$ *(ii)*.

Sei also $h \bmod p$ ein Teiler von $g \bmod p$ und $f = gq$. Da $(h \bmod p)^2$ kein Teiler von $f \bmod p$ ist, kann $h \bmod p$ kein Teiler von $q \bmod p$ sein, also auch h_0 kein Teiler von q. Somit muß h_0 als irreduzibler Teiler von f ein Teiler von g sein und *(iii)* ist bewiesen.

Zum Beweis von *(ii)* beachten wir, daß $h \bmod p$ und $q \bmod p$ teilerfremd sind. Der erweiterte EUKLIDische Algorithmus liefert uns daher eine Darstellung der Eins als Linearkombination dieser beiden Polynome in $\mathbb{F}_p[X]$. Wir liften die Koeffizienten nach $\mathbb{Z}[X]$ und haben dann Polynome $a, b \in \mathbb{Z}[X]$, für die $ah + bq \equiv 1 \bmod p$ ist. Es gibt also ein Polynom $c \in \mathbb{Z}[X]$, für das $ah + bq = 1 - pc$ ist. Da

$$\big(1 - pc\big)\big(1 + pc + (pc)^2 + \cdots + (pc)^{k-1}\big) = 1 - (pc)^k$$

ist, erhalten wir durch Multiplikation dieser Gleichung mit dem zweiten Faktor eine neue Gleichung der Form

$$\tilde{a}h + \tilde{b}q = 1 - (pc)^k \,.$$

Multiplizieren wir diese noch mit g, so folgt

$$\tilde{a}g \cdot h + \tilde{b}q \cdot g = \tilde{a}gh + \tilde{b}f \equiv g \bmod p^k \,.$$

Hier ist die linke Seite modulo p^k durch h teilbar, also auch die rechte Seite g, womit *(ii)* bewiesen wäre.

■

Der sechste Schritt des Algorithmus von ZASSENHAUS kann so interpretiert werden, daß er zu jedem irreduziblen Faktor $h \in \mathbb{F}_p[X]$ von $f^{(p)}$ den zugehörigen Faktor $h_0 \in \mathbb{Z}[X]$ von f bestimmt, indem er nötigenfalls alle Kombinationen aus h und den anderen Faktoren von f mod p durchprobiert. Der Algorithmus von LENSTRA, LENSTRA und LOVÁSZ konstruiert h_0 direkt und ohne Kenntnis der anderen Faktoren, indem er den Vektor der Koeffizienten von h_0 als einen „kurzen“ Gittervektor identifiziert.

Wir fixieren dazu eine natürliche Zahl $m \geq e = \deg h$ und betrachten die Menge Λ aller Polynome aus $\mathbb{Z}[X]$ vom Grad höchstens m, die modulo p^k durch h mod p^k teilbar sind. Λ ist eine Teilmenge des $(m+1)$-dimensionalen $\mathbb{R}$-Vektorraums aller Polynome vom Grad höchstens m, den wir über die Basis $1, X, \ldots, X^m$ mit $\mathbb{R}^{m+1}$ identifizieren. Dabei ist die L^2-Norm $\|f\|_2$ eines Polynoms aus V gleich der üblichen EUKLIDischen Länge $|v|$ seines Koeffizientenvektors $v \in \mathbb{R}^{m+1}$.

Definition: Eine Teilmenge $\Gamma \subset \mathbb{R}^{m+1}$ heißt *Gitter,* wenn es eine Basis $(b_0, \ldots, b_m)$ von $\mathbb{R}^{m+1}$ gibt, so daß

$$\Gamma = \left\{ \sum_{i=0}^{m} \lambda_i b_i \;\middle|\; \lambda_i \in \mathbb{Z} \right\} .$$

Wir bezeichnen diese Basis dann als eine Basis des Gitters Γ und schreiben kurz $\Gamma = \mathbb{Z}b_0 \oplus \cdots \oplus \mathbb{Z}b_m$.

Da h führenden Koeffizienten eins und Grad e hat, bilden die Polynome $p^k X^i$ mit $0 \leq i < e$ und hX^j mit $0 \leq j \leq m - e$ eine Basis der oben definierten Menge Λ; diese ist also ein Gitter.

Gitterbasen sind genauso wenig eindeutig wie Basen von Vektorräumen. Sind $(b_0, \ldots, b_m)$ und $(c_0, \ldots, c_m)$ zwei Basen des Gitters Γ, so sind sie insbesondere beide auch Basen von $\mathbb{R}^{m+1}$; es gibt also Matrizen $M, N \in \mathbb{R}^{(m+1)\times(m+1)}$, die diese beiden Basen ineinander überführen. Am einfachsten läßt sich das dadurch ausdrücken, daß wir die Spaltenvektoren b_i zu einer Matrix B zusammenfassen und die c_j zu einer Matrix C; dann ist $C = MB$ und $B = NC$. Die Einträge von M und N

müssen ganzzahlig sein, denn die c_j müssen ja ganzzahlige Linearkombinationen der b_i sein und umgekehrt. Außerdem ist $MN = NM$ gleich der Einheitsmatrix. Somit sind $\det M$ und $\det N$ ganzzahlig mit Produkt eins, d.h. $\det M = \det N = \pm 1$. Daher unterscheiden sich $\det B$ und $\det C$ höchstens durch das Vorzeichen.

Definition: Der Betrag von $\det B$ heißt Determinante $d(\Gamma)$ des Gitters Γ.

Wie wir gerade gesehen haben, ist $d(\Gamma)$ unabhängig von der gewählten Gitterbasis. Im oben definierten Gitter Λ ist $d(\Lambda) = p^{ke}$, da h den führenden Koeffizienten eins hat und die hinteren Terme der Polynome hX^j durch Zeilenoperationen aus der Determinante entfernt werden können.

Ein Vektorraum hat keinen echten Untervektorraum gleicher Dimension; bei Gittern ist das natürlich anders: Mit Γ ist auch 2Γ ein Gitter und ganz offensichtlich verschieden von Γ. Allgemein sagen wir, ein Gitter $\Gamma \subset \mathbb{R}^{m+1}$ sei ein *Untergitter* des Gitters $\Delta \subset \mathbb{R}^{m+1}$, wenn Γ eine Teilmenge von Δ ist.

Ist in dieser Situation $b_0, \ldots b_n$ eine Gitterbasis von Γ und $c_0, \ldots, c_n$ eine von Δ, so lassen sich die b_i als Linearkombinationen der c_j schreiben. Mit den gleichen Bezeichnungen wie oben ist daher $B = NC$ mit einer ganzzahligen Matrix N. Die inverse Matrix M freilich ist im Falle eines echten Untergitters nicht mehr ganzzahlig, sondern hat nur rationale Einträge. Wir können allerdings die Nenner begrenzen: Die Gleichung NM gleich Einheitsmatrix läßt sich übersetzen in $m + 1$ lineare Gleichungssysteme für die Spalten m_i von M, denn $Nm_i = e_i$ ist der i-te Einheitsvektor des $\mathbb{R}^{m+1}$. Lösen wir dieses Gleichungssystem nach der CRAMERschen Regel, so stehen im Zähler der Formeln für die Einträge von m_i Determinanten ganzzahliger Matrizen und in Nenner steht jeweils die Determinante D von M. Somit kann höchstens diese als Nenner auftreten und $D \cdot \Delta \subseteq \Gamma \subseteq \Delta$.

In Kürze wird es für uns wichtig sein, daß es zu einer gegebenen Basis von Δ spezielle, daran angepaßte Basen von Γ gibt:

Lemma: Ist Γ ein Untergitter von Δ und $(b_0, \ldots, b_m)$ eine Gitterbasis

von Δ, so gibt es eine Gitterbasis $(c_0, \ldots, c_m)$ von Γ derart, daß

$$\begin{aligned} c_0 &= \mu_{00} b_0 \\ c_1 &= \mu_{10} b_0 + \mu_{11} b_1 \\ &\cdots \\ c_m &= \mu_{m0} b_0 + \cdots + \mu_{mm} b_m \end{aligned}$$

mit ganzen Zahlen μ_{ij} und $\mu_{ii} \neq 0$.

Beweis: Da Db_i in Γ liegt, gibt es in Γ auf jeden Fall für jedes i Vektoren der Form $\mu_{i0} b_0 + \cdots + \mu_{ii} b_i$ mit $\mu_{ii} \neq 0$. c_i sei ein solcher Vektor mit minimalem $|\mu_{ii}|$. Wir wollen zeigen, daß diese Vektoren c_i eine Gitterbasis von Γ bilden. Da die lineare Unabhängigkeit trivial ist, muß nur gezeigt werden, daß sich jeder Vektor aus Γ als ganzzahlige Linearkombination der c_i schreiben läßt.

Angenommen, es gibt Vektoren $v \in \Gamma$, für die das nicht der Fall ist. Da v auch in Δ liegt, gibt es auf jeden Fall eine Darstellung $v = \lambda_0 b_0 + \cdots + \lambda_k b_k$ mit ganzen Zahlen λ_i und einem $k \leq m$. Wir wählen einen solchen Vektor v mit kleinstmöglichem k.

Da μ_{kk} nach Voraussetzung nicht verschwindet, gibt es eine ganze Zahl q, so daß $|\lambda_k - q\mu_{kk}|$ kleiner ist als der Betrag von μ_{kk}. Dann kann der Vektor

$$v - qc_k = (\lambda_0 - q\mu_{k0}) b_0 + \cdots + (\lambda_k - q\mu_{kk}) b_k$$

nicht als ganzzahlige Linearkombination der c_i dargestellt werden, denn sonst hätte auch v eine solche Darstellung. Wegen der Minimalität von k kann daher $\lambda_k - q\mu_{kk}$ nicht verschwinden. Da aber Betrag von $\lambda_k - q\mu_{kk}$ kleiner ist als der von μ_{kk}, widerspricht dies der Wahl von v als Vektor mit minimalem $|\mu_{kk}|$. Somit kann es keinen Gittervektor aus Γ geben, der nicht als ganzzahlige Linearkombination der c_i darstellbar ist, und das Lemma ist bewiesen. ∎

Bei der Anwendung von Gittern auf das Faktorisierungsproblem werden die GRAM-SCHMIDT-Orthogonalisierungen von Gitterbasen eine große Rolle spielen; daher sei kurz an diesen Orthogonalisierungsprozeß erinnert. Zunächst die

Definition: *a)* Ein EUKLIDischer Vektorraum ist ein reeller Vektorraum V zusammen mit einer Abbildung

$$\begin{cases} V \times V \to V \\ (v, w) \mapsto \big(v, w\big) \end{cases}$$

mit folgenden Eigenschaften:

1.) $\big(\lambda u + \mu v, w\big) = \lambda \big(u, w\big) + \mu \big(v, w\big)$ für alle $\lambda, \mu \in \mathbb{R}$ und alle $u, v, w \in V$.

2.) $\big(v, w\big) = \big(w, v\big)$ für alle $v, w \in V$.

3.) $\big(v, v\big) \geq 0$ für alle $v \in V$ und $\big(v, v\big) = 0$ genau dann, wenn $v = 0$.

Die Abbildung $V \times V \to V$ wird als Skalarprodukt bezeichnet.

b) Zwei Vektoren $v, w \in V$ heißen orthogonal, wenn $\big(v, w\big) = 0$ ist.

c) Eine Basis $(c_0, \ldots, c_m)$ eines EUKLIDischen Vektorraums heißt *Orthogonalbasis,* wenn $\big(c_i, c_j\big) = 0$ für alle $i \neq j$.

Wichtigstes Beispiel ist der Vektorraum $\mathbb{R}^{m+1}$ mit seinem Standardskalarprodukt

$$\left(\begin{pmatrix} v_0 \\ \vdots \\ v_m \end{pmatrix}, \begin{pmatrix} w_0 \\ \vdots \\ w_m \end{pmatrix} \right) = \sum_{i=0}^{m} v_i w_i \, .$$

Das Produkt eines Vektors v mit sich selbst ist dann das Quadrat seiner EUKLIDischen Länge, und wenn wir ihn als Koeffizientenvektor eines Polynoms vom Grad m aus $\mathbb{R}[X]$ auffassen, ist das auch das Quadrat der L^2-Norm dieses Polynoms.

Das Orthogonalisierungsverfahren von GRAM und SCHMIDT konstruiert aus einer beliebigen Basis $(b_0, \ldots, b_m)$ des $\mathbb{R}^{m+1}$ schrittweise eine Orthogonalbasis $(c_0, \ldots, c_m)$, und zwar so, daß in jedem Schritt der von den Vektoren $b_0, \ldots, b_r$ erzeugte Untervektorraum gleich dem von $c_0, \ldots, c_r$ erzeugten ist.

Der erste Schritt ist der einfachste: Da es noch keine Orthogonalitätsbedingung für c_0 gibt, können wir einfach $c_0 = b_0$ setzen.

Nachdem wir $r \geq 1$ Schritte durchgeführt haben, haben wir r linear unabhängige Vektoren $c_0, \ldots, c_{r-1}$ mit $\big(c_i, c_j\big) = 0$ für $i \neq j$ aus

dem von $b_0, \ldots, b_r$ aufgespannten Untervektorraum. Ist $r = m$, haben wir eine Orthogonalbasis; andernfalls muß ein auf den bisher konstruierten c_i senkrecht stehender Vektor c_r gefunden werden, der zusammen mit diesen den von b_0 bis b_r erzeugten Untervektorraum erzeugt.

Da $c_0, \ldots, c_{r-1}$ und $b_0, \ldots, b_{r-1}$ denselben Untervektorraum erzeugen, gilt dasselbe für $c_0, \ldots, c_{r-1}, b_r$ und $b_0, \ldots, b_r$; das Problem ist, daß b_r im allgemeinen nicht orthogonal zu den c_i sein wird. Wir dürfen b_r aber abändern um einen beliebigen Vektor aus dem von $c_0, \ldots, c_{r-1}$ aufgespannten Untervektorraum; also setzen wir

$$c_r = b_r - \lambda_0 c_0 - \cdots - \lambda_{r-1} c_{r-1}$$

und versuchen, die λ_i so zu bestimmen, daß dieser Vektor orthogonal zu $c_0, \ldots, c_{r-1}$ wird.

Wegen der Orthogonalität der c_i ist

$$\left(c_r, c_i\right) = \left(b_r, c_i\right) - \sum_{j=0}^{r-1} \lambda_j \left(c_j, c_i\right) = \left(b_r, c_i\right) - \lambda_i \left(c_i, c_i\right) \ ;$$

setzen wir daher

$$\lambda_i = \frac{\left(b_{r+1}, c_i\right)}{\left(c_i, c_i\right)} \ ,$$

so ist $\left(v, c_i\right) = 0$ für alle $i = 0, \ldots, r-1$.

Nach dem $m+1$-ten Schritt haben wir eine Orthogonalbasis $(c_0, \ldots, c_m)$ von $\mathbb{R}^{m+1}$ konstruiert.

Der dänische Mathematiker JØRGAN PEDERSEN GRAM (1850–1916) lehrte an der Universität Kopenhagen, war aber gleichzeitig auch noch geschäftsführender Direktor einer Versicherungsgesellschaft und Präsident des Verbands der dänischen Versicherungsunternehmen. Er publizierte anscheinend nur eine einzige mathematische Arbeit *Sur quelque théorèmes fondamentaux de l'algèbre moderne,* die 1874 erschien. Das GRAM-SCHMIDTsche Orthogonalisierungsverfahren, durch das er heute hauptsächlich bekannt ist, stammt wohl von LAPLACE (1749–1827) und wurde auch schon 1836 von CAUCHY verwendet.

ERHARD SCHMIDT (1876–1959) wurde im damals deutschen Ort Dorpat geboren; heute gehört dieser zu Estland und heißt Tartu. Er studierte in Berlin bei SCHWARZ und promovierte 1905 ins Göttingen bei HILBERT mit einer Arbeit über Integralgleichungen. Nach seiner Promotion wechselte er nach Bonn, wo er 1906 habilitierte. Danach lehrte er in Zürich, Erlangen und Breslau, bis er 1917 als Nachfolger von SCHWARZ nach Berlin berufen wurde. Er ist einer der Begründer der modernen Funktionalanalysis; insbesondere geht die Verallgemeinerung EUKLIDischer und HERMITEscher Vektorräume zu sogenannten HILBERT-Räumen auf ihn zurück.

Als Beispiel wollen wir eine Orthogonalbasis des von

$$b_0 = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 4 \end{pmatrix}, \quad b_1 = \begin{pmatrix} -3 \\ 4 \\ 0 \\ 5 \end{pmatrix} \text{ und } b_2 = \begin{pmatrix} -1 \\ -2 \\ 3 \\ 6 \end{pmatrix}$$

aufgespannten Untervektorraums U von $\mathbb{R}^4$ bestimmen.

Als ersten Vektor der Orthogonalbasis wählen wir einfach $c_0 = b_0$.

Für den zweiten Vektor machen wir den Ansatz $c_1 = b_1 - \lambda c_0$, wobei λ so gewählt werden muß, daß $(c_1, c_0) = (b_1, c_0) - \lambda (c_0, c_0) = 0$ ist. Da

$$(c_0, b_1) = -3 + 2 \cdot 4 + 4 \cdot 5 = 25 \quad \text{und} \quad (c_0, c_0) = 1^2 + 2^2 + 2^2 + 4^2 = 25\,,$$

müssen wir $\lambda = 1$ setzen und $c_1 = b_1 - c_0 = \begin{pmatrix} -4 \\ 2 \\ -2 \\ 1 \end{pmatrix}$.

Für den noch fehlenden dritten Vektor der Orthogonalbasis ist der Ansatz entsprechend:

$$c_2 = b_2 - \lambda c_0 - \mu c_1 \quad \text{mit} \quad (c_2, c_0) = (c_2, c_1) = 0\,.$$

$$(c_2, c_0) = (b_2, c_0) - \lambda (c_0, c_0) = (-1 - 4 + 6 + 24) + 25\lambda \Longrightarrow \lambda = 1$$

$$(c_2, c_1) = (b_2, c_1) + \mu (c_1, c_1) = (4 - 4 - 6 + 6) - 25\mu \Longrightarrow \mu = 0$$

$$c_2 = b_2 - c_0 = \begin{pmatrix} -2 \\ -4 \\ 1 \\ 2 \end{pmatrix}.$$

Unsere Orthogonalbasis besteht also aus den drei Vektoren

$$c_0 = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 4 \end{pmatrix}, \qquad c_1 = \begin{pmatrix} -4 \\ 2 \\ -2 \\ 1 \end{pmatrix} \quad \text{und} \quad c_2 = \begin{pmatrix} -2 \\ -4 \\ 1 \\ 2 \end{pmatrix}.$$

Wenn wir den Zusammenhang zwischen der Ausgangsbasis $(b_0, \ldots, b_m)$ und der Orthogonalbasis $(c_0, \ldots, c_m)$ explizit festhalten wollen, müssen wir den oben berechneten Koeffizienten Namen geben, die auch vom Schritt abhängen. Wir schreiben

$$c_i = b_i - \sum_{j=0}^{i-1} \mu_{ij} c_j \quad \text{für } i = 0, \ldots, m \quad \text{mit} \quad \mu_{ij} = \frac{\big(b_i, c_j\big)}{\big(c_j, c_j\big)}.$$

Im gerade durchgerechneten Beispiel etwa ist

$$c_0 = b_0, \quad c_1 = b_1 - c_0 \quad \text{und} \quad c_2 = b_2 - c_0\,,$$

also $\mu_{10} = \mu_{20} = 1$ und $\mu_{21} = 0$.

Lösen wir die obigen Formeln auf nach b_i, kommen wir auf

$$b_i = c_i + \sum_{j=0}^{i-1} \mu_{ij} c_j \quad \text{und} \quad \left(c_i, \sum_{j=0}^{i-1} \mu_{ij} c_j\right) = \sum_{j=0}^{i-1} \mu_{ij} \big(c_i, c_j\big) = 0\,.$$

Geometrisch bedeutet dies, daß c_i der Lotvektor bei der Projektion von b_i auf den von $c_1, \ldots, c_{i-1}$ aufgespannten Untervektorraum ist oder, anders ausgedrückt, die orthogonale Projektion von b_i auf das orthogonale Komplement dieses Raums.

Ist allgemein $w = u + v$ die Summe zweier aufeinander senkrecht stehenden Vektoren, so ist

$$\big(w, w\big) = \big(u + v, u + v\big) = \big(u, u\big) + \big(v, v\big)\,,$$

denn $\big(u, v\big) = 0$. Insbesondere sind daher die Längen von u und v höchstens gleich der Länge von w. In unserer Situation bedeutet dies, daß

$$|c_i| \le \big|b_i\big| \quad \text{für} \quad i = 0, \ldots, m\,,$$

kein Vektor der Orthogonalbasis kann also länger sein als der entsprechende Vektor der Ausgangsbasis.

Im Falle einer Gitterbasis $(b_0, \ldots, b_m)$ ist die nach GRAM-SCHMIDT berechnete Orthogonalbasis zwar eine Basis des $\mathbb{R}^{m+1}$, aber im allgemeinen keine Gitterbasis: Es gibt schließlich keinen Grund, warum die μ_{ij} ganze Zahlen sein sollten, so daß die c_j oft nicht einmal im Gitter liegen, und tatsächlich muß ein Gitter auch keine Orthogonalbasis haben.

Hätte etwa das Gitter

$$\Lambda = \mathbb{Z}\begin{pmatrix}1\\0\end{pmatrix} \oplus \mathbb{Z}\begin{pmatrix}\sqrt{2}\\1\end{pmatrix}$$

eine Orthogonalbasis (u, v), so gäbe es ganze Zahlen a, b, c, d, so daß

$$u = a\begin{pmatrix}1\\0\end{pmatrix} + b\begin{pmatrix}\sqrt{2}\\1\end{pmatrix} = \begin{pmatrix}a + b\sqrt{2}\\b\end{pmatrix}$$

und

$$v = c\begin{pmatrix}1\\0\end{pmatrix} + d\begin{pmatrix}\sqrt{2}\\1\end{pmatrix} = \begin{pmatrix}c + d\sqrt{2}\\d\end{pmatrix}$$

wäre. Das Skalarprodukt dieser beiden Vektoren wäre null, d.h.

$$(a + b\sqrt{2})(c + d\sqrt{2}) + bd = (ac + 3bd) + (ad + bc)\sqrt{2} = 0\,.$$

Wegen der Irrationalität von $\sqrt{2}$ ist dies nur dann möglich, wenn $ad+bc$ verschwindet. Nun wissen wir aber, daß die Determinante der Matrix für einen Wechsel der Gitterbasis ± 1 sein muß, d.h. $ad - bc = \pm 1$. Addition dieser Gleichung zu $ad + bc = 0$ führt auf $2ad = \pm 1$, was mit ganzen Zahlen a, d offensichtlich nicht gelten kann. Somit hat zumindest dieses Gitter keine Orthogonalbasis.

Trotzdem ist die aus einer Gitterbasis konstruierte Orthogonalbasis des $\mathbb{R}^n$ auch nützlich zum Verständnis des Gitters. Als erstes Beispiel dafür wollen wir die Determinante des Gitters geometrisch interpretieren:

Die reelle Matrix M, die den Wechsel von der Ausgangsbasis zur Orthogonalbasis beschreibt, ist wie die obigen Formeln zeigen eine Dreiecksmatrix mit lauter Einsen in der Hauptdiagonale, hat also Determinante eins. Obwohl die c_i keine Gitterbasis bilden, ist daher die Determinante

des Gitters auch gleich dem Betrag der Determinante der Matrix C mit den c_i als Spaltenvektoren. Das ist aber einfach das Produkt der Längen der c_i, denn der ij-Eintrag von $C^T C$ ist $\left(c_i, c_j\right)$. Da die c_i eine Orthogonalbasis bilden, verschwindet dieses Skalarprodukt für $i \neq j$, also ist $C^T C$ eine Diagonalmatrix und ihre Determinante ist das Produkt der Längenquadrate $\left(c_i, c_i\right)$. Der Betrag der Determinante von C selbst ist daher das Produkt der Längen.

Geometrisch entspricht die Orthogonalisierung nach GRAM-SCHMIDT einer Folge von Scherungen, die aus dem von den Vektoren b_i aufgespannten Parallelepiped einen Quader machen. Nach dem Prinzip von CAVALIERI bleibt das Volumen dabei unverändert, und das Volumen eines Quaders ist natürlich einfach das Produkt seiner Seitenlängen. Somit ist die Determinante eines Gitters gleich dem Volumen des von den Basisvektoren aufgespannten Parallelepipeds, dem sogenannten Fundamentalbereich des Gitters. Je nach Wahl der Basis kann dieser sehr verschiedene Formen haben, aber sein Volumen ist stets dasselbe.

Das wollen nun anwenden, um die Determinante eines Gitters abzuschätzen durch die Längen der Vektoren aus einer beliebigen Gitterbasis $(b_0, \ldots, b_m)$. Ist $(c_0, \ldots, c_m)$ die zugehörige Orthogonalbasis, so ist die Determinante des Gitters das Produkt der Längen der Vektoren c_i. Wie wir oben gesehen haben, kann kein Vektor c_i länger sein als der entsprechende Vektor b_i, und damit folgt die

Ungleichung von Hadamard: Im Gitter $\Gamma = \mathbb{Z}b_0 \oplus \cdots \oplus \mathbb{Z}b_m$ ist

$$d(\Gamma) \leq \prod_{i=0}^{m} \left|b_i\right|$$

■

Die Ungleichung von HADAMARD spielt eine wichtige Rolle im Beweis des folgenden Lemmas, das bei der Suche nach dem zu Beginn des Paragraphen definierten Polynoms h_0 nützlich sein wird. Alle Bezeichnungen seien wie dort.

Lemma: Erfüllt ein Polynom $v \in \Lambda$ die Ungleichung $\|f\|_2 \cdot \|v\|_2 < p^{ke}$, so ist v durch h_0 teilbar.

