Wolfgang K. Seiler

Computeralgebra

Vorlesung im Herbstsemester 2011
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 1
Nullstellen von Polynomgleichungen

§0: Was ist Computeralgebra

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 Intel-

ligenz. 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 Computeralgebrasysteme 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 liegt, bescherte *Mathematica* einen großen Erfolg. Da auch Systeme wie MACSYMA und MAPLE mittlerweile in selbständige 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 auf die Bedeutung des Worts *Computeralgebra* 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.

Hier in der Vorlesung wird es in erster Linie um die Algorithmen gehen, die hinter solche System stehen, insbesondere denen, die sich mit der klassischen Aufgabe des symbolischen Rechnens befassen. In den Übungen wird es allerdings zumindest auch teilweise darum gehen,

Computeralgebrasysteme effizient einzusetzen auch zur Visualisierung mathematischer Sachverhalte.

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 Gleichtkommazahlen.

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, null 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

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

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 $g(x) = f(x) - 10^{-9}x^{19}$; wir ersetzen in f also den Koeffizienten -210 durch $-210{,}000000001$. Die neuen Nullstellen sind, auf fünf Nachkommastellen gerundet,

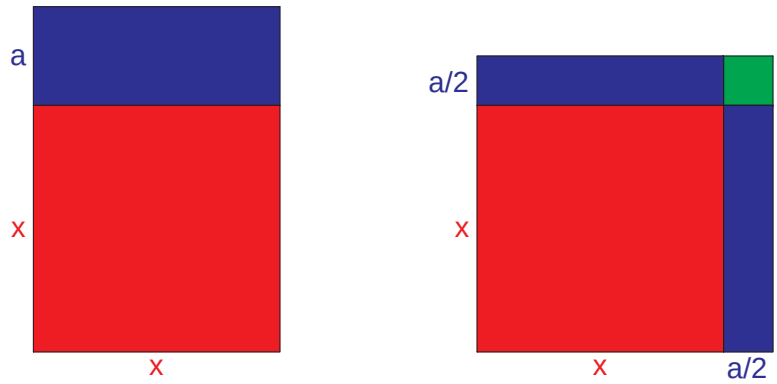
$$\begin{aligned}
&1{,}0000, \quad 2{,}0000, \quad 3{,}0000, \quad 4{,}0000, \quad 5{,}0000,\\
&6{,}0000, \quad 7{,}0000, \quad 8{,}0001, \quad 8{,}9992, \quad 10{,}008,\\
&10{,}957, \quad 12{,}383 \pm 0{,}10867i, \quad 14{,}374 \pm 0{,}77316i,\\
&16{,}572 \pm 0{,}88332i, \quad 18{,}670 \pm 0{,}35064i,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

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.

§1: Quadratische Ergänzung

Um 830 legte der arabische Gelehrte ABU DSCHA'FAR MUHAMMAD IBN MUSA AL-CHWARIZMI sein zweites Buch *Al-Kitāb al-muchtasar fi hisab al-dschabr wa-l-muqabala* (Rechnen durch Ergänzung und Ausgleich) vor; *al-dschabr* gab der Algebra ihren Namen und AL-CHWARIZMI führte zum Wort Algorithmus. Das Buch befaßte sich mit systematischen Lösungsverfahren für lineare und quadratische Gleichungen, die auch geometrisch motiviert und veranschaulicht wurden.

Wenn wir beispielsweise die quadratische Gleichung $x^2 + ax = b$ geometrisch interpretieren wollen, suchen wir nach einem Quadrat einer unbekannten Seitenlänge x derart, daß die Fläche des Quadrats zusammen mit der des Rechtecks mit Seiten x und a gleich b ist.



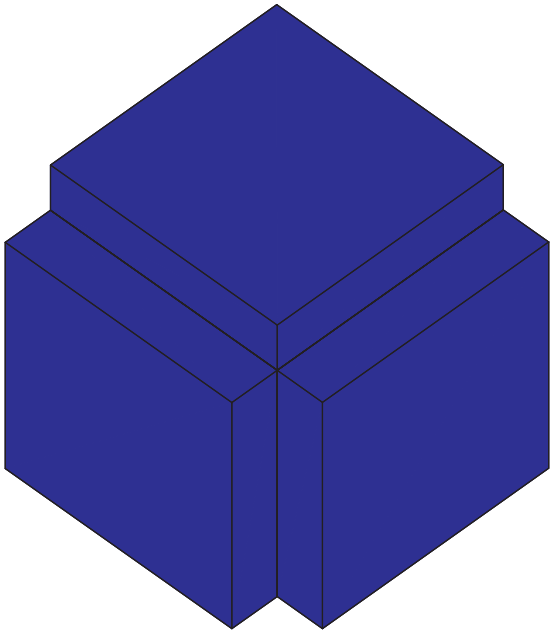
Die linke Zeichnung zeigt dieses Quadrat und darüber das Rechteck; auf der rechten Seite ist die Hälfte des Rechtecks neben das Quadrat gewandert, so daß abgesehen von dem kleinen Quadrat rechts oben

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:

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

Dieses Verfahren war zumindest grundsätzlich in allen frühen Hochkulturen bekannt; die ersten bekannten Hinweise darauf finden sich bereits vor über vier Tausend Jahren bei den Babyloniern.

Wenn wir versuchen, für die Gleichungen $x^3 + ax^2 = b$ ähnlich zu argumentieren, 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, sogar für Gleichungen höheren Grades. Die allgemeine Gleichung n-ten Grades hat die Form

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

wobei wir natürlich voraussetzen, daß a_n nicht verschwindet. Falls wir über einem Körper wie $\mathbb{R}$ oder $\mathbb{C}$ arbeiten, können wir durch a_n dividieren und erhalten die neue Gleichung

$$x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0 = 0$$

mit höchstem Koeffizienten eins.

Geometrisch betrachtet wollen wir einen n-dimensionalen Hyperwürfel bekommen, dessen Seitenlänge $x + \frac{c_{n-1}}{n}$ sein sollte; rechnerisch bedeutet dies, daß wir die neue Variable $y = x + \frac{c_{n-1}}{n}$ betrachten und überall in der Gleichung x durch $y - \frac{c_{n-1}}{n}$ ersetzen:

$$\begin{aligned}
&x^n + c_{n-1} x^{n-1} + c_{n-2} x^{n-2} + \cdots + c_1 x + c_0 \\
&= \left(y - \frac{c_{n-1}}{n}\right)^n + c_{n-1}\left(y - \frac{c_{n-1}}{n}\right)^{n-2} + \cdots + c_1\left(y - \frac{c_{n-1}}{n}\right) + c_0 \\
&= \left(y^n - c_{n-1} y^{n-1} + n c_{n-1}^2 y^{n-2} + \cdots\right) \\
&\quad + c_{n-1}\left(y^{n-1} - \frac{(n-1)c_{n-1}}{n} y^{n-2} + \cdots\right) \\
&\quad + c_{n-2}\left(y^{n-2} - \frac{(n-2)c_{n-1}}{n} y^{n-3} + \cdots\right) \\
&\quad + \cdots \\
&= y^n + \left(n c_{n-1}^2 - \frac{(n-1)c_{n-1}^2}{n} + c_{n-2}\right) y_{n-2} + \cdots\,.
\end{aligned}$$

Wir kommen also auf ein Polynom n-ten Grades in y, das keinen Term mit y^{n-1} hat.

Im Falle $n = 2$ hat dieses Polynom die Form $y^2 + p$, wir können also seine Nullstellen einfach durch Wurzelziehen ermitteln. Für $n > 2$ haben wir immerhin einen Term weniger als in der allgemeinen Gleichung n-ten Grades und müssen sehen, ob uns das bei der Lösung helfen kann.

§2: Polynome in Computeralgebrasystemen

Die Elimination des zweithöchsten Terms läßt sich für Gleichungen nicht allzu hohen Grades problemlos explizit durchführen, ist aber eine eher unangenehme und ziemlich langweilige Rechnung. Gerade in der Computeralgebra sollte man so etwas besser einem Computer überlassen.

Computeralgebrasysteme sind Programme, die genau für solche Zwecke entwickelt wurden – auch wenn ihre Fähigkeiten heute meist weit darüber hinausgehen. Hier soll beispielhaft das Computeralgebrasystem Maple betrachtet werden; in anderen Systemen werden die Befehle zwar im allgemeinen etwas anders aussehen, bezüglich der grundsätzlichen Fähigkeiten gibt es aber keine nennenswerten Unterschiede. Die hier abgedruckten Ein- und Ausgabezeilen sind im Stil des „klassischen“ Maple *worksheets* gehalten; die neueren Versionen benutzen standardmäßig eine Form mit zweidimensionaler Eingabe, für die auch Pfeiltasten benutzt werden müssen, was sich im Druck nur schwer darstellen läßt.

In Maple werden mathematische Ausdrücke, die nur Grundrechenarten enthalten, in der üblichen mathematischen Notation eingegeben, allerdings sollte das Multiplikationszeichen `*` nie weggelassen werden. (In den neuen *worksheets* kann es fehlen, wenn der entstehende Ausdruck nicht als Variablenname oder Funktionsaufruf interpretiert werden kann. So wird beispielsweise $2a$ als $2 \cdot a$ erkannt und $a\ b - b\ a$ als Null; ohne Zwischenräume ist aber $ab-ba$ die Differenz der beiden (verschiedenen) Variablen ab und ba. Auch $(a+b)(a-b)$ führt nicht auf a^2-b^2, sondern auf $a(a-b)+b(a-b)$, wobei $a(a+b)$ und $b(a-b)$ keine Multiplikationen, sondern Anwendungen der Funktionen a und b auf das Argument $a+b$ sind. Wer nicht auf solche Feinheiten achten möchte, sollte besser stets das Multiplikationszeichen mit eingeben.) Für die Exponentiation wird das Zeichen `^` verwendet, für Zuweisungen `:=`. Als Variablen können (unter anderem) alle mit einem Buchstaben beginnenden Folgen aus Buchstaben und Ziffern verwendet werden; abgeschlossen wird ein Befehl im allgemeinen mit einem Strichpunkt. In den neuen *worksheets* führt `^` dazu, daß die nächsten Zeichen im Exponenten geschrieben werden; zum Verlassen des Exponenten dient die Pfeiltaste nach rechts.

Entsprechendes gilt für `/`, was in den Nenner führt.

Die Definition eines Polynoms vom Grad drei kann somit in der Form

```
> P := x^3 + a*x^2 + b*x + c;
```

eingegeben werden. Maple quittiert diesen Befehl mit der Ausgabe

$$P := x^3 + ax^2 + bx + c$$

Wie wir bald sehen werden, ist das nicht immer wünschenswert; Zwischenergebnisse können in der Computeralgebra oft sehr umfangreich sein und mehrere Bildschirmseiten in Anspruch nehmen. Falls die Ausgabe des Ergebnisses nicht erwünscht ist, muß der Befehl einfach mit einem Doppelpunkt anstelle eines Semikolons abgeschlossen werden.

Um den quadratischen Term von P zu eliminieren, müssen wir das Polynom in der neuen Variable $y = x + \frac{a}{3}$ schreiben; dies leistet der Substitutionsbefehl

```
> Q := subs(x = y - a/3, P);
```

$$Q := \left(y - \frac{a}{3}\right)^3 + a\left(y - \frac{a}{3}\right)^2 + b\left(y - \frac{a}{3}\right) + c$$

Einige ältere Computeralgebrasysteme wie beispielsweise REDUCE hätten gleich ausmultipliziert; die meisten heute gebräuchlichen Systeme verzichten darauf, sofern es nicht explizit verlangt wird. Der Befehl zum Ausmultiplizieren heißt `expand`:

```
> expand(Q);
```

$$y^3 - \frac{ya^2}{3} + \frac{2a^3}{27} + by - \frac{ba}{3} + c$$

Das Ergebnis sieht ziemlich nach Kraut und Rüben aus, allerdings war das auch nicht anders zu erwarten: Zwar betrachten *wir* y als die Variable dieses Polynoms und a, b, c als Parameter, aber für Maple sind a, b, c, x und y gleichberechtigte Variablen.

Wenn wir einen Ausdruck f als Polynom in y dargestellt sehen wollen, müssen wir den Befehl `collect(f, y)` eingeben; zum Sortieren der Terme von f nach Potenzen von y dient `sort(f, y)`. Für ein übersichtliches Ergebnis sollten wir hier gleich beides anwenden. Da wir dem letzten Ergebnis keinen Namen gegeben haben, müssen wir

auch noch eine neue Variable kennen lernen: % bezeichnet in Maple stets das Ergebnis der letzten Ausgabe; daneben gibt es noch %% für die vorletzte und %%% für die drittletzte.

Damit ist klar, wie es weitergeht:

```
> R := sort(collect(%, y), y);
```

$$R := y^3 + \left(-\frac{a^2}{3} + b\right) y + \frac{2a^3}{27} + c - \frac{ba}{3}$$

Somit haben wir ein Polynom der Form $y^3 + py + q$ gefunden; mit dem Befehl `coeff(f, y^n)` oder `coeff(f, y, n)` können wir p und q noch isolieren, wobei für q natürlich nur die zweite Form in Frage kommt. Alternativ läßt sich q auch isolieren, indem wir y einfach auf Null setzen:

```
> p := coeff(R, y); q := subs(y=0, R);
```

$$p := -\frac{a^2}{3} + b$$

$$q := \frac{2a^3}{27} + c - \frac{ba}{3}$$

Wir können das Ergebnis für q auch noch etwas sortierter darstellen:

```
> sort(q, [a, b, c]);
```

$$\frac{2a^3}{27} - \frac{ab}{3} + c$$

§3: Kubische Gleichungen

Unser nächstes Ziel ist es, die Gleichung

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

explizit zu lösen. Auch wenn die Griechen geometrische Konstruktionen (jenseits von Zirkel und Lineal) kannten, mit denen sie Lösungen kubischer Gleichungen konstruieren konnten, sollte es noch bis ins 16. Jahrhundert dauern, bevor eine explizite Lösungsformel gefunden

war – ein Zeichen dafür, daß der Lösungsansatz nicht gerade offensichtlich ist.

Der Trick, der schließlich zum Erfolg führte, ist folgender: Wir schreiben die Variable y als Summe zweier neuer Variablen 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 auch u^3 und v^3:

Haben zwei Zahlen h, k das Produkt r und die Summe s, so sind h und k die beiden Nullstellen der Gleichung

$$(z-h)(z-k) = z^2 - (h+k)z + hk = z^2 - sz + r = 0\,;$$

falls wir r und s kennen, erhalten wir h und k also einfach als Lösungen einer quadratischen Gleichung:

$$h = \frac{s}{2} + \sqrt{\frac{s^2}{4} - r} \quad \text{und} \quad k = \frac{s}{2} - \sqrt{\frac{s^2}{4} - r}$$

oder umgekehrt.

In unserem Fall ist daher

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

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

wobei es auf die Reihenfolge natürlich nicht ankommt.

Somit 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 *immer* auch mit komplexen Zahlen rechnen, selbst wenn sowohl Koeffizienten als auch Lösungen allesamt reell sind.

Betrachten wir dazu als einfaches Beispiel die Gleichung

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

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 $x = y + 2$

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, mittleres 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 in seinem Buch *Ars magna* die Lösungsmethoden für kubische und biquadratische Gleichungen.

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

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

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

Die beiden anderen Kubikwurzeln erhalten wir, indem wir die reelle Kubikwurzel mit einer der beiden komplexen dritten Einheitswurzeln multiplizieren, d.h. also mit

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

Im ersten Fall ist

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

und

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

wir erhalten somit die Lösung $y = u + v = -1$.

Die dritte Kubikwurzel

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

schließlich führt auf

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

und liefert so die Lösung $y = u + v = 1$.

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

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

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. Wir können auch Maple nach dem Wert dieser Kubikwurzel fragen: Die imaginäre Einheit i wird dort als großes I eingegeben und Wurzeln (abgesehen von der auch als `sqrt` darstellbaren Quadratwurzel) durch Potenzen mit gebrochenem Exponenten; wir tippen also

```
> (-3 + 10/9*sqrt(3)*I)^(1/3);
```

$$\left(-3 + \frac{10}{9} I\sqrt{3}\right)^{\left(\frac{1}{3}\right)}$$

und erhalten unsere Eingabe unausgerechnet zurück. Mit dem Befehl `evalc` können wir Maple veranlassen, das Ergebnis – falls möglich – in der Form $a + bi$ darzustellen:

```
> evalc(%);
```

$$\frac{1}{9}7^{\left(\frac{1}{3}\right)}9^{\left(\frac{2}{3}\right)}21^{\left(\frac{1}{6}\right)}\left(\cos\left(-\frac{1}{3}\arctan\left(\frac{10\sqrt{3}}{27}\right)+\frac{\pi}{3}\right)\right.$$
$$\left.+\frac{1}{9}i7^{\left(\frac{1}{3}\right)}9^{\left(\frac{2}{3}\right)}21^{\left(\frac{1}{6}\right)}\left(\sin\left(-\frac{1}{3}\arctan\left(\frac{10\sqrt{3}}{27}\right)+\frac{\pi}{3}\right)\right)\right.$$

Dieses Ergebnis ist offensichtlich noch nicht in der bestmöglichen Weise dargestellt; mit dem Kommando `simplify` können wir Maple dazu überreden, sich etwas mehr Mühe zu geben:

```
> u := simplify(%);
```

$$u := \frac{1}{3}\sqrt{7}\sqrt{3}\left(\cos\left(-\frac{1}{3}\arctan\left(\frac{10\sqrt{3}}{27}\right)+\frac{\pi}{3}\right)\right.$$
$$\left.+\sin\left(-\frac{1}{3}\arctan\left(\frac{10\sqrt{3}}{27}\right)+\frac{\pi}{3}\right)i\right)$$

Maple arbeitet hier also mit der Polarkoordinatendarstellung komplexer Zahlen: Jede komplexe Zahl z läßt sich darstellen in der Form

$$z = |z| \left(\cos\varphi + i \sin\varphi\right),$$

und

$$\sqrt[3]{z} = \sqrt[3]{|z|}\left(\cos\tfrac{\varphi}{3} + i \sin\tfrac{\varphi}{3}\right).$$

Leider gibt es 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 $\cos 3\alpha$ und $\sin 3\alpha$ verschaffen; wenn wir das nicht selbst ausrechnen wollen, tut es Maple für uns:

```
> expand(cos(3*alpha));
```

$$4\cos(\alpha)^3 - 3\cos(\alpha)$$

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. Der Maple-Befehl dafür heißt `evalf`, wobei das `f` für *floating point* steht, d.h. also für Gleitkommaarithmetik.

```
> evalf(u);
```

$$1.000000001 + 1.154700538\, I$$

Wie jedes numerische Ergebnis stimmt diese Zahl natürlich nur näherungsweise, und zumindest in diesem Fall ist die Hypothese, daß es sich beim Realteil um eine durch Rundungsfehler verfälschte Eins handeln kann, eine Überlegung wert. Falls dem so sein sollte, ist

$$\cos\left(-\frac{1}{3}\arctan\left(\frac{10}{27}\sqrt{3}\right) + \frac{\pi}{3}\right) = \frac{3}{\sqrt{3}\sqrt{7}} = \sqrt{\frac{3}{7}},$$

und daraus folgt dann über die Beziehung $\sin^2\alpha + \cos^2\alpha = 1$, daß

$$\sin\left(-\frac{1}{3}\arctan\left(\frac{10}{27}\sqrt{3}\right) + \frac{\pi}{3}\right) = \pm\sqrt{1 - \frac{3}{7}} = \pm\sqrt{\frac{4}{7}} = \pm\frac{2\sqrt{7}}{7}$$

und

$$u = 1 \pm \frac{1}{3}\sqrt{3}\sqrt{7} \cdot \frac{2\sqrt{7}}{7} i = 1 \pm \frac{2\sqrt{3}}{3} i$$

ist. Bislang war alles noch Spekulation; nun kommt die Probe:

$$\left(1 \pm \frac{2\sqrt{3}}{3} i\right)^3 = 1 \pm 2\sqrt{3}\, i - 4 \mp \frac{8}{9}\sqrt{3}\, i = -3 \pm \frac{10}{9}\sqrt{3}\, i\,,$$

also ist $u = 1 + \frac{2\sqrt{3}}{3} i$. Das zugehörige v ist

$$v = \frac{7}{3u} = \frac{7}{3 + 2\sqrt{3}\, i} = \frac{7(3 - 2\sqrt{3}\, i)}{3^2 + 2^2 \cdot 3} = 1 - \frac{2\sqrt{3}}{3} i\,.$$

Damit ist die erste Lösung $x = u + v = 2$ gefunden. Die beiden anderen sind nun (etwas langwierige) Routine, können also beruhigt Maple überlassen werden. Wir verwenden dabei den Befehl `evalc`, der eine komplexe Zahl – sofern möglich – auf die Form $a + ib$ bringt und den Befehl `conjugate`, der die konjugiert komplexe Zahl berechnet:

```
> rho := -1/2 + sqrt(3)/2*I; u := 1 + 2/3*sqrt(3)*I;
```

$$\rho := -\frac{1}{2} + \frac{1}{2} I\sqrt{3}$$

$$u := 1 + \frac{2}{3} I\sqrt{3}$$

```
> evalc(u*rho + 7/(3*u*rho));
```

$$-3$$

```
> evalc(u*conjugate(rho) + 7/(3*u*conjugate(rho)));
```

$$1$$

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.

Wenn wir eine reelle Kubikwurzel finden können, ist die Situation auch nicht unbedingt viel besser. 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; wir rechnen zunächst mit dieser und erhalten als erste Lösung

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

Damit haben wir im Prinzip eine Lösung gefunden, die auch Maple nicht weiter vereinfachen kann:

```
> u := (-10 + 6*sqrt(3))^(1/3); simplify(u - 2/u);
```

$$u := (-10 + 6\sqrt{3})^{(\frac{1}{3})}$$
$$\frac{(-10 + 6\sqrt{3})^{(\frac{2}{3})} - 2}{(-10 + 6\sqrt{3})^{(\frac{1}{3})}}$$

Wenn wir das allerdings numerisch auswerten, drängt sich wieder die Hypothese auf, daß hier tatsächlich etwas sehr viel einfacheres steht:

```
> evalf(%);
```

$$-1.999999986$$

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

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

Quadratische Ergänzung macht daraus $(u+1)^2 = 3$, also ist $u = -1 \pm \sqrt{3}$. 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 $u = -1 + \sqrt{3}$ 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 in einfacher Form ausgedrückt haben, lassen sich nun auch die anderen beiden Lösungen berechnen:

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

und

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

Damit ist

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

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

also

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

Entsprechend folgt $y_3 = u\bar{\rho} - \dfrac{2}{u\bar{\rho}} = 1-3i$.