JACQUES SALOMON HADAMARD wurde 1865 in Versailles geboren, lebte aber ab dem Alter von drei Jahren in Paris. Dort studierte er von 1884-1888 an der Ecole Normale Supérieure; während der anschließenden Arbeit an seiner Dissertation verdiente er seinen Lebensunterhalt als Lehrer. Nach seiner Promotion 1892 ging er zunächst als Dozent, ab 1896 als Professor für Astronomie und Theoretische Mechanik an die Universität von Bordeaux. Während dieser Zeit bewies er unter anderem den berühmten Primzahlsatz, wonach sich die Anzahl der Primzahlen $\leq n$ asymptotisch verhält wie $n/\ln n$. Um wieder nach Paris zurückzukommen, akzeptierte er 1897 dort zwei (schlechtere) Stellen an der Sorbonne und am Collège de France; am letzteren erhielt er 1909 einen Lehrstuhl. 1912 wurde er Nachfolger von CAMILLE JORDAN an der Ecole Polytechnique sowie Nachfolger von HENRI POINCAR'E an der Académie des Sciences. 1940 mußte er nach USA emigrieren und lehrte an der Columbia University in New York, kehrte aber sofort nach Kriegsende zurück nach Paris. Unter seinen Arbeiten befinden sich außer dem Primzahlsatz auch fundamentale Beiträge unter anderem zur Theorie der partiellen Differentialgleichung, zu geodätischen Linien und zur Variationsrechnung. Auch politisch war er sehr aktiv, zunächst zugunsten von ALFRED DREYFUS. Nach 1945 engagierte er sich, nachdem drei seiner Söhne in den Weltkriegen gefallen waren, für die Friedensbewegung; zum Internationalen Mathematikerkongress in Cambridge, *Mass.,* dessen Ehrenpräsident er war, erhielt er deshalb nur nach der Intervention zahlreicher amerikanischer Mathematiker ein Einreisevisum für die USA.

Beweis: Für das Nullpolynom ist die Aussage trivial; sei also $v \neq 0$ und $g = \mathrm{ggT}(f, v)$. Nach dem ersten Lemma dieses Paragraphen reicht es zu zeigen, daß h mod p ein Teiler von g mod p ist.

Sollte dies nicht der Fall sein, sind h mod p und g mod p wegen der Irreduzibilität von h mod p teilerfremd; wie oben gibt es also Polynome $a, b, c \in \mathbb{Z}[X]$, so daß gilt

$$ah + bg = 1 - pc\,.$$

Sei $n = \deg g$ und $m' = \deg v$; dann ist $n \leq m' \leq m$. Wir definieren eine neue Teilmenge

$$\mathsf{M} = \left\{\lambda f + \mu v \;\middle|\; \lambda, \mu \in \mathbb{Z}[X],\ \deg\lambda < m' - n,\ \deg\mu < d - n\right\}$$

des Gitters $\mathbb{Z} \oplus \mathbb{Z}X \oplus \cdots \oplus \mathbb{Z}X^{d+m'-n-1}$; ihre natürliche Projektion auf das Untergitter $\mathbb{Z}X^n \oplus \mathbb{Z}X^{n+1} \oplus \cdots \oplus \mathbb{Z}X^{d+m'-n-1}$ sei M'.

Angenommen, das Element $\lambda f + \mu v \in \mathsf{M}$ wird dabei auf das Nullpolynom projiziert. Dann muß einerseits der Grad von $\lambda f + \mu v$ kleiner als n sein, andererseits ist $\lambda f + \mu v$ durch g teilbar und $n = \deg g$. Also muß $\lambda f + \mu v = 0$ sein und $\lambda f = -\mu v$. Division durch $g = \mathrm{ggT}(f, v)$ führt auf

$$\lambda \frac{f}{g} = -\mu \frac{v}{g} \quad \text{und} \quad \mathrm{ggT}\left(\frac{f}{g}, \frac{v}{g}\right) = 1 \,.$$

Somit muß μ ein Vielfaches von f/g sein. Der Grad von μ ist aber nach Definition kleiner als $d - n = \deg f - \deg g$, also ist $\mu = 0$ und damit auch $\lambda = 0$. Daher sind die Projektionen der Polynome

$$X^i f \;\text{ für }\; 0 \le i < m' - n \quad \text{und} \quad X^j v \;\text{ für }\; 0 \le j < d - n$$

nach M' linear unabhängig. Wie die Definition von M zeigt, bilden sie auch ein Erzeugendensystem, also ist M' ein Gitter, und die obigen Polynome bilden eine Gitterbasis. Darauf können wir die Ungleichung von HADAMARD anwenden:

$$d(\mathsf{M}') \le \|f\|_2^{m'-n} \cdot \|v\|_2^{e-n} \le \|f\|_2^m \, \|v\|_2^d < p^{ke} \,,$$

wobei das letzte Kleinerzeichen die Voraussetzung des Lemmas ist.

Im Rest des Beweises wollen wir zeigen, daß $d(\mathsf{M}') \ge p^{ke}$ sein muß, was zusammen mit der gerade gezeigten Ungleichung zu einem Widerspruch führt und damit das Lemma beweist.

Sei dazu $w \in \mathsf{M}$ ein Polynom vom Grad kleiner $n + e$. Als Element von M ist es durch g teilbar. Multiplizieren wir die obige Gleichung $ah + bg = 1 - pc$ mit $1 + pc + \cdots + (pc)^{k-1}$, erhalten wir eine Gleichung der Form $\tilde{a}h + \tilde{b}g = 1 - (pc)^k$ mit $\tilde{a}, \tilde{b} \in \mathbb{Z}[X]$. Multiplikation dieser Gleichung mit dem Polynom w/g führt auf eine neue Gleichung

$$a^* h + b^* w = \frac{w}{g}\big(1 - (pc)^k\big) \equiv \frac{w}{g} \bmod p^k \quad \text{mit} \quad a^*, b^* \in \mathbb{Z}[X] \,.$$

Als Element von M läßt sich w in der Form $w = \lambda f + \mu v$ schreiben, und nach Voraussetzung sind sowohl f als auch v modulo p^k durch h teilbar. Also ist auch w und damit nach der gerade bewiesenen Gleichung w/g modulo p^k durch h teilbar. Der Grad von w ist kleiner als $n+e$, und g hat Grad n, also ist der Grad von w/g kleiner als $n+e-n = e = \deg h$. Da h

führenden Koeffizienten eins hat, wird dieser Grad modulo p^k nicht kleiner; also muß w/g und damit auch w modulo p^k das Nullpolynom sein. Somit ist jedes Polynom aus M mit Grad kleiner $n+e$ durch p^k teilbar.

Das Gitter M' liegt in $\mathbb{Z}X^n \oplus \cdots \oplus \mathbb{Z}X^{d+m'-n-1}$ und hat eine Gitterbasis aus $d+m'-2n$ Elementen, ist also ein Untergitter im Sinne der Definition dieses Paragraphen. Die Polynome X^n bis $X^{d+m'-n-1}$ bilden natürlich eine Basis des größeren Gitters; daher hat M nach dem zweiten Lemma dieses Paragraphen eine Gitterbasis aus Polynomen der Grade $n, n+1, \ldots, d+m'-n-1$. Die ersten e davon müssen, wie wir gerade gesehen haben, durch p^k teilbar sein. Die Determinante von M' ist der Betrag der Determinante der Matrix aus den Basisvektoren; auf Grund der Gradbedingung ist dies eine Dreiecksmatrix, die Determinante ist also einfach das Produkt der führenden Koeffizienten und hat damit mindestens Betrag p^{ke}. Dies liefert den verlangten Widerspruch. ∎

Das gerade bewiesene Lemma legt nahe, daß uns Polynome kleiner L^2-Norm im Gitter Λ zu Faktoren von f verhelfen können, und in der Tat zeigen LENSTRA, LENSTRA und LOVÁSZ, daß wir h_0 konstruieren können als ggT der Polynome aus einer „geeigneten“ Gitterbasis von Λ. Im nächsten Paragraphen soll diese auch für viele andere Aufgaben „geeignete“ Basis allgemein konstruiert werden.

§9: Der LLL-Algorithmus zur Basisreduktion

Der hier vorgestellte Algorithmus wurde zwar in der zu Beginn des vorigen Paragraphen zitierten Arbeit von LENSTRA, LENSTRA und LOVÁSZ speziell für die Faktorisierung von Polynomen aus $\mathbb{Z}[X]$ entwickelt, er fand aber inzwischen zahlreiche weitere Anwendungen in der Kryptographie, der diskreten Optimierung und anderswo. Deshalb wird hier nicht von Polynomen, sondern nur allgemein von Vektoren die Rede sein, und wir werden auch, wie dort üblich, die Nummerierung nicht wie im vorigen Paragraphen bei der für Polynome sinnvollen Null beginnen, sondern bei eins.

Wir gehen daher aus von einem Gitter $\Gamma = \mathbb{Z}b_1 \oplus \cdots \oplus \mathbb{Z}b_n \leq \mathbb{R}^n$ und wollen dort nach kurzen Vektoren suchen.

Falls die Gitterbasis $b_1, \ldots, b_n$ eine Orthogonalbasis des $\mathbb{R}^n$ ist, hat ein Vektor $v = a_1 b_1 + \cdots + a_n b_n$ aus Γ die Länge

$$|v| = \sqrt{a_1^2 \ \left(b_1, b_1\right) + \cdots + a_n^2 \ \left(b_n, b_n\right)} \ ;$$

wenn wir zusätzlich noch annehmen, daß $\left|b_1\right| \leq \cdots \leq \left|b_n\right|$ ist, sind daher $\pm b_1$ kürzeste Vektoren in Γ. Je nach Länge von b_2 gilt eventuell dasselbe auch für $\pm b_2$; falls b_2 aber länger ist als b_1, müssen wir auf der Suche nach zweitkürzesten Vektoren die Längen von b_2 und $2b_1$ miteinander vergleichen und können uns entsprechend weiter hochhangeln, bis wir alle Vektoren unterhalb einer vorgegebenen Länge gefunden haben.

Wie wir bereits im vorigen Paragraphen gesehen haben, hat ein Gitter jedoch im allgemeinen keine Orthogonalbasis; wir müssen wir uns daher mit weniger zufrieden geben. Trotzdem wollen wir eine Basis, die sich zumindest nicht allzu sehr von einer Orthogonalbasis unterscheidet. Letzteres können wir auch so formulieren, daß sich die Basis nicht zu sehr von ihrer GRAM-SCHMIDT-Orthogonalisierung unterscheiden soll, denn wenn wir dieses Verfahren auf eine Orthogonalbasis anwenden, ändert sich ja nichts.

Die nach GRAM-SCHMIDT konstruierten Vektoren der Orthogonalbasis sind

$$c_i = b_i - \sum_{j=1}^{i-1} \mu_{ij} c_j \quad \text{mit} \quad \mu_{ij} = \frac{\left(b_i, c_j\right)}{\left(c_j, c_j\right)} \ ,$$

wobei die μ_{ij} aber im allgemeinen keine ganzen, sondern nur rationale Zahlen sind.

Wenn der Vektor c_i nicht im Gitter liegt, können wir wenigstens versuchen, ihn durch einen möglichst ähnlichen Gittervektor zu ersetzen, um so näher an das orthogonale Komplement zu kommen. Wir können beispielsweise alle Zahlen μ_{ij} ersetzen durch die jeweils nächstgelegene ganze Zahl (oder eine der beiden, falls μ_{ij} die Hälfte einer ungeraden Zahl sein sollte).

Wir können daher eine Basis findet, für die alle Koeffizienten μ_{ij} bei der GRAM-SCHMIDT-Orthogonalisierung höchstens den Betrag $\frac{1}{2}$ haben.

LENSTRA, LENSTRA und LOVÁSZ stellen noch eine weitere Bedingung:

$$\left|c_i + \mu_{i\,i-1} c_{i-1}\right|^2 \geq \tfrac{3}{4} \left|c_{i-1}\right|^2 \quad \text{für alle } i > 1 \,.$$

Um diese sogenannte LOVÁSZ-Bedingung zu verstehen, multiplizieren wir beiden Seiten aus. Wegen der Orthogonalität der c_j ist

$$|c_i|^2 + \mu_{i\,i-1}^2 \left|c_{i-1}\right|^2 \geq \tfrac{3}{4} \left|c_{i-1}\right|^2 \quad \text{oder} \quad |c_i|^2 \geq \left(\tfrac{3}{4} - \mu_{i\,i-1}^2\right) \left|c_{i-1}\right|^2 \,.$$

Da die Beträge der μ_{ij} höchstens $\frac{1}{2}$ sind, folgt insbesondere

$$|c_i|^2 \geq \tfrac{1}{2} \left|c_{i-1}\right|^2 \quad \text{oder} \quad \left|c_{i-1}\right|^2 \leq 2\,|c_i|^2 \,.$$

Diese Bedingung sorgt also dafür, daß sich die Längen der c_i nicht zu stark unterscheiden.

Man könnte sich fragen, warum hier ausgerechnet die Konstante $\frac{3}{4}$ verwendet wird. In der Tat funktioniert die folgende Konstruktion auch, wenn $\frac{3}{4}$ durch irgendeine Konstante α mit $\frac{1}{4} < \alpha < 1$ ersetzt wird. Die Zwei in der gerade bewiesenen Ungleichung wird dann zu $4/(4\alpha - 1)$, d.h. für α knapp unter eins können wir sie auf eine Zahl knapp über $4/3$ herunterdrücken, während α-Werte nahe $\frac{1}{4}$ zu sehr schwachen Schranken führen. Starke Schranken sind zwar besser, allerdings wird dann auch der Aufwand für die Konstruktion einer entsprechenden Basis deutlich größer. Der Wert $\alpha = \frac{3}{4}$ ist ein Kompromiß, der sich bewährt hat und daher – soweit mir bekannt – praktisch überall verwendet wird.

Die formale Definition einer „geeigneten“ Basis ist somit

Definition: Eine Gitterbasis $b_1, \ldots, b_n$ mit GRAM-SCHMIDT-Orthogonalisierung

$$c_i = b_i - \sum_{j=1}^{i-1} \mu_{ij} c_j \,, \;\; i = 1, \ldots, n$$

heißt LLL-reduziert, wenn

$$\left|\mu_{ij}\right| \leq \frac{1}{2} \quad \text{für alle } i, j \qquad \text{und}$$

$$\left|c_i + \mu_{i\,i-1} c_{i-1}\right|^2 \geq \tfrac{3}{4} \left|c_{i-1}\right|^2 \quad \text{für alle } i > 1 \,.$$

Diese Basen können nur dann nützlich sein, wenn sie existieren. Wir wollen uns daher zunächst ansehen, wie LENSTRA, LENSTRA und

LOVÁSZ eine solche Basis konstruieren. Der Algorithmus ist natürlich nahe an der GRAM-SCHMIDT-Orthogonalisierung. Da es für Gitterbasen deutlich weniger Manipulationsmöglichkeiten gibt als für Vektorraumbasen und wir außerdem noch die LOVÁSZ-Bedingung erfüllen müssen, treten aber eine ganze Reihe zusätzlicher Komplikationen auf.

Wir gehen aus von irgendeiner Basis $(b_1, \ldots, b_n)$ eines Gitters $\Gamma \subset \mathbb{R}^n$ und wollen daraus eine LLL-reduzierte Basis konstruieren. Als erstes konstruieren wir dazu nach GRAM-SCHMIDT eine Orthogonalbasis bestehend aus den Vektoren

$$c_i = b_i - \sum_{j=1}^{i-1} \mu_{ij} c_j \in \mathbb{R}^n \quad \text{mit} \quad \mu_{ij} = \frac{(b_i, c_j)}{(c_j, c_j)} \,. \qquad (*)$$

Im Laufe des Algorithmus werden die b_i, c_j und die μ_{ij} in jedem Schritt verändert, allerdings stets so, daß die Gleichungen $(*)$ erfüllt bleiben.

Wie bei GRAM-SCHMIDT hangeln wir uns dimensionsweise hoch, d.h. wir realisieren die Bedingungen aus der Definition einer LLL-reduzierten Basis zunächst für Teilgitter. Dazu fordern wir für eine natürliche Zahl $k \geq 1$ die Bedingungen

$$\left|\mu_{ij}\right| \leq \tfrac{1}{2} \quad \text{für} \quad 1 \leq j < i \leq k \qquad (A_k)$$

und

$$\left|c_i + \mu_{i\,i-1} c_{i-1}\right|^2 \geq \tfrac{3}{4} \left|c_{i-1}\right|^2 \quad \text{für} \quad 1 < i \leq k \qquad (B_k)$$

Für $k = 1$ gibt es keine Indizes i, j, für die die rechtsstehenden Ungleichungen erfüllt sind, die Bedingungen sind also leer und somit trivialerweise erfüllt. Für $k = n$ dagegen besagen diese beiden Bedingungen, daß $(b_1, \ldots, b_n)$ eine LLL-reduzierte Gitterbasis von Γ ist. Wir müssen also k schrittweise erhöhen. Im Gegensatz zur Verfahren von GRAM-SCHMIDT müssen wir hier allerdings k gelegentlich auch *erniedrigen* statt erhöhen. Der Wert von k wird aber stets zwischen Null und n liegen und stets so gewählt sein, daß die Bedingungen (A_k) und (B_k) erfüllt sind.

Für jeden neuen Wert von k, egal ob er größer oder kleiner ist als sein Vorgänger, führen wir die folgenden Schritte durch:

Wir wollen die Gitterbasis und die davon abgeleitete Orthogonalbasis einschließlich der Koeffizienten μ_{ij} so verändern, daß auch (A_{k+1}) und (B_{k+1}) gelten.

Die Bedingung $|\mu_{k+1\,k}| \leq \frac{1}{2}$ ist kein großes Problem: Wir runden einfach $\mu_{k+1\,k}$ zur nächsten ganzen Zahl q (oder einer der beiden nächsten) und ersetzen b_{k+1} durch $b_{k+1} - qb_k$. Damit $(*)$ weiterhin gilt, ersetzen wir $\mu_{k+1\,k}$ durch $\mu_{k+1\,k} - q$ und die $\mu_{k+1\,j}$ mit $j < k$ durch $\mu_{k+1\,j} - q\mu_{kj}$.

Für das weitere Vorgehen müssen wir zwei Fälle unterscheiden:

Fall 1: $k \geq 1$ und $|c_{k+1} + \mu_{k+1\,k}c_k|^2 < \frac{3}{4}\,|c_k|^2$

In diesem Fall vertauschen wir b_k und b_{k+1}. Da die GRAM-SCHMIDT-Orthogonalisierung von der Reihenfolge der Basisvektoren abhängt, müssen wir dann eine neue Orthogonalbasis $(d_1, \ldots, d_n)$ berechnen.

An den c_j mit $j < k$ (so es welche gibt) ändert sich dabei nichts: Sie werden beim GRAM-SCHMIDTschen Orthogonalisierungsverfahren berechnet, bevor die Vektoren b_k und b_{k+1} ins Spiel kommen. Für $j < k$ ist somit $d_j = c_j$.

Auch für $j > k+1$ ist $d_j = c_j$, denn der j-te Vektor der Orthogonalbasis ist die Projektion des j-ten Vektors der Ausgangsbasis auf das orthogonale Komplement des von den ersten $j-1$ Basisvektoren aufgespannten Untervektorraums, und für $j > k+1$ ist

$$[c_1, \ldots, c_j] = [b_1, \ldots, b_j] = [d_1, \ldots, d_j]\,.$$

Bleiben noch die Vektoren d_k und d_{k+1}. Die müssen verschieden sein von den Vektoren c_k und d_{k+1}, denn $[c_1, \ldots, c_k] = [b_1, \ldots, b_k]$, aber $[d_1, \ldots, d_k] = [b_1, \ldots, b_{k-1}, b_{k+1}]$.

Nach den Formeln zur GRAM-SCHMIDT-Orthogonalisierung ist

$$c_k = b_k - \sum_{j=1}^{k-1} \mu_{kj} c_j \qquad \text{mit} \quad \mu_{kj} = \frac{(b_k, c_j)}{(c_j, c_j)}$$

und

$$\begin{aligned} c_{k+1} &= b_{k+1} - \sum_{j=1}^{k} \mu_{k+1\,j} c_j \qquad \text{mit} \quad \mu_{k+1\,j} = \frac{(b_{k+1}, c_j)}{(c_j, c_j)} \\ &= b_{k+1} - \sum_{j=1}^{k-1} \mu_{k+1\,j} c_j - \mu_{k+1\,k} b_k + \sum_{j=1}^{k-1} \mu_{k+1\,k} \mu_{kj} c_j \\ &= b_{k+1} - \mu_{k+1\,k} b_k - \sum_{j=1}^{k-1} (\mu_{k+1\,j} - \mu_{k+1\,k} \mu_{kj}) c_j \; ; \end{aligned}$$

entsprechend ist

$$d_k = b_{k+1} - \sum_{j=1}^{k-1} \nu_{kj} d_j \qquad \text{mit} \quad \nu_{kj} = \frac{(b_{k+1}, d_j)}{(d_j, d_j)}$$

und

$$\begin{aligned} d_{k+1} &= b_k - \sum_{j=1}^{k} \nu_{k+1\,j} d_j \qquad \text{mit} \quad \nu_{k+1\,j} = \frac{(b_k, d_j)}{(d_j, d_j)} \\ &= b_k - \sum_{j=1}^{k-1} \nu_{k+1\,j} c_j - \nu_{k+1\,k} \left(b_{k+1} - \sum_{j=1}^{k-1} \nu_{kj} c_j \right) . \end{aligned}$$

Da $c_j = d_j$ für $j < k$, folgt insbesondere

$$\nu_{kj} = \mu_{k+1\,j} \quad \text{und} \quad \nu_{k+1\,j} = \mu_{kj} \qquad \text{für } j < k$$

$$d_{k+1} = c_k - \nu_{k+1\,k} d_k \qquad \text{und}$$

$$d_k = b_{k+1} - \sum_{j=1}^{k-1} \nu_{kj} d_j = b_{k+1} - \sum_{j=1}^{k-1} \mu_{k+1\,j} c_j = c_{k+1} + \mu_{k+1\,k} c_k \, .$$

Nach der Voraussetzung für das Eintreten von Fall 1 ist das Längenquadrat des letzten Vektors kleiner als $\frac{3}{4}$ des Längenquadrats von c_k;

zumindest ein Vektor der Orthogonalbasis wird also durch die Vertauschung von b_k und b_{k+1} kürzer. Das Quadrat seiner Länge ist

$$\big(d_k, d_k\big) = \big(c_{k+1}, c_{k+1}\big) + \mu_{k+1\,k}^2 \ \big(c_k, c_k\big) \ .$$

Damit können wir nun auch $\nu_{k+1\,k}$ durch Daten der „alten“ Basis ausdrücken: Der Zähler ist

$$\big(b_k, d_k\big) = \left(c_k + \sum_{j=1}^{k-1} \mu_{kj} c_j, d_k\right) = \big(c_k, d_k\big) \ ,$$

denn für $j \le k-1$ steht d_k senkrecht auf $d_j = c_j$. Deshalb ist auch

$$\big(c_k, d_k\big) = \left(c_k, b_{k+1} - \sum_{j=1}^{k-1} \nu_{kj} d_j\right) = \big(c_k, b_{k+1}\big) \ ,$$

also ist

$$\begin{aligned}\nu_{k+1\,k} &= \frac{\big(b_k, d_k\big)}{\big(d_k, d_k\big)} = \frac{\big(c_k, b_{k+1}\big)}{\big(d_k, d_k\big)} = \frac{|c_k|^2}{|d_k|^2} \cdot \frac{\big(c_k, b_{k+1}\big)}{\big(d_k, d_k\big)} = \frac{|c_k|^2}{|d_k|^2} \cdot \mu_{k+1\,k} \\ &= \frac{\mu_{k+1\,k}\,|c_k|^2}{|c_{k+1}|^2 + \mu_{k+1\,k}^2\,|c_k|^2} \ .\end{aligned}$$

Dank dieser Formeln ist daher auch $d_{k+1} = c_k - \nu_{k+1\,k} d_k$ vollständig durch Daten der „alten“ Basis ausdrückbar.

Die Vektoren d_j mit $j > k+1$ sind wieder gleich den entsprechenden c_j, denn der j-te Vektor der Orthogonalbasis ist ja der Lotvektor von b_j auf den von $b_1, \ldots, b_{j-1}$ aufgespannten Untervektorraum, und für $j > k+1$ hat sich durch die Vertauschung von b_k und b_{k+1} weder an b_j noch an diesem Vektorraum etwas geändert. Aus diesem Grund sind auch die Koeffizienten ν_{ij} gleich den entsprechenden μ_{ij}, sofern weder i noch j gleich k oder $k+1$ sind.

Damit fehlen uns nur noch die Koeffizienten ν_{ik} und $\nu_{i\,k+1}$ für $i > k+1$. Um sie zu berechnen, drücken wir zunächst c_k und c_{k+1} aus durch d_k und d_{k+1}: Nach den obigen Formeln ist

$$\begin{aligned} d_k &= c_{k+1} + \mu_{k+1\,k} c_k \\ d_{k+1} &= c_k - \nu_{k+1\,k} d_k = (1 - \nu_{k+1\,k}\mu_{k+1\,k}) c_k - \nu_{k+1\,k} c_{k+1} \ . \end{aligned}$$

Addition von $\nu_{k+1\,k}$-mal der ersten Gleichung zur zweiten eliminiert c_{k+1} und liefert uns die Gleichung

$$c_k = \nu_{k+1\,k} d_k + d_{k+1} .$$

Setzen wir dies ein in die zweite Gleichung, erhalten wir

$$(1 - \nu_{k+1\,k}\mu_{k+1\,k})(\nu_{k+1\,k} d_k + d_{k+1}) - \nu c_{k+1} = d_{k+1}$$

und damit

$$c_{k+1} = (1 - \nu_{k+1\,k}\mu_{k+1\,k}) d_k - \mu_{k+1\,k} d_{k+1} .$$

Da in der Summe $d_k = c_{k+1} + \mu_{k+1\,k} c_k$ die Vektoren c_k und c_{k+1} orthogonal sind, ist $\left|d_k\right|^2 = \left|c_{k+1}\right|^2 + \mu_{k+1\,k}^2 \left|c_k\right|^2$, also

$$\frac{\left|c_{k+1}\right|^2}{\left|d_k\right|^2} = \frac{\left|d_k\right|^2 - \mu_{k+1\,k}^2 \left|c_k\right|^2}{\left|d_k\right|^2} = 1 - \frac{\left|c_k\right|^2}{\left|d_k\right|^2} \mu_{k+1\,k}^2 = 1 - \nu_{k+1\,k}\mu_{k+1\,k} .$$

Somit ist

$$c_{k+1} = \frac{\left|c_{k+1}\right|^2}{\left|d_{k+1}\right|^2} d_k - \mu_{k+1\,k} d_{k+1} .$$

Mit diesen beiden Formel gehen wir nun für $i > k+1$ in die Gleichungen

$$b_i = c_i + \sum_{j=1}^{i-1} \mu_{ij} c_j ,$$

Die Teilsumme $\mu_{ik} c_k + \mu_{i\,k+1} c_{k+1}$ aus den Termen für $j = k$ und $j = k+1$ wird dabei zu

$$\left(\mu_{ik}\nu_{k+1\,k} + \mu_{i\,k+1} \frac{\left|c_{k+1}\right|^2}{\left|d_k\right|^2} \right) d_k + \left(\mu_{ik} - \mu_{i\,k+1}\mu_{k+1\,k}\right) d_{k+1} .$$

Somit ist

$$\nu_{ik} = \mu_{ik}\nu_{k+1\,k} + \mu_{i\,k+1} \frac{\left|c_{k+1}\right|^2}{\left|d_k\right|^2} \quad \text{und} \quad \nu_{i\,k+1} = \mu_{ik} - \mu_{i\,k+1}\mu_{k+1\,k} .$$

Wir ersetzen nun alle c_i durch die entsprechenden d_i und alle μ_{ij} durch die entsprechenden ν_{ij}; dann ist $(*)$ auch für die Gitterbasis mit vertauschten Positionen von b_k und b_{k+1} erfüllt. Die Bedingungen (A_k) und (B_k) sind nun allerdings nur noch für $k-1$ sicher erfüllt; wir

müssen daher k durch $k-1$ ersetzen und einen neuen Iterationsschritt starten.

2. Fall: $k = 0$ oder $|c_{k+1} + \mu_{k+1\,k} c_k|^2 \geq \frac{3}{4} |c_k|^2$

In diesem Fall sorgen wir zunächst dafür, daß alle $\mu_{k+1\,j}$ einen Betrag von höchstens ein halb haben. (Im Fall $k = 0$ gibt es hier natürlich nichts zu tun.)

Für $j = k$ haben wir das bereits zu Beginn des Schritts für k sichergestellt; wir wählen nun den größten Index $\ell < k$, für den $|\mu_{k+1\,\ell}|$ größer ist als $\frac{1}{2}$ und verfahren damit wie oben: Wir ersetzen b_{k+1} durch $b_{k+1} - q b_\ell$ und $\mu_{k+1\,\ell}$ durch $\mu_{k+1\,\ell} - q$, wobei q die nächste ganze Zahl zu $\mu_{k+1\,\ell}$ ist, und wir ersetzen alle $\mu_{k+1\,j}$ mit $j < \ell$ durch $\mu_{k+1\,j} - q\mu_{\ell\,j}$. Sofern es danach immer noch ein $\mu_{k+1\,j}$ vom Betrag größer $\frac{1}{2}$ gibt, wählen wir wieder den größten Index j mit dieser Eigenschaft und so weiter, bis alle $\left|\mu_{k+1\,j}\right| \leq \frac{1}{2}$ sind.

Falls $k = n$ ist, endet der Algorithmus an dieser Stelle, und wir haben eine LLL-reduzierte Basis gefunden. Andernfalls ersetzen wir k durch $k + 1$ und beginnen mit einem neuen Iterationsschritt.

Um zu sehen, daß das Verfahren nach endlich vielen Schritten abbricht, müssen wir uns überlegen, daß der obige Fall 1, in dem der Index k erniedrigt wird, nicht unbegrenzt häufig auftreten kann. Ausgangspunkt dazu ist die Beobachtung, daß zumindest *ein* Vektor der Orthogonalbasis im ersten Fall verkürzt wird: Der k-te Vektor wird ersetzt durch einen neuen, dessen Längenquadrat höchstens gleich $\frac{3}{4}$ mal des alten ist.

Um dies auszunutzen, definieren wir für $k = 1, \ldots, n$ die reelle Zahl D_k als Determinante der $k \times k$-Matrix $\mathcal{B}_k$ mit ij-Eintrag $\left(b_i, b_j\right)$. Für $k = n$ können wir sie leicht auf bekannte Größen zurückführen: Ist $\mathcal{B}$ die $n \times n$-Matrix mit dem Basisvektor b_i als i-ter Spalte, so ist offensichtlich $\mathcal{B}_n = \mathcal{B}^T \mathcal{B}$, also ist $D_n = (\det \mathcal{B})^2 = d(\Gamma)^2$. Insbesondere ist also D_n unabhängig von der Gitterbasis und hängt nur ab vom Gitter.

Entsprechend können wir für D_k das Gitter $\Gamma_k = \mathbb{Z}b_1 \oplus \cdots \oplus \mathbb{Z}b_k$ im Vektorraum $\mathbb{R}b_1 \oplus \cdots \oplus \mathbb{R}b_k$ betrachten. Auch dies ist ein EUKLIDischer Vektorraum. Wenn wir dort eine Orthonormalbasis (d.h. eine Orthogonalbasis, deren sämtliche Vektoren Länge eins haben) auszeichnen, wird

er isomorph zum $\mathbb{R}^k$ mit seinem üblichen Skalarprodukt. Daher hängt auch D_k nur ab von Γ_k, nicht aber von den Vektoren $b_1, \ldots, b_k$.

Solange wir im LLL-Algorithmus nur die μ_{ij} auf Werte vom Betrag höchstens $\frac{1}{2}$ reduzieren, ändert sich nichts an den Gittern Γ_k, also bleiben auch die D_k unverändert.

Wenn wir aber zwei Basisvektoren b_k und b_{k+1} miteinander vertauschen, ändert sich das Gitter Γ_k *und nur dieses;* alle anderen Γ_i bleiben unverändert. D_k ist das Quadrat von $d(\Gamma_k)$, und $d(\Gamma_k)$ können wir auch als Produkt der Längen der Vektoren der zugeordneten Orthogonalbasis berechnen. Von diesen ändert sich nur der k-te, und dessen Längenquadrat wird kleiner als drei Viertel des Längenquadrats des entsprechenden Vektors der vorherigen Orthogonalbasis. Somit wird D_k mit einem Faktor von höchstens $\frac{3}{4}$ multipliziert.

Dasselbe gilt dann auch für das Produkt D aller D_k; wenn wir zeigen können, daß dieses eine nur vom Gitter abhängige untere Schranke hat, folgt also, daß wir nicht unbegrenzt oft im Fall 1 des Algorithmus sein können und dieser daher nach endlich vielen Schritten enden muß. Diese untere Schranke liefert uns der

Gitterpunktsatz von Minkowski: $\Gamma \subset \mathbb{R}^n$ sei ein Gitter und $M \subset \mathbb{R}^n$ sei eine zum Nullpunkt symmetrische beschränkte konvexe Teilmenge von $\mathbb{R}^n$. Falls das Volumen von M größer ist als $2^n d(\Gamma)$, enthält M mindestens einen vom Nullpunkt verschiedenen Punkt des Gitters.

Bevor wir diesen Satz beweisen, überlegen wir uns zunächst, daß er uns wirklich untere Schranke für alle D_k und damit auch für D liefert. Dazu wenden wir ihn an auf das Gitter Γ_k, das wir – siehe oben – als Teilmenge eines Vektorraums $\mathbb{R}^k$ auffassen können, und den Würfel

$$M = \left\{(x_1, \ldots, x_k) \in \mathbb{R}^k \;\middle|\; |x_i| \leq \varepsilon \text{ für alle } i\right\}.$$

Wenn dessen Volumen $(2\varepsilon)^k$ größer ist als $2^n d(\Gamma_k)$, gibt es in Γ_k (und damit erst recht in Γ) einen vom Nullvektor verschiedenen Vektor aus M. Dessen Länge ist höchstens gleich der halben Diagonale von M, also $\varepsilon\sqrt{k}$. Somit gibt es in Γ_k und damit erst recht in Γ einen Vektor $v \neq 0$, dessen Länge höchstens gleich $\varepsilon\sqrt{n}$ ist.

Da Gitter diskrete Mengen sind, gibt es in Γ eine Untergrenze μ für die Länge eines vom Nullvektor verschiedenen Vektors: Andernfalls wäre der Nullvektor ein Häufungspunkt des Gitters.

Falls der gerade betrachtete Würfel die Voraussetzung des Satzes von MINKOWSKI erfüllt, muß daher $\mu \leq \varepsilon\sqrt{n}$ sein, d.h. für $\varepsilon < \mu/\sqrt{n}$ kann die Voraussetzung nicht erfüllt sein. Somit ist

$$\left(\frac{\mu}{\sqrt{n}}\right)^k \leq d(\Gamma_k)\,.$$

Damit haben wir eine nur vom Gitter abhängige Untergrenze für $d(\Gamma_k)$ gefunden, also für D_k und damit auch für das Produkt D aller d_k. Dies zeigt, sofern wir den Gitterpunktsatz von MINKOWSKI voraussetzen, daß der LLL-Algorithmus zur Basisreduktion nach endlich vielen Schritten endet.

Bleibt also noch der Beweis des Satzes von MINKOWSKI:

$(b_1, \ldots, b_n)$ sei eine Gitterbasis und B sei die Matrix mit den b_i als Spaltenvektoren. Dann ist

$$\varphi\colon \begin{cases} \mathbb{R}^n \to \mathbb{R}^n \\ \quad v \mapsto Bv \end{cases}$$

eine bijektive lineare Abbildung, die das Gitter $\mathbb{Z}^n \subset \mathbb{R}^n$ auf Γ abbildet. Volumina werden bei dieser Abbildung mit $\det B = d(\Gamma)$ multipliziert; die Menge $\varphi^{-1}(M)$ hat also mindestens das Volumen 2^n und ist wegen der Linearität von φ beschränkt, symmetrisch und konvex. Es reicht, wenn wir zeigen, daß diese Menge einen vom Nullpunkt verschiedenen Punkt aus $\mathbb{Z}^n$ enthält.

Dies ist die Version des Satzes, die MINKOWSKI selbst bewiesen hat. Hier soll aber nicht sein Beweis wiedergegeben werden, sondern eine später gefundene Alternative von BLICHFELDT. Dieser bewies 1914 den folgenden

Satz: B sei eine beschränkte Teilmenge von $\mathbb{R}^n$ mit einem Volumen größer eins. Dann enthält B zwei verschiedene Punkte P, Q, deren Verbindungsvektor in $\mathbb{Z}^n$ liegt.

Wenden wir diesen Satz an auf die Menge $B = \frac{1}{2}\varphi^{-1}(M)$, so finden wir zwei Punkte $P, Q \in B$, deren Verbindungsvektor in $\mathbb{Z}^n$ liegt. Die Punkte $2P$ und $2Q$ liegen in $\varphi^{-1}(M)$, also –wegen der vorausgesetzten Symmetrie – auch $-2Q$. Wegen der Konvexität von $\varphi(M)$ liegt auch der Mittelpunkt der Verbindungsstrecke von $2P$ und $-2Q$ in $\varphi^{-1}(M)$; in Koordinaten ist dies der Punkt

$$\tfrac{1}{2}\big(2P + (-2Q)\big) = P - Q \in \mathbb{Z}^n \,.$$

Damit ist der Gitterpunktsatz vom MINKOWSKI modulo dem Satz von BLICHFELDT bewiesen.

Als letztes bleibt damit noch der Satz von BLICHFELDT zu zeigen.

Dazu sei $W = \big\{(x_1, \ldots, x_n) \in \mathbb{R}^n \;\big|\; 0 \le x_i < 1 \text{ für alle } i\big\}$ und für $v \in \mathbb{Z}^n$ sein $W_v = W + v$ der um v verschobene Würfel W. Dann sind offensichtlich alle W_v disjunkt und überdecken gemeinsam den $\mathbb{R}^n$. Da B beschränkt ist, können nur endlich viele der W_v einen nichtleeren Durchschnitt B_v mit B haben. Für jeden dieser Durchschnitte betrachten wir seine Verschiebung $B_v - v$ um den Vektor $-v$; das ist offensichtlich eine Teilmenge von W.

Die Summe der Volumina aller dieser Teilmengen ist gleich dem Volumen von B, denn B ist die disjunkte Vereinigung aller B_v. Damit ist diese Summe nach Voraussetzung größer als eins; da W das Volumen eins hat, muß es also zwei Vektoren $v \neq w$ geben, so daß $B_v \cap B_w$ nicht leer ist. Für einen Punkt R aus diesem Durchschnitt sei P seine Translation um v und Q die um w. Dann liegen $P \in B_v$ und $Q \in B_w$ beide in B, und ihr Verbindungsvektor ist $w - v \in \mathbb{Z}^n$. ∎

Als Beispiel für die LLL-Reduktion wollen wir eine LLL-reduzierte Basis des von

$$b_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad b_2 = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} \quad \text{und} \quad b_3 = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$$

erzeugten Gitters $\Gamma \subset \mathbb{R}^3$ bestimmen.

HERMANN MINKOWSKI wurde 1864 als Sohn einer deutsch-jüdischen Kaufmannsfamilie im damals russischen Aleksotas (heute Kaunas in Litauen) geboren. Als er acht Jahre alt war, zog die Familie um nach Königsberg, wo er auch zur Schule und ab 1880 zur Universität ging. Einer seiner Komillitonen war HILBERT. Während seines Studiums ging er auch für drei Semester nach Berlin, promovierte aber 1885 in Königsberg über quadratische Formen. In seiner Habilitationsschrift von 1887, die ihm eine Stelle an der Universität Bonn verschaffte, beschäftigt er sich erstmalig mit seiner *Geometrie der Zahlen,* in der der Gitterpunktsatz ein wichtiges Hilfsmittel ist. 1892 ging er zurück nach Königsberg, 1894 dann an die ETH Zürich, wo EINSTEIN einer seiner Studenten war. 1902 folgte (auf Initiative von HILBERT) der Ruf auf einen Lehrstuhl in GÖTTINGEN, wo er sich vor allem mit mathematischer Physik beschäftigte. Die Geometrie des (in heutiger Terminologie) MINKOWSKI-Raums erwies sich als fundamental für die Entwicklung der Relativitätstheorie. Er starb 1909 im Alter von nur 44 Jahren an einem damals nicht behandelbaren Blinddarmdurchbruch.

HANS FREDERIK BLICHFELDT wurde 1873 in Dänemark geboren, jedoch wanderte die Familie bereits 1888 aus in die USA. Er bestand zwar bereits in Dänemark die Aufnahmeprüfung zur Universität mit Auszeichnung, aber seine Eltern konnten die Studiengebühren nicht aufbringen. So konnte er erst nach vier Jahren Arbeit in Farms und Sägemühlen ab 1894 an der Stanford University in Palo Alto, Kalifornien studieren. Einer seiner dortigen Professoren lieh im das notwendige Geld zum Promotionsstudium bei SOPHUS LIE in Leipzig; seine Promotion beschäftigte sich mit Transformationsgruppen im $\mathbb{R}^3$. 1898 kehrte er zurück nach Stanford, wo er zunächst als *instructor* arbeitete. 1913 erhielt er einen Lehrstuhl, 1927 bis zu seiner Emeritierung 1938 war er Dekan der mathematischen Fakultät. Seine Arbeiten beschäftigen sich mit der Geometrie der Zahlen und der Gruppentheorie. Er starb 1945 in Palo Alto.

Als erstes brauchen wir die zugehörige Orthogonalbasis. Nach GRAM-SCHMIDT setzen wir $c_1 = b_1$ und wählen dann μ_{21} so, daß

$$\big(b_2 - \mu_{21} c_1, c_1\big) = 10 - 14\mu_{21} = 0$$

ist, d.h.

$$\mu_{21} = \frac{5}{7} \quad \text{und} \quad c_2 = \frac{1}{7}\begin{pmatrix} 16 \\ 4 \\ -8 \end{pmatrix} .$$

Der dritte Vektor $c_3 = b_3 - \mu_{31}c_1 - \mu_{32}c_3$ wird so gewählt, daß

$$\big(c_3, c_1\big) = 11 - 14\mu_{21} = 0 \quad \text{und} \quad \big(c_3, c_2\big) = \frac{36}{7} - \frac{48}{7}\mu_{32} = 0\,,$$

wir haben also

$$\mu_{31} = \frac{11}{14}, \quad \mu_{32} = \frac{3}{4} \quad \text{und} \quad c_3 = \frac{1}{2}\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}\,.$$

Wir beginnen die LLL-Reduktion mit $k = 1$ und müssen als erstes testen, ob $\mu_{21} = \frac{5}{7}$ einen Betrag von höchstens $\frac{1}{2}$ hat. Das ist offensichtlich nicht der Fall; die nächste ganze Zahl ist $q = 1$, also ersetzen wir b_2 durch

$$b_2 - b_1 = \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix}$$

und μ_{21} durch $\mu_{21} - 1 = -\frac{2}{7}$.

$$c_2 + \mu_{21}c_1 = \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix}$$

hat Längenquadrat acht, was kleiner ist als $\frac{3}{4}\,|c_1|^2 = \frac{21}{2}$. Daher sind wir in Fall 1 und müssen b_1 und b_2 vertauschen. Der neue erste Vektor der Orthogonalbasis ist der gerade berechnete Vektor $d_1 = c_2 + \mu_{21}c_1$ und

$$\nu_{21} = \mu_{21}\frac{|c_1|^2}{|c_2|^2 + \mu_{21}\,|c_1|^2} = -\frac{1}{2}\,.$$

Damit ist

$$d_2 = c_1 - \nu_{21}d_1 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix};$$

die neuen Koeffizienten sind

$$\nu_{31} = \mu_{31}\nu_{21} + \mu_{32}\frac{|c_2|^2}{\big|d_1\big|^2} = \frac{1}{4} \quad \text{und} \quad \nu_{32} = \mu_{31} - \mu_{32}\mu_{21} = 1\,.$$

Wir ersetzen die c_i durch die d_i und die μ_{ij} durch die ν_{ij}; außerdem müssen wir, da wir Basisvektoren vertauscht haben, k um eins erniedrigen. Wir gehen also mit $k = 0$ in den nächsten Iterationsschritt.

Dort sind wir mit $k = 0$ automatisch im zweiten Fall, und es gibt nichts zu tun; also erhöhen wir k wieder auf eins und starten mit einen neuen Iterationsschritt.

Als erstes muß sichergestellt werden, daß μ_{21} höchstens Betrag $\frac{1}{2}$ hat; da $\mu_{21} = -\frac{1}{2}$, ist dies der Fall.

$$c_2 + \mu_{21} c_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

hat Längenquadrat 14, was größer ist als $\frac{3}{4} |c_1|^2$; wir sind also im zweiten Fall und müssen die μ_{2j} mit $j < 1$ auf Beträge von höchstens $\frac{1}{2}$ reduzieren. Da es keine $j < 1$ gibt, ist diese Bedingung leer; wir können also k auf zwei erhöhen und zum nächsten Schritt gehen.

$\mu_{32} = 1$ hat zu großen Betrag, muß also auf Null reduziert werden; damit Bedingung $(*)$ erfüllt bleibt, müssen wir auch μ_{31} durch $\mu_{31} - \mu_{21} = \frac{3}{4}$ ersetzen und b_3 durch

$$b_3 - b_2 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} .$$

Dann hat

$$c_3 + \mu_{32} c_2 = \frac{1}{2} \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

Längenquadrat $\frac{3}{2}$, während $\frac{3}{4} |c_2|^2 = 9$ ist, wir sind also wieder im ersten Fall und müssen b_2 mit b_3 vertauschen. Neuer zweiter Vektor der Orthogonalbasis wird

$$d_2 = c_3 + \mu_{32} c_2 = \frac{1}{2} \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

und

$$\nu_{32} = \mu_{32} \frac{|c_2|^2}{|c_3|^2 + \mu_{32}^2 |c_2|^2} = 0, \quad \nu_{21} = \mu_{31} = \frac{3}{4}, \quad \nu_{31} = \mu_{21} = -\frac{1}{2} .$$

Damit können wir nun auch den dritten Vektor der neuen Orthogonalbasis berechnen als

$$d_3 = c_2 - \nu_{32} d_2 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} .$$

Wir ersetzen c_2, c_3 durch d_2, d_3 und die μ_{ij} durch die entsprechenden ν_{ij}, erniedrigen k auf eins und beginnen mit einem neuen Iterationsschritt.

Dieser beginnt mit der Reduktion von $\mu_{21} = \frac{3}{4}$. Nächste ganze Zahl ist eins, also setzen wir

$$b_2 \leftarrow b_2 - b_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \mu_{21} \leftarrow \mu_{21} - 1 = -\frac{1}{4} .$$

Um zu sehen, in welchem Fall wir sind, müssen wir das Längenquadrat zwei von

$$c_2 + \mu_{21} c_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$$

mit $\frac{3}{4} |c_1|^2 = 6$ vergleichen: Wir sind wieder in Fall 1 und müssen jetzt b_1 und b_2 miteinander vertauschen. Neuer erster Vektor der Orthogonalbasis wird $d_1 = c_2 + \mu_{21} c_1$; nach GRAM-SCHMIDT muß das natürlich der neue Vektor b_1 sein. Weiter ist

$$\nu_{21} = \mu_{21} \frac{|c_1|^2}{|c_2|^2 + \mu_{21}^2 |c_1|^2} = -1 \quad \text{und} \quad d_2 = c_1 - \nu_{21} d_1 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} .$$

Die restlichen ν_{ij} sind

$$\nu_{31} = \mu_{31}\nu_{21} + \mu_{23} \frac{|c_2|^2}{|d_1|^2} = \frac{1}{2} \quad \text{und} \quad \nu_{32} = \mu_{31} - \mu_{32}\mu_{21} = -\frac{1}{2} .$$

Wir ersetzen c_1, c_2 durch d_1, d_2 und die μ_{ij} durch ν_{ij}, setzen $k = 0$ und beginnen einen neuen Iterationsschritt.

Für $k = 0$ gibt es nichts zu tun, also können wir gleich wieder auf $k = 1$ erhöhen und einen neuen Schritt starten. Hier muß als erstes $\mu_{21} = -1$

reduziert und die Basis entsprechend angepaßt werden:

$$b_2 \leftarrow b_2 + b_1 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \quad \text{und} \quad \mu_{21} \leftarrow \mu_{21} + 1 = 0\,.$$

Da μ_{21} verschwindet, ist $c_2 + \mu_{21}c_1 = c_2 = b_2$, das alte b_1; sein Längenquadrat ist sechs und damit größer als $\frac{3}{4}\,|c_1|^2 = \frac{3}{2}$. Daher sind wir im Fall 2, wo es auch für $k = 1$ nichts zu tun gibt, wir können also gleich mit $k = 2$ weitermachen.

$\mu_{32} = \frac{1}{2}$ ist bereits klein genug, und

$$c_3 + \mu_{32}c_2 = \frac{1}{2}\begin{pmatrix} 3 \\ 3 \\ 6 \end{pmatrix}$$

hat Längenquadrat $\frac{27}{2}$, was größer ist als $\frac{3}{4}\,|c_2|^2 = \frac{9}{2}$. Daher sind wir wieder im Fall 2 und müssen uns daher nur noch um die restlichen μ_{3j} kümmern, d.h. um $\mu_{31} = \frac{1}{2}$. Dessen Betrag ist nicht größer als $\frac{1}{2}$, somit gibt es nichts mehr zu tun. Wir können also auf $k = 3$ erhöhen und der Algorithmus endet mit der LLL-reduzierten Basis aus

$$b_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad b_2 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \quad \text{und} \quad b_3 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}\,.$$

Die zugehörige Orthogonalbasis des $\mathbb{R}^3$ besteht aus $c_1 = b_1$ und $c_2 = b_2$ sowie

$$c_3 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} = b_3 - \frac{1}{2}c_1 + \frac{1}{2}c_2\,.$$

Nachdem wir nun gesehen haben, daß wir aus einer vorgegebenen Basis eine LLL-reduzierte Basis konstruieren können, stellt sich die Frage, was uns so eine Basis nützt bei der Suche nach kurzen Vektoren in einem Gitter. Unter den Vektoren einer LLL-reduzierten Basis muß kein kürzester Vektor des Gitters vorkommen, aber wir können immerhin obere Schranken finden für die Längen der Basisvektoren.