Die Mathematiker des fünfzehnten und sechzehnten Jahrhunderts, auf die die Lösungsformel für kubische Gleichungen zurückgeht, hatten natürlich weder Computer noch Taschenrechner; auch kannten sie weder Dezimalbrüche noch komplexe Zahlen. Trotzdem 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

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

Der numerische Näherungswert 2,414213562 für diese (reelle) Wurzel hilft uns nicht weiter. Wenn wir aber auf gut Glück versuchen, eine Wurzel zu finden, die sich auch in der Form $a + b\sqrt{2}$ mit ganzen Zahlen a und b schreiben läßt, Dann ist

$$(a+b\sqrt{2})^3 = a^3 + 3a^2b\sqrt{2} + 6ab^2 + 2b^3\sqrt{2} = 7 + 5\sqrt{2}\,,$$

also

$$a^3 + 6ab^2 = 7 \quad \text{und} \quad 3a^2b + 2b^3 = 5\,.$$

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 a ausklammern und erhalten $a(a^2 + 6b^2) = 7$. Somit muß a ein Teiler von sieben sein, d.h. $a = \pm 1$ oder $a = \pm 7$. Die negativen Zahlen scheiden aus, da die Klammer nicht negativ werden kann, und auch $a = 7$ ist nicht möglich, denn dann wäre die linke Seite mindestens gleich 7^3. Wenn es eine ganzzahlige Lösung gibt, muß daher $a = 1$ sein; durch Einsetzen folgt, daß dann mit $b = \pm 1$ die erste Gleichung in der Tat erfüllt ist. Die

zweite Gleichung $b(3a^2 + 2b^2) = 5$ zeigt, daß auch b positiv sein muß und $a = b = 1$ 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 und (zumindest gelegentlich) auch für komplexe Zahlen; die beiden noch fehlenden Lösungen können wir nun entweder berechnen als $u\rho + v\rho^2$ und $u\rho^2 + v\rho$, oder aber wir dividieren die Gleichung 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 oben gesehen haben, *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 $u = a + b\sqrt{-3}$ machen, kommen wir auf das Gleichungssystem

$$a^3 - 9ab^2 = 10 \quad \text{und} \quad 3a^2b - 3b^3 = 9\,.$$

Ausklammern von a *bzw.* b und Kürzen der zweiten Gleichung durch drei führt auf

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

Wenn es ganzzahlige Lösungen gibt, muß wegen der zweiten Gleichung $b = \pm 1$ oder $b = \pm 3$ sein. $b = \pm 1$ führt auf $a^2 - 1 = \pm 3$, also $b = 1$ und $a = \pm 2$; für $b = \pm 3$ läßt sich kein ganzzahliges a finden. Einsetzen in die erste Gleichung zeigt, daß $a = -2, b = 1$ das System löst, also ist $-2+\sqrt{-3}$ eine der drei Wurzeln. Die anderen könnten wir finden, indem wir das Problem durch Abdividieren auf eine quadratische Gleichung reduzieren; alternativ – und das war wohl die Methode des 17. Jahrhunderts – können wir die Einschränkung aufheben, daß a und b ganze Zahlen sein müssen und auch Brüche mit kleinen Nennern zulassen.

Der kleinstmögliche Nenner ist zwei; der Ansatz

$$\left(\frac{a}{2} + \frac{b}{2}\sqrt{-3}\right)^3 = 10 + 9\sqrt{-3}$$

führt auf die Gleichungen $a(a^2 - 9b^2) = 80$ und $b(a^2 - b^2) = 24$, wobei mindestens eine der Zahlen a und b ungerade sein muß, da wir ansonsten nichts neues bekommen. Da rechts jeweils gerade Zahlen stehen, sieht man leicht, daß dann beide Zahlen ungerade sein müssen; damit bleiben also für a nur die Möglichkeiten $a = \pm 1$ oder ± 5 und für b entsprechend $b = \pm 1$ oder ± 3. Einsetzen zeigt, daß wir mit $a = 5, b = 1$ und $a = -1, b = -3$ Lösungen bekommen. Die drei Kubikwurzeln von $-10 + 9\sqrt{-3}$ sind somit

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

Zu jedem dieser drei möglichen Werte von u müssen wir jene Zahl v finden, für die $uv = \frac{21}{3} = 7$ ist; in allen drei Fällen erhält man den Summanden $v = 7/u$ dadurch, daß man einfach das Vorzeichen des Koeffizienten von $\sqrt{-3}$ ändert. Die drei mit so großem Aufwand ermittelten Lösungen der kubischen Gleichung sind also einfach die drei ganzen Zahlen

$$(-2 + \sqrt{-3}) + (-2 - \sqrt{-3}) = -4\,,$$

$$\left(\frac{5}{2} + \frac{1}{2}\sqrt{-3}\right) + \left(\frac{5}{2} - \frac{1}{2}\sqrt{-3}\right) = \quad 5 \quad \text{und}$$

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

Wie die Beispiele in diesem Paragraphen zeigen, haben wir beim exakten Lösen kubischer Gleichungen nach 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 der Computeralgebra, zu dem es leider keine allgemeine Lösung gibt: Wie D. RICHARDSON 1968 gezeigt hat, kann es keinen Algorithmus geben, der von zwei beliebigen reellen Ausdrücken entscheidet, ob sie gleich sind oder nicht. Dabei reicht es schon, wenn wir nur Ausdrücke betrachten, die aus ganzen Zahlen, den Grundrechenarten, der Sinus- und der Betragsfunktion sowie der Zahl π aufgebaut werden können. Wir werden allerdings auch sehen, daß wir für wichtige Teilmengen von $\mathbb{R}$ solche Algorithmen haben.

Dies gilt insbesondere im Falle aller in diesem Paragraphen aufgetretenen Ausdrücke; wir werden im Laufe der Vorlesung noch mehrere Strategien kennenlernen, wie wir zumindest bei den hier betrachteten Beispielen die Lösungen erheblich einfacher und schneller gefunden hätten als über die allgemeine Lösungsformel.

§4: Biquadratische Gleichungen

Vorher wollen wir aber noch Gleichungen vom Grad vier betrachten. Auch sie lassen sich auflösen: Hier eliminiert man den kubischen Term von

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

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

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

Zu deren Lösung benutzen wir einen anderen Trick als im kubischen Fall: Wir versuchen, die Quadrate der Nullstellen durch eine geeignete Verschiebung zu Lösungen einer quadratischen Gleichung zu machen, die wir dann mit der bekannten Lösungsformel auflösen können.

Wir nehmen also an, wir hätten eine Lösung z dieser Gleichung und betrachten dazu für eine zunächst noch beliebige Zahl u die Zahl $z^2 + u$.

Da $z^4 + pz^2 + qz + r$ verschwindet, ist $z^4 = -pz^2 - qz - r$, also

$$(z^2 + u)^2 = z^4 + 2uz^2 + u^2 = (2u - p)z^2 - qz + u^2 - r\,.$$

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

$$(z^2 + u)^2 = (sz + t)^2 \Longrightarrow z^2 + u = \pm(sz + t)\,,$$

wir müssen also nur die beiden quadratischen Gleichungen

$$z^2 \mp (sz + t) + u = 0$$

lösen, um die Lösungen der biquadratischen Gleichung zu finden.

Natürlich ist die rechte Seite $(2u - p)z^2 - qz + u^2 - r$ im allgemeinen kein Quadrat eines linearen Polynoms in z; wir können aber hoffen, daß es zumindest für gewisse spezielle Werte der bislang noch willkürlichen Konstante u eines ist.

Ein quadratisches Polynom $\alpha z^2 + \beta z + \gamma$ ist genau dann Quadrat eines linearen, wenn die beiden Nullstellen der quadratischen Gleichung $\alpha z^2 + \beta z + \gamma = 0$ übereinstimmen. Diese Nullstellen können wir einfach berechnen:

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

Die Ausgangsgleichung ist genau dann das Quadrat eines linearen Polynoms, wenn beide Lösungen übereinstimmen, wenn also der Ausdruck unter der Wurzel verschwindet. Bringen wir diesen auf den Hauptnenner, kommen wir auf die Bedingung $\beta^2 - 4\alpha\gamma = 0$. 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 lösen können. Ist u_0 eine der Lösungen, so steht in der Gleichung

$$(z^2 + u_0)^2 = (2u_0 - p)z^2 - qz + u_0^2 - r\,.$$

rechts das Quadrat eines linearen Polynoms in z, das wir – da wir alle Koeffizienten kennen – problemlos hinschreiben können. Dies führt

dann nach Wurzelziehen zu zwei quadratischen Gleichungen für z, deren Wurzeln die Nullstellen der biquadratischen Gleichung 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.

§5: 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 *unmöglich* ist, die Lösungen einer allgemeinen Gleichung fünften (oder höheren) Grades durch Grundrechenarten und Wurzeln auszudrücken.

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. Der Beweis ist so umfangreich, daß er (zusammen mit den dafür notwendigen Definitionen und Sätzen) typischerweise den größten Teil der Vorlesung *Algebra I* einnimmt; über Einzelheiten kann daher hier nichts weiter gesagt werden. Interessenten finden ihn in fast jedem Algebralehrbuch im Kapitel über GALOIS-Theorie. Ein kurzes, gut lesbares Buch, das sich ganz darauf konzentriert, ist etwa

EMIL ARTIN: Galoissche Theorie, 1959 (Neuauflage 2004 bei *Deutsch*)

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 allgemeinen* 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.

§6: Der Wurzelsatz von Viète

In einem Großteil dieser Vorlesung wird es um Methoden gehen, wie man trotz des Fehlens brauchbarer Lösungsformeln Aussagen über die Nullstellen von Polynomgleichungen höheren Grades machen kann. Die folgende Methode führt nur in speziellen Fällen, dann aber mit sehr geringem Aufwand zu Nullstellen:

Angenommen, wir haben ein Polynom

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

mit (nicht notwendigerweise verschiedenen) Nullstellen $z_1, \ldots, z_n$. Dann ist auch $f(x) = (x - z_1)(x - z_2)\cdots(x - z_n)$. Ausmultiplizieren

und Koeffizientenvergleich liefert uns die Gleichungen

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

Allgemein ist a_{n-r} bis aufs Vorzeichen gleich der Summe alle Produkte aus r Werten z_i mit verschiedenem Index. Diese Summen bezeichnet man als die *elementarsymmetrischen Funktionen* in $z_1, \ldots, z_n$ 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.

Für eine quadratische Gleichung $x^2 + px + q = 0$ besagt er einfach, daß die Summe der Lösungen gleich $-p$ und das Produkt gleich q ist. Das hatten wir bereits in bei der Lösung kubischer Gleichungen ausgenutzt, um zwei Zahlen mit vorgegebenen Werten für Summe und Produkt zu berechnen.

Diese Summen, die sogenannten elementarsymmetrischen Funktionen, sind für r-Werte im mittleren Bereich recht umfangreich, die beiden Fälle $r = 0$ und $r = n - 1$ können aber gelegentlich ganz nützlich sein, um Lösungen zu erraten:

Falls wir aus irgendeinem Grund erwarten, daß alle Nullstellen ganzzahlig sind, folgt aus der Tatsache, daß ihr Produkt gleich $(-1)^n a_0$ ist, daß sie allesamt Teiler von a_0 sein müssen. Außerdem ist ihre Summe gleich $-a_{n-1}$.

Bei der Gleichung $f(x) = x^3 - 7x + 6 = 0$ etwa, die uns in §3 so viele Schwierigkeiten machte, ist das Produkt aller Nullstellen gleich -6; *falls* sie alle ganzzahlig sind, kommen also nur $\pm 1, \pm 2, \pm 3$ und ± 6 in Frage. Aus diesen acht Zahlen müssen wir drei (nicht notwendigerweise verschiedene) auswählen mit Produkt -6 und Summe null. Das geht offensichtlich nur mit $1, 2$ und -3; Einsetzen zeigt, daß dies auch tatsächlich Nullstellen sind.

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 $g(-3) = -3$. Hier führt aber die Lösungsformel aus §3 relativ schnell ans Ziel: Einsetzen der Parameter $p = -6$ und $q = 6$ in die Lösungsformel führt zunächst auf

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

für die reelle Wurzel; die erste Lösung ist also

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

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

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

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

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. Natürlich erfüllen auch diese Zahlen den Satz von VIÈTE, jedoch nützt uns dieser nichts, um sie zu erraten.

Auch die ebenfalls aus §3 bekannte Gleichung $f(x) = x^3 - 21x - 20 = 0$ läßt sich nach VIÈTE leicht 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

$$f(x) = x^4 + 14x^3 - 52x^2 - 14x + 51$$

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

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

schließlich 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.

Man beachte, daß diese Vorgehensweise nur funktioniert, wenn das Polynom höchsten Koeffizienten eins hat; andernfalls ist das Produkt der Nullstellen gleich dem Quotienten aus konstantem Koeffizienten und führendem Koeffizienten mal $(-1)^{\text{Grad}}$.

Kapitel 2
Der Euklidische Algorithmus

Angenommen, wir suchen die gemeinsamen Nullstellen zweier Polynome in derselben Veränderlichen x. Dann können wir natürlich versuchen, zunächst die Nullstellen eines der beiden Polynome zu finden und dann durch Einsetzen zu überprüfen, welche davon auch Nullstellen des zweiten sind. Im Extremfall gibt es keinerlei gemeinsame Nullstellen, und wir müssen trotzdem zunächst alle Nullstellen eines Polynoms von möglicherweise hohem Grad berechnen.

Die bessere Alternative besteht darin, sich zunächst zu überlegen, daß es zu zwei Polynomen über einem Körper (und auch noch unter deutlich schwächeren Voraussetzungen an die Koeffizienten) stets einen größten gemeinsamen Teiler gibt, dessen Nullstellen genau die gemeinsamen Nullstellen der beiden Polynome sind. Wenn sich dieser größte gemeinsame Teiler effizient berechnen läßt, müssen wir nur noch seine Nullstellen bestimmen, was aus Gradgründen meist erheblich einfacher sein dürfte.

Die Existenz des größten gemeinsamen Teilers zweier Polynome beweist der EUKLIDische Algorithmus, der gleichzeitig auch zu dessen Berechnung eingesetzt werden kann. Seine Anwendung auf ein Polynom und dessen Ableitung wird uns auf die sogenannte quadratfreie Zerlegung eines Polynoms führen und damit auf einen ersten Schritt zur Zerlegung eines Polynoms in irreduzible Faktoren.

Wir werden allerdings sehen, daß der EUKLIDische Algorithmus in seiner einfachsten Form zu sehr unübersichtlichen und schwer handhabbaren Zwischenergebnissen führen kann. Durch geeignete Modifikationen läßt sich seine Effizienz ganz deutlich steigern. Diese Modifikationen führen

uns allerdings in andere Zahlbereiche; um hier nicht wieder alles neu beweisen zu müssen, wollen wir uns daher nicht darauf beschränken, den EUKLIDischen Algorithmus für Polynome mit rationalen oder reellen Koeffizienten zu betrachten, sondern gleich am Anfang eine algebraische Struktur definieren, die alles bietet, was für den EUKLIDischen Algorithmus benötigt wird, den EUKLIDischen Ring. Wenn wir dann später größte gemeinsame Teiler anderer Objekte suchen, müssen wir uns nur noch überlegen, ob wir uns in einem EUKLIDischen Ring befinden; falls ja, können wir alle Konstruktionen und Sätze einfach übernehmen.

§1: Euklidische Ringe

Wie wir im nächsten Abschnitt sehen werden, beruht der EUKLIDische Algorithmus wesentlich auf der Division mit Rest. Ein EUKLIDischer Ring soll daher definiert werden als eine algebraische Struktur, in der Addition, Subtraktion, Multiplikation und Division mit Rest durchgeführt werden können und den „gewohnten“ Regeln genügen. Konkret heißt das folgendes:

Definition: *a)* Ein Ring ist eine Menge R zusammen mit zwei Rechenoperationen „+“ und „$\cdot$“ von $R \times R$ nach R, so daß 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 Ring; für einen Körper werden schließlich genau dieselben Eigenschaften gefordert und zusätzlich auch noch die Kommutativität der Multiplikation sowie die Existenz multiplikativer Inverser. Ein Körper ist somit insbesondere auch ein Integritätsbereich.

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

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

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 wäre etwa im Polynomring $\mathbb{Z}[x]$ nicht 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; dessen Nichtexistenz 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.

§2: 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 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 $\mathbb{R}[x]$ sind also genau die Einheiten von R. Speziell für Polynomringe über Körpern sind dies genau die von null verschiedenen Konstanten.

Damit wissen wir auch, wann zwei Polynome assoziiert sind:

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 um 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 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 größte 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.

■

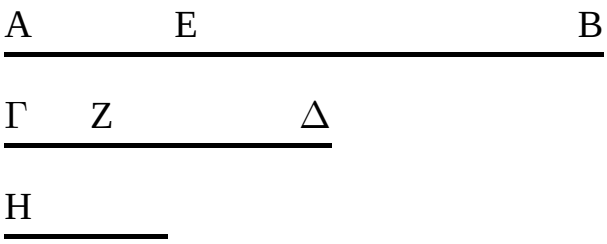
§3: Berechnung des größten gemeinsamen Teilers

Hier kommen wir endlich zum EUKLIDischen Algorithmus.

Bei EUKLID, in Proposition 2 des siebten Buchs seiner *Elemente,* wird er so beschrieben (nach der Übersetzung von CLEMENS THAER in Oswalds Klassiker der exakten Wissenschaften, Band 235):

> *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.
>
> A B
>
> Γ Δ
>
> 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.

Es ist nicht ganz sicher, ob EUKLID wirklich gelebt hat; das nebenstehende Bild aus dem 18. Jahrhundert ist mit Sicherheit reine Phantasie. EUKLID ist vor allem bekannt als Autor der *Elemente,* in denen er u.a. die Geometrie seiner Zeit systematisch darstellte und (in gewisser Weise) auf wenige Definitionen sowie die berühmten fünf Postulate zurückführte. Diese Elemente entstanden um 300 v. Chr. und waren zwar nicht der erste, aber doch der erfolgreichste Versuch einer solchen Zusammenfassung. EUKLID arbeitete wohl am Museion in Alexandrien; außer den Elementen schrieb er noch ein Buch über Optik und weitere, teilweise verschollene Bücher.

Was hier als erstes überrascht, ist die Beschränkung auf nicht zueinander teilerfremde Zahlen. Der Grund dafür liegt darin, daß die klassische griechische Philosophie und Mathematik die Eins nicht als Zahl betrachtete: Zahlen begannen erst bei der Zwei, und auch Mengen mußten mindestens zwei Elemente haben. Auch bei den Aristotelischen Syllogismen mußte sich ein Prädikat auf mindestens zweielementige Klassen beziehen: Die oft als klassischer Syllogismus zitierte Schlußweise

> Alle Menschen sind sterblich
> SOKRATES ist ein Mensch
> Also ist SOKRATES sterblich

wäre von ARISTOTELES nicht anerkannt worden, denn es gab schließlich nur einen SOKRATES. Erst bei seinen Nachfolgern, den Peripatetikern, setzte sich langsam auch die Eins als Zahl durch; ihr Zeitgenosse EUKLID macht noch brav eine Fallunterscheidung: In Proposition 1, unmittelbar vor der hier abgedruckten Proposition 2, führt er praktisch dieselbe Konstruktion durch für teilerfremde Zahlen.

Als zweites fällt auf, daß EUKLID seine Konstruktion rein geometrisch durchführt; wenn er von einer Strecke eine andere Strecke abträgt solange es geht, ist das natürlich in unserer heutigen arithmetischen Sprache gerade die Konstruktion des Divisionsrests bei der Division der beiden Streckenlängen durcheinander.

Die wesentliche Operation beim EUKLIDischen Algorithmus ist somit die Division mit Rest, und die haben wir (nach Definition) in jedem EUKLIDischen Ring. Tatsächlich funktioniert der so modifizierte EUKLIDische Algorithmus in jedem EUKLIDischen Ring und berechnet dort den größten gemeinsamen Teiler.

In heutiger Sprache ausgedrückt beruht der EUKLIDische Algorithmus auf 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 solcher Zahlen gibt,

muß nach endlich vielen Schritten $r_i = 0$ sein, womit der Algorithmus abbricht. ∎

Als erstes Beispiel wollen wir den EUKLIDischen Algorithmus anwenden auf zwei ganze Zahlen: Um den ggT von 200 und 148 zu Berechnen, müssen wir als erstes 200 durch 148 dividieren:

$$200 : 148 = 1 \text{ Rest } 52$$

Als nächstes wird 148 durch 52 dividiert:

$$148 : 52 = 2 \text{ Rest } 44$$

Weiter geht es mit der Division von 52 durch 44:

$$52 : 44 = 1 \text{ Rest } 8$$

Im nächsten Schritt dividieren wir

$$44 : 8 = 5 \text{ Rest } 4$$

und kommen schließlich mit

$$8 : 4 = 2 \text{ Rest } 0$$

zu einer Division,die aufgeht. Somit haben 200 und 148 den größten gemeinsamen Teiler vier.

Als zweites Beispiel wollen wir den größten gemeinsamen Teiler der beiden Polynome

$$f = x^8 + x^6 - 3x^4 - 3x^3 + 8x^2 + 2x - 5$$

und

$$g = 3x^6 + 5x^4 - 4x^2 - 9x + 21$$

aus $\mathbb{Q}[x]$ berechnen. Da Polynomdivision aufwendiger ist als die obigen Rechnungen, wollen wir die Rechenarbeit von Maple erledigen lassen. Wir brauchen dazu im wesentlichen nur den Befehl `rem(f, g, x)`, der den Rest bei der Division von f durch g berechnet, wobei f und g als Polynome in x aufgefaßt werden. Falls uns auch der Quotient interessiert, können wir den durch `quo(f, g, x)` berechnen lassen. Alternativ können wir aber auch dem Befehl `rem` noch ein viertes Argument geben: Die Eingabe `rem(f, g, x, 'q')` führt auf dasselbe Ergebnis

wie `rem(f, g, x)`, weist aber zusätzlich noch der Variablen q den Wert des Quotienten zu. Das q muß dabei in Hochkommata stehen, weil auf der linken Seite einer Zuweisung eine Variable stehen muß. Falls der Quotient etwa das Polynom $x^2 + x + 1$ wäre und die Variable q aus einer vorigen Rechnung den Wert $x-3$ hätte, würde `rem(f, g, x, q)` versuchen, die Zuweisung $x - 3 := x^2 + x + 1$ auszuführen, was natürlich Unsinn ist und auf eine Fehlermeldung führt. Die Hochkommata in `'q'` sorgen dafür, daß unabhängig von einem etwaigen vorigen Wert von q in jedem Fall nur der Variablenname q verwendet wird, so daß die sinnvolle Anweisung $q := x^2 + x + 1$ ausgeführt wird.

```
> f := x^8 + x^6 - 3*x^4 - 3*x^3 + 8*x^2 + 2*x - 5;
```

$$f := x^8 + x^6 - 3x^4 - 3x^3 + 8x^2 + 2x - 5$$

```
> g := 3*x^6 + 5*x^4 - 4*x^2 - 9*x + 21;
```

$$g := 3x^6 + 5x^4 - 4x^2 - 9x + 21$$

```
> r2 := rem(f, g, x, 'q'); q;
```

$$r2 := -\frac{5}{9}x^4 + \frac{1}{9}x^2 - \frac{1}{3}$$

$$\frac{x^2}{3} - \frac{2}{9}$$

```
> r3 := rem(g, r2, x);
```

$$r3 := -\frac{117}{25}x^2 - 9x + \frac{441}{25}$$

```
> r4 := rem(r2, r3, x);
```

$$r4 = \frac{233150}{6591}x - \frac{102500}{2197}$$

```
> r5 := rem(r3, r4, x);
```

$$r5 := \frac{1288744821}{543589225}$$

```
> r6 := rem(r4, r5, x);
```

$$r6 := 0$$

Der ggT von f und g ist somit $r_5 = \frac{1288744821}{543589225}$. Da der ggT nur bis auf eine multiplikative Konstante bestimmt ist, können wir freilich genauso

gut sagen, der ggT von f und g sei eins. In der Tat liefert uns Maple auch diese Antwort, wenn wir direkt nach dem ggT von f und g fragen:

```
> gcd(f, g);
```

$$1$$

Die Frage ist nun: Müssen wir wirklich mit so riesigen Brüchen wie r_5 rechnen, um auf diese einfache Antwort zu kommen?