$(b_1, \ldots, b_n)$ sei eine LLL-reduzierte Basis eines Gitters $\Gamma \subset \mathbb{R}^n$ und $(c_1, \ldots, c_n)$ sei die dazu nach GRAM-SCHMIDT berechnete Orthogonalbasis. Dann ist

$$b_i = c_i + \sum_{j=1}^{i-1} \mu_{ij} c_j \Longrightarrow \left|b_i\right|^2 = \left|c_i\right|^2 + \sum_{j=1}^{i-1} \mu_{ij}^2 \left|c_j\right|^2 \leq \left|c_i\right|^2 + \frac{1}{4} \sum_{j=1}^{i-1} \left|c_j\right|^2,$$

denn für eine LLL-reduzierte Basis sind alle $\left|\mu_{ij}\right| \leq \frac{1}{2}$. Außerdem ist wegen der LOVÁSZ-Bedingung $\left|c_{j-1}\right|^2 \leq 2\left|c_j\right|^2$; mehrfache Anwendung dieser Ungleichung führt auf $\left|c_j\right|^2 \leq 2^{i-j} \left|c_i\right|^2$ für alle $j < i$ und

$$\begin{aligned}\left|b_i\right|^2 &\leq \left|c_i\right|^2 + \frac{1}{4} \sum_{j=1}^{i-1} 2^{i-j} \left|c_i\right|^2 = \left(1 + \frac{1}{4} \sum_{j=1}^{i-1} 2^j\right) \left|c_i\right|^2 = \frac{1 + 2^{i-1}}{2} \left|c_i\right|^2 \\ &\leq 2^{i-1} \left|c_i\right|^2 .\end{aligned}$$

Ebenfalls wegen der LOVÁSZ-Bedingung gilt für $j < i$ die Ungleichung $\left|c_j\right|^2 \leq 2^{i-j} \left|c_i\right|^2$ und damit auch

$$\left|b_j\right|^2 \leq 2^{j-1} \left|c_j\right|^2 \leq 2^{j-1} 2^{i-j} \left|c_i\right|^2 = 2^{i-1} \left|c_i\right|^2 .$$

Nun seien $v_1, \ldots, v_m$ irgendwelche linear unabhängige Vektoren aus dem Gitter Γ und k sei die kleinste Zahl mit der Eigenschaft, daß alle v_i im von b_1 bis b_k aufgespannten Teilgitter liegen. Dann gibt es Koeffizienten λ_{ij}, ν_{ij} derart, daß

$$v_i = \sum_{j=1}^{k} \lambda_{ij} b_j = \sum_{j=1}^{k} \nu_{ij} c_j$$

ist. Die λ_{ij} müssen dabei ganze Zahlen sein, die ν_{ij} natürlich nicht. Da wir die ν_{ij} aus den λ_{ij} berechnen können, indem wir die Darstellung der b_j als Linearkombinationen der c_j einsetzen, muß aber $\nu_{ik} = \lambda_{ik}$ und somit ganzzahlig sein, denn außer b_k liefert kein anderes b_j einen Beitrag mit c_k.

Wir wählen ein i, für das $\lambda_{ik} = \nu_{ik}$ nicht verschwindet; wegen der Minimalität von k muß es das geben. Dann ist

$$\left|b_k\right|^2 \leq 2^{k-1} \left|c_k\right|^2 \leq 2^{k-1} \lambda_{ik}^2 \left|c_k\right|^2 \leq 2^{k-1} \left|v_i\right|^2$$

und für $j < k$ ist

$$\left|b_j\right|^2 \leq 2^{k-1} \left|c_k\right|^2 \leq 2^{k-1} \left|v_i\right|^2 .$$

Über die Zahlen i und k wissen wir nur, daß k zwischen m und n liegen muß (sonst wären die v_j linear abhängig) und $i \leq m$. Daher haben wir für alle $j \leq m$ die Abschätzung

$$\left|b_j\right|^2 \leq 2^{n-1} \max\left\{\left|v_1\right|^2, \ldots, \left|v_m\right|^2\right\}$$

oder

$$\left|b_j\right| \leq 2^{(n-1)/2} \max\left\{\left|v_1\right|, \ldots, \left|v_m\right|\right\} .$$

Speziell für $m = 1$ erhalten wir die Abschätzung

$$\left|b_1\right| \leq 2^{(n-1)/2} \left|v_1\right|$$

für jeden vom Nullvektor verschiedenen Gittervektor v_1. Dies gilt insbesondere für den kürzesten solchen Vektor; die Länge von b_1 übersteigt dessen Länge also höchstens um den Faktor $2^{(n-1)/2}$.

§10: Anwendung auf Faktorisierungsprobleme

Wie in §8 betrachten wir wieder ein Polynom $f \in \mathbb{Z}[X]$ vom Grad d sowie ein Polynom $h \in \mathbb{Z}[X]$ vom Grad e mit führendem Koeffizienten eins, das modulo einer Primzahl p irreduzibel ist und modulo einer gewissen p-Potenz p^k Teiler von f. Wir nehmen außerdem an, daß h^2 modulo p kein Teiler von f mod p ist; wenn wir von einem quadratfreien Polynom f ausgehen und p die Diskriminante nicht teilt, ist letzteres automatisch erfüllt.

Nach dem ersten Lemma aus §8 hat f einen bis aufs Vorzeichen eindeutig bestimmten irreduziblem Faktor h_0, der modulo p durch h teilbar ist. Diesen Faktor wollen wir berechnen.

Dazu betrachten wir wieder das Gitter Λ aller Polynome aus $\mathbb{Z}[X]$ vom Grad höchstens einer gewissen Schranke m, die modulo p^k durch h teilbar sind; nach dem letzten Lemma aus §8 ist ein Polynom $v \in \Lambda$ mit $\|f\|_2 \cdot \|v\|_2 < p^{ke}$ ein Vielfaches von h_0. Wenn wir genügend viele

kurze Vektoren aus Λ finden, können wir daher hoffen, daß deren ggT gleich h_0 ist.

Als erstes wollen wir eine Schranke für die L^2-Norm eines Teilers von f finden. Aus den Überlegungen zur LANDAU-MIGNOTTE-Schranke in Kapitel 2, §8 folgt leicht

Lemma: Ist $g \in \mathbb{Z}[X]$ ein Teiler vom Grad e des Polynoms $f \in \mathbb{Z}[X]$, so ist

$$\|g\|_2 \leq \sqrt{\binom{2e}{e}}\, \|f\|_2 \ .$$

Beweis: Wenn g Teiler von f ist, muß auch der führende Koeffizient von g Teiler des führenden Koeffizienten von f sein. Daher ist das Maß $\mu(g)$ kleiner oder gleich $\mu(f)$, und letzteres wiederum ist nach Lemma 4 aus Kapitel 2, §8 kleiner oder gleich $\|f\|_2$. Nach dem dortigen Lemma 2 ist außerdem der Betrag des i-ten Koeffizienten von g kleiner oder gleich $\binom{e}{i}\mu(g)$, also kleiner oder gleich $\binom{e}{i}\|f\|_2$. Somit ist

$$\|g\|_2^2 \leq \sum_{i=0}^{e} \binom{e}{i}^2 \|f\|_2^2 \ .$$

Das Lemma ist bewiesen, wenn wir zeigen können, daß die Summe der $\binom{e}{i}^2$ gleich $\binom{2e}{e}$ ist. Dies läßt sich am einfachsten kombinatorisch einsehen: $\binom{2e}{e}$ ist die Anzahl von Möglichkeiten, aus einer Menge $\mathcal{M}$ mit $2e$ Elementen e auszuwählen. Wir zerlegen $\mathcal{M}$ in zwei disjunkte Teilmengen $\mathcal{M}_1$ und $\mathcal{M}_2$ mit je e Elementen. Die Wahl von e Elementen aus $\mathcal{M}$ ist gleichbedeutend damit, daß wir für irgendein i zwischen 0 und e zunächst i Elemente von $\mathcal{M}_1$ auswählen und dann $e-i$ Elemente aus $\mathcal{M}_2$. Die Anzahl der Möglichkeiten dafür ist $\binom{e}{i}\binom{e}{e-i} = \binom{e}{i}^2$, die Summe aller dieser Quadrate ist also $\binom{2e}{e}$.

■

Damit können wir nun zunächst eine Schranke für den Grad von h_0 finden:

Lemma: $(b_0, \ldots, b_m)$ sei eine LLL-reduzierte Basis des Gitters Λ und $p^{ke} < 2^{md/2} \binom{2m}{m}^{d/2} \|f\|_2^{m+d}$. Genau dann hat h_0 höchstens den Grad m, wenn

$$\|b_0\|_2 < \sqrt[d]{\frac{p^{ke}}{\|f\|_2^m}} \, .$$

Beweis: Falls b_0 diese Ungleichung erfüllt, ist $\|f\|_2^m \, \|b_0\|_2^d < p^{ke}$, nach dem letzten Lemma aus §8 ist also b_0 ein Vielfaches von h_0, und damit kann h_0 höchstens den Grad m haben.

Umgekehrt sei $\deg h_0 \leq m$. Nach der oben bewiesenen LANDAU-MIGNOTTE-Schranke für die L^2-Norm eines Teilers von f ist

$$\|h_0\|_2 \leq \sqrt{\binom{2m}{m}} \, \|f\|_2 \, .$$

Kombinieren wir dies mit der Ungleichung am Ende des vorigen Paragraphen, erhalten wir die Abschätzung

$$\|b_0\|_2 \leq 2^{m/2} \, \|h_0\|_2 \leq 2^{m/2} \sqrt{\binom{2m}{m}} \, \|f\|_2 \, .$$

Nach Voraussetzung ist

$$p^{ke} < 2^{md/2} \binom{2m}{m}^{d/2} \|f\|_2^{m+d} \Longrightarrow \frac{p^{ke}}{\|f\|_2^m} < 2^{md/2} \binom{2m}{m}^{d/2} \|f\|_2^d \, .$$

Ziehen wir auf beiden Seiten der letzten Ungleichung die d-te Wurzel und kombinieren dies mit der obigen Abschätzung für $\|b_0\|_2$, folgt die Behauptung.

∎

Lemma: Angenommen, zusätzlich zu den Voraussetzungen des vorigen Lemmas existieren Indizes j, für die

$$\|b_j\|_2 < \sqrt[d]{\frac{p^{ke}}{\|f\|_2^m}}$$

ist, und t ist der größte solche Index. Dann ist

$$\deg h_0 = m - t \quad \text{und} \quad h_0 = \text{ggT}(b_0, \ldots, b_t) .$$

Beweis: Wir betrachten die Menge J aller Indizes j mit $\|b_j\| < \sqrt[d]{\frac{p^{ke}}{\|f\|_2^m}}$. Das Lemma am Ende von §8 sagt uns, daß h_0 jedes b_j mit $j \in J$ teilt, also auch

$$h_1 = \text{ggT}\big(\{b_j \mid j \in J\}\big) .$$

Da jedes dieser b_j durch h_1 teilbar ist und höchstens den Grad m hat, liegt es im Gitter

$$\mathbb{Z} \cdot h_1 \oplus \mathbb{Z} \cdot X h_1 \oplus \cdots \oplus \mathbb{Z} \cdot X^{m-\deg h_1} h_1 .$$

Da die b_j als Elemente einer Basis linear unabhängig sind, ist die Elementanzahl von J höchstens gleich $m+1-\deg h_1$. Nach der zu Beginn dieses Paragraphen bewiesenen LANDAU-MIGNOTTE-Schranke für die L^2-Norm eines Teilers ist außerdem

$$\big\|X^i h_0\big\|_2 = \|h_0\|_2 \le \sqrt{\binom{2m}{m}}\, \|f\|_2$$

für jedes i. Da die verschiedenen $X^i h_0$ linear unabhängig sind, ist daher nach der Abschätzung am Ende von §8

$$\big\|b_j\big\|_2 \le 2^{m/2} \sqrt{\binom{2m}{m}}\, \|f\|_2 \quad \text{für } j = 0, \ldots, m - \deg h_0 .$$

Wegen der vom vorigen Lemma übernommenen Voraussetzung

$$p^{ke} > 2^{md/2} \binom{2m}{m}^{d/2} \|f\|_2^{m+d}$$

ist die rechte Seite kleiner als die d-te Wurzel aus $p^{ke} / \|f\|_2^m$, so daß alle $j \le m+1-\deg h_0$ in J liegen. J hat daher mindestens $m+1-\deg h_0$ Elemente und höchstens $m+1-\deg h_1$ Elemente. Da h_0 ein Teiler von h_1 ist, muß also $\deg h_0 = \deg h_1$ sein. Um zu sehen, daß sich die beiden Polynome höchstens durch das Vorzeichen unterscheiden, genügt es zu zeigen, daß auch h_1 primitiv ist.

h_1 ist der ggT von b_0 bis b_t; wäre h_1 nicht primitiv, könnten also auch diese b_j nicht primitiv sein. Wenn aber eine ganze Zahl d eines der Polynome b_j teilt, ist wegen der Primitivität von h_0 auch b_j/d ein Vielfaches von h_0, d.h. b_j/d liegt im Gitter Λ. Da $(b_0, \ldots, b_m)$ eine Gitterbasis ist, geht das nur für $d = \pm 1$. Also ist b_j und damit auch h_1 primitiv.

■

Damit ist klar, wie wir den Algorithmus von ZASSENHAUS zur Faktorisierung eines primitiven Polynoms $f \in \mathbb{Z}[X]$ vom Grad d so abändern können, daß nicht mehr im Extremfall alle Kombinationen der Faktorisierung modulo p miteinander kombiniert werden müssen: Die ersten drei Schritte des Algorithmus von ZASSENHAUS bleiben unverändert. Danach schreiben wir $f = f_1 f_2$ mit zwei Polynomen $f_1, f_2 \in \mathbb{Z}[X]$, wobei wir von f_1 die Faktorisierung in $\mathbb{Z}[X]$ kennen und von f_2 nur die modulo p. Zunächst ist natürlich $f_1 = 1$ und $f_2 = f$.

Solange f_2 positiven Grad hat, betrachten wir einen der irreduziblen Faktoren von $f_2 \bmod p$ aus $\mathbb{F}_p[X]$; sein Grad sei e. Mit Hilfe des HENSELschen Lemmas liften wir den Faktor zu einem Polynom $h \in \mathbb{Z}[X]$, der auch noch modulo einer p-Potenz $p^k \geq M$ Teiler von f_2 ist.

Wir wählen einen Grad $m \geq e$ und wollen entweder einen modulo p durch h teilbaren irreduziblen Faktor h_0 von f_2 mit Grad höchstens m konstruieren oder aber beweisen, daß es keinen solchen Faktor gibt.

Dazu stellen wir zunächst sicher, daß

$$p^{ke} > 2^{md/2} \binom{2m}{m}^{d/2} \|f\|_2^{m+d}$$

ist und betrachten dann zum Gitter Λ mit Basis

$$p^k X^i \quad \text{für } 0 \leq i < e \quad \text{und} \quad X^j h \quad \text{für } 0 \leq j \leq m - e$$

die LLL-Reduktion $b_0, \ldots, b_m$ dieser Basis. Falls

$$\|b_0\|_2 \geq \sqrt[d]{\frac{p^{ke}}{\|f\|_2^m}},$$

gibt es keinen Faktor h_0 vom Grad höchstens m. Wenn $m \geq [\frac{1}{2} \deg f_2]$ ist, wissen wir, daß f_2 irreduzibel, also $h_0 = f_2$ ist; andernfalls müssen wir entweder m erhöhen oder einen anderen irreduziblen Faktor von f_2 mod p betrachten.

Wenn obige Ungleichung nicht gilt, wählen wir für t den größten Index j mit

$$\|b_j\|_2 < \sqrt[d]{\frac{p^{ke}}{\|f\|_2^m}}$$

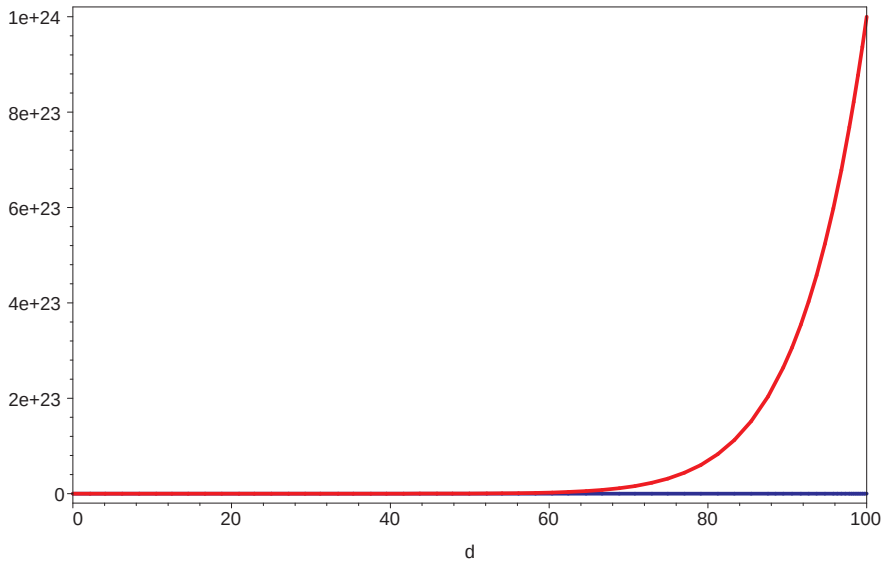
und können h_0 berechnen als ggT von $b_0, \ldots, b_t$.

Wir ersetzen dann f_1 durch $f_1 h_0$ und f_2 durch f_2/h_0. Zur Faktorisierung des neuen f_2 modulo p brauchen wir natürlich keinen BERLEKAMP-Algorithmus, sondern können einfach testen, welche Faktoren des alten f_2 mod p in h_0 mod p (oder im neuen f_2 mod p) stecken.

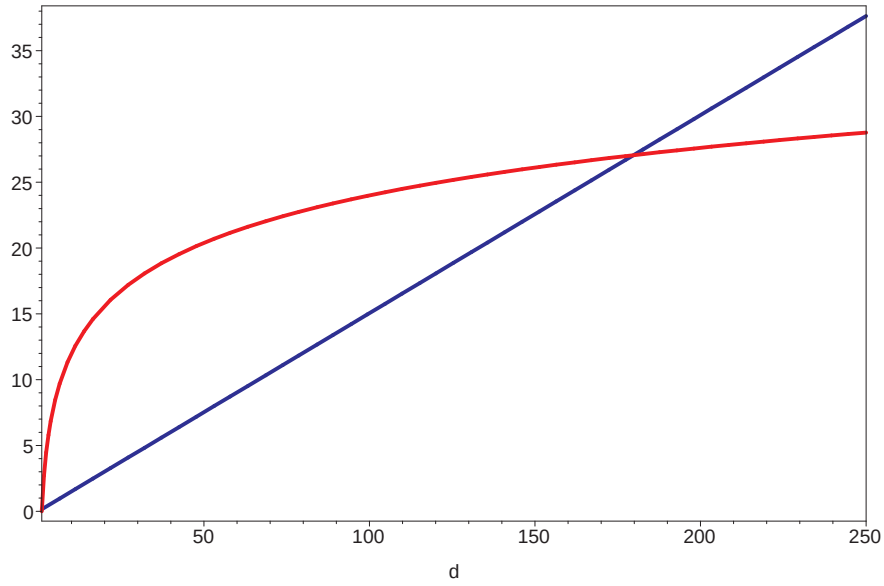
Was haben wir durch diese mathematisch deutlich kompliziertere Modifikation gewonnen? Theoretisch sehr viel: Wie das Beispiel der SWINNERTON-DYER-Polynome zeigte, kann es uns im sechsten Schritt des Algorithmus von ZASSENHAUS passieren, daß wir bei der Faktorisierung eines Polynoms vom Grad d bis zu $2^{[d/2]}$ Kombinationen ausprobieren müssen, bevor wir erkennen, daß das zu faktorisierende Polynom irreduzibel ist. Alle anderen Schritte erfordern einen Zeitaufwand, der als Funktion von n sehr viel langsamer ansteigt als $2^{[d/2]}$, so daß asymptotisch betrachtet dieser Term dominiert.

Ein guter Teil der zitierten Arbeit von LENSTRA, LENSTRA und LOVASZ widmet sich der Frage, wie groß der entsprechende Aufwand mit LLL-Reduktion ist; sie zeigen, daß der dominierende Term hier nur d^{12} ist, was – wie aus der Analysis bekannt – deutlich schwächer ansteigt.

Läßt man zur Illustration die beiden Kurven im Bereich von $d = 0$ bis $d = 100$ von Maple zeichnen, so ist eine der beiden praktisch ununterscheidbar von der d-Achse, während die andere steil ansteigt:



Nachrechnen zeigt allerdings, daß die steil ansteigende rote Kurve der Graph von $d \mapsto d^{12}$ ist: Für $d = 100$ etwa ist $2^{50} \approx 1{,}125899907 \cdot 10^{15}$ und $50^{12} \approx 2{,}441406250 \cdot 10^{20}$ ist mehr als 200 000 mal so groß. Erst zwischen $d = 179$ und $d = 180$ schneiden sich die beiden Kurven, und ab dort dominiert dann die Exponentialkurve. Damit man etwas sehen kann, sind in der folgenden Abbildung die (dekadischen) Logarithmen der beiden Funktionen aufgezeichnet. Wie man sieht, liegt im unteren Bereich, mit dem wir es meist zu tun haben, die Kurve zu d^{12} deutlich über der für $2^{d/2}$; erst ab $d = 180$ wird $2^{d/2}$ größer.



Für Polynome in den Größenordnungen, mit denen wir es üblicher-

weise zu tun haben, sollte also der klassische ZASSENHAUS-Algorithmus schneller sein. Theoretisch könnte man den Exponenten 12 noch für jedes $\varepsilon > 0$ auf $9 + \varepsilon$ reduzieren, indem man asymptotisch schnellere Algorithmen zur Multiplikation ganzer Zahlen einsetzt, in der Praxis wird der Algorithmus dadurch allerdings deutlich langsamer: Die entsprechenden Methoden sind zwar nützlich für Zahlen mit Millionen von Dezimalstellen, nicht aber bei „nur“ ein paar hundert oder Tausend.

Man muß auch bedenken, daß sich alle hier angegebenen Schranken auf den schlechtestmöglichen Fall beziehen, der nur selten eintritt. In der Praxis sind beide Algorithmen deutlich schneller als es die asymptotischen Schranken vermuten lassen.

Maple benutzt anscheinend zur Faktorisierung keine Gittermethoden, obwohl die LLL-Reduktion für Gitterbasen als Funktion `lattice` zur Verfügung steht. Bislang scheint Faktorisierung mit LLL auf experimentelle zahlentheoretische Systeme beschränkt zu sein, in denen nicht über $\mathbb{Q}$, sondern über einem Erweiterungskörper faktorisiert wird. LLL-Basisreduktion hat heute Anwendungen in vielen Teilen der Mathematik, unter anderem bei der ganzzahligen Optimierung, der Lösung diophantischer Gleichungssysteme, in der Kryptographie und vor allem auch in der Kryptanalyse. bei der ursprünglich vorgesehenen Anwendung der Faktorisierung von Polynomen aus $\mathbb{Z}[X]$ spielt er aber bislang in der Praxis nur eine ziemlich untergeordnete Rolle,

§11: Das Henselsche Lemma für Polynomringe

Bei der Berechnung des größten gemeinsamen Teilers zweier Polynome in mehreren Veränderlichen hatten wir das Problem zurückgeführt auf die ggT-Berechnung in einer Veränderlichen weniger. Diese Strategie funktioniert hier nicht, da wir für eine Liftung nach dem HENSELschen Lemma den erweiterten EUKLIDischen Algorithmus benötigen, und der steht nur für EUKLIDische Ringe zur Verfügung. Unter den Polynomringen sind nur die in einer Veränderlichen über einem Körper EUKLIDisch; wenn wir den erweiterten EUKLIDischen Algorithmus anwenden wollen, müssen wir also in einem Schritt von n Veränderlichen absteigen zu einer

Veränderlichen, und selbst dann haben wir im Falle ganzzahliger Polynome das Problem, das wir im vorigen Kapitel unbedingt vermeiden wollten: Wir müssen den erweiterten EUKLIDischen Algorithmus anwenden auf Polynome aus $\mathbb{Z}[X]$, und bekommen dabei im allgemeinen Koeffizienten aus $\mathbb{Q}[X]$ mit sehr schnell ansteigenden Nennern.

Der erste auch für nicht triviale Probleme einsetzbare Algorithmus zur Faktorisierng multivariater Polynome wurde 1975 veröffentlicht und folgte weitgehend dem Algorithmus von ZASSENHAUS, den wir in den vorigen Paragraphen betrachteten:

PAUL S. WANG, LINDA PREIS ROTHSCHILD: Factoring Multivariate Polynomials Over the Integers, *Math. Comp.* **29**, 935–950

(Die Zeitrschrift *Mathematics of Computation* der *American Mathematical Society* wird von der Mannheimer Universitätsbibliothek abonniert und im Uninetz auch digital zu Verfügung gestellt.)

In den fast vierzig Jahren seit Erscheinen dieser Arbeit gab es natürlich gewaltige Forschritte auch bei Faktorisierungsalgorithmen für multivariate Polynome, genau wie es auch bei der Berechnung von GRÖBNER-Basen gewaltige Fortschritte gegenüber dem einfachen BUCHBERGER-Algorithmus gab; auch die neueren Algorithmen bauen aber auf auf dem, was WANG und ROTHSCHILD den *extended* ZASSENHAUS-Algorithmus nannten. Inzwischen wird ihre Grundidee sogar gelegentlich als Alternative zur modularen Methode bei der ggT-Berechnung für multivariate Polynome benutzt. Für diese einführende Vorlesung möchte ich mich daher auf ihren Algorithmus beschränken.

Beim zu faktorisierenden Polynom $f \in R = \mathbb{Z}[X_1, \ldots, X_n]$ entscheiden sie sich zunächst für eine *Hauptvariable* X_i; für uns sei das o.B.d.A. die Variable X_1. Für alle anderen Variablen werden Werte eingesetzt; wir betrachten also das Polynom

$$f(X_1, a_2, \ldots, a_n] \in \mathbb{Z}[X_1] \,.$$

Es ist kongruent zu f modulo dem Ideal

$$I = (X_2 - a_2, \ldots, X_n - a_n) \lhd R$$

und kann mit den Methoden aus den vorigen Abschnitten faktorisiert werden.

Eine solche Faktorisierung soll dann schrittweise hochgehoben werden zu einer Faktorisierung modulo immer höherer Potenzen von I. Da die Potenzen von I schnell unübersichtlich werden, müssen wir dazu zunächst unser Koordinatensystem geeignet wählen: Wir betrachten für $i = 2, \ldots, n$ die neuen Variablen $Y_i = X_i - a_i$ und das Polynom

$$F(X_1, Y_2, \ldots, Y_n) = f(X_1, Y_2 + a_2, \ldots, Y_n + a_n) \in \mathbb{Z}[X_1, Y_2, \ldots, Y_n] .$$

In den neuen Koordinaten ist $I = (Y_2, \ldots, Y_n)$, und die Potenzen I^k werden erzeugt von den Monomen vom Grad k in den Y_i.

Das HENSELschen Lemma für diese Situation besagt

Lemma: $g, h \in \mathbb{Z}[X_1]$ seien zwei teilerfremde Polynome derart, daß $f(X_1, a_2, \ldots, a_n) = g(X_1) \cdot h(X_1)$. Dann gibt es zu jedem $k \in \mathbb{N}$ Polynome $G_k, H_k \in \mathbb{Z}[X_1, \ldots, X_n]$ derart, daß

$$f(X_1, \ldots, X_n) \equiv G_k(X_1, \ldots, X_n) \cdot H_k(X_1, \ldots, X_n) \bmod I^k$$

und $G_k(X_1, a_2, \ldots, a_n) = g(X_1)$ und $H_k(X_1, a_2, \ldots, a_n) = h(X_1)$.

Beweis: Für $n = 1$ setzen wir einfach $G_1 = g$ und $H_1 = h$. Mit den neuen Koordinaten $Y_2, \ldots, Y_n$ ist dann $F \equiv G_1 H_1 \bmod I$.

Da g und h teilerfremd sind, gibt es für jedes $i \in \mathbb{N}_0$ zwei Polynome $\alpha_i, \beta_i \in \mathbb{Q}[X_1]$ derart, daß

$$\deg \alpha_i < \deg h, \quad \deg \beta_i < \deg g \quad \text{und} \quad \alpha_i g + \beta_i h = X^i .$$

Da die Berechnung dieser Polynome wegen der zu erwartenden Nenner aufwendig sein kann, werden die Polynome α_i, β_i nur dann berechnet, wenn sie wirklich gebraucht werden; die berechneten Polynome werden dann zur weiteren Verwendung gespeichert.

Nach Voraussetzung ist $F - G_1 H_1 \in I$.

§ 12: Faktorisierung von Polynomen mehrerer Veränderlicher

Wie beim ggT können wir uns auch bei der Faktorisierung von Polynomen in mehrerer Veränderlichen anlehnen an die in den vorigen Paragraphen betrachtete Vorgehensweise für Polynome einer Veränderlichen über $\mathbb{Z}$. Eine einfache rekursive Vorgehensweise wäre die folgende:

Wir fassen den Polynomring $R_n = k[X_1, \ldots, X_n]$ in n Veränderlichen über einem Ring oder Körper k auf als Polynomring in der einen Veränderlichen X_1 über $k[X_2, \ldots, X_n]$ und führen die Faktorisierung eines Polynoms eines Polynoms in n Variablen wie folgt zurück auf Faktorisierungsprobleme in $n-1$ Variablen:

Erster Schritt: Berechne den Inhalt des Polynoms über $k[X_1, \ldots, X_{n-1}]$. Da dieser ein Polynom in $n-1$ Veränderlichen ist, kann er faktorisiert werden. Für die folgenden Schritte können wir daher annehmen, daß das zu faktorisierende Polynom primitiv ist.

Zweiter Schritt: Wir setzen für eine der übrigen Variablen, beispielsweise X_{n-1}, einen festen Wert $c \in k$ ein derart, daß der Koeffizient der führenden X_n-Potenz dabei nicht verschwindet. Dadurch erhalten wir ein Polynom in $n-1$ Veränderlichen, das wir faktorisieren können.

Dritter Schritt: Die Faktorisierung wird nach einem Analogon des HENSELschen Lemmas hochgehoben zu einer Faktorisierung modulo $(X_{n-1} - c)^d$, wobei d den Grad des zu faktorisierenden Polynoms in der Variablen X_{n-1} bezeichnet.

Vierter Schritt: Setze $m = 1$ und teste für jeden der gefundenen Faktoren, ob er das zu faktorisierende Polynom teilt. Falls ja, kommt er in die Liste $\mathcal{L}_1$ der Faktoren, andernfalls in eine Liste $\mathcal{L}_2$.

Fünfter Schritt: Falls die Liste $\mathcal{L}_2$ keine Einträge hat, endet der Algorithmus, und das Polynom ist das Produkt der Faktoren aus $\mathcal{L}_1$. Andernfalls setzen wir $m = m + 1$ und testen für jedes Produkt aus m verschiedenen Polynomen aus $\mathcal{L}_2$, ob ihr Produkt modulo $(X_{n-1} - c)^d$ ein Teiler von g ist. Falls ja, entfernen wir die m Faktoren aus $\mathcal{L}_2$ und fügen ihr Produkt in die Liste $\mathcal{L}_1$ ein. Dieser Schritt wird wiederholt, bis die Liste $\mathcal{L}_2$ leer ist.

Kapitel 5
Gröbner-Basen

Die klassische Aufgabe der Algebra besteht in der Lösung von Gleichungen und Gleichungssystemen. Im Falle eines Systems von Polynomgleichungen in mehreren Veränderlichen kann die Lösungsmenge sehr kompliziert sein und, sofern sie unendlich ist, möglicherweise nicht einmal explizit angebbar: Im Gegensatz zum Fall linearer Gleichungen können wir hier im allgemeinen keine endliche Menge von Lösungen finden, durch die sich alle anderen Lösungen ausdrücken lassen. Trotzdem gibt es Algorithmen, mit denen sich nichtlineare Gleichungssysteme deutlich vereinfachen lassen, und zumindest bei endlichen Lösungsmengen lassen sich diese auch konkret angeben – sofern wir die Nullstellen von Polynomen einer Veränderlichen explizit angeben können.

§1: Algebraische Vorbereitungen

Wenn wir lineare Gleichungssysteme mit dem GAUSS-Algorithmus lösen, verändern wir das Gleichungssystem sukzessive, indem wir Gleichungen so durch Linearkombinationen mit anderen Gleichungen ersetzen, daß sich an der Lösungsmenge nichts ändert. Indem wir eine lineare Gleichung

$$a_1X_1 + \cdots + a_nX_n = b$$

über einem Körper k mit dem $(n+1)$-Tupel $(a_1, \ldots, a_n, b) \in k^{n+1}$ identifizieren, sehen wir leicht, daß die sämtlichen linearen Gleichungen in n Unbekannten über einem Körper k einen $(n+1)$-dimensionalen Vektorraum bilden; die Gleichungen eines konkreten linearen Gleichungssystems erzeugen darin einen Untervektorraum. Dieser besteht aus allen

Linearkombinationen der gegebenen Gleichungen, und das sind gleichzeitig alle linearen Gleichungen, die auf der Lösungsmenge des linearen Gleichungssystems verschwinden. Zwei lineare Gleichungssysteme haben somit genau dann die gleiche Lösungsmenge, wenn sie den gleichen Untervektorraum erzeugen.

Wenn wir Systeme nichtlinearer Gleichungen betrachten, ist es sinnvoll, die Menge aller möglicher Gleichungen nicht mehr nur als Vektorraum zu betrachten, sondern auch die Multiplikation mit Polynomen zuzulassen: Zur Lösung des Gleichungssystems

$$X^2Y^2 + 2X^3 - 3X^2 - X = 0 \quad \text{und} \quad Y^2 + X - 3 = 0$$

bietet sich etwa an, die zweite Gleichung mit X^2 zu multiplizieren und das Produkt $X^2Y^2 + X^3 - 3X^2 = 0$ von der ersten Gleichung zu subtrahieren; die Differenz $X^3 - X$ hängt nur noch von X ab und verschwindet bei 0 und ± 1. Setzen wir dies in die zweite Gleichung ein, erhalten wir die Lösungsmenge

$$\left\{ \left(0, \sqrt{3}\right), \left(0, -\sqrt{3}\right), \left(1, \sqrt{2}\right), \left(1, -\sqrt{2}\right), (-1, 2), (-1, -2) \right\}.$$

Wir sollten die Menge aller möglicher Gleichungen daher nicht mehr nur als einen Vektorraum betrachten, sondern als einen *Ring,* und wir sollten nicht nur skalare Linearkombinationen der Ausgangsgleichungen betrachten, sondern solche mit beliebigen Polynomen als Koeffizienten. Dies führt zum Begriff des *Ideals:*

Definition: Eine nichtleere Teilmenge I eines Rings R heißt *Ideal,* in Zeichen $I \lhd R$, wenn gilt:
1.) Für je zwei Elemente $f, g \in I$ ist auch $f + g \in I$
2.) Für jedes $f \in I$ und jedes $r \in R$ liegt auch rf in I.

Bei den Produkten verlangen wir also, daß sie bereits dann in I liegen, wenn nur *ein* Faktor in I liegt.

Die Bedingung, daß ein Ideal mindestens ein Element enthalten muß, können wir auch ersetzen durch die Bedingung, daß es die Null von R enthalten muß, denn wenn es irgendein Element $f \in R$ enthält, muß es gemäß der zweiten Bedingung auch $0 \cdot f = 0$ enthalten.

Um mit dem Idealbegriff vertraut zu werden, betrachten wir zunächst Ideale im Ring der ganzen Zahlen:

Lemma: Zu jedem Ideal $I \lhd \mathbb{Z}$ gibt es eine ganze Zahl $n \in \mathbb{Z}$, so daß $I = \{nq | q \in \mathbb{Z}\}$.

Beweis: I ist nach Definition nicht leer, enthält also mindestens ein Element. Falls I nur aus der Null besteht, können wir $n = 0$ setzen und sind fertig. Wenn es ein Element $m \neq 0$ gibt, enthält das Ideal auch dessen sämtliche ganzzahlige Vielfachen, insbesondere also gibt es in I dann positive Zahlen. Die kleinste dieser Zahlen sei n. Wir wollen uns überlegen, daß I genau aus den ganzzahligen Vielfachen von n besteht.

Dazu sei $m \in I$ ein beliebiges Element von I. Wir dividieren m mit Rest durch n; das Ergebnis sei

$$m : n = q \quad \text{Rest } r \quad \text{mit} \quad 0 \leq r < n \,.$$

Dann liegt mit m und n auch $r = m - qn$ in I und ist echt kleiner als n. Da n die kleinste positive Zahl in I ist, muß daher $r = 0$ sein, d.h. $m = qn$ ist ein ganzzahliges Vielfaches von n.

■

Definition: *a)* Ist R ein Ring und $f \in R$ so bezeichnen wir

$$(f) \underset{\text{def}}{=} \{rf \mid r \in R\}$$

als das von f erzeugte *Hauptideal.*
b) R heißt *Hauptidealring,* wenn jedes Ideal von R ein Hauptideal ist.

Das gerade bewiesene Lemma zeigt also, daß $\mathbb{Z}$ ein Hauptidealring ist.

Allgemeiner definieren wir

Definition: Ist R ein Ring und ist $M \subset R$ eine Teilmenge von R, so ist das *von M erzeugte Ideal* (M) das kleinste Ideal von R, das M enthält, d.h. der Durchschnitt aller Ideale, die M enthalten. Für eine endliche Menge $M = \{f_1, \ldots, f_m\}$ schreiben wir (M) kurz als $(f_1, \ldots, f_m)$. Die Menge M bezeichnen wir als ein *Erzeugendensystem* des Ideals I.

Diese Definition macht nicht wirklich klar, wie das von M erzeugte Ideal aussieht. Da uns in der Computeralgebra nur endlich erzeugte Ideale interessieren (und wir bald auch sehen werden, daß jedes Ideal im Polynomring $k[X_1, \ldots, X_n]$ über einem Körper ein endliches Erzeugendensystem hat), möchte ich mich auf diesen Fall beschränken. Die Verallgemeinerung auf beliebige Mengen M sollte für jeden, der den folgenden Beweis verstanden hat, offensichtlich sein.

Lemma: $\left(f_1, \ldots, f_m\right) = \left\{ \sum_{i=1}^{m} r_i f_i \;\middle|\; r_i \in R \right\}$

Beweis: Da jedes Ideal, das $f_1, \ldots, f_m$ enthält, auch für $r_1, \ldots, r_m \in R$ die Elemente $r_i f_i$ enthält und damit auch deren Summe, ist klar, daß die rechte Seite in jedem Ideal enthalten ist, das die f_i enthält. Außerdem ist die rechtsstehende Menge selbst ein Ideal: Da sie die f_i enthält, ist sie nicht leer; die Summe zweier Elemente ist offensichtlich wieder ein Element, da wir einfach die Koeffizienten addieren müssen, und wenn wir ein Element mit einem beliebigen Element $r \in R$ multiplizieren, werden einfach alle Koeffizienten mit r multipliziert. Somit ist die rechte Seite in der Tat das kleinste Ideal, das alle f_i enthält. ∎

Sei nun $R = k[X_1, \ldots, X_n]$ der Polynomring in n Variablen über einem Körper k,und $f_1, \ldots, f_m \in R$ seien Polynome. Wir interessieren uns für die Lösungsmenge des durch die f_i gegebenen Gleichungssystems, also die Menge aller $(x_1, \ldots, x_n) \in k^n$, für die alle f_i verschwinden. Wir definieren gleich allgemein

Definition: Die Nullstellenmenge einer Teilmenge $M \subseteq k[X_1, \ldots, X_n]$ ist

$$V(M) \underset{\text{def}}{=} \left\{ (x_1, \ldots, x_n) \in k^n \;\middle|\; f(x_1, \ldots, x_n) = 0 \text{ für alle } f \in M \right\}.$$

Im Falle eine endlichen Menge $M = \{f_1, \ldots, f_m\}$ schreiben wir kurz $V(f_1, \ldots, f_m)$.

(In der algebraischen Geometrie bezeichnet man Mengen dieser Art als Varietäten; daher der Buchstabe V.)

Lemma: Ist $I = (f_1, \ldots, f_m)$ das von den f_i erzeugte Ideal, so ist

$$V(I) = V(f_1, \ldots, f_m)\,.$$

Beweis: Da alle f_i in I liegen, ist natürlich $V(I) \subseteq V(f_1, \ldots, f_m)$. Umgekehrt sei $(x_1, \ldots, x_n)$ ein Element von $V(f_1, \ldots, f_m)$ und g irgendein Element von I. Nach dem vorigen Lemma gibt es Polynome $r_i \in R$; so daß $g = \sum_{i=1}^{m} r_i f_i$ ist. Damit ist auch

$$g(x_1, \ldots, x_n) = \sum_{i=1}^{m} r_i(x_1, \ldots, x_m) f_i(x_1, \ldots, x_m) = 0\,,$$

so daß $(x_1, \ldots, x_n)$ in $V(I)$ liegt. Damit ist das Lemma bewiesen. ∎

Dieses Lemma zeigt, daß zwei Gleichungssysteme

$$f_1(x_1, \ldots, x_n) = 0, \quad \ldots, \quad f_m(x_1, \ldots, x_n) = 0$$

und

$$g_1(x_1, \ldots, x_n) = 0, \quad \ldots, \quad g_r(x_1, \ldots, x_n) = 0$$

die gleiche Lösungsmenge haben, wenn die Ideale $(f_1, \ldots, f_m)$ und $(g_1, \ldots, g_r)$ übereinstimmen.

Die Umkehrung dieser Aussage ist allerdings falsch. Ein einfaches Gegenbeispiel haben wir bereits bei nur einer Gleichung in einer Variablen: Die Gleichungen

$$x = 0, \qquad x^2 = 0, \qquad x^3 = 0, \qquad \ldots$$

haben allesamt nur die Null als Lösung, aber natürlich sind die Ideale $\left(x^d\right) \lhd k[X]$ für verschiedene Werte von d verschieden. Auch gibt es beispielsweise in $\mathbb{Q}[X]$ und auch in $\mathbb{R}[X]$ zahlreiche Polynome f mit $V(f) = \emptyset$, und natürlich erzeugen die nicht alle das gleiche Hauptideal.

Zum Abschluß dieses Paragraphen soll noch kurz festgehalten werden, wie sich Ideale und Nullstellenmengen zueinander verhalten. Dazu müssen wir zunächst die Summe und das Produkt zweier Ideale definieren:

Definition: *a)* Die Summe $I + J$ zweier Ideale I, J eines Rings R ist das kleinste Ideal, das sowohl I als auch J enthält.

b) Das Produkt IJ dieser Ideale ist das kleinste Ideal, das alle Produkte fg mit $f \in I$ und $g \in J$ enthält.

Man überlegt sich leicht (mit dem gleichen Argument, mit dem wir das Ideal $(f_1, \ldots, f_m)$ oben explizit bestimmt haben), daß $I + J$ gerade die Menge aller $f + g$ mit $f \in I$ und $g \in J$ ist. IJ dagegen enthält im allgemeinen auch Elemente, die sich *nicht* in der Form fg mit $f \in I$ und $g \in J$ schreiben lassen: Ist etwa $I = J = (X, Y) \lhd \mathbb{R}[X, Y]$, so enthält IJ mit $X^2 = X \cdot X$ und $Y^2 = Y \cdot Y$ auch deren Summe $X^2 + Y^2$, die sich nicht als Produkt zweier Polynome aus $\mathbb{R}[X, Y]$ schreiben läßt. Wenn wir $\mathbb{R}$ durch $\mathbb{C}$ ersetzen, läßt sich $X^2 + Y^2$ zwar zerlegen als $(X + iY)(X - iY)$, aber auch in $\mathbb{C}[X, Y]$ gibt es in $(X, Y) \cdot (X, Y)$ irreduzible Polynome, und die lassen sich natürlich nicht als Produkte darstellen. Um IJ als Menge zu beschreiben, müssen wir daher auch alle (endlichen) Summen der Form $\sum f_i g_i$ mit $f_i \in I$ und $g_i \in J$ betrachten. Die Gesamtheit dieser Summen bildet offensichtlich ein Ideal, und somit besteht IJ genau aus diesen Summen.

Satz: Für zwei Ideale I, J im Polynomring $R = k[X_1, \ldots, X_n]$ gilt

a) Ist $I \subseteq J$, so ist $V(J) \subseteq V(I)$

b) $V(I + J) = V(I) \cap V(J)$

c) $V(IJ) = V(I) \cup V(J)$

Beweis: a) Sei $(x_1, \ldots, x_n) \in V(J)$. Dann verschwindet $f(x_1, \ldots, x_n)$ für alle $f \in J$, erst recht also für alle $f \in I$, d.h. $(x_1, \ldots, x_n) \in V(I)$.

b) Da $I + J$ das kleinste Ideal ist, das sowohl I als auch J enthält, liegt $V(I + J)$ nach *a)* sowohl in $V(I)$ als auch in $V(J)$, also auch in deren Durchschnitt. Liegt umgekehrt ein Punkt $(x_1, \ldots, x_n)$ sowohl in $V(I)$ als auch in $V(J)$, so liegt er auch in $V(I + J)$, denn wie wir gerade gesehen haben, läßt sich jedes Element von $I + J$ schreiben als $f + g$ mit $f \in I$ und $g \in J$, und sowohl f als auch g verschwinden im Punkt $(x_1, \ldots, x_n)$.

c) Da IJ erzeugt wird von den Produkten fg mit $f \in I$ und $g \in J$ und jedes dieser Produkte sowohl in I als auch in J liegt, ist IJ eine Teilmenge sowohl von I als auch von J. Somit liegt $V(I) \cup V(J)$ nach *a)* in $V(IJ)$. Umgekehrt sei $(x_1, \ldots, x_n) \in V(IJ)$, liege aber nicht in

$V(I)$. Dann gibt es ein $f \in I$ mit $f(x_1, \ldots, x_n) \neq 0$. Für jedes $g \in J$ liegt aber fg in IJ, so daß das Produkt $f(x_1, \ldots, x_n)g(x_1, \ldots, x_n)$ verschwinden muß. Da die Funktionswerte im Körper k liegen und der Faktor $f(x_1, \ldots, x_n)$ nicht verschwindet, muß $g(x_1, \ldots, x_n) = 0$ sein für alle $g \in J$; der Punkt liegt also in $V(J)$. Somit liegt er in jedem Fall in $V(I) \cup V(J)$.

■

§2: Gauß und Euklid

Zur (exakten) Lösung eines linearen Gleichungssystems in mehreren Veränderlichen verwenden wir üblicherweise den GAUSS-Algorithmus. Für die Lösung eines System von Polynomgleichungen höheren Grades in nur einer Veränderlichen können wir den EUKLIDischen Algorithmus verwenden, denn die gemeinsamen Nullstellen zweier Polynome in einer Veränderlichen sind gerade die Nullstellen ihres größten gemeinsamen Teilers, so daß wir das System durch mehrfache Anwendung des EUKLIDischen Algorithmus reduzieren können auf eine einzige Polynomgleichung.

Der um 1966 von BRUNO BUCHBERGER vorgestellte Ansatz zur Lösung nichtlinearer Gleichungssysteme in mehreren Veränderlichen kann als eine Kombination von Ideen hinter dem GAUSSschen Eliminationsverfahren und dem EUKLIDischen Algorithmus aufgefaßt werden; er hat Anwendungen, die weit über das Problem der Lösung nichtlinearer Gleichungssysteme hinausgehen. In der Tat wurde die Grundidee des Verfahrens bereits knapp vor BUCHBERGER, und ohne daß dieser davon wußte, von dem japanischen Mathematiker HEISUKE HIRONAKA entdeckt, der es für ein klassisches Problem der algebraischen Geometrie entwickelte: Für die damit bewiesene sogenannte Auflösung der Singularitäten einer algebraischen Varietät über einem Körper der Charakteristik Null erhielt HIRONAKA 1970 die Fields-Medaille, die damals höchste Auszeichnung der Mathematik. (Seit 2003 gibt es zusätzlich den von der norwegischen Akademie der Wissenschaften vergebenen und wie ein Nobelpreis dotierten Abelpreis. Während die Fields-Medaille satzungsgemäß an jüngere Mathematiker, typischerweise unter vierzig, geht und

nur gering dotiert ist, würdigt der Abelpreis das Lebenswerk eines Mathematikers, so daß die Preisträger meist recht alt sind.)

Wenn wir ein lineares Gleichungssystem durch GAUSS-Elimination lösen, bringen wir es zunächst auf eine Treppengestalt, indem wir die erste vorkommende Variable aus allen Gleichungen außer der ersten eliminieren, die zweite aus allen Gleichungen außer den ersten beiden, und so weiter, bis wir schließlich Gleichungen haben, deren letzte entweder nur eine Variable enthält oder aber eine Relation zwischen Variablen, für die es sonst keine weiteren Bedingungen mehr gibt. Konkret sieht ein Eliminationsschritt folgendermaßen aus: Wenn wir im Falle der beiden Gleichungen

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = u \quad \text{mit} \quad a_1 \neq 0 \tag{1}$$

$$b_1x_1 + b_2x_2 + \cdots + b_nx_n = v \tag{2}$$

die Variable x_1 mit Hilfe von (1) aus (2) eliminieren wollen, ersetzen wir die zweite Gleichung durch ihre Summe mit $-b_1/a_1$ mal der ersten. Die theoretische Rechtfertigung für diese Umformung besteht darin, daß das Gleichungssystem bestehend aus (1) und (2) sowie das neue Gleichungssystem dieselbe Lösungsmenge haben, und daran ändert sich auch dann nichts, wenn noch weitere Gleichungen dazukommen.

Ähnlich können wir vorgehen, wenn wir ein nichtlineares Gleichungssystem in nur einer Variablen betrachten: Am schwersten sind natürlich die Gleichungen vom höchsten Grad, also versuchen wir, die zu reduzieren auf Polynome niedrigeren Grades. Das kanonische Verfahren dazu ist die Polynomdivision: Haben wir zwei Polynome

$$f = a_dX^d + a_{d-1}X^{d-1} + \cdots + a_1X + a_0 \quad \text{und}$$

$$g = b_eX^e + b_{e-1}X^{e-1} + \cdots + b_1X + b_0$$

mit $e \leq d$, so dividieren wir f durch g, d.h. wir berechnen einen Quotienten q und einen Rest r derart, daß $f = qg + r$ ist und r entweder verschwindet oder kleineren Grad als g hat. Konkret: Bei jedem Divisionsschritt haben wir ein Polynom

$$f = c_\delta X^\delta + c_{\delta-1}X^{\delta-1} + \cdots + c_1X + c_0 \quad \text{mit} \quad c_\delta \neq 0\,,$$

das wir für $\delta \geq e$ mit Hilfe des Divisors

$$g = b_e X^e + b_{e-1} X^{e-1} + \cdots + b_1 X + b_0$$

reduzieren, indem wir es ersetzen durch

$$f - \frac{b_e}{c_\delta} X^{\delta - e} g \,.$$

Das führen wir so lange fort, bis f auf Null oder ein Polynom von kleinerem Grad als e reduziert ist: Das ist dann der Divisionsrest r. Auch hier ist klar, daß sich nichts an der Lösungsmenge ändert, wenn man die beiden Gleichungen f, g ersetzt durch g, r, denn

$$f = qg + r \quad \text{und} \quad r = f - qg \,,$$

d.h. f und g verschwinden genau dann für einen Wert x, wenn g und r an der Stelle x verschwinden.

In beiden Fällen ist die Vorgehensweise sehr ähnlich: Wir vereinfachen das Gleichungssystem schrittweise, indem wir eine Gleichung ersetzen durch ihre Summe mit einem geeigneter Vielfachen einer anderen Gleichung.

Dieselbe Strategie wollen wir auch anwenden Systeme von Polynomgleichungen in mehreren Veränderlichen. Erstes Problem dabei ist, daß wir nicht wissen, wie wir die Monome eines Polynoms anordnen sollen und damit, was der führende Term ist. Dazu gibt es eine ganze Reihe verschiedener Strategien, von denen je nach Anwendung mal die eine, mal die andere vorteilhaft ist.

§3: Monomordnungen und der Divisionsalgorithmus

Wir betrachten Polynome in n Variablen $X_1, \ldots, X_n$ über einem Körper k und setzen zur Abkürzung

$$X^\alpha = X_1^{\alpha_1} \cdots X_n^{\alpha_n} \quad \text{mit} \quad \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0 \,.$$

Die Terme X^α bezeichnen wir als *Monome,* und die Summe der α_i als den Grad von X^α.