Da der größte gemeinsame Teiler ohnehin nur bis auf eine multiplikative Konstante bestimmt ist, bestünde ein einfacher Ausweg darin, vor jeder Polynomdivision den Dividenden mit einer geeigneten Konstanten zu multiplizieren um so sicherzustellen, daß beim Dividieren keine Nenner auftreten. Bei der Division eines Polynoms vom Grad n durch ein Polynom vom Grad $m \leq n$ wird bis zu $n - m + 1$ mal durch den führenden Koeffizienten a des Divisors dividiert; wir müssen als den Dividenden vorher mit a^{n-m+1} multiplizieren. Im obigen Beispiel führt das auf folgende Rechnung:

```
> r2 := rem(3^3*f, g, x);
```

$$r2 := -15x^4 + 3x^2 - 9$$

```
> r3 := rem((-15)^3*g, r2, x);
```

$$r3 := 15795x^2 + 30375x - 59535$$

```
> r4 := rem(15795^3*r2, r3, x);
```

$$r4 := 1254542875143750x - 1654608338437500$$

```
> r5 := rem(1254542875143750^2*r3, r4, x);
```

$$r5 := 1259333879550074310093114199218750 0$$

Verglichen mit der Größe der Ausgangsdaten und des Ergebnisses entstehen auch hier wieder riesige Zahlen. Das ist leider kein Einzelfall: Auch wenn es sich hier um ein (von DONALD E. KNUTH für sein Buch *The Art of Computer Programming*, Abschnitt 4.6.1) konstruiertes besonders extremes Beispiel handelt, zeigt die Erfahrung, daß wir es beim EUKLIDischen Algorithmus für Polynome über den rationalen Zahlen oft mit einer Explosion der Koeffizienten zu tun haben, die in keiner

Weise der Komplexität des Ergebnisses entspricht. Wenn wir den Algorithmus ernsthaft auf größere Polynome anwenden wollen, sollten wir nach Wegen suchen, um dieses Problem zu umschiffen.

Solche Wege gibt es in der Tat; für ihr Verständnis ist allerdings einiges mehr an Theorie erforderlich als für den hier behandelten einfachen EUKLIDischen Algorithmus.

§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

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

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.

Obwohl es keinerlei Anhaltspunkt dafür gibt, daß diese Erweiterung EUKLID bekannt gewesen sein könnte, bezeichnet man sie als den *erweiterten* EUKLIDischen Algorithmus, Vor allem in der französischen Literatur wird die Darstellung des ggT als Linearkombination auch als Identität von BÉZOUT bezeichnet, da dieser sie 1766 in einem Lehrbuch beschrieb und als erster auch auf Polynome anwandte. Für Zahlen ist die Erweiterung jedoch bereits 1624 zu finden in der zweiten Auflage des Buchs *Problèmes plaisants et délectables qui se fonts par les nombres* von BACHET DE MÉZIRIAC. (Eine vereinfachte Ausgabe dieses Buchs von 1874 wurde 1993 bei Blanchard neu aufgelegt; sie ist auch online verfügbar unter `cnum.cnam.fr/DET/8PY45.html`.)

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.

Als Beispiel wollen wir den ggT von 200 und 148 als Linearkombination darstellen. Der Rechengang ist natürlich genau derselbe wie in §3, nur daß wir jetzt noch in jedem Schritt den Divisionsrest als ganzzahlige Linearkombination von 200 und 148 darstellen.

Im nullten Schritt haben wir 200 und 148 als die trivialen Linearkombinationen

$$200 = 1 \cdot 200 + 0 \cdot 148 \quad \text{und} \quad 148 = 0 \cdot 200 + 1 \cdot 148 .$$

Im ersten Schritt dividieren wir, da 148 nicht verschwindet, 200 mit Rest durch 148:

$$200 = 1 \cdot 148 + 52 \Longrightarrow 52 = 1 \cdot 200 - 1 \cdot 148$$

Da auch $52 \neq 0$ ist, dividieren wir im zweiten Schritt 148 durch 52 mit Ergebnis $148 = 2 \cdot 52 + 44$, d.h.

$$44 = 148 - 2 \cdot (1 \cdot 200 - 1 \cdot 148) = 3 \cdot 148 - 2 \cdot 200$$

Auch $44 \neq 0$, wir dividieren also weiter: $52 = 1 \cdot 44 + 8$ und

$$\begin{aligned} 8 = 52 - 44 &= (1 \cdot 200 - 1 \cdot 148) - (3 \cdot 148 - 2 \cdot 200) \\ &= 3 \cdot 200 - 4 \cdot 148 . \end{aligned}$$

Im nächsten Schritt erhalten wir $44 = 5 \cdot 8 + 4$ und

$$\begin{aligned} 4 = 44 - 5 \cdot 8 &= (3 \cdot 148 - 2 \cdot 200) - 5 \cdot (3 \cdot 200 - 4 \cdot 148) \\ &= 23 \cdot 148 - 17 \cdot 200 . \end{aligned}$$

Bei der Division von acht durch vier schließlich erhalten wir Divisionsrest Null; damit ist vier der ggT von 148 und 200 und kann in der angegebenen Weise linear kombiniert werden. Diese Darstellung ist freilich nicht eindeutig: Beispielsweise können wir beliebige Vielfache der trivialen Darstellung $200 \cdot 148 - 148 \cdot 200 = 0$ der Null addieren. Tatsächlich können wir diese auch noch durch den ggT kürzen zu $50 \cdot 148 - 36 \cdot 200 = 0$; wir haben also für jede ganze Zahl k eine Darstellung $1 = (23 + 50k) \cdot 148 - (17 + 36k) \cdot 200$.

Wir können den erweiterten EUKLIDischen Algorithmus natürlich auch auf die beiden Polynome f und g aus dem vorigen Paragraphen anwenden, allerdings ist das Ergebnis alles andere als schön: $1 = \alpha f + \beta g$, wobei α ein Polynom vom Grad fünf und β eines vom Grad sieben ist. Der Hauptnenner der Koeffizienten ist in beiden Fällen 130 354. Wir können den Algorithmus allerdings verwenden, um ein negatives Resultat zu beweisen:

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: $1 = 0 \cdot x + \frac{1}{2} \cdot 2$. Allerdings ist dort 2 ohnehin ein Teiler von x.)

§5: Die endlichen Primkörper

Die Explosion der Koeffizienten bei der Rechnung in §3 hängt natürlich damit zusammen, daß wir mit rationalen Zahlen gerechnet haben. Über

einem endlichen Körper gibt es nur endlich viele Möglichkeiten für jeden Koeffizienten, er kann also nicht unbegrenzt wachsen.

In der Computeralgebra führt man deshalb Probleme mit ganzzahligen Koeffizienten gerne zurück auf solche über endlichen Körpern, und auch das gängigste Verfahren zur effizienten Berechnung des größten gemeinsamen Teilers zweier Polynome verwendet diese Strategie.

Für eine zusammengesetzte Zahl $n = pq$ mit $p, q > 1$ bilden die Zahlen modulo n offensichtlich keinen Körper, denn $p \bmod n$ und $q \bmod n$ sind beide von Null verschieden, aber ihr Produkt $pq \bmod n = n \bmod n = 0$ verschwindet. Daher müssen wir uns auf Primzahlen beschränken, und hier gilt

Satz: Für jede Primzahl p ist die Menge $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$ mit den Operationen

$$a \oplus b = (a + b) \bmod p \quad \text{und} \quad a \otimes b = (a \cdot b) \bmod p$$

ein Körper. Alle vier Grundrechenarten in $\mathbb{F}_p$ können algorithmisch ausgeführt werden.

Beweis: Es ist klar, daß alle die Addition betreffenden Körperaxiome sowie die Distributivgesetze erfüllt sind und daß sich Addition, Subtraktion und Multiplikation problemlos durchführen lassen. Auch mit Assoziativ- und Kommutativgesetz der Multiplikation gibt es keinerlei Probleme, und natürlich ist $1 \in \mathbb{F}_p$ Neutralelement bezüglich der Multiplikation. Bis hierher ist es auch völlig egal, ob p eine Primzahl ist oder nicht.

Zu zeigen bleibt die Existenz von multiplikativen Inversen.

Seien also $b \in \mathbb{F}_p \setminus \{0\}$. Dann ist $1 \leq b \leq p-1$, d.h. b ist teilerfremd zu p, da die Primzahl p keine echten Teiler hat. Der größte gemeinsame Teiler von b und p ist also die Eins, und mit Hilfe des erweiterten EUKLIDischen Algorithmus können wir sie darstellen als ganzzahlige Linearkombination von b und p:

$$1 = \alpha b + \beta p \quad \text{mit} \quad \alpha, \beta \in \mathbb{Z}.$$

Somit ist $\alpha b \equiv 1 \bmod p$, d.h. $\alpha \bmod p$ ist ein multiplikatives Inverses von a und läßt sich mit Hilfe des erweiterten EUKLIDischen Algorithmus

auch effektiv berechnen. Damit können alle Quotienten algorithmisch berechnet werden, denn $a/b = a \cdot b^{-1}$.

■

Im folgenden werden wir auf die Symbole $\oplus$ und $\odot$ verzichten und Addition sowie Multiplikation auch in $\mathbb{F}_p$ einfach mit dem gewöhnlichen Plus- und Malzeichen schreiben.

Da $\mathbb{F}_p$ ein Körper ist, bilden die Polynome über $\mathbb{F}_p$ einen EUKLIDischen Ring, wir können also mit dem EUKLIDischen Algorithmus größte gemeinsame Teiler berechnen. Das Problem explodierender Koeffizienten, mit dem wir in §3 zu kämpfen hatten, existiert hier nicht, denn jedes Divisionsergebnis ist einfach wieder ein Element von $\mathbb{F}_p$, also eine ganze Zahl zwischen 0 und $p-1$.

Zur Illustration betrachten wir die beiden Polynome aus §3 über dem Körper mit elf Elementen. Der Operator `mod` sorgt dafür, daß Maple etwas modulo dem zweiten Argument des Operators betrachtet; wir erhalten also

```
> f mod 11;
```

$$x^8 + x^6 + 8x^4 + 8x^3 + 8x^2 + 2x + 6$$

```
> g mod 11;
```

$$3x^6 + 5x^4 + 7x^2 + 2x + 10$$

Bei der Berechnung von Quotienten und Divisionsresten dürfen wir allerdings nicht einfach `rem(f, g, x) mod 11` schreiben, denn dann würde ja erst modulo 11 reduziert, wenn der Rest bereits über $\mathbb{Q}$ berechnet wurde. Um dies zu vermeiden, bietet Maple die Operatoren `rem` und `quo` auch in einer trägen Form `Rem` *bzw.* `Quo` an: Diese Operatoren führen zu keiner Polynomdividision, sondern bleiben einfach unausgewertet stehen:

```
> Rem(f, g, x);
```

$$\mathrm{Rem}(x^8 + x^6 - 3x^4 - 3x^3 + 8x^2 + 2x - 5, 3x^6 + 5x^4 - 4x^2 - 9x + 21, x)$$

Wendet man darauf nun allerdings den Operator `mod p` an, so sorgt dieser dafür, daß die Polynomdivision modulo p durchgeführt und der Divisionsrest aus $\mathbb{F}_p[x]$ zurückgegeben wird:

```
> r2 := Rem(f, g, x) mod 11;
```

$$r2 := 8x^4 + 5x^2 + 7$$

```
> r3 := sort(Rem(g, r2, x) mod 11, x);
```

$$r3 := 5x^2 + 2x + 4$$

```
> r4 := sort(Rem(r2, r3, x) mod 11, x);
```

$$r4 := 10x + 10$$

```
> r5 := Rem(r3, r4, x) mod 11;
```

$$r5 := 7$$

```
> r6 := Rem(r4, r5, x) mod 11;
```

$$r6 := 0$$

Der ggT von f mod 11 und g mod 11 ist somit gleich der Konstanten sieben und damit ist auch die Eins ein ggT, denn der ggT im Polynomring über einem Körper ist nur bis auf eine multiplikative Konstante bestimmt.

Folgt damit, daß auch die Ausgangspolynome f und g aus $\mathbb{Z}[x]$ teilerfremd sind? Mit unseren derzeitigen Kenntnissen können wir das nicht sagen, denn bislang wissen war ja noch nicht einmal, ob es in $\mathbb{Z}[x]$ größte gemeinsame Teiler gibt. Um zu sehen, daß es durchaus Unterschiede zwischen ganzzahligen Polynomen und solchen über endlichen Körpern gibt, wollen wir die Rechnung auch modulo sieben durchführen:

```
> f mod 7;
```

$$x^8 + x^6 + 4x^4 + 4x^3 + x^2 + 2x + 2$$

```
> g mod 7;
```

$$3x^6 + 5x^4 + 3x^2 + 5x$$

```
> r2 := Rem(f, g, x) mod 7;
```

$$r2 := x^4 + 4x^2 + 2$$

```
> r3 := sort(Rem(g, r2, x) mod 7, x);
```

$$r3 := 4x^2 + 5x$$

```
> r4 := sort(Rem(r2, r3, x) mod 7, x);
```

$$r4 := 3x + 2$$

```
> r5 := Rem(r3, r4, x) mod 7;
```

$$r5 := 0$$

Hier ist der ggT also das lineare Polynom $3x + 2$ oder, wenn wir durch drei dividieren, $x + 3$.

§6: Faktorielle Ringe

Wenn wir größte gemeinsame Teiler für Polynome mit rationalen Koeffizienten in Verbindung bringen wollen mit solchen für Polynome über endlichen Körpern, kommen eigentlich nur Polynome mit ganzzahligen Koeffizienten als Bindeglied in Frage: Durch Multiplikation mit dem Hauptnenner können wir die Koeffizienten eines rationalen Polynoms ganzzahlig machen, und das entstehende Polynom können wir dann modulo einer Primzahl p betrachten.

Das so erhaltene Polynom muß nicht unbedingt viel mit dem Ausgangspolynom zu tun haben: Falls wir modulo einer Primzahl rechnen, die den führenden Koeffizienten teilt, unterscheiden sich selbst die Grade, und wenn wir eine Primzahl nehmen, die *alle* Koeffizienten teilt, erhalten wir einfach das Nullpolynom.

Wir müssen uns daher genau überlegen, womit wir erweitern und modulo welcher Primzahlen wir reduzieren, und vor allem müssen wir auch etwas mehr wissen über den Polynomring $\mathbb{Z}[x]$: Bislang wissen wir schließlich nur, daß er *kein* EUKLIDischer Ring ist,

Definition: *a)* Ein Element f eines Integritätsbereichs R heißt *irreduzibel*, falls gilt: f ist keine Einheit, und ist $f = gh$ das Produkt zweier Elemente aus R, so muß g oder h eine Einheit sein.
b) Ein Integritätsbereich R heißt *faktoriell* oder *ZPE*-Ring, wenn gilt: Jedes Element $f \in R$ läßt sich bis auf Reihenfolge und Assoziiertheit eindeutig schreiben als Produkt $f = u \prod_{i=1}^{n} p_i^{e_i}$ mit einer Einheit $u \in R^\times$, irreduziblen Elementen $p_i \in R$ und natürlichen Zahlen e_i.
(*ZPE* steht für **Z**erlegung in **P**rimfaktoren **E**indeutig.)

Lemma: In einem faktoriellen Ring R gibt es zu je zwei Elementen $f, g \in R$ einen größten gemeinsamen Teiler.

Beweis: Sind $f = u\prod_{i=1}^{n} p_i^{e_i}$ und $g = v\prod_{j=1}^{m} q_j^{d_j}$ mit $u, v \in R^\times$ und p_i, q_j irreduzibel die Zerlegungen von f und g in Primfaktoren, so können wir, indem wir nötigenfalls Exponenten null einführen, o.B.d.A. annehmen, daß $n = m$ ist und $p_i = q_i$ für alle i. Dann ist offenbar $\prod_{i=1}^{n} p_i^{\min(e_i,d_i)}$ ein ggT von f und g, denn $h = \prod_{i=1}^{r} p_i^{a_i}$ ist genau dann Teiler von f, wenn $a_i \leq e_i$ für alle i, und Teiler von g, wenn $a_i \leq d_i$. ∎

Wie wir bald sehen werden, bedeutet dies keineswegs, daß jeder faktorielle Ring EUKLIDisch wäre. Umgekehrt gilt allerdings

Satz: Jeder EUKLIDische Ring ist faktoriell.

Beweis: Wir müssen zeigen, daß jedes Element $f \neq 0$ aus R bis auf Reihenfolge und Assoziiertheit eindeutig als Produkt aus einer Einheit und geeigneten Potenzen irreduzibler Elemente geschrieben werden kann. Wir beginnen damit, daß sich f überhaupt so darstellen läßt.

Dazu benutzen wir die Funktion $\nu\colon R \setminus \{0\} \to \mathbb{N}_0$ des EUKLIDischen Rings R und beweisen induktiv, daß für $N \in \mathbb{N}_0$ alle $f \neq 0$ mit $\nu(f) \leq N$ in der gewünschten Weise darstellbar sind.

Ist $\nu(f) = 0$, so ist f eine Einheit: Bei der Division $1 : f = g$ Rest r ist nämlich entweder $r = 0$ oder aber $\nu(r) < \nu(f) = 0$. Letzteres ist nicht möglich, also ist $gf = 1$ und f eine Einheit. Diese kann als sich selbst mal dem leeren Produkt von Potenzen irreduzibler Elemente geschrieben werden.

Für $N > 1$ unterscheiden wir zwei Fälle: Ist f irreduzibel, so ist $f = f$ eine Darstellung der gewünschten Form, und wir sind fertig.

Andernfalls läßt sich $f = gh$ als Produkt zweier Elemente schreiben, die beide keine Einheiten sind. Nach Definition eines EUKLIDischen Rings sind dann $\nu(g), \nu(h) \leq \nu(f)$. Wir wollen uns überlegen, daß sie tatsächlich sogar echt kleiner sind.

Dazu dividieren wir g mit Rest durch f; das Ergebnis sei q Rest r, d.h. $g = qf + r$ mit $r = 0$ oder $\nu(r) < \nu(f)$. Wäre $r = 0$, wäre g ein Vielfaches

von f, es gäbe also ein $u \in R$ mit $g = uf = u(gh) = (uh)g$. Damit wäre $uh = 1$, also h eine Einheit, im Widerspruch zur Annahme. Somit ist $\nu(r) < \nu(f)$. Außerdem ist g ein Teiler von $r = g - qf = g(1 - qh)$, also muß gelten $\nu(g) \leq \nu(r) < \nu(f)$.

Genauso folgt die strikte Ungleichung $\nu(h) < \nu(f)$.

Nach Induktionsvoraussetzung lassen sich daher g und h als Produkte von Einheiten und Potenzen irreduzibler Elemente schreiben, und damit läßt sich auch $f = gh$ so darstellen.

Als nächstes müssen wir uns überlegen, daß diese Darstellung bis auf Reihenfolge und Einheiten eindeutig ist. Das wesentliche Hilfsmittel hierzu ist die folgende Zwischenbehauptung:

Falls ein irreduzibles Element p ein Produkt fg teilt, teilt es mindestens einen der beiden Faktoren.

Zum *Beweis* betrachten wir den ggT von f und p. Dieser ist insbesondere ein Teiler von p, also bis auf Assoziiertheit entweder p oder 1. Im ersten Fall ist p Teiler von f, und wir sind fertig; andernfalls können wir

$$1 = \alpha p + \beta f$$

als Linearkombination von p und f schreiben. Multiplikation mit g macht daraus $g = \alpha pf + \beta fg$, und hier sind beide Summanden auf der rechten Seite durch p teilbar: Bei αpf ist das klar, und bei βfg folgt es daraus, daß nach Voraussetzung p ein Teiler von fg ist. Also ist p Teiler von g, und die Zwischenbehauptung ist bewiesen.

Induktiv folgt sofort:

Falls ein irreduzibles Element $p \in R$ ein Produkt $\prod_{i=1}^{n} f_i$ teilt, teilt es mindestens einen der Faktoren.

Um den Beweis des Satzes zu beenden, zeigen wir induktiv, daß für jedes $N \in \mathbb{N}_0$ alle Elemente mit $\nu(f) \leq N$ eine bis auf Reihenfolge und Einheiten *eindeutige* Primfaktorzerlegung haben.

Für $N = 0$ haben wir oben gesehen, daß f eine Einheit sein muß, und hier ist die Zerlegung $f = f$ eindeutig.

Für $N \geq 1$ betrachten wir ein Element

$$f = u \prod_{i=1}^{n} p_i^{e_i} = v \prod_{j=1}^{m} q_j^{d_j}$$

mit zwei Zerlegungen, wobei wir annehmen können, daß alle $e_i, f_j \geq 1$ sind. Dann ist p_1 trivialerweise Teiler des ersten Produkts, also auch des zweiten. Wegen der Zwischenbehauptung teilt p_1 daher mindestens eines der Elemente q_j, d.h. $p_1 = wq_j$ ist bis auf eine Einheit w gleich q_j. Da p_i keine Einheit ist, ist $\nu(f/p_i) < \nu(f)$; nach Induktionsannahme hat also $f/p_i = x/(wq_j)$ eine bis auf Reihenfolge und Einheiten eindeutige Zerlegung in irreduzible Elemente. Damit hat auch x diese Eigenschaft.

∎

Als nächstes wollen wir Produktzerlegungen in $\mathbb{Q}[x]$ vergleichen mit solchen in $\mathbb{Z}[x]$. Das entsprechende Argument von GAUSS wird uns auch nützlich sein für den Beweis der Faktorialität von Polynomringen in mehreren Veränderlichen; wir fassen es daher gleich etwas allgemeiner.

Dazu brauchen wir als erstes für einen beliebigen Integritätsbereich eine Entsprechung für die rationalen Zahlen, den sogenannten Quotientenkörper, der genauso konstruiert wird wie die rationalen Zahlen aus den ganzen:

Wir betrachten für einen Integritätsbereich R auf der Menge aller Paare (f, g) mit $f, g \in R$ und $g \neq 0$ die Äquivalenzrelation

$$(f, g) \sim (r, s) \iff fs = gr\,;$$

die Äquivalenzklasse von (f, g) bezeichnen wir als den *Bruch* $\dfrac{f}{g}$.

Verknüpfungen zwischen diesen Brüchen werden nach den üblichen Regeln der Bruchrechnung definiert:

$$\frac{f}{g} + \frac{r}{s} = \frac{fs + rg}{gs} \quad \text{und} \quad \frac{f}{g} \cdot \frac{r}{s} = \frac{fr}{gs}\,.$$

Dies ist wohldefiniert, denn sind $(f, g) \sim (\tilde{f}, \tilde{g})$ und $(r, s) \sim (\tilde{r}, \tilde{s})$, so ist

$$\frac{\tilde{f}}{\tilde{g}} + \frac{\tilde{r}}{\tilde{s}} = \frac{\tilde{f}\tilde{s} + \tilde{r}\tilde{g}}{\tilde{g}\tilde{s}} \quad \text{und} \quad \frac{\tilde{f}}{\tilde{g}} \cdot \frac{\tilde{r}}{\tilde{s}} = \frac{\tilde{f}\tilde{r}}{\tilde{g}\tilde{s}}\,.$$

Wegen $f\tilde{g} = \tilde{f}g$ und $r\tilde{s} = \tilde{r}s$ ist

$$(\tilde{f}\tilde{s} + \tilde{r}\tilde{g}) \cdot gs = \tilde{f}\tilde{s}gs + \tilde{r}\tilde{g}gs = \tilde{f}gs\tilde{s} + \tilde{r}sg\tilde{g}$$
$$= g\tilde{g}s\tilde{s} + r\tilde{s}g\tilde{g} = (gs + ry)\tilde{g}\tilde{s}$$

und $(\tilde{f}\tilde{r})(gs) = \tilde{f}g\tilde{r}s = g\tilde{g}r\tilde{s} = (gr)(\tilde{g}\tilde{s})$, d.h. auch die Ergebnisse sind äquivalent.

Man rechnet leicht nach (wie bei $\mathbb{Q}$), daß diese Äquivalenzklassen einen Ring bilden mit $\frac{0}{1}$ als Null und $\frac{1}{1}$ als Eins; er ist sogar ein Körper, denn für $f, g \neq 0$ ist $\frac{g}{s}$ ein multiplikatives Inverses zu $\frac{f}{g}$, da $(fg, fg) \sim (1, 1)$. Identifizieren wir schließlich ein Element $f \in R$ mit dem Bruch $\frac{f}{1}$, so können wir R in den Körper K einbetten.

Definition: Der so konstruierte Körper K heißt Quotientenkörper von R, in Zeichen $K = \text{Quot}\, R$.

Das Standardbeispiel ist natürlich $\mathbb{Q} = \text{Quot}\, \mathbb{Z}$, aber auch der Quotientenkörper $k(x) \underset{\text{def}}{=} \text{Quot}\, k[x]$ eines Polynomrings über einem Körper k ist wichtig: $k(x)$ heißt rationaler Funktionenkörper in einer Veränderlichen über k. Seine Elemente sind rationale Funktionen in x, d.h. Quotienten von Polynomen in x, wobei der Nenner natürlich nicht das Nullpolynom sein darf.

Für Polynome, die statt über einem Körper nur über einem faktoriellen Ring definiert sind, sind die beiden folgenden Begriffe sehr wesentlich:

Definition: *a)* Der *Inhalt* eines Polynoms $f = a_n x^n + \cdots + a_0 \in R[x]$ ist der größte gemeinsame Teiler $I(f)$ seiner Koeffizienten a_i.
b) f heißt *primitiv,* wenn die a_i zueinander teilerfremd sind.

Indem wir die sämtlichen Koeffizienten eines Polynoms durch deren gemeinsamen ggT dividieren sehen wir, daß sich jedes Polynom aus $R[x]$ als Produkt seines Inhalts mit einem primitiven Polynom schreiben läßt. Diese Zerlegung bleibt bei der Multiplikation zweier Polynome erhalten:

Lemma: R sei ein faktorieller Ring. Für zwei Polynome

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

und

$$g = b_m x^m + b_{m-1} x^{m-1} + \ldots + b_1 x + b_0$$

aus $R[x]$ ist $I(fg) = I(f) \cdot I(g)$. Insbesondere ist das Produkt zweier primitiver Polynome wieder primitiv.

Beweis: Wir schreiben $f = I(f) \cdot f^*$ und $g = I(g) \cdot g^*$ mit primitiven Polynomen f^* und g^*; dann ist $fg = I(f) \cdot I(g) \cdot (f^* g^*)$. Falls $f^* g^*$ wieder ein primitives Polynom ist, folgt, daß $I(fg) = I(f) \cdot I(g)$ sein muß.

Es genügt daher, zu zeigen, daß das Produkt zweier primitiver Polynome wieder primitiv ist. Sei

$$fg = c_{n+m} x^{n+m} + c_{n+m-1} x^{n+m-1} + \ldots + c_1 x + c_0 \,;$$

dann ist $c_r = \sum_{i,j \text{ mit } i+j=r} a_i b_j$.

Angenommen, diese Koeffizienten c_r haben einen gemeinsamen Teiler, der keine Einheit ist. Wegen der Faktorialität von R gibt es dann auch ein irreduzibles Element p, das alle Koeffizienten c_r teilt.

Insbesondere ist p ein Teiler von $c_0 = a_0 b_0$; da p irreduzibel ist, muß mindestens einer der beiden Faktoren a_0, b_0 durch p teilbar sein. Da es im Lemma nicht auf die Reihenfolge von f und g ankommt, können wir o.B.d.A. annehmen, daß a_0 Vielfaches von p ist.

Da f ein primitives Polynom ist, kann nicht jeder Koeffizient a_i durch p teilbar sein; ν sei der kleinste Index, so daß a_ν kein Vielfaches von p ist. Genauso gibt es auch einen kleinsten Index $\mu \geq 0$, für den b_μ nicht durch p teilbar ist. In

$$c_{\mu+\nu} = \sum_{i,j \text{ mit } i+j=\mu+\nu} a_i b_j$$

ist dann der Summand $a_\nu b_\mu$ nicht durch p teilbar, denn für jeden anderen Summanden $a_i b_j$ ist entweder $i < \nu$ oder $j < \mu$, so daß mindestens einer der Faktoren und damit auch das Produkt durch p teilbar ist. Insgesamt ist daher $c_{\mu+\nu}$ nicht durch p teilbar, im Widerspruch zur Annahme.

Somit muß fg ein primitives Polynom sein.

■

Satz von Gauß: R sei ein faktorieller Ring und $K = \operatorname{Quot} R$. Falls sich ein Polynom $f \in R[x]$ in $K[x]$ als Produkt zweier Polynome $g, h \in K[x]$ schreiben läßt, gibt es ein $\lambda \in K$, so daß $\widetilde{g} = \lambda g$ und $\widetilde{h} = \lambda^{-1} h$ in $R[x]$ liegen und $f = \widetilde{g} \cdot \widetilde{h}$.

Beweis: Durch Multiplikation mit einem gemeinsamen Vielfache aller Koeffizienten können wir aus einem Polynom mit Koeffizienten aus K eines mit Koeffizienten aus R machen. Dieses wiederum ist gleich seinem Inhalt mal einem primitiven Polynom. Somit läßt sich jedes Polynom aus $K[x]$ schreiben als Produkt eines Elements von K mit einem primitiven Polynom aus $R[x]$. Für g und h seien dies die Zerlegungen

$$g = cg^* \quad \text{und} \quad h = dh^* .$$

Dann ist $f = (cd)g^*h^*$, und nach dem Lemma ist g^*h^* ein primitives Polynom. Daher liegt $cd = I(f)$ in R, und wir können beispielsweise $\widetilde{g} = I(P)g^*$ und $\widetilde{h} = h^*$ setzen. ∎

Korollar: Ein primitives Polynom $f \in R[x]$ ist genau dann irreduzibel in $R[x]$, wenn es in $K[x]$ irreduzibel ist. ∎

CARL FRIEDRICH GAUSS (1777–1855) leistete wesentliche Beiträge zur Zahlentheorie, zur nichteuklidischen Geometrie, zur Differentialgeometrie und Kartographie, zur Fehlerrechnung und Statistik, zur Astronomie und Geophysik *usw.* Als Direktor der Göttinger Sternwarte baute er zusammen mit dem Physiker Weber den ersten Telegraphen. Er leitete die erste Vermessung und Kartierung des Königreichs Hannover, was sowohl seine Methode der kleinsten Quadrate als auch sein *Theorema egregium* motivierte, und zeitweise auch den Witwenfond der Universität Göttingen. Seine hierbei gewonnene Erfahrung nutzte er für erfolgreiche Spekulationen mit Aktien.

Aus dem Satz von GAUSS folgt induktiv sofort, daß seine Aussage auf für Produkte von mehr als zwei Polynomen gilt, und daraus folgt

Satz: Der Polynomring über einem faktoriellen Ring R ist faktoriell.

Beweis: Wir müssen zeigen, daß sich jedes $f \in R[x]$ bis auf Reihenfolge und Einheiten eindeutig als Produkt von Potenzen irreduzibler Elemente aus $R[x]$ und einer Einheit schreiben läßt. Dazu schreiben wir $f = I(f) \cdot f^*$ mit einem primitiven Polynom $f^* \in R[x]$ und zerlegen zunächst den Inhalt $I(f)$ in R. Da R nach Voraussetzung faktoriell ist, ist diese Zerlegung eindeutig bis auf Reihenfolge und Einheiten in R, und wie wir aus §2 wissen, sind die Einheiten von $R[x]$ gleich denen von R.

Als nächstes zerlegen wir das primitive Polynom f^* über dem Quotientenkörper K von R; dies ist möglich, da $K[x]$ als EUKLIDischer Ring faktoriell ist. Jedes der irreduziblen Polynome q_i, die in dieser Zerlegung vorkommen, läßt sich schreiben als $q_i = \lambda_i p_i$ mit einem $\lambda_i \in K^\times$ und einem primitiven Polynom $p_i \in R[x]$. Wir können daher annehmen, daß in der Zerlegung von f nur primitive Polynome aus $R[x]$ auftreten sowie eine Einheit aus K. Diese muß, da f^* Koeffizienten aus R hat und ein Produkt primitiver Polynome primitiv ist, in R liegen; da auch f^* primitiv ist, muß sie dort sogar eine Einheit sein.

Kombinieren wir diese Primzerlegung von f^* mit der Primzerlegung des Inhalts, haben wir eine Primzerlegung von f gefunden; sie ist (bis auf Reihenfolge und Einheiten) eindeutig, da entsprechendes für die Zerlegung des Inhalts, die Zerlegung von f^* sowie die Zerlegung eines Polynoms in Inhalt und primitiven Anteil gilt. ∎

Da wir einen Polynomring $R[x_1, \ldots, x_n]$ in n Veränderlichen als Polynomring $R[x_1, \ldots, x_{n-1}][x_n]$ in einer Veränderlichen über dem Polynomring $R[x_1, \ldots, x_{n-1}]$ in $n-1$ Veränderlichen auffassen können, folgt induktiv sofort:

Satz: Der Polynomring $R[x_1, \ldots, x_n]$ in n Veränderlichen über einem faktoriellen Ring R ist selbst faktoriell. Insbesondere sind $\mathbb{Z}[x_1, \ldots, x_n]$ sowie $k[x_1, \ldots, x_n]$ für jeden Körper k faktoriell. ∎

Damit wissen wir also, daß auch Polynome in mehreren Veränderlichen über $\mathbb{Z}$ oder über einem Körper in Produkte irreduzibler Polynome zer-

legt werden können; insbesondere existieren daher auch in dieser Ringen größte gemeinsame Teiler.

Der Beweis des obigen Satzes ist allerdings nicht konstruktiv; wir werden im nächsten Kapitel noch viel Arbeit investieren müssen, bevor wir Polynome über $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{F}_p$ oder anderen Körpern, in denen wir rechnen können, wirklich in ihre irreduziblen Faktoren zerlegen können.

§7: Resultanten

Wir wissen inzwischen, daß es auch in $\mathbb{Z}[x]$ größte gemeinsame Teiler gibt, und wir wissen auch, daß wir sie über den EUKLIDischen Algorithmus in $\mathbb{Q}[x]$ berechnen können. Wir wissen allerdings auch, daß der EUKLIDische Algorithmus in $\mathbb{Q}[x]$ bei der praktischen Durchführung seine Tücken hat und wollen daher eine Alternative finden, die stattdessen den EUKLIDischen Algorithmus in einem oder mehreren Polynomringen über endlichen Körpern benutzt. Hier haben wir allerdings auch Beispiele gesehen, wonach der ggT zweier Polynome aus $\mathbb{Z}[x]$ nicht einmal denselben Grad hat wie der der beiden Reduktionen modulo einer vorgegebenen Primzahl; insbesondere können Polynome, die in $\mathbb{Z}[x]$ und damit auch in $\mathbb{Q}[x]$ teilerfremd sind, modulo gewisser Primzahlen gemeinsame Teiler positiven Grades haben.

Um dieses Phänomen genauer zu untersuchen, wollen wir in diesem Paragraphen Kriterien dafür entwickeln, daß zwei Polynome einen größten gemeinsamen Teiler vom Grad d haben.

Was wir auf einfache Weise erhalten, ist ein Kriterium dafür, daß der ggT *mindestens* den Grad d hat. Wie üblich betrachten wir das Problem gleich über einem beliebigen faktoriellen Ring R; das wird uns später auch nützlich sein für die Lösung nichtlinearer Gleichungssysteme.

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

und

$$g = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0$$

seien also zwei Polynome über $R[x]$ und nehmen an, es gebe ein Polynom $h \in R[x]$ vom Grad mindestens $d \geq 1$, das sowohl f als auch g

teilt. Dann ist

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

ein gemeinsames Vielfaches von f und q, dessen Grad

$$\deg f + \deg g - \deg h = n + m - \deg h$$

höchstens gleich $n + m - d$.

Haben umgekehrt f und g ein gemeinsames Vielfaches vom Grad höchstens $n + m - d$, so hat auch ihr kleinstes gemeinsames Vielfaches S höchstens den Grad $n + m - d$. (Ein kleinstes gemeinsames Vielfaches existiert, da mit R auch $R[x]$ faktoriell ist.)

Zu S gibt es einerseits Polynome $u, v \in R[x]$, für die $S = uf = vg$ ist, andererseits ist p als *kleinstes* gemeinsames Vielfaches von f und g Teiler von fg, es gibt also ein Polynom $h \in R[x]$ mit $fg = Sh$. Für dieses ist

$$hv = \frac{fg}{S} \cdot v = f \cdot \frac{vg}{S} = f \quad \text{und} \quad hu = \frac{fg}{S} \cdot u = g \cdot \frac{uf}{S} = g\,,$$

es teilt also sowohl f als auch g und sein Grad $n + m - \deg S$ ist mindestens d. Damit ist gezeigt:

Lemma: Zwei Polynome $f, g \in R[x]$ haben genau dann einen gemeinsamen Teiler vom Grad mindestens d, wenn es nichtverschwindende Polynome $u, v \in R[x]$ gibt mit $\deg u \leq \deg g - d$ und $\deg v \leq \deg f - d$, so daß $uf = vg$ ist.

■

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

$$u = u_{m-d}x^{m-d} + u_{m-d-1}x^{m-d-1} + \cdots + u_1 x + u_0$$

und

$$v = v_{n-d}x^{n-d} + v_{n-d-1}x^{n-d-1} + \cdots + v_1 x + v_0\,.$$

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 d, wenn es nicht allesamt verschwindende Körperelemente $u_0, \ldots, u_{m-d}$ und $v_0, \ldots, v_{n-d}$ gibt, so daß

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

ist. Dies ist ein homogenes lineares Gleichungssystem aus $n+m+1-d$ Gleichungen für die $n+m+2-2d$ Unbekannten $u_0, \ldots, u_{m-d}$ und $v_0, \ldots, v_{n-d}$; es hat genau dann eine nichttriviale Lösung, wenn seine Matrix kleineren Rang als $n+m+2-2d$ hat. Für $d = 1$, wenn die Matrix quadratisch ist, bedeutet dies einfach, daß ihre Determinante verschwindet; für $d > 1$ müssen wir d Determinanten von quadratischen Untermatrizen betrachten

Ausgeschrieben wird dieses Gleichungssystem, wenn wir mit dem Koeffizienten von x^{m+n-d} anfangen, zu

$$\begin{aligned}
a_n u_{m-d} - b_m v_{n-d} &= 0 \\
a_{n-1} u_{m-d} + a_n u_{m-d-1} - b_{m-1} v_{n-d} - b_m v_{n-d-1} &= 0 \\
a_{n-2} u_{m-d} + a_{n-1} u_{m-d-1} + a_n u_{m-d-2} & \\
- b_{m-2} v_{n-d} - b_{m-1} v_{n-d-1} b_m v_{n-d-2} &= 0 \\
\ldots & \\
a_0 u_3 + a_1 u_2 + a_2 u_1 + a_3 u_0 - b_0 v_3 - b_1 v_2 - b_2 v_1 - b_3 v_0 &= 0 \\
a_0 u_2 + a_1 u_1 + a_2 u_0 - b_0 v_2 - b_1 v_1 - b_2 v_0 &= 0 \\
a_0 u_1 + a_1 u_0 - b_0 v_1 - b_1 v_0 &= 0 \\
a_0 u_0 - b_0 v_0 &= 0
\end{aligned}$$

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

$(n+m+2-2d) \times (n+m+1-d)$-Matrix

$$\begin{pmatrix}
a_n & a_{n-1} & a_{n-2} & \dots & a_1 & a_0 & 0 & 0 & \dots & 0 \\
0 & a_n & a_{n-1} & \dots & a_2 & a_1 & a_0 & 0 & \dots & 0 \\
0 & 0 & a_n & \dots & a_3 & a_2 & a_1 & a_0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \dots & a_n & a_{n-1} & a_{n-2} & a_{n-3} & \dots & a_0 \\
b_m & b_{m-1} & b_{m-2} & \dots & b_2 & b_1 & b_0 & 0 & \dots & 0 \\
0 & b_m & b_{m-1} & \dots & b_3 & b_2 & b_1 & b_0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \dots & 0 & b_m & b_{m-1} & b_{m-2} & \dots & b_0
\end{pmatrix},$$

in der $m+1-d$ Zeilen aus Koeffizienten von f stehen und $n+1-d$ Zeilen aus Koeffizienten von g.

Für $d = 1$ ist diese Matrix quadratisch; man bezeichnet sie als SYLVESTER-Matrix und ihre Determinante als die *Resultante* $\mathrm{Res}(f,g)$ der beiden Polynome f und g. Falls man, etwa bei späteren Anwendungen auf Polynome mehrerer Veränderlicher, auf die Variable x hinweisen möchte, schreibt man auch $\mathrm{Res}_x(f,g)$. Die beiden Polynome f und g haben somit genau dann einen gemeinsamen Faktor positiven Grades, wenn $\mathrm{Res}(f,g)$ verschwindet.

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 mit die wichtigste mathematische Fachzeitschrift Amerikas ist.

Für $d > 1$ betrachten wir für $j = n + m + 1 - d, \ldots, n + m + 2 - 2d$ jene quadratische Matrix M_j, die aus den ersten $n+m+2-d$ Spalten sowie der j-ten Spalte der obigen Matrix besteht. Offensichtlich hat letztere genau dann einen kleineren Rang als $n + m + 1 - d$, wenn alle d Matrizen M_j singulär sind, wenn also deren Determinanten verschwinden.

Sowohl die Resultante als auch die Determinanten der M_j sind Polynome in den Koeffizienten von f und g; wir haben daher den

Satz: Zwei Polynome $f, g \in R[X]$ über dem faktoriellen Ring R haben genau dann einen gemeinsamen Faktor vom Grad mindestens d, wenn gewisse Polynome in ihren Koeffizienten verschwinden. Insbesondere gibt es genau dann einen gemeinsamen Faktor positiven Grades, wenn die Resultante der beiden Polynome verschwindet.

■

Die Resultante zweier Polynome der Grade 30 und 40 ist eine 70×70-Determinante – nichts, was man mit den aus der Linearen Algebra bekannten Algorithmen leicht und schnell ausrechnen könnte. Im Augenblick braucht uns das noch nicht sonderlich zu kümmern, da uns für den Algorithmus zur modularen Berechnung des ggT die bloße Existenz der Resultante genügt. Später, wenn wir Resultanten ernsthaft anwenden, werden wir sehen, daß sie sich erheblich effizienter berechnen lassen als andere Determinanten vergleichbarer Größe. Der Grund dafür liegt – wie eigentlich fast zu erwarten war – in der engen Beziehung zwischen Resultante und größtem gemeinsamen Teiler, die auch eine Beziehung zwischen Resultantenberechnung und EUKLIDischem Algorithmus liefert.

Für die ggT-Berechnung in $\mathbb{Z}[x]$ (und damit auch $\mathbb{Q}[x]$) sind Resultanten aus folgendem Grund für uns wichtig: 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 $\overline{f}, \overline{g} \in \mathbb{F}_p[x]$ 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 $\overline{h}$, $\overline{f_1}$, $\overline{g_1}$ die Reduktionen von h, f_1 und g_1 modulo p, so ist offensichtlich $\overline{f} = \overline{h}\overline{f_1}$ und $\overline{g} = \overline{h}\overline{g_1}$. Somit ist $\overline{h}$ auf jeden Fall ein gemeinsamer Teiler von $\overline{f}$ und $\overline{g}$, 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 $\overline{h}$ kleineren Grad als h. Ein Beispiel dafür können wir uns leicht mit Maple konstruieren:

```
> h := 3*x+1: f1 := (x^3 - x^2 + 2): g1 := (x^2+x+1):

> f := expand(h*f1); g := expand(h*g1);
```

$$f := 3x^4 - 2x^3 + 6x - x^2 + 2$$

$$g := 3x^3 + 4x^2 + 4x + 1$$

```
> gcd(f, g);
```

$$3x + 1$$

```
> Gcd(f, g) mod 3;
```

$$1$$

Das Kommando `Gcd` mit großem `G` ist die „träge“ Form des `gcd`-Kommandos, die erst vom `mod`-Operator ausgewertet wird, so daß die ggT-Berechnung über $\mathbb{F}_3$ erfolgt. Im vorliegenden Beispiel freilich hätten wir auch mit `gcd` dasselbe Ergebnis bekommen, denn hier ist h mod p der ggT von f und $\overline{g}$. Wir werden gleich sehen, daß dies nicht immer so sein muß.

Zunächst aber wollen wir uns überlegen, wie wir ausschließen können, daß der Grad von h mod p kleiner ist als der von h. Das ist offensichtlich dann und nur dann der Fall, wenn der führende Koeffizient von h durch p teilbar ist. Da wir h erst ausrechnen wollen, hat dieses Kriterium freilich keinen großen praktischen Nutzen.

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 mod p und g mod p in $\mathbb{F}_p[x]$ mindestens denselben Grad wie h = ggT$(f, g) \in \mathbb{Z}[x]$ und ist ein Vielfaches von h mod p.

■

Falls unter diesen Bedingungen der ggT in $\mathbb{F}_p[x]$ denselben Grad hat wie der in $\mathbb{Z}[x]$, ist somit h mod p ein ggT in $\mathbb{F}_p[x]$. Wann das der Fall ist, sagen uns die Resultante *bzw.* die Determinanten der Matrizen M_j:

Angenommen, der ggT h von f und g in $\mathbb{Z}[x]$ hat den Grad $d \geq 0$. Dann haben f und g keinen gemeinsamen Teiler vom Grad mindestens $d + 1$; folglich ist im Falle $d = 0$ die Resultante eine von Null verschiedene ganze Zahl und für $d > 0$ hat mindestens eine der Matrizen M_j eine von null verschiedene Determinante.

Nun betrachten wir dasselbe Problem über $\mathbb{F}_p$. Da die Resultante und die Determinanten der M_j Polynome in den Koeffizienten der beiden Ausgangspolynome sind, führt ihre Berechnung über $\mathbb{F}_p$ zum selben Ergebnis wie die Berechnung über $\mathbb{Z}$ mit anschließender Reduktion modulo p. Somit gibt es modulo p genau dann einen gemeinsamen Teiler positiven Grades, wenn die Resultante durch p teilbar ist, und es gibt einen gemeinsamen Teiler vom Grad $d + 1$ mit $d > 0$, wenn die Determinanten *aller* Matrizen M_j durch p teilbar sind. Da eine ganze Zahl nur endlich viele Teiler hat, gibt es höchstens endlich viele solche Primzahlen.