Eine Anordnung der Monome ist offensichtlich äquivalent zu einer Anordnung auf $\mathbb{N}_0^n$, und es gibt sehr viele Möglichkeiten, diese Menge

anzuordnen. Für uns sind allerdings nur Anordnungen interessant, die einigermaßen kompatibel sind zur algebraischen Struktur des Polynomrings $k[X_1, \ldots, X_n]$. Beispielsweise wollen wir sicherstellen, daß der führende Term des Produkts zweier Polynome das Produkt der führenden Terme der Faktoren ist – wie wir es auch vom Eindimensionalen her gewohnt sind. Daher definieren wir

Definition: *a)* Eine Monomordnung ist eine Ordnungsrelation „$<$“ auf $\mathbb{N}_0^n$, für die gilt

1. „$<$“ ist eine Linear- oder Totalordnung, d.h. für zwei Elemente $\alpha, \beta \in \mathbb{N}_0^n$ ist entweder $\alpha < \beta$ oder $\beta < \alpha$ oder $\alpha = \beta$.
2. Für $\alpha, \beta, \gamma \in \mathbb{N}_0^n$ gilt: $\alpha < \beta \Longrightarrow \alpha + \gamma < \beta + \gamma$.
3. „$<$“ ist eine Wohlordnung, d.h. jede Teilmenge $I \subseteq \mathbb{N}_0^n$ hat ein kleinstes Element.

b) Für ein Polynom $f = \sum_{\alpha \in I} c_\alpha X^\alpha \in k[X_1, \ldots, X_n]$ mit $c_\alpha \neq 0$ für alle $\alpha \in I \subset \mathbb{N}_0$ sei γ das größte Element von I bezüglich einer fest gewählten Monomordnung. Dann bezeichnen wir bezüglich dieser Monomordnung

- γ = multideg f als Multigrad von f
- X^γ = FM(f) als führendes Monom von f
- c_γ = FK(f) als führenden Koeffizienten von f
- $c_\gamma X^\gamma$ = FT(f) als führenden Term von f

Der Grad deg f von f ist, wie in der Algebra üblich, der höchste Grad eines Monoms von f. Je nach gewählter Monomordnung muß das nicht unbedingt der Grad des führenden Monoms sein.

Beispiele von Monomordnungen sind

a) Die lexikographische Ordnung: Hier ist $\alpha < \beta$ genau dann, wenn für den ersten Index i, in dem sich α und β unterscheiden, $\alpha_i < \beta_i$ ist. Betrachtet man Monome X^α als Worte über dem (geordneten) Alphabet $\{X_1, \ldots, X_n\}$, kommt hier ein Monom X^α genau dann vor X^β, wenn die entsprechenden Worte im Lexikon in dieser Reihenfolge gelistet werden. Die ersten beiden Forderungen an eine Monomordnung sind klar, und auch die Wohlordnung macht keine großen Probleme:

Man betrachtet zunächst die Teilmenge aller Exponenten $\alpha \in I$ mit kleinstmöglichem α_1, unter diesen die Teilmenge derer mit kleinstmöglichem α_2, *usw.,* bis man bei α_n angelangt ist. Spätestens hier ist die verbleibende Teilmenge einelementig, und ihr einziges Element ist das gesuchte kleinste Element von I.

b) Die graduierte lexikographische Ordnung: Hier ist der Grad eines Monoms erstes Ordnungskriterium: Ist $\deg X^\alpha < \deg X^\beta$, so definieren wir $\alpha < \beta$. Falls beide Monome gleichen Grad haben, soll $\alpha < \beta$ genau dann gelten, wenn α im lexikographischen Sinne kleiner als β ist. Auch hier sind offensichtlich alle drei Forderungen erfüllt.

c) Die inverse lexikographische Ordnung: Hier ist $\alpha < \beta$ genau dann, wenn $\alpha_i < \beta_i$ für den *letzten* Index i, in dem sich α und β unterscheiden. Das entspricht offensichtlich gerade der lexikographischen Anordnung bezüglich des rückwärts gelesenen Alphabets $X_n, \ldots, X_1$. Entsprechend läßt sich natürlich auch bezüglich jeder anderen Permutation des Alphabets eine Monomordnung definieren, so daß diese Ordnung nicht sonderlich interessant ist – außer als Bestandteil der im folgenden definierten Monomordnung:

d) Die graduierte inverse lexikographische Ordnung: Wie bei der graduierten lexikographischen Ordnung ist hier der Grad eines Monoms erstes Ordnungskriterium: Falls $\deg X^\alpha < \deg X^\beta$, ist $\alpha < \beta$, und nur falls beide Monome gleichen Grad haben, soll $\alpha < \beta$ genau dann gelten, wenn α im Sinne der inversen lexikographischen Ordnung *größer* ist als β. Man beachte, daß wir hier also nicht nur die Reihenfolge der Variablen invertieren, sondern auch die Ordnungsrelation im Fall gleicher Grade. Es ist nicht schwer zu sehen, daß auch damit eine Monomordnung definiert wird: Mit den ersten beiden Forderungen gibt es wie üblich keine Probleme, und wenn wir eine Menge M von Monomen haben, gibt es darin eine Teilmenge bestehend aus den Monomen kleinsten Grades. Da es für jeden Grad nur endlich viele Monome gibt, ist diese Menge endlich, hat also bezüglich der inversen lexikographischen Ordnung nicht nur ein kleinstes, sondern auch ein größtes Element. Dieses ist das kleinste Element von M bezüglich der graduierten inversen lexikographischen Ordnung.

Für das folgende werden wir noch einige Eigenschaften einer Monomordnung benötigen, die in der Definition nicht erwähnt sind.

Als erstes wollen wir uns überlegen, daß bezüglich jeder Monomordnung auf $\mathbb{N}_0^n$ kein Element kleiner sein kann als $(0, \ldots, 0)$: Wäre nämlich $\alpha < (0, \ldots, 0)$, so wäre wegen der zweiten Eigenschaft auch

$$2\alpha = \alpha + \alpha < \alpha + (0, \ldots, 0) = \alpha$$

und so weiter, so daß wir eine unendliche absteigende Folge

$$\alpha > 2\alpha > 3\alpha > \cdots$$

hätten, im Widerspruch zur dritten Forderung.

Daraus folgt nun sofort, daß das Produkt zweier Monome größer ist als jeder der beiden Faktoren und damit auch, daß ein echter Teiler eines Monoms immer kleiner ist als dieses. Außerdem folgt, daß für ein Produkt von Polynomen stets $\mathrm{FM}(fg) = \mathrm{FM}(f) \cdot \mathrm{FM}(g)$ ist.

Die Eliminationsschritte beim GAUSS-Algorithmus können auch als Divisionen mit Rest verstanden werden, und beim EUKLIDischen Algorithmus ist ohnehin alles Division mit Rest. Für ein Verallgemeinerung der beiden Algorithmen auf Systeme nichtlinearer Gleichungssysteme brauchen wir also auch einen Divisionsalgorithmus für Polynome in mehreren Veränderlichen, der die eindimensionale Polynomdivision mit Rest und die Eliminationsschritte beim GAUSS-Algorithmus verallgemeinert.

Beim GAUSS-Algorithmus brauchen wir im allgemeinen mehr als nur einen Eliminationsschritt, bis wir eine Gleichung auf eine Variable reduziert haben. Entsprechend wollen wir auch hier einen Divisionsalgorithmus betrachten, der gegebenenfalls mehrere Divisoren gleichzeitig durchführen kann.

Wir gehen also aus von einem Polynom $f \in R = k[X_1, \ldots, X_n]$, wobei k irgendein Körper ist, in dem wir rechnen können, meistens also $k = \mathbb{Q}$ oder $k = \mathbb{F}_p$ oder eine endliche Erweiterung davon. Dieses Polynom wollen wir dividieren durch die Polynome $f_1, \ldots, f_m \in R$, d.h. wir suchen Polynome $a_1, \ldots, a_m, r \in R$, so daß

$$f = a_1 f_1 + \cdots + a_m f_m + r$$

ist, wobei r in irgendeiner noch zu präzisierenden Weise kleiner als die f_i sein soll.

Da es sowohl bei GAUSS als auch bei EUKLID auf die Anordnung der Terme ankommt, legen wir als erstes eine Monomordnung fest. Wenn im folgenden von führenden Termen *etc.* die Rede ist, soll es sich stets um die führenden Terme *etc.* bezüglich dieser Ordnung handeln.

Mit dieser Konvention geht der Algorithmus dann folgendermaßen:

Gegeben sind $f, f_1, \ldots, f_m \in R$
Berechnet werden $a_1, \ldots, a_m, r \in R$ mit $f = a_1 f_1 + \cdots + a_m f_m + r$, wobei r kein Monom enthält, das durch das führende Monom eines der f_i teilbar ist.

1. Schritt (Initialisierung): Setze $a_1 = \cdots = a_m = r = 0$ und $p = f$.

2. Schritt (Endebedingung): Im Falle $p = 0$ endet der Algorithmus.

3. Schritt (Divisionsschritt): Falls keiner der führenden Terme FT f_i den führenden Term FT p teilt, wird p ersetzt durch $p - \mathrm{FT}\, p$ und r durch $r + \mathrm{FT}\, p$. Andernfalls sei i der kleinste Index, für den FT f_i Teiler von FT p ist; der Quotient sei q. Dann wird a_i ersetzt durch $a_i + q$ und p durch $p - q f_i$. Weiter geht es mit dem 2. Schritt.

Offensichtlich ist die Bedingung $f - p = a_1 f_1 + \cdots + a_m f_m + r$ nach der Initialisierung im ersten Schritt erfüllt, und sie bleibt auch bei jeder Anwendung des Divisionsschritts erfüllt. Außerdem endet der Algorithmus nach endlich vielen Schritten: Bei jedem Divisionsschritt wird der führende Term von p eliminiert, und alle Monome, die eventuell neu dazukommen, sind kleiner oder gleich dem führenden Monom von f_i. Da letzteres das (alte) führende Monom von p teilt, kann es nicht größer sein als dieses, d.h. der führende Term des neuen p ist kleiner als der des alten. Wegen der Wohlordnungseigenschaft einer Monomordnung kann es keine unendliche absteigende Kette von Monomen geben; daher muß der Algorithmus nach endlich vielen Schritten abbrechen.

Bei der klassischen Polynomdivision für Polynome in einer Variablen über einem Körper wissen wir, daß der Rest kleineren Grad hat als der Divisor. Das muß hier nicht der Fall sein; wir können nur sagen, daß der

Rest keine Monome enthält, die durch den führenden Term eines der Divisoren f_i teilbar sind.

Um den Algorithmus besser zu verstehen, betrachten wir zunächst zwei Beispiele:

Als erstes dividieren wir $f = X^2Y + XY^2 + Y^2$ durch $f_1 = XY - 1$ und $f_2 = Y^2 - 1$.

Zur Initialisierung setzen wir $a_1 = a_2 = r = 0$ und $p = f$. Wir verwenden die lexikographische Ordnung; bezüglich derer ist der führende Term von p gleich X^2Y und der von f_1 gleich XY. Letzteres teilt X^2Y, wir setzen also

$$p \leftarrow p - Xf_1 = XY^2 + X + Y^2 \quad \text{und} \quad a_1 \leftarrow a_1 + X = X\,.$$

Neuer führender Term von p ist XY^2; auch das ist ein Vielfaches von XY, also setzen wir

$$p \leftarrow p - Yf_1 = X + Y^2 + Y \quad \text{und} \quad a_1 \leftarrow a_1 + Y = X + Y\,.$$

Nun ist X der führende Term von p, und der ist weder durch XY noch durch Y^2 teilbar, also kommt er in den Rest:

$$p \leftarrow p - X = Y^2 + Y \quad \text{und} \quad r \leftarrow r + X = X\,.$$

Der nun führende Term Y^2 von p ist gleichzeitig der führende Term von f_2 und nicht teilbar durch XY, also wird

$$p \leftarrow p - f_2 = Y + 1 \quad \text{und} \quad a_2 \leftarrow a_2 + 1 = 1\,.$$

Die verbleibenden Terme von p sind weder durch XY noch durch Y^2 teilbar, kommen also in den Rest, so daß wir als Ergebnis erhalten

$$f = a_1f_1 + a_2f_2 + r \quad \text{mit} \quad a_1 = X + Y, \quad a_2 = 1 \quad \text{und} \quad r = X + Y + 1\,.$$

Wenn wir statt durch das Paar (f_1, f_2) durch (f_2, f_1) dividiert hätten, hätten wir im ersten Schritt zwar ebenfalls X^2Y durch XY dividiert, denn durch Y^2 ist es nicht teilbar. Der neue führende Term XY^2 ist aber durch beides teilbar, und wenn f_2 an erster Stelle steht, nehmen wir im Zweifelsfall dessen führenden Term. Man rechnet leicht nach, daß man hier mit folgendem Ergebnis endet:

$$f = a_1f_1 + a_2f_2 + r \quad \text{mit} \quad a_1 = X + 1, \quad a_2 = X \quad \text{und} \quad r = X + 1\,.$$

Wie wir sehen, sind also sowohl die „Quotienten“ a_i als auch der „Rest“ r von der Reihenfolge der f_i abhängig. Sie hängen natürlich im allgemeinen auch ab von der verwendeten Monomordnung; deshalb haben wir die schließlich eingeführt.

Als zweites Beispiel wollen wir $f = XY^2 - X$ durch die beiden Polynome $f_1 = XY + 1$ und $f_2 = Y^2 - 1$ dividieren. Im ersten Schritt dividieren wir XY^2 durch XY mit Ergebnis Y, ersetzen also f durch $-X - Y$. Diese beiden Terme sind weder durch XY noch durch Y^2 teilbar, also ist unser Endergebnis

$$f = a_1 f_1 + a_2 f_2 + r \quad \text{mit} \quad a_1 = Y, \quad a_2 = 0 \quad \text{und} \quad r = -X - Y\,.$$

Hätten wir stattdessen durch (f_2, f_1) dividiert, hätten wir als erstes XY^2 durch Y^2 dividiert mit Ergebnis X; da $f = X f_2$ ist, geht die Division hier ohne Rest auf. Der Divisionsalgorithmus erlaubt uns also nicht einmal die sichere Feststellung, ob f als Linearkombination der f_i darstellbar ist oder nicht; als alleiniges Hilfsmittel zur Lösung nichtlinearer Gleichungssysteme reicht er offenbar nicht aus. Daher müssen wir in den folgenden Paragraphen noch weitere Werkzeuge betrachten.

§4: Der Hilbertsche Basissatz

Die Grundidee des Algorithmus von BUCHBERGER besteht darin, das Gleichungssystem so abzuändern, daß möglichst viele seiner Eigenschaften bereits an den führenden Termen der Gleichungen ablesbar sind.

Angenommen, wir haben ein nichtlineares Gleichungssystem

$$f_1(X_1, \ldots, X_n) = \cdots = f_m(X_1, \ldots, X_n) = 0$$

mit $f_i \in R = k[X_1, \ldots, X_n]$; seine Lösungsmenge sei $\mathcal{L} \subseteq k^n$.

Wie wir aus §1 wissen, hängt $\mathcal{L}$ nur ab von dem Ideal $I = (f_1, \ldots, f_m)$; zur Lösung des Systems sollten wir daher versuchen, ein möglichst „einfaches“ Erzeugendensystem für dieses Ideal zu finden.

Ganz besonders einfach (wenn auch selten ausreichend) sind Ideale, die von Monomen erzeugt werden:

Definition: Ein Ideal $I \lhd R = k[X_1, \ldots, X_n]$ heißt *monomial,* wenn es ein (nicht notwendigerweise endliches) Erzeugendensystem aus Monomen hat.

Nehmen wir an, I werde erzeugt von den Monomen X^α mit α aus einer Indexmenge A. Ist dann X^β irgendein Monom aus I, kann es als endliche Linearkombination

$$X^\beta = \sum_{i=1}^{r} f_i X^{\alpha_i} \quad \text{mit} \quad \alpha_i \in A$$

geschrieben werden, wobei die f_i irgendwelche Polynome aus R sind. Da sich jedes Polynom als Summe von Monomen schreiben läßt, können wir f_i als k-Linearkombination von Monomen X^γ schreiben und bekommen damit eine neue Darstellung von X^β als Summe von Termen der Form $cX^\gamma X^\alpha$ mit $\alpha \in A, \gamma \in \mathbb{N}_0^n$ und $c \in k$. Sortieren wir diese Summanden nach den resultierenden Monomen $X^{\gamma+\alpha}$ und fassen alle Summanden mit gleichem Monom zusammen, so entsteht eine k-Linearkombination verschiedener Monome, die insgesamt gleich X^β ist. Das ist aber nur möglich, wenn diese Summe aus dem einen Summanden X^β besteht, d.h. β läßt sich schreiben in der Form $\beta = \alpha + \gamma$ mit einem $\alpha \in A$ und einem $\gamma \in \mathbb{N}_0^n$.

Dies zeigt, daß ein Monom X^β genau dann in I liegt, wenn $\beta = \alpha + \gamma$ ist mit einem $\alpha \in A$ und einem $\gamma \in \mathbb{N}_0^n$, d.h. X^β ist das Produkt eines der erzeugenden Monome mit *irgendeinem* Monom. Das Ideal I besteht genau aus den Polynomen f, die sich als k-Linearkombinationen solcher Monome schreiben lassen.

Damit folgt insbesondere, daß ein Polynom f genau dann in einem monomialen Ideal I liegt, wenn jedes seiner Monome dort liegt.

Lemma von Dickson: Jedes monomiale Ideal in $R = k[X_1, \ldots, X_n]$ kann von endlich vielen Monomen erzeugt werden.

Der *Beweis* wird durch vollständige Induktion nach n geführt. Im Fall $n = 1$ ist alles klar, denn da sind die Monome gerade die Potenzen der einzigen Variable, und natürlich erzeugt jede Menge von Potenzen genau dasselbe Ideal wie die Potenz mit dem kleinsten Exponenten aus

dieser Menge. Hier kommt man also sogar mit einem einzigen Monom aus.

Im Fall $n > 1$ und $\alpha \in \mathbb{N}_0^n$ setzen wir $X'^{\alpha} = X_1^{\alpha_1} \cdots X_{n-1}^{\alpha_{n-1}}$ und betrachten das Ideal

$$J = \left(X'^{\alpha} \;\middle|\; X^{\alpha} \in I \right) \lhd k[X_1, \ldots, X_{n-1}] .$$

Nach Induktionsvoraussetzung wird J erzeugt von endlich vielen Monomen X'^{α}.

Jedes Monom aus aus dem endlichen Erzeugendensystem von J läßt sich in der Form X'^{α} schreiben mit einem $\alpha \in \mathbb{N}_0^n$, für das X^{α} in I liegt. Unter den Indizes α_n, die wir dabei jeweils an das $(n-1)$-Tupel $(\alpha_1, \ldots, \alpha_{n-1})$ anhängen, sei r der größte. Dann liegt $X'^{\alpha'} X_n^r$ für jedes Monom aus dem Erzeugendensystem von J in I und damit für jedes Monom aus J. Die endlich vielen Monome $X^{\alpha'} X_n^r$ erzeugen also zumindest ein Teilideal von I.

Es kann aber natürlich auch noch Monome in I geben, in denen X_n mit einem kleineren Exponenten als r auftritt. Um auch diese Elemente zu erfassen, betrachten wir für jedes $s < r$ das Ideal $J_s \lhd k[X_1, \ldots, X_{n-1}]$, das von allen jeden Monomen X'^{α} erzeugt wird, für die $X'^{\alpha} X_n^s$ in I liegt. Auch jedes der J_s wird nach Induktionsannahme erzeugt von endlich vielen Monomen X'^{α}, und wenn wir die sämtlichen Monome $X'^{\alpha} X_n^s$ zu unserem Erzeugendensystem hinzunehmen (für alle $s = 0, 1, \cdots, r-1$), haben wir offensichtlich ein Erzeugendensystem von I aus endlich vielen Monomen gefunden.

∎

Beliebige Ideale sind im allgemeinen nicht monomial; schon das von $X + 1$ erzeugte Ideal in $k[X]$ ist ein Gegenbeispiel, denn es enthält weder das Monom X noch das Monom 1. Dies widerspricht der oben gezeigten Tatsache, daß ein monomiales Ideal, zu jedem seiner Elemente auch dessen sämtliche Monome enthält.

Um monomiale Ideale auch für die Untersuchung solcher Ideale nützlich zu machen, wählen wir eine Monomordnung auf R und definieren für ein beliebiges Ideal $I \lhd R = k[X_1, \ldots, X_n]$ das monomiale Ideal

$$\mathrm{FM}(I) = \left(\mathrm{FM}(f) \;\middle|\; f \in I \setminus \{0\} \right) ,$$

LEONARD EUGENE DICKSON (1874–1954) wurde in Iowa geboren, wuchs aber in Texas auf. Seinen Bachelor- und Mastergrad bekam er von der University of Texas, danach ging er an die University of Chicago. Mit seiner 1896 dort eingereichte Dissertation *Analytic Representation of Substitutions on a Power of a Prime Number of Letters with a Discussion of the Linear Group* wurde er der erste dort promovierte Mathematiker. Auch die weiteren seiner 275 wissenschaftlichen Arbeiten, darunter acht Bücher, beschäftigen sich vor allem mit der Algebra und Zahlentheorie. Den größten Teil seines Berufslebens verbrachte er als Professor an der Universität von Chicago, dazu kommen regelmäßige Besuche in Berkeley.

das von den führenden Monomen *aller* Elemente von I erzeugt wird – außer natürlich dem nicht existierenden führenden Monom der Null.

Nach dem Lemma von DICKSON ist FM(I) erzeugt von endlich vielen Monomen. Jedes dieser Monome ist, wie wir eingangs gesehen haben, ein Vielfaches eines der erzeugenden Monome, also eines führenden Monoms eines Elements von I. Ein Vielfaches des führenden Monoms ist aber das führende Monom des entsprechenden Vielfachen des Elements von I, denn FM($X^\gamma f$) = X^γ FM(f), da für jede Monomordnung gilt: $\alpha < \beta \Longrightarrow \alpha + \beta < \alpha + \gamma$. Somit wird FM($I$) erzeugt von endlich vielen Monomen der Form FM(f_i), wobei die f_i Elemente von I sind. Wir wollen sehen, daß die Elemente f_i das Ideal I erzeugen; damit folgt insbesondere

Hilbertscher Basissatz: Jedes Ideal $I \lhd R = k[X_1, \ldots, X_n]$ hat ein endliches Erzeugendensystem.

Beweis: Wie wir bereits wissen, gibt es Elemente $f_1, \ldots, f_m \in I$, so daß FM(I) von den Monomen FM(f_i) erzeugt wird. Um zu zeigen, daß die Elemente f_i das Ideal I erzeugen, betrachten wir ein beliebiges Element $f \in I$ und versuchen, es als R-Linearkombination der f_i zu schreiben. Division von f durch $f_1, \ldots, f_m$ zeigt, daß es Polynome $a_1, \ldots, a_m$ und r in R gibt derart, daß

$$f = a_1 f_1 + \cdots + a_m f_m + r$$

ist. Wir sind fertig, wenn wir zeigen können, daß der Divisionsrest r verschwindet.

Falls r *nicht* verschwindet, zeigt der Divisionsalgorithmus, daß das führende Monom FM(r) von r durch kein führendes Monom FM(f_i) eines der Divisoren f_i teilbar ist. Andererseits ist aber

$$r = f - (a_1 f_1 + \cdots + a_m f_m)$$

ein Element von I, und damit liegt FM(r) im von den FM(f_i) erzeugten Ideal FM(I). Somit muß FM(r) Vielfaches eines FM(f_i) sein, ein Widerspruch. Also ist $r = 0$.

■

DAVID HILBERT (1862–1943) wurde in Königsberg geboren, wo er auch zur Schule und zur Universität ging. Er promovierte dort 1885 mit einem Thema aus der Invariantentheorie, habilitierte sich 1886 und bekam 1893 einen Lehrstuhl. 1895 wechselte er an das damalige Zentrum der deutschen wie auch internationalen Mathematik, die Universität Göttingen, wo er bis zu seiner Emeritierung im Jahre 1930 lehrte. Seine Arbeiten umfassen ein riesiges Spektrum aus unter anderem Invariantentheorie, Zahlentheorie, Geometrie, Funktionalanalysis, Logik und Grundlagen der Mathematik sowie auch zur Relativitätstheorie. Er gilt als einer der Väter der modernen Algebra.

HILBERT veröffentlichte seinen Basissatz 1890, als DICKSON gerade sechzehn Jahre alt war. HILBERTs Beweis verwendet daher nicht das Lemma von DICKSON, sondern arbeitet ausschließlich mit den Polynomen aus I. Er war für die damalige Zeit vor allem deshalb eine Sensation, weil die Existenz eines endlichen Erzeugendensystems bewiesen wurde *ohne* Angabe eines Verfahren zu dessen Konstruktion. Die damalige Algebra beschäftigte sich vor allem mit der sogenannten Invariantentheorie und suchte nach konkreten Erzeugendensystemen von speziellen Invariantenringen, was nur in wenigen Fällen gelang. HILBERT konnte in seiner Arbeit zeigen, daß alle der damals betrachteten klassischen Invariantenringe endliche Erzeugendensysteme hatten, konnte aber keine solchen Systeme angeben.

§5: Gröbner-Basen und der Buchberger-Algorithmus

Die Strategie, möglichst viele Eigenschaften eines Ideals über führenden Monome seiner Elemente zu untersuchen, erwies sich als äuërst erfolgreich. Dies motiviert die folgende

Definition: Eine endliche Teilmenge $G = \{g_1, \ldots, g_m\} \subset I$ eines Ideals $I \lhd R = k[X_1, \ldots, X_n]$ heißt Standardbasis oder GRÖBNER-Basis von I bezüglich einer vorgegebenen Monomordnung, falls die Monome FM(g_i) das Ideal FM(I) erzeugen.