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 ggT$(f, g) = h$ und ihre Reduktionen $\overline{f}, \overline{g} \in \mathbb{F}_p[x]$ mit ggT$(\overline{f}, \overline{g}) = h^*$ gilt:
a) Falls p nicht die führenden Koeffizienten von sowohl f als auch g teilt, ist die Reduktion $\overline{h}$ von h ein Teiler von h^* und deg $h^* \geq$ deg h.

b) Es gibt höchstens endlich viele Primzahlen p, für die $\overline{h}$ nicht gleich dem ggT von $\overline{f}$ und $\overline{g}$ 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 $\mathbb{Z}[x]$ den 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. Wenn wir von zwei primitiven Polynomen ausgehen, können wir trotzdem den modulo p berechneten ggT so liften, daß er c als führenden Koeffizienten hat; im obigen Beispiel bekämen wir also das Polynom $4x + 2$.

Da wir von zwei primitiven Polynomen f und g ausgehen, muß nach den Ergebnissen aus §5 auch deren ggT h primitiv sein. 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 $2x + 1$.

§8: Die Landau-Mignotte-Schranke

In gewisser Weise leistet der vorige Paragraph das genaue Gegenteil dessen, was wir wollen: Wir wollen die schwierig zu berechnenden größten gemeinsamen Teiler in $\mathbb{Z}[x]$ zurückführen auf die leichter zu berechnenden über endlichen Körpern; stattdessen haben wir den ggT

der modulo p reduzierten Polynome zurückgeführt auf den in $\mathbb{Z}[x]$ berechneten. Leider gibt es zu einem Polynom $h_p \in \mathbb{F}_p[x]$ unendlich viele Polynome $h \in \mathbb{Z}[x]$, die modulo p gleich h sind.

Trotzdem können wir die obigen Resultate zur Berechnung von ggTs in $\mathbb{Z}[x]$ verwenden, *falls* wir eine Schranke für die Beträge der Koeffizienten des ggT finden: Wenn wir wissen, daß alle Koeffizienten von h ganze Zahlen vom Betrag höchstens M sind und wir eine Primzahl $p \geq 2M + 1$ betrachten, so ist h durch h mod p eindeutig bestimmt. Da wir tatsächlich, wie wir gerade gesehen haben, nicht h_p liften, sondern ch_p und dabei auch ein Vielfaches von ggT(f, g) bekommen können mit einem Faktor, der nur durch den ggT c der führenden Koeffizienten von f und g beschränkt ist, müssen wir sogar fordern, daß $p \geq 2cM + 1$ ist.

Sofern wir eine solche Schranke M für den ggT zweier Polynome $f, g \in \mathbb{Z}[x]$ haben, können wir also eine Primzahl $p \geq 2cM + 1$ wählen, die nicht beide führende Koeffizienten teilt, und den ggT $\bar{h} \in \mathbb{F}_p[x]$ von f mod p und g mod p berechnen. Zu $\bar{h}$ gibt es höchstens ein Polynom $h \in \mathbb{Z}[x]$, so daß h mod $p = \bar{h}$ ist. Falls es keines gibt, wissen wir, daß p eine der endlich vielen Ausnahmeprimzahlen ist; andernfalls müssen wir testen, ob h Teiler von f und g ist. Wenn ja, haben wir einen ggT von f und g gefunden, andernfalls folgt wieder, daß p eine Ausnahmeprimzahl war.

Im letzteren Fall müssen wir die Rechnung mit einer neuen Primzahl wiederholen; da es nur endlich viele Ausnahmeprimzahlen gibt, kann uns das höchstens endlich viele Male passieren.

Tatsächlich werden wir diesen Algorithmus im übernächsten Paragraphen noch etwas optimieren; aber damit er überhaupt funktioniert, brauchen wir auf jeden Fall eine Schranke für die Koeffizienten.

Eine solche Schranke brauchen wir auch im nächsten Kapitel für die Faktorisierung von Polynomen; deshalb fragen wir allgemeiner gleich nach einer Schranke für die Koeffizienten eines beliebigen Teilers eines vorgegebenen Polynoms $f \in \mathbb{Z}[x]$.

$f \in \mathbb{Z}[x]$ 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: Am wichtigsten ist natürlich

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

die sogenannte *Höhe* des Polynoms. 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

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

und das letzte schließlich ist klar, denn

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

■

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; den Zusammenhang zwischen Nullstellen und Koeffizienten liefert uns der aus Kapitel 1, §6 bekannte Wurzelsatz von Viète, der die Koeffizienten a_i eines Polynoms

$$x^d + a_{d-1}x^{d-1} + \cdots + a_0 = \prod_{i=1}^{d}(x - z_i)$$

durch die Nullstellen z_i ausdrückt: Bis aufs Vorzeichen ist a_k die Summe aller Produkte aus jeweils $d-k$ der z_i.

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:

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

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 ist bis aufs Vorzeichen die Summe aller Produkte von $d - k$ Nullstellen, und jedes einzelne solche Produkt 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 $f = \sum_{k=0}^{d} a_d x^d$ 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 $\|(x-z)f\|_2 = \left\| a_d x^{d+1} + \sum_{k=1}^{d} (z a_k - a_{k-1}) x^k - a_0 z \right\|_2$ ist die Summe aller Koeffizientenquadrate, also

$$a_d\,\overline{a_d} + \sum_{k=1}^{d} (z a_k - a_{k-1})\,\overline{(z a_k - a_{k-1})} + a_0 z \overline{a_0 z}$$

$$= |a_d|^2 + \sum_{k=1}^{d} \left(|a_k|^2 |z|^2 - 2\,\mathfrak{Re}(z a_k \overline{a_{k-1}}) + |a_{k-1}|^2 \right) + |a_0|^2 |z|^2$$

$$= (1 + |z|^2) \sum_{k=0}^{d} |a_k|^2 - 2 \sum_{k=1}^{d} \mathfrak{Re}(z a_k \overline{a}_{k-1})\,.$$

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

$$a_d \overline{z} \cdot \overline{a_d} z + \sum_{k=1}^{d} (\overline{z} a_{k-1} - a_k)(z \overline{a_{k-1}} - \overline{a_k}) + a_0 \overline{a_0}$$

$$= |z a_d|^2 + \sum_{k=1}^{d} \left(|z a_{k-1}|^2 - 2\,\mathfrak{Re}(z a_k \overline{a_{k-1}}) + |a_k|^2 \right) + |a_0|^2$$

$$= (1 + |z|^2) \sum_{k=0}^{d} |a_k|^2 - 2 \sum_{k=1}^{d} \mathfrak{Re}(z a_k \overline{a}_{k-1})\,.$$

■

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 $f, g \in \mathbb{Z}[x]$

$$H\big(\mathrm{ggT}(f,g)\big) \le \|\mathrm{ggT}(f,g)\|_1$$
$$\le \mathrm{LM}(f,g) \underset{\text{def}}{=} 2^{\min(d,e)}\,\mathrm{ggT}(a_d,b_e)\min\left(\frac{\|f\|_2}{|a_d|},\frac{\|g\|_2}{|b_e|}\right).$$

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 arbeitet 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 aus §3.

$$f = x^8 + x^6 - 3x^4 - 3x^3 + 8x^2 + 2x - 5$$

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; für

$$g = 3x^6 + 5x^4 - 4x^2 - 9x + 21$$

haben wir führenden Koeffizienten drei und

$$\|g\|_2 = \sqrt{3^2 + 5^2 + 4^2 + 9^2 + 21^2} = \sqrt{572} = 2\sqrt{143}\,.$$

Da $3^2 \cdot 113 > 900$ größer ist als $2^2 \cdot 143 < 600$, ist die LANDAU-MIGNOTTE-Schranke für diese beiden Polynome

$$\mathrm{LM}(f,g) = 2^6 \cdot \frac{2}{3}\sqrt{143} \approx 510{,}2191249\,.$$

Da die Koeffizienten des ggT ganze Zahlen sind, kann der Betrag eines jeden Koeffizienten also höchstens gleich 510 sein. Da die führenden Koeffizienten von f und g ggT eins haben, wissen wir außerdem, daß auch der führende Koeffizient von ggT(f, g) gleich eins ist.

Wir suchen daher eine Primzahl $p \geq 2 \cdot 510 + 1$, d.h. $p > 2 \cdot 510$:

```
> p := nextprime(2*510);
```

$$p := 1021$$

```
> f := x^8+x^6-3*x^4-3*x^3+8*x^2+2*x-5 mod p;
```

$$f := x^8 + x^6 + 1018 * x^4 + 1018 * x^3 + 8 * x^2 + 2 * x + 1016$$

```
> g := 3*x^6 + 5*x^4 - 4*x^2 - 9*x + 21 mod p;
```

$$g := 3x^6 + 5x^4 + 1017x^2 + 1012x + 21$$

```
> r2 := Rem(f, g, x) mod p;
```

$$r2 := 907x^4 + 227x^2 + 340$$

```
> r3 := Rem(g, r2, x) mod p;
```

$$r3 := 77x^2 + 1012x + 181$$

```
> r4 := Rem(r2, r3, x) mod p;
```

$$r4 := 405x + 581$$

```
> r5 := Rem(r3, r4, x) mod p;
```

$$r5 := 956$$

```
> r6 := Rem(r4, r5, x) mod p;
```

$$r6 := 0$$

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 Rechengang in §3, hat sich abgesehen von den modularen Polynomdivisionen nichts wesentliches geändert, jedoch sind die Zwischenergebnisse erheblich angenehmer geworden.

§9: Der chinesische Restesatz

Zu Beginn des vorigen Paragraphen haben wir uns kurz überlegt, wie man grundsätzlich mit Hilfe der LANDAU-MIGNOTTE-Schranke und des modularen ggT größte gemeinsame Teiler in $\mathbb{Z}[x]$ berechnen kann. Bei nicht allzu großer LANDAU-MIGNOTTE-Schranke ist dies wahrscheinlich die schnellste Art und Weise, den ggT zu berechnen.

Bei großer LANDAU-MIGNOTTE-Schranke M wird allerdings eine Primzahl $p \geq 2M + 1$ nicht mehr in ein Maschinenwort passen, so daß alle Rechnungen in Langzahlarithmetik und damit recht langsam ausgeführt werden müssen.

Der Ausweg besteht darin, nicht modulo einer, sondern gleich modulo mehrerer Primzahlen zu rechnen. Die Idee dazu ist einfach: Wenn wir beispielsweise ein Zahl sowohl modulo zwei als auch modulo fünf kennen, kennen wir sie auch modulo zehn. Entsprechendes gilt allgemein:

Chinesischer Restesatz: Sind m, n zwei zueinander teilerfremde und a, b zwei beliebige Elemente eines EUKLIDischen Rings R, so gibt es Elemente $r \in R$ mit

$$r \equiv a \bmod m \quad \text{und} \quad r \equiv b \bmod n\,.$$

r ist modulo mn eindeutig bestimmt.

Beweis: Ausgangspunkt ist der erweiterte EUKLIDische Algorithmus: Da m und n teilerfremd sind, liefert er uns zwei Elemente $\alpha, \beta \in R$, so daß

$$\alpha m + \beta n = \text{ggT}(m, n) = 1$$

ist. Somit ist

$$1 - \alpha m = \beta n \equiv \begin{cases} 1 & \bmod m \\ 0 & \bmod n \end{cases} \quad \text{und} \quad 1 - \beta n = \alpha m \equiv \begin{cases} 0 & \bmod m \\ 1 & \bmod n \end{cases},$$

also löst $r = \beta n a + \alpha m b \equiv \begin{cases} a & \bmod m \\ b & \bmod n \end{cases}$ das Problem.

Für jede andere Lösung s ist $r - s$ sowohl durch m als auch durch n teilbar; da beide teilerfremd sind, also durch mn. Umgekehrt ist klar, daß für jedes $\lambda \in R$ auch $r + \lambda mn$ eine Lösung ist. Die allgemeine

Lösung ist somit $r = (\beta a + \lambda m)n + (\alpha b - \lambda n)m$; insbesondere ist sie eindeutig modulo nm.

■

Bei der Lösung eines Systems

$$r \equiv a_i \bmod m_i \quad \text{für } i = 1, \ldots, N$$

können wir rekursiv vorgehen: Wir lösen die ersten beiden Kongruenzen $r \equiv a_1 \bmod m_1$ und $r \equiv a_2 \bmod m_2$ wie gerade besprochen; das Ergebnis ist eindeutig modulo $m_1 m_2$. Ist r_2 eine feste Lösung, so läßt sind die sämtlichen Lösung dieser beiden Kongruenzen gerade die Lösungen der einen Kongruenz

$$r \equiv r_2 \bmod m_1 m_2 \,.$$

Da die m_i paarweise teilerfremd sind, ist auch $m_1 m_2$ teilerfremd zu m_3. Mit dem erweiterten EUKLIDischen Algorithmus können wir daher wie oben das System

$$r \equiv r_2 \bmod m_1 m_2 \quad \text{und} \quad r \equiv a_3 \bmod m_3$$

lösen und zusammenfassen in einer Kongruenz

$$r \equiv r_3 \bmod m_1 m_2 m_3$$

und so weiter, bis wir schließlich ein r gefunden haben, das modulo aller m_i den gewünschten Wert hat und das modulo dem Produkt aller m_i eindeutig bestimmt ist.

Alternativ läßt sich die Lösung auch in einer geschlossenen Formel darstellen, allerdings um den Preis einer N-maligen statt $(N-1)$-maligen Anwendung des EUKLIDischen Algorithmus und größeren Zahlen schon von Beginn an: Um das System

$$r \equiv a_i \bmod m_i \quad \text{für} \quad i = 1, \ldots, N$$

zu lösen, berechne man zunächst für jedes i das Produkt

$$\widehat{m}_i = \prod_{j \neq i} m_j$$

der sämtlichen übrigen m_j und bestimme dazu Elemente $\alpha_i, \beta_i \in R$, für die gilt $\alpha_i m_i + \beta_i \widehat{m}_i = 1$ Dann ist

$$r = \sum_{j=1}^{N} \beta_j \widehat{m}_j a_j \equiv \beta_i \widehat{m}_i a_i = (1 - \alpha_i m_i) a_i \equiv a_i \bmod m_i \,.$$

Natürlich wird r hier – wie auch bei den obigen Formel – oft größer sein als das Produkt der m_i; um (in welchem Sinne auch immer) „kleine“ Lösungen zu finden, müssen wir also noch modulo diesem Produkt reduzieren.

Der chinesische Restesatz hat seinen Namen daher, daß angeblich chinesische Generäle ihre Truppen in Zweier-, Dreier-, Fünfer-, Siebenerreihen *usw.* antreten ließen und jeweils nur die (i.a. unvollständige) letzte Reihe abzählten. Aus den Ergebnissen ließ sich die Gesamtzahl der Soldaten berechnen, wenn das Produkt der verschiedenen Reihenlängen größer war als diese Anzahl.

Es ist zwar fraglich, ob es in China wirklich Generäle gab, die diesen Satz kannten und anwendeten, aber Beispiele dazu finden sich bereits in einem chinesischen Lehrbuch des dreizehnten Jahrhunderts, den 1247 erschienenen *Mathematischen Abhandlungen in neun Bänden* von CH'IN CHIU-SHAO (1202–1261). Dort geht es allerdings nicht um Soldaten, sondern um Reiskörner.

§10: Die modulare Berechnung des ggT

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 $\mathrm{ggT}(\lambda, \mu) \cdot \mathrm{ggT}(f, g)$.

Seien 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. Nach dem Satz am Ende von §7 wissen wir dann, daß der ggT h_p von f mod p und g mod p mindestens denselben Grad hat wie $\mathrm{ggT}(f, g)$ und daß es nur endlich viele Primzahlen gibt, für die sich die beiden Grade unterscheiden. Für alle anderen p ist $h_p = \mathrm{ggT}(f, g) \bmod p$.

b) Die endlich vielen „schlechten“ Primzahlen, für die 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 p eine schlechte Primzahl sein. 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:

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 $M = 2c\big\lceil \mathrm{LM}(f,g)\big\rceil + 1$. Setze außerdem $\mathcal{P} = \emptyset$ und $N = 1$.

Da der Betrag eines jeden Koeffizienten des ggT höchstens gleich $\big\lceil \mathrm{LM}(f,g)\big\rceil$ 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 mod p und g mod 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 mod p und g mod p. Falls dieser gleich eins ist, endet der Algorithmus und 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 schlechte Primzahlen p gibt, für die der in $\mathbb{F}_p[x]$ berechnete ggT nicht einfach die Reduktion von 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 ggT(f, g) haben kann, ist auch klar, daß der Algorithmus mit einem korrekten Ergebnis abbricht.

Betrachten wir dazu ein Beispiel:

```
> f := x^6-124*x^5-125*x^4-2*x^3+248*x^2+249*x+125:
> g := x^5+127*x^4+124*x^3-255*x^2-381*x-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

```
> Gcd(f, g) mod 107;
```

$$x^3 + 90x^2 + 90x + 89$$

Damit ist $\mathcal{P} = \{107\}$ und $N = 107 < M$. Also wählen wir eine weitere Primzahl, etwa $p = 271$:

```
> Gcd(f, g) mod 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:

```
> chrem([90, 127], [107, 271]);
```

$$5547$$

```
> chrem([89, 126], [107, 271]);
```

$$5546$$

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:

```
> rem(f, x^3 + 5547*x^2 + 5547*x + 5546, x);
```

$$967384732340761x^2 + 967384732340761x + 967384732340761$$

Offensichtlich nicht; somit sind 107 und 271 für dieses Problem schlechte Primzahlen. Versuchen wir unser Glück als nächstes mit $p = 367$:

```
> Gcd(f, g) mod 367;
```

$$x^2 + x + 1$$

Also 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. $p = 73$.

```
> Gcd(f, g) mod 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.

```
> Gcd(f, g) mod 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:

```
> rem(f, x^2+x+1, x);
```

$$0$$

```
> rem(g, x^2+x+1, x);
```

$$0$$

Damit ist 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.

§11: Polynome in mehreren Veränderlichen

Wie wir aus §6 wissen, ist auch der Polynomring in mehreren Veränderlichen über den ganzen Zahlen oder über einem Körper faktoriell; somit existieren auch dort größte gemeinsame Teiler. Im vorigen Paragraphen

haben wir 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 über faktoriellen Ring k; wichtig sind vor allem die Fälle $k = \mathbb{Z}$, $k = \mathbb{Q}$ und $k = \mathbb{F}_p$. Wie beim GAUSSschen Lemma betrachten wir die Polynome aus $R_n = k[x_1, \ldots, x_n]$ als Polynome in der einer Veränderlichen x_n über dem Ring $R_{n-1} = k[x_1, \ldots, x_{n-1}]$, 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.

In Kapitel II 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 in $\mathbb{F}_p[x]$ berechnet. Formal können wir das auch so ausdrücken, daß wir auf die Koeffizienten die Abbildung

$$\varphi_p\colon \begin{cases} \mathbb{Z} \to \mathbb{F}_p \\ a \mapsto a \bmod p \end{cases}$$

angewendet haben. Entsprechend können wir im Polynomring R_{n-1} noch einmal eine Variable auszeichnen, etwa x_{n-1}, und für diese einen festen Wert $c \in k$ einsetzen, d.h. wir wenden auf alle Koeffizienten die Abbildung

$$\varphi_c\colon \begin{cases} R_{n-1} \to R_{n-2} \\ a(x_1, \ldots, x_{n-2}, x_{n-1}) \mapsto a(x_1, \ldots, x_{n-2}, c) \end{cases}$$

an. Die entstehenden Polynome $\overline{f}$ und $\overline{g}$ aus $R_{n-2}[x]$ haben wieder insgesamt $n-1$ Variable, wir können ihren ggT also mit dem Algorithmus für Polynome in $n-1$ Variablen berechnen.

Auch hier stellt sich die Frage, was der ggT von $\overline{f}$ und $\overline{g}$ mit dem von f und g zu tun hat. Im folgenden bezeichne $\overline{h}$ für jedes Polynom $h \in R_{n-1}[x_n]$ das Polynom aus $R_{n-2}[x]$, das durch Anwendung von φ_c auf die Koeffizienten von h entsteht.

Ist $h \in R_{n-1}[x]$ ein Teiler von f, etwa $f = qh$, so ist $\overline{f} = \overline{q}\overline{h}$, d.h. auch $\overline{h}$ ist ein Teiler von $\overline{f}$. Dieser Teiler könnte aber einen kleineren Grad haben als h; dies passiert offensichtlich genau dann, wenn der führende Koeffizient von h im Kern von φ_c liegt, durch Einsetzen von $x_{n-1} = c$ also zur Null wird. Da der führende Koeffizient von f das Produkt der führenden Koeffizienten von $\overline{h}$ und $\overline{q}$ 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 φ_c nicht auf die Null abgebildet wird. Wenn wir das für f *oder* g sicherstellen, wissen wir daher, daß $\overline{\text{ggT}(f,g)}$ ein Teiler von $\overline{f}$ und $\overline{g}$, also auch von $\text{ggT}(\overline{f},\overline{g})$ 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 $\overline{h} = \overline{\text{ggT}(f,g)}$ ein Teiler von $\text{ggT}(\overline{f},\overline{g})$ ist. $\overline{h}$ ist genau dann ein echter Teiler, wenn $\overline{f}/\overline{h}$ und $\overline{g}/\overline{h}$ einen gemeinsamen Faktor haben, der keine Einheit ist, wenn also die Resultante von $\overline{f}/\overline{h}$ und $\overline{g}/\overline{h}$ bezüglich x_n verschwindet. Bezeichnet h den ggT von f und g, so entsteht diese Resultante aus $\text{Res}_{x_n}(f/h, g/h) \in R_{n-1}$ durch Anwendung von φ_c; da diese Resultante als Polynom in x_{n-1} geschrieben werden kann, gibt es also 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

ggT($\overline{f}, \overline{g}$) für 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 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 mathematischer Direktor der dortigen Akademie 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

$$F = \sum_{i=0}^{d} a_i(x_1, \ldots, x_{n-1}) x_n^i \quad \text{und} \quad G = \sum_{j=0}^{e} b_j(x_1, \ldots, x_{n-1}) x_n^j ,$$

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 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

$$\overline{f} = \sum_{i=0}^{d} a_i(x_1,\ldots,x_{n-2},c)x_n^i \quad \text{und} \quad \overline{g} = \sum_{j=0}^{e} b_j(x_1,\ldots,x_{n-2},c)x_n^j\,.$$

Falls $h_c = 1$, endet der Algorithmus mit dem Ergebnis $\mathrm{ggT}(f,g) = 1$. Andernfalls wird $\mathcal{M} = \{c\}$ und $N = \deg_{x_n} h_c$ und m wird eins mehr als das Maximum der Grade der a_i und der b_j in der Variablen x_{n-1}.

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

$$\overline{f} = \sum_{i=0}^{d} a_i(x_1,\ldots,x_{n-2},c)x_n^i \quad \text{und} \quad \overline{g} = \sum_{j=0}^{e} b_j(x_1,\ldots,x_{n-2},c)x_n^j\,.$$

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 $c = 1, 2, 3, 4$ 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.

Kapitel 3
Faktorisierung von Polynomen

Die Lösung sowohl einzelner Polynomgleichungen als auch von Systemen solcher Gleichungen wird mit steigendem Grad der Polynome sehr schnell sehr viel schwieriger; falls man einzelne der Polynome in Faktoren zerlegen kann, ist es meist effizienten, mit diesen zu arbeiten – obwohl die Anzahl der betrachteten Fälle bei Systemen von hinreichend vielen Polynomgleichungen auch da ziemlich groß werden kann.