WOLFGANG GRÖBNER wurde 1899 im damals noch österreichischen Südtirol geboren. Nach Ende des ersten Weltkriegs, in dem er an der italienischen Front kämpfte, studierte er zunächst an der TU Graz Maschinenbau, beendete dieses Studium aber nicht, sondern begann 1929 an der Universität Wien ein Mathematikstudium. Nach seiner Promotion ging er zu EMMY NOETHER nach Göttingen, um dort Algebra zu lernen. Aus materiellen Gründen mußte er schon bald nach Österreich zurück, konnte aber auch dort zunächst keine Anstellung finden, so daß er Kleinkraftwerke baute und im Hotel seines Vaters aushalf. Ein italienischen Mathematiker, der dort seinen Urlaub verbrachte, vermittelte ihm eine Stelle an der Universität Rom, die er 1939 wieder verlassen mußte, nachdem er sich beim Anschluß Südtirols an Italien für die deutsche Staatsbürgerschaft entschieden hatte. Während des zweiten Weltkriegs arbeitete er größtenteils an einem Forschungsinstitut der Luftwaffe, nach Kriegsende als Extraordinarius in Wien, dann als Ordinarius in Innsbruck, wo er 1980 starb. Seine Arbeiten beschäftigen sich mit der Algebra und algebraischen Geometrie sowie mit Methoden der Computeralgebra zur Lösung von Differentialgleichungen.

Die Theorie der GRÖBNER-Basen wurde von seinem Studenten BRUNO BUCHBERGER in dessen Dissertation entwickelt. BUCHBERGER wurde 1942 in Innsbruck geboren, wo er auch Mathematik studierte und 1966 bei GRÖBNER promovierte mit der Arbeit *Ein Algorithmus zum Auffinden der Basiselemente des Restklassenrings nach einem nulldimensionalen Polynomideal.* Er arbeitete zunächst als Assistent, nach seiner Habilitation als Dozent an der Universität Innsbruck, bis er 1974 einen Ruf auf den Lehrstuhl für Computermathematik an der Universität Linz erhielt. Dort gründete er 1987 das Research Institute for Symbolic Computation (RISC), dessen Direktor er bis 1999 war. 1989 initiierte er in Hagenberg (etwa 20 km nordöstlich von Linz) die Gründung eines Softwareparks mit angeschlossener Fachhochschule; er hat mittlerweile fast Tausend Mitarbeiter. Außer mit Computeralgebra beschäftigt er sich auch im Rahmen des Theorema-Projekts mit dem automatischen Beweisen mathematischer Aussagen.

Wie der Beweis des HILBERTschen Basissatzes im vorigen Paragraphen zeigt, erzeugen Elemente $g_1, \ldots, g_r$ ein Ideal I, falls ihre führenden

Monome das Ideal FM(I) erzeugen. Somit ist eine GRÖBNER-Basis automatisch ein Erzeugendensystem. Außerdem hat jedes Ideal I im Polynomring eine GRÖBNER-Basis, denn nach dem Lemma von DICKSON hat das Ideal der führenden Monome ein endliches Erzeugendensystem, und jedes Monom aus diesem Erzeugendensystem ist führendes Monom eines Polynoms $f_i \in I$. Die Menge der Polynome f_i ist offensichtlich eine GRÖBNER-Basis im Sinne der obigen Definition.

Bevor wir uns damit beschäftigen, wie man diese berechnen kann, wollen wir zunächst eine wichtige Eigenschaft betrachten.

$\{g_1, \ldots, g_m\}$ sei eine GRÖBNER-Basis eines Ideals $I \lhd R$. Wir wollen ein beliebiges Element $f \in R$ durch $g_1, \ldots, g_m$ dividieren. Dies liefert als Ergebnis

$$f = a_1 g_1 + \cdots + a_m g_m + r\,,$$

wobei kein Monom von r durch eines der Monome FM(g_i) teilbar ist. Wie wir wissen, sind allerdings bei der Polynomdivision im allgemeinen weder der Divisionsrest r noch die Koeffizienten a_i auch nur im entferntesten eindeutig. Wir wollen untersuchen, wie sich das hier verhält.

Angenommen. wir haben zwei Darstellungen

$$f = a_1 g_1 + \cdots + a_m g_m + r = b_1 g_1 + \cdots + b_m g_m + s$$

der obigen Form. Dann ist

$$(a_1 - b_1) g_1 + \cdots + (a_m - b_m) g_m = s - r\,.$$

Links steht ein Element von I, also auch rechts. Andererseits enthält aber weder r noch s ein Monom, das durch eines der Monome FM(g_i) teilbar ist, d.h. $r - s = 0$, da die FM(g_i) ja das Ideal FM(I) erzeugen. Somit ist bei der Division durch die Elemente einer GRÖBNER-Basis der Divisionsrest eindeutig bestimmt. Insbesondere ist f genau dann ein Element von I, wenn der Divisionsrest verschwindet. Wenn wir eine GRÖBNER-Basis haben, können wir also leicht entscheiden, ob ein gegebenes Element $f \in R$ im Ideal I liegt.

Nachdem im Fall einer GRÖBNER-Basis der Divisionsrest nicht von der Reihenfolge der Basiselemente abhängt, können wir ihn durch ein

Symbol bezeichnen, das nur von der *Menge* $G = \{g_1, \ldots, g_m\}$ abhängt; wir schreiben $\overline{f}^G$.

Als nächstes wollen wir uns mit der Frage beschäftigen, wie wir für ein vorgegebenes Ideal I eine GRÖBNER-Basis bestimmen können.

Dazu müssen wir uns als erstes überlegen, *wie* das Ideal vorgegeben sein soll. Wenn wir damit rechnen wollen, müssen wir irgendeine Art von endlicher Information haben; was sich anbietet ist natürlich ein endliches Erzeugendensystem.

Wir gehen also aus von einem Ideal $I = (f_1, \ldots, f_m)$ und suchen eine GRÖBNER-Basis. Das Problem ist, daß die Monome $\mathrm{FM}(f_i)$ im allgemeinen nicht ausreichen, um das monomiale Ideal $\mathrm{FM}(I)$ zu erzeugen, denn dieses enthält ja *jedes* Monom eines jeden Elements von I und nicht nur das führende. Wir müssen daher neue Elemente produzieren, deren führende Monome in den gegebenen Elementen f_i oder auch anderen Elementen von I erst weiter hinten vorkommen.

BUCHBERGERs Idee dazu war die Konstruktion sogenannter S-Polynome: Seien $f, g \in R$ zwei Polynome; $\mathrm{FM}(f) = X^\alpha$ und $\mathrm{FM}(g) = X^\beta$ seien ihre führenden Monome, und X^γ sei das kgV von X^α und X^β, d.h. $\gamma_i = \max(\alpha_i, \beta_i)$ für $i = 1, \ldots, n$. Das S-Polynom von f und g ist

$$S(f,g) = \frac{X^\gamma}{\mathrm{FT}(f)} \cdot f - \frac{X^\gamma}{\mathrm{FT}(g)} \cdot g\,.$$

Da $\frac{X^\gamma}{\mathrm{FT}(f)} \cdot f$ und $\frac{X^\gamma}{\mathrm{FT}(g)} \cdot g$ beide nicht nur dasselbe führende Monom X^γ haben, sondern es wegen der Division durch den führenden *Term* statt nur das führende Monom auch beide mit Koeffizient eins enthalten, fällt es bei der Bildung von $S(f,g)$ weg. Daher ist das führende Monom von $S(f,g)$ kleiner als X^γ. Das folgende Lemma ist der Kern des Beweises, daß S-Polynome alles sind, was wir brauchen, um GRÖBNER-Basen zu berechnen.

Lemma: Für die Polynome $f_1, \ldots, f_m \in R$ sei

$$S = \sum_{i=1}^{m} \lambda_i X^{\alpha_i} f_i \quad \text{mit} \quad \lambda_i \in k \quad \text{und} \quad \alpha_i \in \mathbb{N}_0^n$$

eine Linearkombination, zu der es ein $\delta \in \mathbb{N}_0^n$ gebe, so daß alle Summanden X^δ als führendes Monom haben, d.h. α_i + multideg $f_i = \delta_i$ für $i = 1, \ldots, m$. Falls multideg $S < \delta$ ist, gibt es Elemente $\lambda_{ij} \in k$, so daß

$$S = \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_{ij} X^{\gamma_{ij}} S(f_i, f_j)$$

ist mit $X^{\gamma_{ij}} = \mathrm{kgV}\big(\mathrm{FM}(f_i), \mathrm{FM}(f_j)\big)$.

Beweis: Der führende Koeffizient von f_i sei μ_i; dann ist $\lambda_i \mu_i$ der führende Koeffizient von $\lambda_i X^{\alpha_i} f_i$. Somit ist multideg S genau dann kleiner als δ, wenn $\sum_{i=1}^{m} \lambda_i \mu_i$ verschwindet. Wir normieren alle $X^{\alpha_i} f_i$ auf führenden Koeffizienten eins, indem wir $p_i = X^{\alpha_i} f_i / \mu_i$ betrachten; dann ist

$$\begin{aligned} S = \sum_{i=1}^{m} \lambda_i \mu_i p_i = {} & \lambda_1 \mu_1 (p_1 - p_2) + (\lambda_1 \mu_1 + \lambda_2 \mu_2)(p_2 - p_3) + \cdots \\ & + (\lambda_1 \mu_1 + \cdots + \lambda_{m-1} \mu_{m-1})(p_{m-1} - p_m) \\ & + (\lambda_1 \mu_1 + \cdots + \lambda_m \mu_m) p_m \,, \end{aligned}$$

wobei der Summand in der letzten Zeile genau dann verschwindet, wenn multideg $S < \delta$.

Da alle p_i denselben Multigrad δ und denselben führenden Koeffizienten eins haben, kürzen sich in den Differenzen $p_i - p_j$ die führenden Terme weg, genau wie in den S-Polynomen. In der Tat: Bezeichnen wir den Multigrad von $\mathrm{kgV}\big(\mathrm{FM}(f_i), \mathrm{FM}(f_j)\big)$ mit γ_{ij}, so ist

$$p_i - p_j = X^{\delta - \gamma_{ij}} S(f_i, f_j) \,.$$

Damit hat die obige Summendarstellung von S die gewünschte Form. ∎

Daraus folgt ziemlich unmittelbar

Satz: Ein Erzeugendensystem $f_1, \ldots, f_m$ eines Ideals I im Polynomring $R = k[X_1, \ldots, X_n]$ ist genau dann eine GRÖBNER-Basis, wenn jedes S-Polynom $S(f_i, f_j)$ bei der Division durch $f_1, \ldots, f_m$ Rest Null hat.

Beweis: Als R-Linearkombination von f_i und f_j liegt das S-Polynom $S(f_i, f_j)$ im Ideal I; falls $f_1, \ldots, f_m$ eine GRÖBNER-Basis von I ist, hat es also Rest Null bei der Division durch $f_1, \ldots, f_m$.

Umgekehrt sei $f_1, \ldots, f_m$ ein Erzeugendensystem von $I \lhd R$ mit der Eigenschaft, daß alle $S(f_i, f_j)$ bei der Division durch $f_1, \ldots, f_m$ (in irgendeiner Reihenfolge) Divisionsrest Null haben. Wir wollen zeigen, daß $f_1, \ldots, f_m$ dann eine GRÖBNER-Basis ist, daß also die führenden Monome $\mathrm{FM}(f_1), \ldots, \mathrm{FM}(f_m)$ das Ideal $\mathrm{FM}(I)$ erzeugen.

Sei also $f \in I$ ein beliebiges Element; wir müssen zeigen, daß $\mathrm{FM}(f)$ im von den $\mathrm{FM}(f_i)$ erzeugten Ideal liegt.

Da f in I liegt, gibt es eine Darstellung

$$f = h_1 f_1 + \cdots + h_m f_m \quad \text{mit} \quad h_i \in R\,.$$

Falls sich hier bei den führenden Termen nichts wegkürzt, ist der führende Term von f die Summe der führenden Terme gewisser Produkte $h_i f_i$, die allesamt dasselbe führende Monom $\mathrm{FM}(f)$ haben. Wegen $\mathrm{FM}(h_i f_i) = \mathrm{FM}(h_i)\,\mathrm{FM}(f_i)$ liegt $\mathrm{FM}(f)$ daher im von den $\mathrm{FM}(f_i)$ erzeugten Ideal.

Falls sich die maximalen unter den führenden Termen $\mathrm{FT}(h_i f_i)$ gegenseitig wegkürzen, läßt sich die entsprechende Teilsumme der $h_i f_i$ nach dem vorigen Lemma auch als eine Summe von S-Polynomen schreiben. Diese wiederum lassen sich nach Voraussetzung durch den Divisionsalgorithmus als Linearkombinationen der f_i darstellen. Damit erhalten wir eine neue Darstellung

$$f = \widetilde{h}_1 f_1 + \cdots + \widetilde{h}_m f_m \quad \text{mit} \quad \widetilde{h}_i \in R\,,$$

in der der maximale Multigrad eines Summanden echt kleiner ist als in der obigen Darstellung, denn in der Darstellung als Summe von S-Polynomen sind die Terme mit dem maximalem Multigrad verschwunden.

Mit dieser Darstellung können wir wie oben argumentieren: Falls sich bei den führenden Termen nichts wegkürzt, haben wir $\mathrm{FM}(f)$ als Element des von den $\mathrm{FM}(f_i)$ erzeugten Ideals dargestellt, andernfalls erhalten wir wieder via S-Polynome und deren Reduktion eine neue Darstellung von f als Linearkombination der f_i mit noch kleinerem maximalem Multigrad der Summanden, und so weiter. Das Verfahren muß schließlich mit einer Summe ohne Kürzungen bei den führenden Termen

enden, da es nach der Wohlordnungseigenschaft einer Monomordnung keine unendliche absteigende Folge von Multigraden geben kann. ∎

Der BUCHBERGER-Algorithmus in seiner einfachsten Form macht aus diesem Satz ein Verfahren zur Berechnung einer GRÖBNER-Basis aus einem vorgegebenen Erzeugendensystem eines Ideals:

Gegeben sind m Elemente $f_1, \ldots, f_m \in R = k[X_1, \ldots, X_n]$.

Berechnet wird eine GRÖBNER-Basis $g_1, \ldots, g_r$ des davon erzeugten Ideals $I = (f_1, \ldots, f_m)$ mit $g_i = f_i$ für $i \leq m$.

1. Schritt (Initialisierung): Setze $g_i = f_i$ für $i = 1, \ldots, m$; die Menge $\{g_1, \ldots, g_m\}$ werde mit G bezeichnet.

2. Schritt: Setze $G' = G$ und teste für jedes Paar $(f, g) \in G' \times G'$ mit $f \neq g$, ob der Rest r bei der Division von $S(f, g)$ durch die Elemente von G' (in irgendeiner Reihenfolge angeordnet) verschwindet. Falls nicht, wird G ersetzt durch $G \cup \{r\}$.

3. Schritt: Ist $G = G'$, so endet der Algorithmus mit G als Ergebnis; andernfalls geht es zurück zum zweiten Schritt.

Wenn der Algorithmus im dritten Schritt endet, ist der Rest bei der Division von $S(f, g)$ durch die Elemente von G stets das Nullpolynom; nach dem gerade bewiesenen Satz ist G daher eine GRÖBNER-Basis. Da sowohl die S-Polynome als auch ihre Divisionsreste in I liegen und G ein Erzeugendensystem von I enthält, ist auch klar, daß es sich dabei um eine GRÖBNER-Basis von I handelt. Wir müssen uns daher nur noch überlegen, daß der Algorithmus nach endlich vielen Iterationen abbricht.

Wenn im zweiten Schritt ein nichtverschwindender Divisionsrest r auftaucht, ist dessen führendes Monom durch kein führendes Monom eines Polynoms $g \in G$ teilbar. Das von den führenden Monomen der $g \in G$ erzeugte Ideal von R wird daher größer, nachdem G um r erweitert wurde. Wenn dies unbeschränkt möglich wäre, erhielten wir daher eine unendliche aufsteigende Folge von monomialen Idealen J_i, von denen jedes echt größer ist als sein Vorgänger:

$$J_1 < J_2 < \cdots J_i < J_{i+1} < \cdots .$$

Natürlich ist auch die Vereinigung J aller J_i ein monomiales Ideal, hat also nach dem Lemma von DICKSON ein endliches Erzeugendensystem $\{M_1, \ldots, M_q\}$. Da jedes M_j in einem J_i und damit auch in allen folgenden liegen muß, gibt es ein m, so daß alle M_j in J_m liegen. Damit ist $J = (M_1, \ldots, M_q) \subseteq J_m$, im Widerspruch zur Annahme, daß J_{m+1} und damit auch J echt größer als M_j ist.

Der Algorithmus kann natürlich auf mehrere offensichtliche Weisen optimiert werden: Beispielsweise stößt man beim wiederholten Durchlaufen des zweiten Schritts immer wieder auf dieselben S-Polynome, die daher nicht jedes Mal neu berechnet werden müssen, und wenn eines dieser Polynome einmal Divisionsrest Null hatte, hat es auch bei jedem weiteren Durchgang Divisionsrest Null, denn dann wird ja wieder durch dieselben Polynome (plus einiger neuer) dividiert. Es gibt inzwischen auch zahlreiche nicht offensichtliche Verbesserungen und Optimierungen; wir wollen uns aber mit dem Prinzip begnügen und stattdessen später noch einige andere Themen behandeln.

Der BUCHBERGER-Algorithmus hat den Nachteil, daß er das vorgegebene Erzeugendensystem in jedem Schritt größer macht ohne je ein Element zu streichen. Dies ist weder beim GAUSS-Algorithmus noch beim EUKLIDischen Algorithmus der Fall, bei denen jeweils eine Gleichung durch eine andere *ersetzt* wird. Obwohl wir sowohl die Eliminationsschritte des GAUSS-Algorithmus als auch die einzelnen Schritte der Polynomdivisionen beim EUKLIDischen Algorithmus durch S-Polynome ausdrücken können, *müssen* wir im allgemeinen Fall zusätzlich zu g und $S(f,g)$ auch noch das Polynom f beibehalten; andernfalls kann sich die Lösungsmenge ändern:

Als Beispiel können wir das Gleichungssystem

$$f(X,Y) = X^2Y + XY^2 + 1 = 0 \quad \text{und} \quad g(X,Y) = X^3 - XY - Y = 0$$

betrachten. Wenn wir mit der lexikographischen Ordnung arbeiten, sind hier die einzelnen Monome bereits der Größe nach geordnet, insbesondere stehen also die führenden Monome an erster Stelle und

$$S(f,g) = Xf(X,Y) - Yg(X,Y) = X^2Y^2 + XY^2 + X + Y^2 .$$

Der führende Term X^2Y^2 ist durch den führenden Term X^2Y von f teilbar; subtrahieren wir Yf vom S-Polynom, erhalten wir das nicht weiter reduzierbare Polynom

$$h(X,Y) = -XY^3 + XY^2 + X + Y^2 - Y\,.$$

Sowohl $g(X,Y)$ als auch $h(X,Y)$ verschwinden im Punkt $(0,0)$; dieser ist aber keine Lösung des Ausgangssystems, da $f(0,0) = 1$ nicht verschwindet.

Aus diesem Grund werden die nach dem BUCHBERGER-Algorithmus berechneten GRÖBNER-Basen oft sehr groß und unhandlich. Betrachten wir dazu als Beispiel das System aus den beiden Gleichungen

$$f_1 = X^3 - 2XY \quad \text{und} \quad f_2 = X^2Y - 2Y^2 + X$$

und berechnen eine GRÖBNER-Basis bezüglich der graduiert lexikographischen Ordnung.

$$S(f_1, f_2) = Yf_1 - Xf_2 = -X^2$$

ist weder durch den führenden Term von f_1 noch den von f_2 teilbar, muß also als neues Element f_3 in die Basis aufgenommen werden.

$$S(f_1, f_3) = f_1 + Xf_3 = -2XY$$

kann wieder mit keinem der f_i reduziert werden, muß also als neues Element f_4 in die Basis. Genauso ist es mit

$$f_5 = S(f_2, f_3) = f_2 + Yf_3 = -2Y^2 + X\,.$$

Für das so erweiterte Erzeugendensystem, bestehend aus den Polynomen

$$f_1 = X^3 - 2XY, \quad f_2 = X^2Y - 2Y^2 + X, \quad f_3 = -X^2,$$
$$f_4 = -2XY \quad \text{und} \quad f_5 = -2Y^2 + X\,,$$

sind die S-Polynome

$$S(f_1, f_2) = f_3, \quad S(f_1, f_3) = f_4 \quad \text{und} \quad S(f_2, f_3) = f_5$$

trivialerweise auf Null reduzierbar, die anderen Kombinationen müssen wir nachrechnen:

$$S(f_1, f_4) = Yf_1 + \frac{X^2}{2} f_4 = -2XY^2 = Yf_4$$

$$S(f_1, f_5) = Y^2 f_1 + \frac{X^3}{2} f_5 = -2XY^3 + \frac{X^4}{2} = \frac{X}{2} f_1 + f_2 + Y^2 f_4 - f_5$$

$$S(f_2, f_4) = f_2 + \frac{X}{2} f_4 = -2Y^2 + X = f_5$$

$$S(f_2, f_5) = Yf_2 + \frac{X^2}{2} f_5 = \frac{X^3}{2} + XY - 2Y^3 = \frac{1}{2} f_1 - \frac{1}{2} f_4 + Yf_5$$

$$S(f_3, f_4) = -Yf_3 - \frac{X}{2} f_4 = 0$$

$$S(f_3, f_5) = -Y^2 f_3 - \frac{X^2}{2} f_5 = \frac{1}{2} f_1 - \frac{1}{2} f_4$$

$$S(f_4, f_5) = -\frac{Y}{2} f_4 - \frac{X}{2} f_5 = \frac{X^2}{2} = -\frac{1}{2} f_3$$

Somit bilden diese fünf Polynome eine GRÖBNER-Basis des von f_1 und f_2 erzeugten Ideals.

Zum Glück brauchen wir aber nicht alle fünf Polynome. Das folgende Lemma gibt ein Kriterium, wann man auf ein Erzeugendes verzichten kann, und illustriert gleichzeitig das allgemeine Prinzip, wonach bei einer GRÖBNER-Basis alle wichtigen Eigenschaften anhand der führenden Termen ablesbar sein sollten:

Lemma: G sei eine GRÖBNER-Basis des Ideals $I \lhd k[X_1, \ldots, X_n]$, und $g \in G$ sei ein Polynom, dessen führendes Monom im von den führenden Monomen der restlichen Basiselemente erzeugten monomialen Ideal liegt. Dann ist auch $G \setminus \{g\}$ eine GRÖBNER-Basis von I.

Beweis: $G \setminus \{g\}$ ist nach Definition genau dann eine GRÖBNER-Basis von I, wenn die führenden Terme der Basiselemente das Ideal FM(I) erzeugen. Da G eine GRÖBNER-Basis von I ist und die führenden Terme egal ob mit oder ohne FT(g) dasselbe monomiale Ideal erzeugen, ist das klar.

■

Man beachte, daß sich dieses Lemma nur anwenden läßt, wenn G eine GRÖBNER-Basis von I ist; wir können nicht schon während des Rechengangs im BUCHBERGER-Algorithmus Elemente streichen. Im obigen Beispiel etwa wird das Ideal $I = (f_1, f_2)$ natürlich auch erzeugt von f_1, f_2 und f_3; dabei ist $\mathrm{FM}(f_1) = X^3$, $\mathrm{FM}(f_2) = X^2Y$, und $\mathrm{FM}(f_3) = X^2$ teilt beide dieser Monome. Wenn das Lemma auf die Basis f_1, f_2, f_3 anwendbar wäre, könnten wir also f_1 und f_2 streichen und f_3 wäre für sich allein eine GRÖBNER-Basis von I. Natürlich ist aber $I \neq (-X^2)$, denn weder f_1 noch f_2 sind Vielfache von X^2.

Von der Menge $\{f_1, f_2, f_3, f_4, f_5\}$ haben wir mit Hilfe des Kriteriums von BUCHBERGER verifiziert, daß sie eine GRÖBNER-Basis von I ist; deshalb können wir das Lemma darauf anwenden und f_1, f_2 streichen. Wir können das aber erst jetzt tun, denn im Verlauf der Berechnungen wurden f_1 und f_2 noch gebraucht um $f_4 = S(f_1, f_3)$ und $f_5 = S(f_2, f_3)$ zu konstruieren. Somit ist $I = (f_3, f_4, f_5)$, und darauf können wir das Lemma nicht weiter anwenden, denn

$$\mathrm{FM}(f_3) = X^2, \quad \mathrm{FM}(f_4) = XY \quad \text{und} \quad \mathrm{FM}(f_5) = Y^2,$$

und keines dieser drei Monome ist Vielfaches eines der anderen.

Zur weiteren Normierung können wir noch durch die führenden Koeffizienten teilen und erhalten dann die *minimale* GRÖBNER-Basis

$$\widetilde{f}_3 = X^2, \quad \widetilde{f}_4 = XY \quad \text{und} \quad \widetilde{f}_5 = Y^2 - \frac{X}{2}.$$

Definition: Eine *minimale* GRÖBNER-Basis von I ist eine GRÖBNER-Basis von I mit folgenden Eigenschaften:

1.) Alle $g \in G$ haben den führenden Koeffizienten eins

2.) Für kein $g \in G$ liegt $\mathrm{FM}(g)$ im von den führenden Monomen der übrigen Elemente erzeugten Ideal.

Da ein Monom X^α genau dann im von einer Menge M von Monomen erzeugten Ideal liegt, wenn es durch eines dieser Monome teilbar ist, können wir die zweite Bedingung auch so ausdrücken, daß es keine zwei Elemente $g \neq g'$ in G geben darf, für die $\mathrm{FM}(g)$ ein Teiler von $\mathrm{FM}(g')$ ist.

Es ist klar, daß jede GRÖBNER-Basis zu einer minimalen GRÖBNER-Basis verkleinert werden kann: Durch Division können wir alle führenden Koeffizienten zu eins machen ohne etwas an der Erzeugung zu ändern, und nach obigem Lemma können wir nacheinander alle Elemente eliminieren, die die zweite Bedingung verletzen.