Wie wir aus Kapitel 2, §6 wissen, läßt sich jedes Polynom über einem faktoriellen Ring als Produkt irreduzibler Polynome schreiben. Wir sollten daher bei der Suche nach den Nullstellen eines Polynoms zunächst einmal versuchen, es in irreduzible Faktoren zu zerlegen. Dazu brauchen wir allerdings neue Methoden, denn der Beweis von GAUSS ist nicht konstruktiv.

Der erste zumindest im Prinzip konstruktive Beweis für Polynome über den ganzen Zahlen geht zurück auf KRONECKER; wirklich effiziente Verfahren gibt es erst seit der zweiten Hälfte des zwanzigsten Jahrhunderts.

Auch sie schlagen wie der modulare Algorithmus zur ggT-Berechnung den Umweg über endliche Körper ein, funktionieren allerdings dort nur für Polynome ohne mehrfache Nullstellen, so daß wir zunächst die mehrfachen Nullstellen eliminieren müssen.

Der Aufbau dieses Kapitels ist daher folgender: Als erstes betrachten wir den klassischen Algorithmus von KRONECKER, danach beschäftigen wir uns mit der sogenannten quadratfreien Faktorisierung, durch die wir auf Polynome ohne mehrfache Nullstellen kommen. Als nächstes geht es um den Faktorisierung von Polynomen über endlichen Körpern, und im

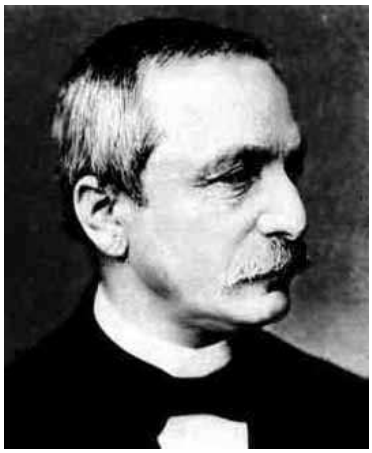
Rest des Kapitels dann darum, wie diese Faktorisierung hochgehoben werden kann für Polynome über $\mathbb{Z}$ und damit auch $\mathbb{Q}$.

§1: Der Algorithmus von Kronecker

KRONECKER führt das Problem der Faktorisierung eines Polynoms über $\mathbb{Z}$ zurück auf die 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 n ist eindeutig bestimmt durch seine Werte an $n+1$ verschiedenen Stellen $a_0, \ldots, a_n$ und kann mit Hilfe wohlbekannter Interpolationsformeln leicht aus den $n+1$ 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 n für $n+1$ beliebig gewählte ganzzahlige Werte $a_0, \ldots, a_n$ die Funktionswerte $f(a_0), \ldots, f(a_n)$, und konstruiert für jedes $(n+1)$-tupel $(b_0, \ldots, b_n)$ 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 n, andernfalls wird einer gefunden.



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

Über den Grad n 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 daher alle diese Grade ausprobiert

werden 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. Zum Glück gibt es deutlich effizientere Alternativen, so daß der Algorithmus von KRONECKER in der Praxis nie eingesetzt wird.

Als Beispiel wollen wir versuchen, das Polynom

$$f = 8x^7 - 16x^6 - 20x^5 + 15x^4 + 13x^3 + 9x^2 + 10x + 2$$

in $\mathbb{Z}[x]$ zu faktorisieren. Falls alle Nullstellen ganzzahlig sind, müssen die linearen Faktoren von der Form $(x-a)$ sein, wobei a den konstanten Term teilt; wegen

$$f(-2) = -1254, \quad f(-1) = -1, \quad f(1) = 21 \quad \text{und} \quad f(2) = 238$$

gibt es keinen Faktor dieser Art.

Dies bedeutet freilich nicht, daß es keine lineare Faktoren gibt; ein solcher Faktor könnte ja auch die Form $(bx + c)$ haben. Um solche Faktoren 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 daher $g(x_0) = \pm 1$ 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. Der Maple-Befehl zur Konstruktion des Interpolationspolynoms durch die Punkte (x_1,y_1) bis (x_n,y_n) ist `interp([x1, ..., xn], [y1, ..., yn], x)`; wir schreiben also

```
> for y0 in [-2, -1, 1, 2] do
> g := interp([-1, 0], [1, y0], x);
> if rem(f, g, x) = 0 then print(g) fi od:
```

Die einzige „Lösung“ die wir bekommen, ist das konstante Polynom 1, das natürlich auf keine Faktorisierung führt. 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 haben muß. Nun müssen also schon $4 \times 8 = 32$ Interpolationspolynome konstruiert und durchprobiert werden:

```
> for y0 in [-2, -1, 1, 2] do
> for y1 in [-21, -7, -3, -1, 1, 3, 7, 21] do
> g := interp([-1, 0, 1], [1, y0, y1], x);
> if rem(f, g, x) = 0 then print(g) fi od od:
```

$$1$$

Es gibt also auch keine quadratischen Teiler.

Für kubische Faktoren brauchen wir einen weiteren Interpolationspunkt. Wir kennen bereits die beiden Funktionswerte $f(2) = -238$ und $f(-2) = -1254$; Versuche mit anderen betragskleinen x-Werten liefern nicht besseres. Da $238 = 2 \cdot 7 \cdot 17$ nur drei Primteiler hat, $1254 = 2 \cdot 3 \cdot 11 \cdot 19$ aber vier, versuchen wir unser Glück mit dem Punkt $(2, -238)$, wobei wir nun für $g(2)$ schon 16 Werte betrachten müssen, insgesamt also $4 \times 8 \times 16 = 512$ Interpolationspolynome:

```
> for y0 in [-2, -1, 1, 2] do
> for y1 in [-21, -7, -3, -1, 1, 3, 7, 21] do
> for y2 in [-238, -119, -34, -17, -14, -7, -2, -1,
>            1, 2, 7, 14, 17, 34, 119, 238] do
> g := interp([-1, 0, 1, 2], [1, y0, y1, y2], x);
> if rem(f, g, x) = 0 then print(g) fi od od od:
```

$$1$$

$$-2x^3 + 3x^2 + 5x + 1$$

Somit gibt es bis aufs Vorzeichen genau einen kubischen Faktor, und die Zerlegung von f ist

$$\begin{aligned} f &= (-2x^3 + 3x^2 + 5x + 1)(-4x^4 + 2x^3 + 3x^2 + 2) \\ &= (2x^3 - 3x^2 - 5x - 1)(4x^4 - 2x^3 - 3x^2 - 2)\,. \end{aligned}$$

§2: Die quadratfreie Zerlegung eines Polynoms

Wir betrachten ein Polynom f über einem Körper 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. 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 erklären läßt.

Wir beschränken uns hier auf Polynome über einem Körper k und definieren die Ableitung eines Polynoms

$$f = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \in k[x]$$

als das Polynom

$$f' = n a_n x^{n-1} + (n-1) a_{n-1} x^{n-2} + \cdots + a_1 \in k[x] \,.$$

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

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

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 $a^p = a$. ∎

Speziell für den Körper $\mathbb{F}_p$ vereinfacht sich daher das obige Lemma zum folgenden

Korollar: Für ein Polynom mit Koeffizienten aus $\mathbb{F}_p$ ist

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

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.

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 positiver Charakteristik kann es allerdings vorkommen, daß $\operatorname{ggT}(f, f') = f$ ist. Da $\deg f' < \deg f$, ist dies genau dann der Fall, wenn $f' = 0$ ist. Dies ist in Charakteristik Null genau dann der Fall, wenn f konstant ist; in Charakteristik $p > 0$ verschwindet aber auch die Ableitung eines jeden Polynoms in x^p. Somit ist hier $f' = 0$ genau dann, wenn alle in f vorkommenden x-Potenzen einen durch p teilbaren Exponenten haben. Für $f \in \mathbb{F}_p[x]$ ist dann, wie wir oben gesehen haben,

$$\begin{aligned} f &= a_{np}x^{np} + a_{(n-1)p}x^{(n-1)p} + \cdots + a_p x^p + a_0 \\ &= \left(a_{np}x^p + a_{(n-1)p}x^{(n-1)} + \cdots + a_p x + a_0\right)^p , \end{aligned}$$

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

$$f = \left(b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0\right)^p .$$

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.

Man kann zeigen, daß jeder endliche Körper vollkommen ist: Im Körper mit p^n Elementen ist $x^{p^n} = x$ für alle $x \in \mathbb{F}_{p^n}$, und damit ist x die p-te Potenz von $y = x^{p^{n-1}}$. Ein Beispiel eines nicht vollkommenen Körpers wäre $\mathbb{F}_p(x)$, 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 $f_1, \ldots, f_N$. 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 $\mathrm{ggT}(g - s_i, f)$ 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 *usw.*

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 daher 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 $\deg f = n$, so können wir jedes $g \in V$ schreiben als

$$g = g_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + g_1 x + g_0$$

und

$$g^p = g_{n-1}x^{(n-1)p} + g_{n-2}x^{(n-2)p} + \cdots + g_1 x^p + g_0$$

mit geeigneten Koeffizienten $g_i \in \mathbb{F}_p$.

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} mit Rest durch f:

$$x^{ip} \equiv \sum_{j=0}^{n-1} b_{ij} x^j \bmod f\,.$$

Dann muß gelten

$$\sum_{i=0}^{n-1} g_i x^{ip} \bmod f = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} b_{ij} g_i x^j = \sum_{j=0}^{n-1} g_j x^j\,.$$

Koeffizientenvergleich führt auf das homogene lineare Gleichungssystem

$$\sum_{i=0}^{n-1} b_{ij} g_i = g_j \quad \text{für } j = 0, \ldots, n-1\,.$$

V ist also auch beschreibbar als der Lösungsraum dieses 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 er also genau dann eindimensional, wenn f irreduzibel ist.

Andernfalls wählen wir irgendein Element $g \in V$ und berechnen die Polynome ggT$(g - \lambda, f)$ für alle $\lambda \in \mathbb{F}_p$. 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 v linear unabhängiges neues Polynom $h \in V$ und verfahren damit genauso. Indem wir für jedes nichtkonstante Polynom ggT$(h - \lambda, f)$ 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($h - \lambda, f$) 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. Seine dortige home page ist `math.berkeley.edu/~berlek/`

Als Beispiel wollen wir das Polynom $f = x^6 + 2x^5 + 4x^4 + x^3 - x^2 - x - 1$ aus $\mathbb{F}_7[x]$ faktorisieren. Wir setzen ein Polynom vom Grad fünf mit unbestimmten Koeffizienten an:

```
> G := x -> add(g[i]*x^i, i=0..5);
```

$$G := x -> add(g_i x^i, i = 0..5)$$

```
> G(x);
```

$$g_0 + g_1 x + g_2 x^2 + g_3 x^3 + g_4 x^4 + g_5 x^5$$

Die siebte Potenz davon ist

```
> G(x^7);
```

$$g_0 + g_1 x^7 + g_2 x^{14} + g_3 x^{21} + g_4 x^{28} + g_5 x^{35}$$

Um sie modulo f auszudrücken, müssen wir die Divisionsreste h_i bei der Division von x^{7i} durch f berechnen:

```
> for i to 5 do h[i] := Rem(x^(7*i), f, x) mod 7; od;
```

$$h_1 := 3x^3 + 6x^2 + 6x + 5$$
$$h_2 := 4x^5 + x^4 + 2x^3 + 6x + 6$$
$$h_3 := x^5 + 4x^4 + 6x^3 + 6x^2 + 2x + 5$$
$$h_4 := x^5 + 6x^4 + 5x^3 + x^2 + x + 2$$
$$h_5 := 2x^5 + 4x^4 + 6x^3 + 4x^2 + 2x + 5$$

Damit können wir g^7 mod f explizit hinschreiben:

```
> G7 := sort(collect(expand(
>             g0 + add(g[i]*h[i], i=1..5)), x), x);
```

$$G7 := (4g_2 + g_3 + 2g_5 + g_4)x^5 + (6g_4 + g_2 + 4g_3 + 4g_5)x^4$$
$$+(5g_4 + 6g_5 + 3g_1 + 6g_3 + 2g_2)x^3 + (g_4 + 4g_5 + 6g_1 + 6g_3)x^2$$
$$+(6g_2 + g_4 + 6g_1 + 2g_3 + 2g_5)x + 5g_5 + g0 + 6g_2 + 5g_3 + 2g_4 + 5g_1$$

Das soll gleich g sein, was auf ein lineares Gleichungssystem für die sechs Variablen $g[i]$ führt:

```
> LGS := seq(coeff(G7, x, i) = g[i], i=0..5);
```

$$LGS := \{5g_5 + g0 + 6g_2 + 5g_3 + 2g_4 + 5g_1 = g_0,$$
$$6g_2 + g_4 + 6g_1 + 2g_3 + 2g_5 = g_1,$$
$$g_4 + 4g_5 + 6g_1 + 6g_3 = g_2,$$
$$5g_4 + 6g_5 + 3g_1 + 6g_3 + 2g_2 = g_3,$$
$$6g_4 + g_2 + 4g_3 + 4g_5 = g_4,$$
$$4g_2 + g_3 + 2g_5 + g_4 = g_5\}$$

Zur Lösung eines Gleichungssystems über einem Körper $\mathbb{F}_p$ können wir den Befehl `msolve` verwenden; sein erstes Argument ist eine Menge von Gleichungen, das zweite, das auch fehlen kann, eine Menge von Variablen, und das dritte die Primzahl p. Da wir hier nach *allen* Variablen auflösen wollen, können wir auf das zweite Argument verzichten:

```
> msolve(LGS, 7);
```

$$g_0 = \_Z1, g_5 = \_Z2, g_4 = 3\_Z2, g_3 = 5\_Z2, g_1 = 6\_Z2, g_2 = 3\_Z2$$

Das bedeutet folgendes: Die Lösungen hängen ab von zwei Parametern; für diese führt Maple die Bezeichnungen `_Z1` und `_Z2` ein, wobei der Buchstabe „Z“ für ganze Zahl stehen soll. Insbesondere ist also der Lösungsraum zweidimensional, das Polynom f hat also zwei irreduzible Faktoren. Wir können uns eine Basis des Lösungsraums verschaffen, indem wir für das erste Basispolynom $\_Z1 = 1$ und $\_Z2 = 0$ setzen und für das zweite $\_Z1 = 0$ und $\_Z2 = 1$. Mit dem Befehl

```
> assign(%);
```

können wir aus obigen Gleichungen Zuweisungen machen und dann substituieren:

```
> G_1 := subs(_Z1 = 1, _Z2 = 0, G(x)) mod 7;
```

$$G\_1 := 1$$

Das bringt offensichtlich nichts. Beim zweiten Versuch

```
> G_2 := subs(_Z1 = 0, _Z2 = 1, G(x)) mod 7;
```

$$G\_2 := x^5 + 3x^4 + 5x^3 + 3x^2 + 6x$$

haben wir mehr Glück und müssen nun für alle $i \in \mathbb{F}_7$ die größten gemeinsamen Teiler von $G\_2 - i$ und f berechnen:

```
> for i from 0 to 6 do Gcd(G_2-i, f) mod 7; od;
```

$$x^3 + 5x + 2$$
$$1$$
$$x^3 + 2x^2 + 6x + 3$$
$$1$$
$$1$$
$$1$$
$$1$$

Somit ist $f = (x^3 + 5x + 2)(x^3 + 2x^2 + 6x + 3)$ die Zerlegung von f in irreduzible Faktoren. Zur Vorsicht können wir das noch von Maple verifizieren lassen:

```
> expand((x^3+2*x^2+6*x+3)*(x^3+5*x+2)) mod 7;
```

$$x^6 + 2x^5 + 4x^4 + x^3 + 6x^2 + 6x + 6$$

Da $6 = -1$ in $\mathbb{F}_7$, ist das in der Tat unser Ausgangspolynom f.

§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 benutzen, 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 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 $\overline{f} \in \mathbb{F}_p[x]$ aber nicht, so haben $\overline{f}$ und seine Ableitung einen gemeinsamen Faktor, ihre Resultante verschwindet also. Da diese Resultante die Reduktion modulo p der Resultante von f und f' ist, bedeutet dies einfach, daß p ein Teiler der über $\mathbb{Z}$ berechneten Resultante ist, 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*

$$D(f) = \frac{(-1)^{\frac{1}{2}n(n-1)}}{a_n} \operatorname{Res}_x(f, f'),$$

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 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, bleibt f auch modulo p quadratfrei, jedoch können sich die Zerlegungen in irreduzible Faktoren in $\mathbb{Z}[x]$ und $\mathbb{F}_p[x]$ 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 $p \equiv 1 \bmod 4$ schreiben wir $p = 4k + 1$. Nach dem kleinen Satz von FERMAT ist für alle $x \in \mathbb{F}_p^\times$

$$(x^{p-1} - 1) = (x^{2k} + 1)(x^{2k} - 1) = 0\,,$$

das linksstehende Polynom hat also $p - 1 = 4k$ Nullstellen und zerfällt damit über $\mathbb{F}_p$ in Linearfaktoren. Damit gilt dasselbe für die beiden rechtsstehenden Faktoren; insbesondere gibt es also 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 $p \equiv 3 \bmod 4$ ist $\mathrm{ggT}(4, p - 1) = 2$ als Linearkombination von 2 und $p - 1$ darstellbar, also ist auch $a^2 = 1$, im Widerspruch zu Annahme $a^2 = -1$. Somit gibt es in $\mathbb{F}_p$ keine Elemente mit Quadrat -1.

■

Damit ist $x^2 + 1$ genau dann irreduzibel über $\mathbb{F}_p$, wenn $p \equiv 3 \bmod 4$; 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.

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 genau wie oben im Falle $x^2 + 1$ 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 also 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 beiden Elemente a, b nicht als Quadrate darstellbar, so ist ab ein Quadrat.

Beweis: Für $p = 2$ ist jedes Element ein Quadrat und nicht 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 y^2, denn sonst wäre $a = y^2x^{-2}$ selbst 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 schlechte Primzahlen gibt: Hier können *alle* Primzahlen schlecht sein in dem Sinne, daß ein irreduzibles Polynom aus $\mathbb{Z}[x]$ modulo p reduzibel wird, und oft wird zumindest die Hälfte aller Primzahlen schlecht sein. Der Ansatz über den chinesischen Restesatz empfiehlt sich hier also definitiv nicht: 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 Weg über endliche Körper verfolgt daher im Falle der Faktorisierung eine andere Strategie als beim EUKLIDischen Algorithmus: Wir beschränken uns auf eine einzige Primzahl – unabhängig davon, ob diese nun gut oder schlecht dafür geeignet ist.

Wir kennen bereits aus dem vorigen Kapitel Schranken für die Koeffizienten der Faktoren eines Polynoms; wir könnten also eine Primzahl wählen, die größer ist als das Doppelte dieser Schranke und modulo dieser rechnen.

Der Nachteil dabei ist, daß das Rechnen modulo einer Primzahl p umso teurer wird, je größer die Primzahl ist: Die Kosten für Multiplikationen wachsen quadratisch mit der Stellenzahl von p, die Kosten für Divisionen modulo p nach dem erweiterten EUKLIDischen Algorithmus können sogar bis zu kubisch ansteigen.

Die Alternative bietet ein für völlig andere Zwecke bewiesenes Resultat des deutschen Zahlentheoretikers HENSEL, das es erlaubt eine Faktorisierung modulo p fortzusetzen zu einer Faktorisierung modulo jeder beliebiger p-Potenz und, was HENSEL wirklich interessierte, zu den sogenannten p-adischen Zahlen, mit denen wir uns in Rahmen dieser Vorlesung allerdings nicht beschäftigen werden.

§5: Das Henselsche Lemma

Lemma: f, g, h seien Polynome aus $\mathbb{Z}[x]$ derart, daß $f \equiv gh \bmod p$; dabei 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

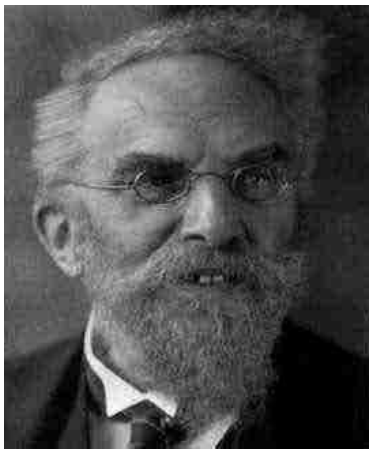
$$p^n f^* \equiv p^n (g_n h^* + h_n g^*) \bmod p^{n-1} .$$

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 $f^* = gh^* + hg^*$ in $\mathbb{F}_p[x]$ finden mit $\deg g^* \le \deg g$ und $\deg h^* \le \deg h$. 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.

§6: Der Algorithmus von Zassenhaus

Die Werkzeuge aus den vorigen Paragraphen erlauben uns, gemeinsam eingesetzt, nun die Faktorisierung von Polynomen f aus $\mathbb{Z}[x]$ oder $\mathbb{Q}[x]$. Das einzige, was wir noch nicht explizit formuliert haben, ist eine Schranke für die Koeffizienten eines Faktors. Aus Kapitel 2, §8, wissen wir, daß für einen Teiler $g \in \mathbb{C}[z]$ eines Polynoms $f \in \mathbb{C}[z]$ gilt:

$$H(g) \le \binom{e}{[e/2]} \left|\frac{b_e}{a_d}\right| \|f\|_2\,,$$

wobei e den Grad von g bezeichnet und a_d, b_e die führenden Koeffizienten von f und g. 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 $\mathbb{Q}[x]$ nun 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 $L = \binom{\deg g}{[\frac{1}{2} \deg g]} \|g\|_2$ und $M = 2L + 1$. Dann wählen wir eine Primzahl p, die weder den führenden Koeffizienten noch die Diskriminante von g teilt. Damit ist auch g mod p quadratfrei.

Dritter Schritt: Wir faktorisiere g mod 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 haben als $\frac{1}{2}\deg g$, denn falls f reduzibel ist, gibt es einen Faktor, der höchstens diesen Grad hat.

HANS JULIUS ZASSENHAUS wurde 1912 in Koblenz geboren, ging aber in Hamburg zur Schule und zur Universität. Er promovierte 1934 mit einer Arbeit 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.

§7: Berechnung von Resultanten und Diskriminanten

Im zweiten Schritt des Algorithmus von ZASSENHAUS wählten wir eine Primzahl, die kein Teiler der Diskriminante des zu faktorisierenden Polynoms sein darf. Um dies zu testen, müssen wir die Diskriminante berechnen oder – was äquivalent ist – die Resultante des Polynoms und seiner Ableitung. Für ein Polynom vom Grad zwanzig ist das eine 39×39-Determinante. Nach dem LAPLACEschen Entwicklungssatz ist dies eine Summe von $39! \approx 2 \cdot 10^{47}$ Summanden, die jeweils Produkte von 39 Zahlen sind. Eine solche Summe zu berechnen liegt weit jenseits der Möglichkeiten heutiger Computer.

Tatsächlich verwendet natürlich niemand den Entwicklungssatz von LAGRANGE um eine Determinante zu berechnen – außer vielleicht bei

einigen kleineren Spielzeugdeterminanten in Mathematikklausuren. In allen anderen Fällen wird man die Matrix durch Zeilen- und/oder Spaltenoperationen auf Dreiecksform bringen und dann die Determinante einfach als Produkt der Diagonaleinträge berechnen. 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 der EUKLIDische Algorithmus.

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

$$f = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \quad \text{und}$$
$$g = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0 \quad \text{mit} \quad n \leq m\,.$$

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

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

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

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 $\mathrm{Res}_x(f, h)$ kleiner und einfacher als $\mathrm{Res}_x(f, g)$.