Wir können aber noch mehr erreichen: Wenn nicht das führende, sondern einfach *irgendein* Monom eines Polynoms $g \in G$ im von den führenden Termen der übrigen Elemente erzeugten Ideal liegt, ist dieses Monom teilbar durch das führende Monom eines anderen Polynoms $h \in G$. Wir können den Term mit diesem Monom daher zum Verschwinden bringen, indem wir g ersetzen durch g minus ein Vielfaches von h. Da sich dabei nichts an den führenden Termen der Elemente von G ändert, bleibt G eine GRÖBNER-Basis. Wir können somit aus den Elementen einer minimalen GRÖBNER-Basis Terme eliminieren, die durch den führenden Term eines anderen Elements teilbar sind. Was dabei schließlich entstehen sollte, ist eine *reduzierte* GRÖBNER-Basis:

Definition: Eine reduzierte GRÖBNER-Basis von I ist eine GRÖBNER-Basis von I mit folgenden Eigenschaften:
1.) Alle $g \in G$ haben den führenden Koeffizienten eins
2.) Für kein $g \in G$ liegt ein Monom von g im von den führenden Monomen der übrigen Elemente erzeugten Ideal.

Die minimale Basis im obigen Beispiel ist offenbar schon reduziert, denn außer $\widetilde{f_5}$ bestehen alle Basispolynome nur aus dem führendem Term, und bei $\widetilde{f_5}$ ist der zusätzliche Term linear, kann also nicht durch die quadratischen führenden Monome der anderen Polynome teilbar sein.

Reduzierte GRÖBNER-Basis haben eine für das praktische Rechnen mit Idealen sehr wichtige zusätzliche Eigenschaft:

Satz: Jedes Ideal $I \lhd k[X_1, \ldots, X_n]$ hat (bei vorgegebener Monomordnung) eine eindeutig bestimmte reduzierte GRÖBNER-Basis.

Beweis: Wir gehen aus von einer minimalen GRÖBNER-Basis G und ersetzen nacheinander jedes Element $g \in G$ durch seinen Rest bei der

Polynomdivision durch $G \setminus \{g\}$. Da bei einer minimalen GRÖBNER-Basis kein führendes Monom eines Element das führende Monom eines anderen teilen kann, ändert sich dabei nichts an den führenden Termen, G ist also auch nach der Ersetzung eine minimale GRÖBNER-Basis. In der schließlich entstehenden Basis hat kein $g \in G$ mehr einen Term, der durch den führenden Term eines Elements von $G \setminus \{g\}$ teilbar wäre, denn auch wenn wir bei der Reduktion der einzelnen Elemente durch eine eventuell andere Menge geteilt haben, hat sich doch an den führenden Termen der Basiselemente nichts geändert. Also gibt es eine reduzierte GRÖBNER-Basis.

Nun seien G und G' zwei reduzierte GRÖBNER-Basen von I. Jedes Element $f \in G'$ liegt insbesondere in I, also ist $\overline{f}^{G} = 0$. Insbesondere muß der führende Term von f durch den führenden Term eines $g \in G$ teilbar sein. Umgekehrt ist aber auch $\overline{g}^{G'} = 0$, d.h. der führende Term von g muß durch den führenden Term eines Elements von $f' \in G'$ teilbar sein. Dieser führende Term teilt dann insbesondere den führenden Term von f, und da G' als reduzierte GRÖBNER-Basis minimal ist, muß $f' = f$ sein. Somit gibt es zu jedem $g \in G$ genau ein $f \in G'$ mit FM(f) = FM(g); insbesondere haben G und G' dieselbe Elementanzahl. Tatsächlich muß sogar $f = g$ sein, denn $f - g$ liegt in I, enthält aber keine Term, der durch den führenden Term irgendeines Elements von G teilbar wäre. Also ist $f - g = 0$.

■

Bemerkung: Die Forderung in den Definitionen von minimalen und reduzierten GRÖBNER-Basen, daß alle führenden Koeffizienten eins sein müssen, ist zwar nützlich für theoretische Diskussionen, führt aber im Falle von Polynomen mit rationalen Koeffizienten oft dazu, daß die Koeffizienten Nenner haben. Computeralgebrasysteme können zwar mit rationalen Zahlen rechnen, indem sie diese durch Paare teilerfremder ganzer Zahlen darstellen, aber diese Rechnungen sind erheblich aufwendiger als solche mit ganzen Zahlen. Daher liefern einige Computeralgebrasysteme beim Kommando zur Berechnung einer reduzierten GRÖBNER-Basis anstelle von Polynomen mit führendem Koeffizienten eins solche mit teilerfremden ganzzahligen Koeffizienten.

§6: Gröbner-Basen für nichtlineare Gleichungssysteme

GRÖBNER-Basen haben eine Vielzahl von Anwendungen in der Algebra; wir wollen uns hier vor allem damit beschäftigen, wie sie direkt oder im Zusammenspiel mit anderen Methoden zur expliziten Lösung nichtlinearer Gleichungssysteme führen können. Explizit angebbar sind die Lösungen meist nur, wenn die Lösungsmenge endlich ist; daher werden wir uns meist auf solche Systeme beschränken und interessieren uns daher auch für Kriterien, wie wir einem Gleichungssystem die Endlichkeit seiner Lösungsmenge ansehen können.

Wir gehen aus von m Polynomgleichungen

$$f_i(x_1, \ldots, x_n) = 0 \quad \text{mit} \quad f_i \in k[X_1, \ldots, X_n] \quad \text{für} \quad i = 1, \ldots, m$$

und suchen die Lösungsmenge

$$\left\{(x_1, \ldots, x_n) \in k^n \;\middle|\; f_i(x_1, \ldots, x_n) = 0 \text{ für } i = 1, \ldots, m\right\}.$$

Diese wird allerdings oft leer sein; für $f_1 = X^2 - 2$ und $f_2 = Y^2 - 3$ aus $\mathbb{Q}[X]$ etwa ist diese Menge leer, da die Lösungen $(\pm\sqrt{2}, \pm\sqrt{3})$ nicht in $\mathbb{Q}^2$ liegen. Wir betrachten daher meist noch einen zweiten Körper K, der k enthält, und interessieren uns allgemeiner für die Lösungsmenge in K^n:

Definition: *a)* Ist I ein Ideal in $k[X_1, \ldots, X_n]$, und ist K ein Körper, der k enthält, setzen wir

$$V_K(I) = \left\{(x_1, \ldots, x_n) \in K^n \;\middle|\; f(x_1, \ldots, x_n) = 0 \text{ für alle } f \in I\right\}.$$

b) Für $I = (f_1, \cdots, f_m)$ schreiben wir auch kurz $V_K(f_1, \ldots, f_m)$ an Stelle von $V_K(I)$.

Der Körper k sollte dabei möglichst klein sein, denn mit den Elementen dieses Körpers müssen wir rechnen, und je größer der Körper, desto aufwendiger sind seine Rechenoperationen. In konkreten Beispielen werden wir uns meist auf $k = \mathbb{Q}$ beschränken und – soweit möglich – sogar versuchen, unsere Konstruktionen in $\mathbb{Z}[X]$ durchzuführen.

Der Körper K hingegen sollte so groß sein, daß er für ein Gleichungssystem, daß in irgendeinem Körper eine nichtleere endliche Lösungsmenge hat, diese Lösungsmenge enthält. Wir werden meist $K = \mathbb{C}$ betrachten.

Wie wir bereits aus §1 des vorigen Kapitels wissen, hängt die Lösungsmenge des Gleichungssystems nur ab vom Ideal $I = (f_1, \ldots, f_m)$; wir suchen ein Erzeugendensystem $\{g_1, \ldots, g_r\}$ dieses Ideals, aus dem wir mehr über die Mengen

$$V_K(I) = V_K(f_1, \ldots, f_m) = V_K(g_1, \ldots, g_r)$$

ablesen können. Wir erwarten natürlich, daß wir hier vor allem im Falle einer geeigneten GRÖBNER-Basis $\{g_1, \ldots, g_r\}$ eventuell Erfolg haben.

Viele Lösungsansätze für Gleichungssysteme in mehreren Veränderlichen beruhen auf der Elimination von Variablen: Im ℓ-ten Schritt suchen wir nach Bedingungen, die ein $(n - \ell)$-Tupel $(x_{\ell+1}, \ldots, x_n)$ erfüllen muß, wenn es ein ℓ-Tupel $(x_1, \ldots, x_\ell)$ gibt, so daß $(x_1, \ldots, x_n)$ in $V(I)$ liegt. Eine solche Bedingung ist trivial: Für jedes Polynom $f \in I$, in dem die Variablen $X_1, \ldots, X_\ell$ nicht vorkommen, muß $f(x_{\ell+1}, \ldots, x_n) = 0$ sein.

Definition: *a)* Das ℓ-te *Eliminationsideal* eines Ideal $I \lhd k[X_1, \ldots, X_n]$ ist $I_\ell = I \cap k[X_{\ell+1}, \ldots, X_n]$.
b) Eine Monomordnung $<$ heißt *Eliminationsordnung* für $X_1, \ldots, X_\ell$, wenn jedes Monom, das mindestens eine der Variablen $X_1, \ldots, X_\ell$ enthält, größer ist als alle Monome, die nur $X_{\ell+1}, \ldots, X_n$ enthalten.

Die lexikographische Ordnung mit $X_1 > X_2 > \cdots > X_{n-1} > X_n$ ist offensichtlich für jedes ℓ eine Eliminationsordnung für $X_1, \ldots, X_\ell$, die graduiert lexikographische aber nicht, da bezüglich dieser beispielsweise $X_1 < X_n^2$ ist.

Satz: Ist G eine GRÖBNER-Basis von I bezüglich einer Eliminationsordnung für $X_1, \ldots, X_\ell$, so ist $G \cap I_\ell$ eine GRÖBNER-Basis von I_ℓ.

Beweis: Die Elemente von $G = \{g_1, \ldots, g_m\}$ seien so angeordnet, daß $G \cap I_\ell = \{g_1, \ldots, g_r\}$ ist. Wir müssen zeigen, daß sich jedes $f \in I_\ell$ als Linearkombination von $g_1, \ldots, g_r$ mit Koeffizienten aus $k[X_{\ell+1}, \ldots, X_n]$ darstellen läßt.

Der Divisionsalgorithmus bezüglich der gewählten Ordnung gibt uns eine Darstellung $f = h_1 g_1 + \cdots + h_m g_m$ von f als Element von I.

Die Polynome $g_{r+1}, \ldots, g_m$ enthalten jeweils mindestens eine der Variablen $X_1, \ldots, X_\ell$, und da wir eine Eliminationsordnung verwenden, muß auch das führende Monom eine dieser Variablen enthalten. Da kein Monom von f eine dieser Variablen enthält, kann im Divisionsalgorithmus das führende Monom eines dieser Polynome nie Teiler des führenden Monoms des jeweils betrachteten Polynoms p sein, Somit ist $h_{r+1} = \cdots = h_m = 0$, und in keinem der Polynome $h_1, \ldots, h_r$ kann eine der Variablen $X_1, \ldots, X_\ell$ auftreten. Dies zeigt, daß f im von $g_1, \ldots, g_r$ erzeugten Ideal von $k[X_{\ell+1}, \ldots, X_n]$ liegt, d.h. dieses Ideal wird von $g_1, \ldots, g_r$ erzeugt.

Um zu zeigen, daß es sich dabei sogar um eine GRÖBNER-Basis handelt, können wir zum Beispiel zeigen, daß alle $S(g_i, g_j)$ mit $i, j \leq r$ ohne Rest durch $g_1, \ldots, g_r$ teilbar sind. Da G nach Voraussetzung eine GRÖBNER-Basis ist, sind sie auf jeden Fall ohne Rest durch G teilbar, und wieder kann bei der Division nie der führende Term eines Dividenden durch den eines g_i mit $i > r$ teilbar sein, d.h. $S(g_i, g_j)$ ist als Linearkombination von $g_1, \ldots, g_r$ mit Koeffizienten aus $k[g_1, \ldots, g_r]$ darstellbar. ∎

Daraus ergibt sich eine Strategie zur Lösung nichtlinearer Gleichungssysteme nach Art des GAUSS-Algorithmus: Wir gehen aus von der lexikographischen Ordnung, die ja für jedes ℓ eine Eliminationsordnung für $X_1, \ldots, X_\ell$ ist, und bestimmen eine (reduzierte) GRÖBNER-Basis für das von den Gleichungen erzeugte Ideal des Polynomrings $k[X_1, \ldots, X_n]$. Dann betrachten als erstes das Eliminationsideal I_{n-1}. Dieses besteht nur aus Polynomen in X_n; falls wir mit einer reduzierten GRÖBNER-Basis arbeiten, gibt es darin höchstens ein solches Polynom.

Falls es ein solches Polynom gibt, muß jede Lösung des Gleichungssystem als letzte Komponente eine von dessen Nullstellen haben. Wir bestimmen daher diese Nullstellen (in K) und setzen sie nacheinander in das restliche Gleichungssystem ein. Dadurch erhalten wir Gleichungssysteme in $n-1$ Unbekannten, wo wir nach Gleichungen nur in X_{n-1} suchen können. Diese erhalten wir, indem wir bei allen Erzeugenden des Eliminationsideals I_{n-2} für X_n nacheinander die Werte aus $V_K(I_{n-1}) \subset k$ einsetzen. Nachdem wir so $V_K(I_{n-2}) \subset K^2$ bestimmt haben, können wir analog die Mengen $V_K(I_{n-3}) \subset K^3$ und so weiter

bis $V_K(I) \subset K^n$ bestimmen.

Betrachten wir noch einmal das Beispiel gegen Ende von §5 des vorigen Kapitels mit

$$f_1 = X^3 - 2XY \quad \text{und} \quad f_2 = X^2Y - 2Y^2 + X\,.$$

Dort hatten wir die reduzierte GRÖBNER-Basis bezüglich der graduiert lexikographischen Ordnung berechnet; sie besteht aus

$$g_1 = X^2, \quad g_2 = XY \quad \text{und} \quad g_3 = Y^2 - \frac{X}{2}\,.$$

Da die graduiert lexikographische Ordnung keine Eliminationsordnung für X ist, können wir nicht erwarten, daß $\{g_1, g_2, g_3\} \cap k[Y]$ ein Erzeugendensystem des Eliminationsideals $(f_1, f_2) \cap k[Y]$ liefert, und in der Tat liegt keines der g_i in $k[Y]$. Zufälligerweise liegt aber $g_1 = X^2$ in $k[X]$, wir wissen also, daß für jede Lösung (x, y) des Gleichungssystems $x = 0$ sein muß. $g_2 = XY$ verschwindet für alle solche Punkte automatisch, und $g_3 = Y^2 - X/2$ verschwindet genau dann, wenn auch $y = 0$ ist. Somit ist $V(f_1, f_2) = \{(0, 0)\}$.

Wenn wir das Gleichungssystem mit dem hier vorgestellten Verfahren lösen wollen. müssen wir mit der lexikographischen Ordnung arbeiten. Da die führenden Terme von f_1 und f_2 bei beiden Ordnungen gleich sind und viele der zu berechnenden S-Polynome nur aus einem Term bestehen, ändert sich zunächst nichts: Wie bei der graduiert lexikographischen Ordnung kommen wir auf

$$f_3 = S(f_1, f_2) = -X^2, \quad f_4 = S(f_1, f_3) = -2XY \quad \text{und}$$
$$f_5 = S(f_2, f_3) = X - 2Y^2\,.$$

Auch $S(f_1, f_4) = -2XY^2 = Yf_4$ kann wie dort auf Null reduziert werden, bei der Berechnung von $S(f_1, f_5)$ ist jetzt aber nicht mehr Y^2, sondern X das führende Monom. Somit ist

$$S(f_1, f_5) = f_1 - X^2 f_5 = 2X^2Y^2 - 2XY = 2Yf_2 + 2f_4 + 4Y^3\,,$$

das S-Polynom läßt sich also modulo $\{f_1, f_2, f_3, f_4, f_5\}$ nicht auf Null reduzieren und wir müssen $f_6 = 4Y^3$ als neues Element in die Basis aufnehmen. Erst jetzt zeigt eine mühsame Rechnung, die man am besten

seinem Computer überläßt, daß $S(f_i, f_j)$ für alle $1 \leq i < j \leq 6$ modulo $\{f_1, f_2, f_3, f_4, f_5, f_6\}$ auf Null reduziert werden kann, womit wir eine GRÖBNER-Basis gefunden haben.

Die führenden Monome der sechs Basiselemente bezüglich der lexikographischen Ordnung sind

$$\begin{array}{lll} \text{FM}(f_1) = X^3\,, & \text{FM}(f_2) = X^2Y\,, & \text{FM}(f_3) = -X^2\,, \\ \text{FM}(f_4) = -2XY\,, & \text{FM}(f_5) = X,\,, & \text{FM}(f_6) = 4Y^3\,; \end{array}$$

wir können also f_1 bis f_4 eliminieren. Die reduzierte GRÖBNER-Basis bedeutet besteht somit aus $g_1 = X - 2Y^2$ und $g_2 = Y^3$.

Das Eliminationsideal I_1 wird daher erzeugt von $g_2 = Y^3$, d.h. für jede Lösung (x, y) muß y verschwinden. Setzen wir $y = 0$ in g_1 ein, so sehen wir, daß auch x verschwinden muß, der Nullpunkt ist also die einzige Lösung.

Es war ein Zufall, daß wir dieses Ergebnis auch der GRÖBNER-Basis bezüglich der graduiert lexikographischen Ordnung ansehen konnten; bei komplizierteren Systemen wird dort oft jedes Basiselement alle Variablen enthalten, so daß wir nichts sehen können. Trotzdem kann die graduiert lexikographische Ordnung zur Lösung nichtlinearer Gleichungssysteme nützlich sein: 1993 publizierten J.C. FAUGÈRE, P. GIANNI, D. LAZARD und T. MORA einen heute nach ihren Anfangsbuchstaben als FGLM benannten Algorithmus, der für ein Ideal I mit endlicher Nullstellenmenge $V(I)$ effizient eine GRÖBNER-Basis bezüglich der lexikographischen Ordnung bestimmt auf dem Umweg über die graduiert lexikographische Ordnung. Wir werden später sehen, daß wir im Falle einer endlichen Lösungsmenge diese auch ausgehend von einer beliebigen GRÖBNER-Basis mit alternativen Techniken bestimmen können.

Nun kann es beim obigen Verfahren für nichtlineare Gleichungssysteme natürlich vorkommen, daß I_{n-1} das Nullideal ist; falls unter den Lösungen des Systems unendlich viele Werte für die letzte Variable vorkommen, muß das sogar so sein. Es kann sogar vorkommen, daß *alle* Eliminationsideale außer $I_0 = I$ das Nullideal sind. In diesem Fall führt die gerade skizzierte Vorgehensweise zu nichts.

Bevor wir uns darüber wundern, sollten wir uns überlegen, was wir überhaupt unter der Lösung eines nichtlinearen Gleichungssystems verstehen wollen. Im Falle einer endlichen Lösungsmenge ist das klar: Dann wollen wir eine Auflistung der sämtlichen Lösungstupel. Bei einer unendlichen Lösungsmenge ist das aber nicht mehr möglich. Im Falle eines linearen Gleichungssystems wissen wir, daß die Lösungsmenge ein affiner Raum ist; wir können sie daher auch wenn sie unendlich sein sollte durch endlich viele Daten eindeutig beschreiben, zum Beispiel durch eine spezielle Lösung und eine Basis des Lösungsraums des zugehörigen homogenen Gleichungssystems.

Bei nichtlinearen Gleichungssystemen gibt es im allgemeinen keine solche Beschreibung unendlicher Lösungsmengen: Die Lösungsmenge des Gleichungssystems

$$X^2 + 2Y^2 + 3Z^2 = 100 \quad \text{und} \quad 2X^2 + 3Y^2 - Z^2 = 0$$

etwa ist die Schnittmenge eines Ellipsoids mit einem elliptischen Kegel; sie besteht aus zwei ovalen Kurven höherer Ordnung. Die GRÖBNER-Basis besteht in diesem Fall aus den beiden Polynomen

$$X^2 - 11Z^2 + 300 \quad \text{und} \quad Y^2 + 7Z^2 - 200\,,$$

stellt uns dieselbe Menge also dar als Schnitt eines hyperbolischen und eines elliptischen Zylinders. Eine explizitere Beschreibung der Lösungsmenge ist schwer vorstellbar.

Auf der Basis von STURMschen Ketten, dem Lemma von THOM und Verallgemeinerungen davon hat die semialgebraische Geometrie Methoden entwickelt, wie man auch allgemeinere Lösungsmengen nichtlinearer Gleichungssysteme durch eine sogenannte zylindrische Zerlegung qualitativ beschreiben kann; dazu wird der $\mathbb{R}^n$ in Teilmengen zerlegt, in denen die Lösungsmenge entweder ein einfaches qualitatives Verhalten hat oder aber leeren Durchschnitt mit der Teilmenge. Dadurch kann man insbesondere feststellen, in welchen Regionen des $\mathbb{R}^n$ Lösungen zu finden sind; diese Methoden sind Gegenstand der reell-algebraischen Geometrie.

In manchen Fällen lassen sich Lösungsmengen parametrisieren. Wie man mit Methoden der algebraischen Geometrie zeigen kann, ist das

aber im allgemeinen nur bei Gleichungen kleinen Grades der Fall und kommt daher für allgemeine Lösungsalgorithmen nicht in Frage.

Stets möglich ist das umgekehrte Problem, d.h. die Beschreibung einer parametrisch gegebenen Menge in impliziter Form. Hier gehen wir aus von Gleichungen der Form

$$x_1 = \varphi_1(t_1, \ldots, t_m), \quad \ldots, \quad x_n = \varphi_n(t_1, \ldots, t_m)\,,$$

und wir suchen Polynome $f_1, \ldots, f_r$ aus $k[X_1, \ldots, X_n]$, die auf der Menge aller jener $(x_1, \ldots, x_n)$ verschwinden, für die es eine solche Darstellung gibt (und eventuell noch auf Grenzwerten davon).

Dazu wählen wir eine lexikographische Ordnung auf dem Polynomring $k[T_1, \ldots, T_m, X_1, \ldots, X_n]$, bei der alle T_i größer sind als die X_j, und bestimmen eine GRÖBNER-Basis für das von den Polynomen $X_i - \varphi_i(T_1, \ldots, T_m)$ erzeugte Ideal. Dessen Schnitt mit $k[X_1, \ldots, X_n]$ ist ein Eliminationsideal, hat also als Basis genau die Polynome aus der GRÖBNER-Basis, in denen keine T_i vorkommen.

Fast genauso können wir auch zu einer vorgegebenen endlichen Menge von Punkten ein Gleichungssystem konstruieren, das genau diese Menge als Lösungsmenge hat; dies spielt beispielsweise in der algebraischen Statistik eine Rolle, wenn zu einem vorgegebenen Design die damit schätzbaren Modelle identifiziert werden sollen.

Wir gehen aus von r Punkten

$$P_i = \left(x_1^{(i)}, \ldots, x_n^{(i)}\right) \in k^n, \quad i = 1, \ldots, r,$$

und suchen ein Ideal $I \lhd k[X_1, \ldots, X_n]$, dessen Elemente genau in den Punkten P_i verschwinden. Im Falle nur eines Punktes P_i können wir einfach das Ideal

$$I_i = \left(X_1 - x_1^{(i)}, \ldots, X_n - x_n^{(i)}\right)$$

nehmen; bei mehreren Punkten brauchen wir den Durchschnitt der Ideale I_1 bis I_r, für den wir kein offensichtliches Erzeugendensystem haben.

Betrachten wir stattdessen die Punkte

$$Q_i = \left(t_1^{(i)}, \ldots, t_r^{(i)}, x_1^{(i)}, \ldots, x_n^{(i)}\right) \in k^{r+n} \quad \text{mit} \quad t_j^{(i)} = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{sonst} \end{cases},$$

so erzeugen die Polynome

$$\left(X_j - x_j^{(i)}\right)T_i \in k[T_1, \ldots, T_r, X_1, \ldots, X_n]$$

für $i = 1, \ldots, n$ und $j = 1, \ldots, r$ zusammen mit dem Polynom $T_1 + \cdots + T_r - 1$ ein Ideal, das alle Punkte Q_i als Nullstellen hat: Die Polynome $\left(X_j - x_j^{(i)}\right)T_i$ verschwinden in Q_i, da $x_j^{(i)}$ die j-te Koordinate von Q_i ist, und für $\ell \neq i$ verschwindet $\left(X_j - x_j^{(i)}\right)T_\ell$, da $t_\ell^{(i)}$ verschwindet.

Ist umgekehrt $Q = (t_1, \ldots, t_r, x_1, \ldots, x_n) \in k^{r+n}$ keiner der Punkte Q_i, so gibt es für jedes i mindestens eine Koordinate, in der sich Q von Q_i unterscheidet. Ist dies etwa die j-te Koordinate, so ist $X_j - x_j^{(i)}$ in Q von Null verschieden; $\left(X_j - x_j^{(i)}\right)T_i$ kann daher nur verschwinden, wenn $t_i = 0$ ist. Dies kann aber nicht für alle i der Fall sein, denn die Summe der t_i ist eins, da $T_1 + \cdots + T_r - 1$ verschwindet. Somit liegt Q nicht in $V(J)$.

Damit haben wir ein Ideal $J \lhd k[T_1, \ldots, T_r, X_1, \ldots, X_n]$ gefunden, dessen Nullstellen genau die Punkte $Q_1, \ldots, Q_r \in k^{r+n}$ sind. Die Punkte $P_1, \ldots, P_r$ sind die Projektionen der Q_i von k^{n+r} nach k^n; deshalb ist klar, daß alle Polynome aus

$$I \underset{\text{def}}{=} J \cap k[X_1, \ldots, X_n]$$

in den Punkten P_i verschwinden. Wir erhalten ein Erzeugendensystem dieses Ideals, indem wir bezüglich einer Eliminationsordnung für $T_1, \ldots, T_r$ eine GRÖBNER-Basis von J berechnen und davon nur die Polynome betrachten, die keine der Variablen T_i enthalten.