Überlegen wir uns, was bei der Polynomdivision mit den Koeffizienten passiert.

Wir berechnen 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 $\deg g - \deg f = m - n$.

Die Zeilen der SYLVESTER-Matrix sind Vektoren in R^{n+m}; die ersten m sind die Koeffizientenvektoren von $x^{m-1}f, \ldots, xf, f$, danach folgen die von $x^{n-1}g, \ldots, xg, g$.

Im ersten Divisionschritt subtrahieren wir von g ein Vielfaches $\lambda x^j f$ mit $j = m - n$; damit subtrahieren wir auch von jeder Potenz $x^i g$ das Polynom $\lambda x^{i+j} f$. Für $0 \leq i < n$ und $0 \leq j \leq m + n$ ist $0 \leq i + j < m$, 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 ersetzt durch die entsprechende Zeilen mit Koeffizienten von h, wobei h als ein Polynom vom Grad m behandelt wird, dessen führende Koeffizienten verschwinden. Ist $h = c_s x^s + \cdots + c_0$, so ist also $\mathrm{Res}_x(f, g)$ gleich

$$\begin{vmatrix}
a_n & a_{n-1} & a_{n-2} & \ldots & a_1 & a_0 & 0 & 0 & \ldots & 0 \\
0 & a_n & a_{n-1} & \ldots & a_2 & a_1 & a_0 & 0 & \ldots & 0 \\
0 & 0 & a_n & \ldots & a_3 & a_2 & a_1 & a_0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & a_n & a_{n-1} & a_{n-2} & a_{n-3} & \ldots & a_0 \\
c_m & c_{m-1} & c_{m-2} & \ldots & c_2 & c_1 & c_0 & 0 & \ldots & 0 \\
0 & c_m & c_{m-1} & \ldots & c_3 & c_2 & c_1 & c_0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 0 & c_m & c_{m-1} & c_{m-2} & \ldots & c_0
\end{vmatrix},$$

wobei die Koeffizienten $c_m, \ldots, c_{s+1}$ alle verschwinden.

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

In den ersten $m - s$ Spalten der Matrix stehen daher nur noch Koeffizienten von f: In der ersten ist dies ausschließlich der führende Koeffizient a_n von f in der ersten Zeile. Entwickeln wir nach der ersten Zeile, können wir also einfach die erste Zeile und die erste Zeile streichen; die Determinante ist dann a_n mal der Determinante der übrigbleibenden Matrix. Diese hat (falls $m > s + 1$) wieder dieselbe Gestalt, wir können also wieder einen Faktor a_n ausklammern und bekommen eine Determinante mit einer Zeile und einer Spalte weniger *usw.;* das Ganze funktioniert $m - s$ 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 $\text{Res}_x(f, g) = a_n^{m-s} \text{Res}_x(f, h)$, 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 $\text{Res}(f, g) = a_n^{m-s} \text{Res}(f, h)$. ∎

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 $\text{Res}_x(f, g)$ durch Multiplikation mit Potenzen der führenden Koeffizienten der Divisoren zurückführen auf die viel kleineren Resultanten $\text{Res}_x(r_i, r_{i+1})$. 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 n und ein Polynom g vom Grad n ist $\text{Res}(f, g) = (-1)^{nm} \text{Res}(g, f)$.

Beweis: Wir müssen in der SYLVESTER-Matrix m Zeilen zu f mit den n 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 n Vertauschungen schließlich unten steht. Dies

müssen wir wiederholen, bis alle f-Zeilen unten stehen, wir haben also insgesamt nm Zeilenvertauschungen. Somit ändert sich das Vorzeichen der Determinante um den Faktor $(-1)^{nm}$.

■

§8: Swinnerton-Dyer Polynome

Der potentiell problematischste Schritt des obigen Algorithmus ist der sechste: 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.

Die GALOIS-Gruppe von K_n über $\mathbb{Q}$ hat somit 2^n Elemente; da sie offensichtlich die Abbildungen $\sqrt{p_i} \mapsto -\sqrt{p_i}$ enthält, ist sie die von diesen

Automorphismen erzeugte elementarabelsche Gruppe. Sie läßt die Nullstellenmenge von f als ganzes betrachtet 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 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 Faktoren 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} (oder sogar nur 2^{100}) Re-

chenoperationen in realistischer Zeit (d.h. wenigen Jahren) auszuführen.

SIR HENRY PETER FRANCIS SWINNERTON-DYER, 16th Baronet, wurde 1927 geboren; er studierte und lehrte an der Universität von Cambridge, wo er unter anderem Dean des Trinity College, Master von St. Catherine College und Vizekanzler der Universität war; heute ist er Professor emeritus. Obwohl er hauptsächlich für seine Beiträge zur Zahlentheorie bekannt ist, beschäftigte er sich zunächst mit 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 von deren ζ-Funktion.

§9: Faktoren und Gittervektoren

In diesem Paragraphen soll eine Methode vorgestellt werden, die für Polynome einer Veränderlichen über $\mathbb{Z}$ (der $\mathbb{Q}$), 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, die ehemaligen Bell Labs; seit 2006 ist er Professor für Kryptologie an der 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, seither nur noch in Leiden. Seine Arbeiten beschäftigen sich hauptsächlich mit der algorithmischen Seite der Zahlentheorie; 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.

Ausgangspunkt der Methode von LENSTRA, LENSTRA und LOVÁSZ 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 ist die Umkehrung.

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 also eine Darstellung der Eins als Linearkombination dieser beiden Polynome in $\mathbb{F}_p[x]$. Wir liften die Koeffizienten nach $\mathbb{Z}[x]$ und haben somit Polynome $a, b \in \mathbb{Z}[x]$, für die $ah + bq \equiv 1 \bmod p$ ist, d.h. es gibt ein Polynom $c \in \mathbb{Z}[x]$, so daß $ah + bq = 1 - pc$ ist. Da

$$\left(1 - pc\right)\left(1 + pc + (pc)^2 + \cdots + (pc)^{k-1}\right) = 1 - (pc)^k$$

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 $f \bmod p$ den zugehörigen Faktor $h_0 \in \mathbb{Z}[x]$ von f bestimmt, indem er nötigenfalls alle Kombinationen aus h und den anderen Faktoren von $f \bmod p$ 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 $\vec{h}_0$ 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 $h \bmod p^k$ 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 $|\vec{v}|$ seines Koeffizientenvektors $\vec{v} \in \mathbb{R}^{m+1}$.

Definition: Eine Teilmenge $\Gamma \subset \mathbb{R}^{m+1}$ heißt *Gitter*, wenn es eine Basis $(\vec{b}_0, \ldots, \vec{b}_m)$ von $\mathbb{R}^{m+1}$ gibt, so daß

$$\Gamma = \left\{ \sum_{i=0}^{m} \lambda_i \vec{b}_i \;\middle|\; \lambda_i \in \mathbb{Z} \right\} .$$

Wir bezeichnen diese Basis dann als eine Basis des Gitters Γ und schreiben kurz $\Gamma = \mathbb{Z}\vec{b}_0 \oplus \cdots \oplus \mathbb{Z}\vec{b}_m$.

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 $(\vec{b}_0, \ldots, \vec{b}_m)$ und $(\vec{c}_0, \ldots, \vec{c}_m)$ zwei Basen des Gitters Γ, so sind beide insbesondere 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 $\vec{b}_i$ zu einer Matrix B zusammenfassen und die $\vec{c}_j$ zu einer Matrix C; dann ist $C = MB$ und $B = NC$. Die Einträge von M und N müssen ganzzahlig sein, denn die $\vec{c}_j$ müssen ja ganzzahlige Linearkombinationen der $\vec{b}_i$ sein und umgekehrt. Außerdem ist $MN = NM$ gleich der Einheitsmatrix. Somit sind $\det M$ und $\det N$ ganzzahlig mit Produkt eins, d.h. $\det M = \det N = \pm 1$. Insbesondere 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 $\vec{b}_0, \ldots \vec{b}_n$ eine Gitterbasis von Γ und $\vec{c}_0, \ldots, \vec{c}_n$ eine von Δ, so lassen sich die $\vec{b}_i$ als Linearkombinationen der $\vec{c}_j$ 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 $\vec{m}_i$ von M, denn $N\vec{m}_i = \vec{e}_i$ 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 $\vec{m}_i$ 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 $(\vec{b}_0, \ldots, \vec{b}_m)$ eine Gitterbasis von Δ, so gibt es eine Gitterbasis $(\vec{c}_0, \ldots, \vec{c}_m)$ von Γ derart, daß

$$\begin{aligned}
\vec{c}_0 &= \mu_{00}\vec{b}_0 \\
\vec{c}_1 &= \mu_{10}\vec{b}_0 + \mu_{11}\vec{b}_1 \\
&\ldots \\
\vec{c}_m &= \mu_{m0}\vec{b}_0 + \cdots + \mu_{mm}\vec{b}_m
\end{aligned}$$

mit ganzen Zahlen μ_{ij} und $\mu_{ii} \neq 0$.

Beweis: Da $D\vec{b}_i$ in Γ liegt, gibt es in Γ auf jeden Fall für jedes i Vektoren der Form $\mu_{i0}\vec{b}_0 + \cdots + \mu_{ii}\vec{b}_i$ mit $\mu_{ii} \neq 0$. $\vec{c}_i$ sei ein solcher Vektor mit minimalem $|\mu_{ii}|$. Wir wollen zeigen, daß diese Vektoren $\vec{c}_i$ 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 $\vec{c}_i$ schreiben läßt.

Angenommen, es gibt Vektoren $\vec{v} \in \Gamma$, für die das nicht der Fall ist. Da $\vec{v}$ auch in Δ liegt, gibt es auf jeden Fall eine Darstellung

$\vec{v} = \lambda_0\vec{b}_0 + \cdots + \lambda_k\vec{b}_k$ mit ganzen Zahlen λ_i und einem $k \leq m$. Wir wählen einen solchen Vektor $\vec{v}$ 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 auch der Vektor

$$\vec{v} - q\vec{c}_k = (\lambda_0 - q\mu_{k0})\vec{b}_0 + \cdots + (\lambda_k - z\mu_{kk})\vec{b}_k$$

nicht als ganzzahlige Linearkombination der $\vec{c}_i$ dargestellt werden, denn sonst hätte auch $\vec{v}$ 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 $\vec{v}$ als Vektor mit minimalem $|\mu_{kk}|$. Somit kann es keinen Gittervektor aus Γ geben, der nicht als ganzzahlige Linearkombination der $\vec{c}_i$ 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

$$\begin{cases} V \times V \to V \\ (\vec{v}, \vec{w}) \mapsto \vec{v} \cdot \vec{w} \end{cases}$$

mit folgenden Eigenschaften:

1.) $(\lambda\vec{u} + \mu\vec{v}) \cdot \vec{w} = \lambda(\vec{u} \cdot \vec{w}) + \mu(\vec{v} \cdot \vec{w})$ für alle $\lambda, \mu \in \mathbb{R}$ und alle $\vec{u}, \vec{v}, \vec{w} \in V$.

2.) $\vec{v} \cdot \vec{w} = \vec{w} \cdot \vec{v}$ für alle $\vec{v}, \vec{w} \in V$.

3.) $\vec{v} \cdot \vec{v} \geq 0$ für alle $\vec{v} \in V$ und $\vec{v} \cdot \vec{v} = 0$ genau dann, wenn $\vec{v} = 0$.

Die Abbildung $V \times V \to V$ wird als Skalarprodukt bezeichnet.

b) Zwei Vektoren $\vec{v}, \vec{w} \in V$ heißen orthogonal, wenn $\vec{v} \cdot \vec{w} = 0$ ist.

c) Eine Basis $(\vec{c}_0, \ldots, \vec{c}_m)$ eines EUKLIDischen Vektorraums heißt *Orthogonalbasis*, wenn $\vec{c}_i \cdot \vec{c}_j = 0$ für alle $i \neq j$.

Wichtigstes Beispiel ist der Vektorraum $\mathbb{R}^{m+1}$ mit seinem Standardskalarprodukt

$$\begin{pmatrix} v_0 \\ \vdots \\ v_m \end{pmatrix} \cdot \begin{pmatrix} w_0 \\ \vdots \\ w_m \end{pmatrix} = \sum_{i=0}^{m} v_i w_i \,.$$

Das Produkt eines Vektors $\vec{v}$ 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 $(\vec{b}_0, \ldots, \vec{b}_m)$ des $\mathbb{R}^{m+1}$ schrittweise eine Orthogonalbasis $(\vec{c}_0, \ldots, \vec{c}_m)$, und zwar so, daß in jedem Schritt der von den Vektoren $\vec{b}_0, \ldots, \vec{b}_r$ erzeugte Untervektorraum gleich dem von $\vec{c}_0, \ldots, \vec{c}_r$ erzeugten ist.

Der erste Schritt ist der einfachste: Da es noch keine Orthogonalitätsbedingung für $\vec{c}_0$ gibt, können wir einfach $\vec{c}_0 = \vec{b}_0$ setzen.

Nachdem wir $r \geq 1$ Schritte durchgeführt haben, haben wir r linear unabhängige Vektoren $\vec{c}_0, \ldots, \vec{c}_{r-1}$ mit $\vec{c}_i \cdot \vec{c}_j = 0$ für $i \neq j$ aus dem von $\vec{b}_0, \ldots, \vec{b}_r$ aufgespannten Untervektorraum. Ist $r = m$, haben wir eine Orthogonalbasis; andernfalls muß ein auf den bisher konstruierten $\vec{c}_i$ senkrecht stehender Vektor $\vec{c}_r$ gefunden werden, der zusammen mit diesen den von $\vec{b}_0$ bis $\vec{b}_r$ erzeugten Untervektorraum erzeugt.

Da $\vec{c}_0, \ldots, \vec{c}_{r-1}$ und $\vec{b}_0, \ldots, \vec{b}_{r-1}$ denselben Untervektorraum erzeugen, gilt dasselbe für $\vec{c}_0, \ldots, \vec{c}_{r-1}, \vec{b}_r$ und $\vec{b}_0, \ldots, \vec{b}_r$; das Problem ist, daß $\vec{b}_r$ im allgemeinen nicht orthogonal zu den $\vec{c}_i$ sein wird. Wir dürfen $\vec{b}_r$ aber abändern um einen beliebigen Vektor aus dem von $\vec{c}_0, \ldots, \vec{c}_{r-1}$ aufgespannten Untervektorraum; also setzen wir

$$\vec{c}_r = \vec{b}_r + \lambda_0 \vec{c}_1 - \cdots - \lambda_{r-1} \vec{c}_{r-1}$$

und versuchen, die λ_i so zu bestimmen, daß dieser Vektor orthogonal zu $\vec{c}_0, \ldots, \vec{c}_{r-1}$ wird.

Wegen der Orthogonalität der $\vec{c}_i$ ist

$$\vec{c}_r \cdot \vec{c}_i = \vec{b}_r \cdot \vec{c}_i - \sum_{j=0}^{r-1} \lambda_j (\vec{c}_j \cdot \vec{c}_i) = \vec{b}_r \cdot \vec{c}_i - \lambda_i (\vec{c}_i \cdot \vec{c}_i)\,;$$

setzen wir daher

$$\lambda_i = \frac{\vec{b}_{r+1} \cdot \vec{c}_i}{\vec{c}_i \cdot \vec{c}_i}\,,$$

so ist $\vec{v} \cdot \vec{c}_i = 0$ für alle $i = 0, \ldots, r-1$.

Nach dem $m+1$-ten Schritt haben wir eine Orthogonalbasis $(\vec{c}_0, \ldots, \vec{c}_m)$ 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

$$\vec{b}_0 = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 4 \end{pmatrix}, \quad \vec{b}_1 = \begin{pmatrix} -3 \\ 4 \\ 0 \\ 5 \end{pmatrix} \text{ und } \vec{b}_2 = \begin{pmatrix} -1 \\ -2 \\ 3 \\ 6 \end{pmatrix}$$

aufgespannten Untervektorraums U von $\mathbb{R}^4$ bestimmen.

Als ersten Vektor der Orthogonalbasis wählen wir einfach $\vec{c}_0 = \vec{b}_0$.

Für den zweiten Vektor machen wir den Ansatz $\vec{c}_1 = \vec{b}_1 - \lambda\vec{c}_0$, wobei λ so gewählt werden muß, daß $\vec{c}_1 \cdot \vec{c}_0 = \vec{b}_1 \cdot \vec{c}_0 - \lambda\vec{c}_0 \cdot \vec{c}_0 = 0$ ist. Da

$$\vec{c}_0 \cdot \vec{b}_1 = -3 + 2 \cdot 4 + 4 \cdot 5 = 25 \quad \text{und} \quad \vec{c}_0 \cdot \vec{c}_0 = 1^2 + 2^2 + 2^2 + 4^2 = 25\,,$$

müssen wir $\lambda = 1$ setzen und $\vec{c}_1 = \vec{b}_1 - \vec{c}_0 = \begin{pmatrix} -4 \\ 2 \\ -2 \\ 1 \end{pmatrix}$.

Für den noch fehlenden dritten Vektor der Orthogonalbasis ist der Ansatz entsprechend:

$$\vec{c}_2 = \vec{b}_2 - \lambda\vec{c}_0 - \mu\vec{c}_1 \quad \text{mit} \quad \vec{c}_2 \cdot \vec{c}_0 = \vec{c}_2 \cdot \vec{c}_1 = 0\,.$$

$$\vec{c}_2 \cdot \vec{c}_0 = \vec{b}_2 \cdot \vec{c}_0 - \lambda\vec{c}_0 \cdot \vec{c}_0 = (-1 - 4 + 6 + 24) + 25\lambda \Longrightarrow \lambda = 1$$

$$\vec{c}_2 \cdot \vec{c}_1 = \vec{b}_2 \cdot \vec{c}_1 + \mu\vec{c}_1 \cdot \vec{c}_1 = (4 - 4 - 6 + 6) - 25\mu \Longrightarrow \mu = 0$$

$$\vec{c}_2 = \vec{b}_2 - \vec{c}_0 = \begin{pmatrix} -2 \\ -4 \\ 1 \\ 2 \end{pmatrix}.$$

Unsere Orthogonalbasis besteht also aus den drei Vektoren

$$\vec{c}_0 = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 4 \end{pmatrix}, \qquad \vec{c}_1 = \begin{pmatrix} -4 \\ 2 \\ -2 \\ 1 \end{pmatrix} \quad \text{und} \quad \vec{c}_2 = \begin{pmatrix} -2 \\ -4 \\ 1 \\ 2 \end{pmatrix}.$$

Wenn wir den Zusammenhang zwischen der Ausgangsbasis $(\vec{b}_0, \ldots, \vec{b}_m)$ und der Orthogonalbasis $(\vec{c}_0, \ldots, \vec{c}_m)$ explizit festhalten wollen, müssen wir den oben berechneten Koeffizienten Namen geben, die auch vom Schritt abhängen. Wir schreiben

$$\vec{c}_i = \vec{b}_i - \sum_{j=0}^{i-1} \mu_{ij} \vec{c}_j \quad \text{für } i = 0, \ldots, m \quad \text{mit} \quad \mu_{ij} = \frac{\vec{b}_i \cdot \vec{c}_j}{\vec{c}_j \cdot \vec{c}_j} \,.$$

Im gerade durchgerechneten Beispiel etwa ist

$$\vec{c}_0 = \vec{b}_0, \quad \vec{c}_1 = \vec{b}_1 - \vec{c}_0 \quad \text{und} \quad \vec{c}_2 = \vec{b}_2 - \vec{c}_0 \,,$$

also $\mu_{10} = \mu_{20} = 1$ und $\mu_{21} = 0$.

Lösen wir die obigen Formeln auf nach $\vec{b}_i$, kommen wir auf

$$\vec{b}_i = \vec{c}_i + \sum_{j=0}^{i-1} \mu_{ij} \vec{c}_j \quad \text{und} \quad \vec{c}_i \cdot \sum_{j=0}^{i-1} \mu_{ij} \vec{c}_j = \sum_{j=0}^{i-1} \mu_{ij} \vec{c}_i \cdot \vec{c}_j = 0 \,.$$

Geometrisch bedeutet dies, daß $\vec{c}_i$ der Lotvektor bei der Projektion von $\vec{b}_i$ auf den von $\vec{c}_1, \ldots, \vec{c}_{i-1}$ aufgespannten Untervektorraum ist oder, anders ausgedrückt, die orthogonale Projektion von $\vec{b}_i$ auf das orthogonale Komplement dieses Raums.

Ist allgemein $\vec{w} = \vec{u} + \vec{v}$ die Summe zweier aufeinander senkrecht stehenden Vektoren, so ist

$$\vec{w} \cdot \vec{w} = (\vec{u} + \vec{v}) \cdot (\vec{u} + \vec{v}) = \vec{u} \cdot \vec{u} + \vec{v} \cdot \vec{v} \,,$$

denn $\vec{v} \cdot \vec{w} = 0$. Insbesondere sind daher die Längen von $\vec{v}$ und $\vec{w}$ höchstens gleich der Länge von $\vec{w}$. In unserer Situation bedeutet dies, daß

$$|\vec{c}_i| \leq \left|\vec{b}_i\right| \quad \text{für} \quad i = 0, \ldots, m \,,$$

kein Vektor der Orthogonalbasis kann also länger sein als der entsprechende Vektor der Ausgangsbasis.

Im Falle einer Gitterbasis $(\vec{b}_0, \ldots, \vec{b}_m)$ 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 $\vec{c}_j$ 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 $(\vec{u}, \vec{v})$, so gäbe es ganze Zahlen a, b, c, d, so daß

$$\vec{u} = a\begin{pmatrix}1\\0\end{pmatrix} + b\begin{pmatrix}\sqrt{2}\\1\end{pmatrix} = \begin{pmatrix}a + b\sqrt{2}\\b\end{pmatrix}$$

und

$$\vec{v} = c\begin{pmatrix}1\\0\end{pmatrix} + d\begin{pmatrix}\sqrt{2}\\1\end{pmatrix} = \begin{pmatrix}c + d\sqrt{2}\\d\end{pmatrix}$$

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 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 $\vec{c}_i$ keine Gitterbasis bilden, ist daher die Determinante des Gitters auch gleich dem Betrag der Determinante der Matrix C mit den $\vec{c}_i$ als Spaltenvektoren. Das ist aber einfach das Produkt der Längen der $\vec{c}_i$, denn der ij-Eintrag von $C^T C$ ist $\vec{c}_i \cdot \vec{c}_j$. Da die $\vec{c}_i$ 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 $\vec{c}_i \cdot \vec{c}_i$. 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 $\vec{b}_i$ 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 $(\vec{b}_0, \ldots, \vec{b}_m)$. Ist $(\vec{c}_0, \ldots, \vec{c}_m)$ die zugehörige Orthogonalbasis, so ist die Determinante des Gitters das Produkt der Längen der Vektoren $\vec{c}_i$. Wie wir oben gesehen haben, kann kein Vektor $\vec{c}_i$ länger sein als der entsprechende Vektor $\vec{b}_i$, und damit folgt die

Ungleichung von Hadamard: Im Gitter $\Gamma = \mathbb{Z}\vec{b}_0 \oplus \cdots \oplus \mathbb{Z}\vec{b}_m$ ist

$$d(\Gamma) \leq \prod_{i=0}^{m} |\vec{b}_i| \qquad \blacksquare$$

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 POINCARÉ 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.

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.

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

$$a^* h + b^* w = \frac{w}{g}\left(1 - (pc)^p\right) \equiv \frac{w}{g} \bmod p^k \quad \text{mit} \quad a^*, b^* \in \mathbb{Z}[x] .$$

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.

§10: 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 Numerierung nicht wie im vorigen Paragraphen bei der für Polynome sinnvollen Null beginnen, sondern bei eins.

Wir gehen daher aus von einem Gitter $\Gamma = \mathbb{Z}\vec{b}_1 \oplus \cdots \oplus \mathbb{Z}\vec{b}_n \leq \mathbb{R}^n$ und wollen dort nach kurzen Vektoren suchen.

Falls die Gitterbasis $\vec{b}_1, \ldots, \vec{b}_n$ eine Orthogonalbasis des $\mathbb{R}^n$ ist, hat ein

Vektor $\vec{v} = a_1\vec{b}_1 + \cdots + a_n\vec{b}_n$ aus Γ die Länge

$$|\vec{v}| = \sqrt{a_1^2\ \vec{b}_1 \cdot \vec{b}_1 + \cdots + a_n^2\ \vec{b}_n \cdot \vec{b}_n}\,;$$

wenn wir zusätzlich noch annehmen, daß $\left|\vec{b}_1\right| \leq \cdots \leq \left|\vec{b}_n\right|$ ist, sind daher $\pm\vec{b}_1$ kürzeste Vektoren in Γ. Je nach Länge von $\vec{b}_2$ gilt eventuell dasselbe auch für $\pm\vec{b}_2$; falls $\vec{b}_2$ aber länger ist als $\vec{b}_1$, müssen wir auf der Suche nach zweitkürzesten Vektoren die Längen von $\vec{b}_2$ und $2\vec{b}_1$ 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

$$\vec{c}_i = \vec{b}_i - \sum_{j=1}^{i-1} \mu_{ij}\vec{c}_j \quad \text{mit} \quad \mu_{ij} = \frac{\vec{b}_i \cdot \vec{c}_j}{\vec{c}_j \cdot \vec{c}_j}\,,$$

wobei die μ_{ij} aber im allgemeinen keine ganzen, sondern nur rationale Zahlen sind.

Wenn der Vektor $\vec{c}_i$ 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:

$$\left|\vec{c}_i + \mu_{i\,i-1}\vec{c}_{i-1}\right|^2 \geq \tfrac{3}{4}\left|\vec{c}_{i-1}\right|^2 \quad \text{für alle } i > 1\,.$$

Um diese sogenannte LOVÁSZ-Bedingung zu verstehen, multiplizieren wir beiden Seiten aus. Wegen der Orthogonalität der $\vec{c}_j$ ist

$$|\vec{c}_i|^2 + \mu_{i\,i-1}^2\left|\vec{c}_{i-1}\right|^2 \geq \tfrac{3}{4}\left|\vec{c}_{i-1}\right|^2 \quad \text{oder} \quad |\vec{c}_i|^2 \geq \left(\tfrac{3}{4} - \mu_{i\,i-1}^2\right)\left|\vec{c}_{i-1}\right|^2\,.$$

Da die Beträge der μ_{ij} höchstens $\frac{1}{2}$ sind, folgt insbesondere

$$|\vec{c}_i|^2 \geq \tfrac{1}{2}\left|\vec{c}_{i-1}\right|^2 \quad \text{oder} \quad \left|\vec{c}_{i-1}\right|^2 \leq 2\,|\vec{c}_i|^2\,.$$

Diese Bedingung sorgt also dafür, daß sich die Längen der $\vec{c}_i$ 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 $\vec{b}_1, \ldots, \vec{b}_n$ mit GRAM-SCHMIDT-Orthogonalisierung

$$\vec{c}_i = \vec{b}_i - \sum_{j=1}^{i-1} \mu_{ij}\vec{c}_j$$

heißt LLL-reduziert, wenn

$$\left|\mu_{ij}\right| \leq \frac{1}{2} \quad \text{für alle } i, j \qquad \text{und}$$

$$\left|\vec{c}_i + \mu_{i\,i-1}\vec{c}_{i-1}\right|^2 \geq \tfrac{3}{4}\left|\vec{c}_{i-1}\right|^2 \quad \text{für alle } i > 1\,.$$

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 $(\vec{b}_1, \ldots, \vec{b}_n)$ 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

$$\vec{c}_i = \vec{b}_i - \sum_{j=1}^{i-1} \mu_{ij} \vec{c}_j \in \mathbb{R}^n \quad \text{mit} \quad \mu_{ij} = \frac{\vec{b}_i \cdot \vec{c}_j}{\vec{c}_j \cdot \vec{c}_j} \,. \qquad (*)$$

Im Laufe des Algorithmus werden die $\vec{b}_i, \vec{c}_j$ 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

$$\left|\vec{c}_i + \mu_{i\,i-1}\vec{c}_{i-1}\right|^2 \geq \tfrac{3}{4}\left|\vec{c}_{i-1}\right|^2 \quad \text{für} \quad 1 < i \leq k \qquad (B_k)$$

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 $(\vec{b}_1, \ldots, \vec{b}_n)$ 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 $\vec{b}_{k+1}$ durch $\vec{b}_{k+1} - q\vec{b}_k$. 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 $|\vec{c}_{k+1} + \mu_{k+1\,k}\vec{c}_k|^2 < \frac{3}{4}\,|c_k|^2$

In diesem Fall vertauschen wir $\vec{b}_k$ und $\vec{b}_{k+1}$. Da die GRAM-SCHMIDT-Orthogonalisierung von der Reihenfolge der Basisvektoren abhängt, müssen wir dann eine neue Orthogonalbasis $(\vec{d}_1, \ldots, \vec{d}_n)$ berechnen.

An den $\vec{c}_j$ mit $j < k$ (so es welche gibt) ändert sich dabei nichts: Sie werden beim GRAM-SCHMIDTschen Orthogonalisierungsverfahren berechnet, bevor die Vektoren $\vec{b}_k$ und $\vec{b}_{k+1}$ ins Spiel kommen. Für $j < k$ ist somit $\vec{d}_j = \vec{c}_j$.

Auch für $j > k+1$ ist $\vec{d}_j = \vec{c}_j$, 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

$$[\vec{c}_1, \ldots, \vec{c}_j] = [\vec{b}_1, \ldots, \vec{b}_j] = [\vec{d}_1, \ldots, \vec{d}_j]\,.$$

Bleiben noch die Vektoren $\vec{d}_k$ und $\vec{d}_{k+1}$. Die müssen verschieden sein von den Vektoren $\vec{c}_k$ und $\vec{d}_{k+1}$, denn $[\vec{c}_1, \ldots, \vec{c}_k] = [\vec{b}_1, \ldots, \vec{b}_k]$, aber $[\vec{d}_1, \ldots, \vec{d}_k] = [\vec{b}_1, \ldots, \vec{b}_{k-1}, \vec{b}_{k+1}]$.

Nach den Formeln zur GRAM-SCHMIDT-Orthogonalisierung ist

$$vc_k = \vec{b}_k - \sum_{j=1}^{k-1} \mu_{kj}\vec{c}_j \qquad \text{mit} \quad \mu_{kj} = \frac{\vec{b}_k \cdot \vec{c}_j}{\vec{c}_j \cdot \vec{c}_j}$$

und

$$\vec{c}_{k+1} = \vec{b}_{k+1} - \sum_{j=1}^{k} \mu_{k+1\,j}\vec{c}_j \qquad \text{mit} \quad \mu_{k+1\,j} = \frac{\vec{b}_{k+1}\cdot\vec{c}_j}{\vec{c}_j\cdot\vec{c}_j}$$

$$= \vec{b}_{k+1} - \sum_{j=1}^{k-1} \mu_{k+1\,j}\vec{c}_j - \mu_{k+1\,k}\vec{b}_k + \sum_{j=1}^{k-1} \mu_{k+1\,k}\mu_{kj}\vec{c}_j$$

$$= \vec{b}_{k+1} - \mu_{k+1\,k}\vec{b}_k - \sum_{j=1}^{k-1}(\mu_{k+1\,j} - \mu_{k+1\,k}\mu_{kj})\vec{c}_j\ ;$$

entsprechend ist

$$\vec{d}_k = \vec{b}_{k+1} - \sum_{j=1}^{k-1} \nu_{kj}\vec{d}_j \qquad \text{mit} \quad \nu_{kj} = \frac{\vec{b}_{k+1}\cdot\vec{d}_j}{\vec{d}_j\cdot\vec{d}_j}$$

und

$$\vec{d}_{k+1} = \vec{b}_k - \sum_{j=1}^{k} \nu_{k+1\,j}\vec{d}_j \qquad \text{mit} \quad \nu_{k+1\,j} = \frac{\vec{b}_k\cdot\vec{d}_j}{\vec{d}_j\cdot\vec{d}_j}$$

$$= \vec{b}_k - \sum_{j=1}^{k-1} \nu_{k+1\,j}\vec{c}_j - \nu_{k+1\,k}\left(\vec{b}_{k+1} - \sum_{j=1}^{k-1}\nu_{kj}\vec{c}_j\right)\ .$$

Da $\vec{c}_j = \vec{d}_j$ für $j < k$, folgt insbesondere

$$\nu_{kj} = \mu_{k+1\,j} \quad \text{und} \quad \nu_{k+1\,j} = \mu_{kj} \qquad \text{für } j < k$$

$$\vec{d}_{k+1} = \vec{c}_k - \nu_{k+1\,k}\vec{d}_k \qquad \text{und}$$

$$\vec{d}_k = \vec{b}_{k+1} - \sum_{j=1}^{k-1} \nu_{kj}\vec{d}_j = \vec{b}_{k+1} - \sum_{j=1}^{k-1} \mu_{k+1\,j}\vec{c}_j = \vec{c}_{k+1} + \mu_{k+1\,k}\vec{c}_k\ .$$

Nach der Voraussetzung für das Eintreten von Fall 1 ist das Längenquadrat des letzten Vektors kleiner als $\frac{3}{4}$ des Längenquadrats von $\vec{c}_k$; zumindest ein Vektor der Orthogonalbasis wird also durch die Vertauschung von $\vec{b}_k$ und $\vec{b}_{k+1}$ kürzer. Das Quadrat seiner Länge ist

$$\vec{d}_k\cdot\vec{d}_k = \vec{c}_{k+1}\cdot\vec{c}_{k+1} + \mu_{k+1\,k}^2\,\vec{c}_k\cdot\vec{c}_k\ .$$

Damit können wir nun auch $\nu_{k+1\,k}$ durch Daten der „alten“ Basis ausdrücken: Der Zähler ist

$$\vec{b}_k \cdot \vec{d}_k = \big(\vec{c}_k + \sum_{j=1}^{k-1} \mu_{kj}\vec{c}_j\big) \cdot \vec{d}_k = \vec{c}_k \cdot \vec{d}_k \,,$$

denn für $j \le k-1$ steht $\vec{d}_k$ senkrecht auf $\vec{d}_j = \vec{c}_j$. Deshalb ist auch

$$\vec{c}_k \cdot \vec{d}_k = \vec{c}_k \cdot \big(\vec{b}_{k+1} - \sum_{j=1}^{k-1} \nu_{kj}\vec{d}_j\big) = \vec{c}_k\vec{b}_{k+1} \,,$$

also ist

$$\nu_{k+1\,k} = \frac{\vec{b}_k \cdot \vec{d}_k}{\vec{d}_k \cdot \vec{d}_k} = \frac{\vec{c}_k \cdot \vec{b}_{k+1}}{\vec{d}_k \cdot \vec{d}_k} = \frac{|\vec{c}_k|^2}{\big|\vec{d}_k\big|^2} \cdot \frac{\vec{c}_k \cdot \vec{b}_{k+1}}{\vec{d}_k \cdot \vec{d}_k} = \frac{|\vec{c}_k|^2}{\big|\vec{d}_k\big|^2} \cdot \mu_{k+1\,k}$$

$$= \frac{\mu_{k+1\,k}\,|\vec{c}_k|^2}{|c_{k+1}|^2 + \mu_{k+1\,k}^2\,|\vec{c}_k|^2} \,.$$

Dank dieser Formeln ist daher auch $\vec{d}_{k+1} = \vec{c}_k - \nu_{k+1\,k}\vec{d}_k$ vollständig durch Daten der „alten“ Basis ausdrückbar.

Die Vektoren $\vec{d}_j$ mit $j > k+1$ sind wieder gleich den entsprechenden $\vec{c}_j$, denn der j-te Vektor der Orthogonalbasis ist ja der Lotvektor von $\vec{b}_j$ auf den von $\vec{b}_1, \ldots, \vec{b}_{j-1}$ aufgespannten Untervektorraum, und für $j > k+1$ hat sich durch die Vertauschung von $\vec{b}_k$ und $\vec{b}_{k+1}$ weder an $\vec{b}_j$ 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 $\vec{c}_k$ und $\vec{c}_{k+1}$ aus durch $\vec{d}_k$ und $\vec{d}_{k+1}$: Nach den obigen Formeln ist

$$\vec{d}_k = \vec{c}_{k+1} + \mu_{k+1\,k}\vec{c}_k$$

$$\vec{d}_{k+1} = \vec{c}_k - \nu_{k+1\,k}\vec{d}_k = (1 - \nu_{k+1\,k}\mu_{k+1\,k})\vec{c}_k - \nu_{k+1\,k}\vec{c}_{k+1} \,.$$

Addition von $\nu_{k+1\,k}$-mal der ersten Gleichung zur zweiten eliminiert $\vec{c}_{k+1}$ und liefert uns die Gleichung

$$\vec{c}_k = \nu_{k+1\,k}\vec{d}_k + \vec{d}_{k+1} \,.$$

Setzen wir dies ein in die zweite Gleichung, erhalten wir

$$(1 - \nu_{k+1\,k}\mu_{k+1\,k})(\nu_{k+1\,k}\vec{d}_k + \vec{d}_{k+1}) - \nu\vec{c}_{k+1} = \vec{d}_{k+1}$$

und damit

$$\vec{c}_{k+1} = (1 - \nu_{k+1\,k}\mu_{k+1\,k})\vec{d}_k - \mu_{k+1\,k}\vec{d}_{k+1}\,.$$

Da in der Summe $\vec{d}_k = \vec{c}_{k+1} + \mu_{k+1\,k}\vec{c}_k$ die Vektoren $\vec{c}_k$ und $\vec{c}_{k+1}$ orthogonal sind, ist $\left|\vec{d}_k\right|^2 = |\vec{c}_{k+1}|^2 + \mu_{k+1\,k}^2\,|\vec{c}_k|^2$, also

$$\frac{|\vec{c}_{k+1}|^2}{\left|\vec{d}_k\right|^2} = \frac{\left|\vec{d}_k\right|^2 - \mu_{k+1\,k}^2\,|\vec{c}_k|^2}{\left|\vec{d}_k\right|^2} = 1 - \frac{|\vec{c}_k|^2}{\left|\vec{d}_k\right|^2}\mu_{k+1\,k}^2 = 1 - \nu_{k+1\,k}\mu_{k+1\,k}\,.$$

Somit ist

$$\vec{c}_{k+1} = \frac{|\vec{c}_{k+1}|^2}{\left|\vec{d}_{k+1}\right|^2}\vec{d}_k - \mu_{k+1\,k}\vec{d}_{k+1}\,.$$

Mit diesen beiden Formel gehen wir nun für $i > k+1$ in die Gleichungen

$$\vec{b}_i = \vec{c}_i + \sum_{j=1}^{i-1}\mu_{ij}\vec{c}_j\,,$$

Die Teilsumme $\mu_{ik}\vec{c}_k + \mu_{i\,k+1}\vec{c}_{k+1}$ aus den Termen für $j = k$ und $j = k+1$ wird dabei zu

$$\left(\mu_{ik}\nu_{k+1\,k} + \mu_{i\,k+1}\frac{|\vec{c}_{k+1}|^2}{\left|\vec{d}_k\right|^2}\right)\vec{d}_k + \left(\mu_{ik} - \mu_{i\,k+1}\mu_{k+1\,k}\right)\vec{d}_{k+1}\,.$$

Somit ist

$$\nu_{ik} = \mu_{ik}\nu_{k+1\,k} + \mu_{i\,k+1}\frac{|\vec{c}_{k+1}|^2}{\left|\vec{d}_k\right|^2} \quad \text{und} \quad \nu_{i\,k+1} = \mu_{ik} - \mu_{i\,k+1}\mu_{k+1\,k}\,.$$

Wir ersetzen nun alle $\vec{c}_i$ durch die entsprechenden $\vec{d}_i$ und alle μ_{ij} durch die entsprechenden ν_{ij}; dann ist $(*)$ auch für die Gitterbasis mit vertauschten Positionen von $\vec{b}_k$ und $\vec{b}_{k+1}$ 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 $|\vec{c}_{k+1} + \mu_{k+1\,k}\vec{c}_k|^2 \geq \frac{3}{4}\,|\vec{c}_k|^2$

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 $\vec{b}_{k+1}$ durch $\vec{b}_{k+1} - q\vec{b}_\ell$ 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 B_k mit ij-Eintrag $\vec{b}_i \cdot \vec{b}_j$. Für $k = n$ können wir sie leicht auf bekannte Größen zurückführen: Ist B die $n \times n$-Matrix mit dem Basisvektor $\vec{b}_i$ als i-ter Spalte, so ist offensichtlich $B_n = B^T B$, also ist $d_n = (\det B)^2 = d(\Gamma)^2$. Insbesondere ist also d_n unabhängig von der Gitterbasis und hängt nur ab vom Gitter.

Entsprechend können wir für d_k das Gitter $\Gamma_k = \mathbb{Z}\vec{b}_1 \oplus \cdots \oplus \mathbb{Z}\vec{b}_k$ im Vektorraum $\mathbb{R}\vec{b}_1 \oplus \cdots \oplus \mathbb{R}\vec{b}_k$ 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 $\vec{b}_1, \ldots, \vec{b}_k$.

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 $\vec{b}_k$ und $\vec{b}_{k+1}$ 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 $\vec{v} \neq \vec{0}$, 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 $d(\Gamma_i)$ gefunden, also auch für d_i und damit auch für das Produkt D aller d_i. 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:

$(\vec{b}_1, \ldots, \vec{b}_n)$ sei eine Gitterbasis und B sei die Matrix mit den $\vec{b}_i$ als Spaltenvektoren. Dann ist

$$\varphi\colon \begin{cases} \mathbb{R}^n \to \mathbb{R}^n \\ \quad \vec{v} \mapsto B\vec{v} \end{cases}$$

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 auch 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 $\vec{v} \in \mathbb{Z}^n$ sein $W_{\vec{v}} = W + \vec{v}$ der um $\vec{v}$ verschobene Würfel W. Dann sind offensichtlich alle $W_{\vec{v}}$ disjunkt und überdecken gemeinsam den $\mathbb{R}^n$. Da B beschränkt ist, können nur endlich viele der $W_{\vec{v}}$ einen nichtleeren Durchschnitt $B_{\vec{v}}$ mit B haben. Für jeden dieser Durchschnitte betrachten wir seine Verschiebung $B_{\vec{v}} - \vec{v}$ um den Vektor $-\vec{v}$; 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 $B_{\vec{v}}$. Damit ist diese Summe nach Voraussetzung größer als eins; da W das Volumen eins hat, muß es also zwei Vektoren $\vec{v} \neq \vec{w}$ geben, so daß $B_{\vec{v}} \cap B_{\vec{w}}$ nicht leer ist. Für einen Punkt R aus diesem Durchschnitt sei P seine Translation um $\vec{v}$ und Q die um $\vec{w}$. Dann liegen $P \in B_{\vec{v}}$ und $Q \in B_{\vec{w}}$ beide in B, und ihr Verbindungsvektor ist $\vec{w} - \vec{v} \in \mathbb{Z}^n$.

■

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 Beispiel für die LLL-Reduktion wollen wir eine LLL-reduzierte Basis des von

$$\vec{b}_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \vec{b}_2 = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} \quad \text{und} \quad \vec{b}_3 = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$$

erzeugten Gitters $\Gamma \subset \mathbb{R}^3$ bestimmen.

Als erstes brauchen wir die zugehörige Orthogonalbasis. Nach GRAM-SCHMIDT setzen wir $\vec{c}_1 = \vec{b}_1$ und wählen dann μ_{21} so, daß

$$(\vec{b}_2 - \mu_{21}\vec{c}_1) \cdot \vec{c}_1 = 10 - 14\mu_{21} = 0$$

ist, d.h.

$$\mu_{21} = \frac{5}{7} \quad \text{und} \quad \vec{c}_2 = \frac{1}{7}\begin{pmatrix} 16 \\ 4 \\ -8 \end{pmatrix}.$$

Der dritte Vektor $\vec{c}_3 = \vec{b}_3 - \mu_{31}\vec{c}_1 - \mu_{32}\vec{c}_3$ wird so gewählt, daß

$$\vec{c}_3 \cdot \vec{c}_1 = 11 - 14\mu_{21} = 0 \quad \text{und} \quad \vec{c}_3 \cdot \vec{c}_2 = \frac{36}{7} - \frac{48}{7}\mu_{32} = 0,$$

wir haben also

$$\mu_{31} = \frac{11}{14}, \quad \mu_{32} = \frac{3}{4} \quad \text{und} \quad \vec{c}_3 = \frac{1}{2}\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix} .$$

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 $\vec{b}_2$ durch

$$\vec{b}_2 - \vec{b}_1 = \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix}$$

und μ_{21} durch $\mu_{21} - 1 = -\frac{2}{7}$.

$$\vec{c}_2 + \mu_{21}\vec{c}_1 = \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix}$$

hat Längenquadrat acht, was kleiner ist als $\frac{3}{4}\left|\vec{c}_1\right|^2 = \frac{21}{2}$. Daher sind wir in Fall 1 und müssen $\vec{b}_1$ und $\vec{b}_2$ vertauschen. Der neue erste Vektor der Orthogonalbasis ist der gerade berechnete Vektor $\vec{d}_1 = \vec{c}_2 + \mu_{21}\vec{c}_1$ und

$$\nu_{21} = \mu_{21}\frac{\left|\vec{c}_1\right|^2}{\left|\vec{c}_2\right|^2 + \mu_{21}\left|\vec{c}_1\right|^2} = -\frac{1}{2} .$$

Damit ist

$$\vec{d}_2 = \vec{c}_1 - \nu_{21}\vec{d}_1 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} ;$$

die neuen Koeffizienten sind

$$\nu_{31} = \mu_{31}\nu_{21} + \mu_{32}\frac{\left|\vec{c}_2\right|^2}{\left|\vec{d}_1\right|^2} = \frac{1}{4} \quad \text{und} \quad \nu_{32} = \mu_{31} - \mu_{32}\mu_{21} = 1 .$$

Wir ersetzen die $\vec{c}_i$ durch die $\vec{d}_i$ 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.

$$\vec{c}_2 + \mu_{21}\vec{c}_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

hat Längenquadrat 14, was größer ist als $\frac{3}{4}\,|\vec{c}_1|^2$; 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 $\vec{b}_3$ durch

$$\vec{b}_3 - \vec{b}_2 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}\,.$$

Dann hat

$$\vec{c}_3 + \mu_{32}\vec{c}_2 = \frac{1}{2}\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

Längenquadrat $\frac{3}{2}$, während $\frac{3}{4}\,|\vec{c}_2|^2 = 9$ ist, wir sind also wieder im ersten Fall und müssen $\vec{b}_2$ mit $\vec{b}_3$ vertauschen. Neuer zweiter Vektor der Orthogonalbasis wird

$$\vec{d}_2 = \vec{c}_3 + \mu_{32}\vec{c}_2 = \frac{1}{2}\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

und

$$\nu_{32} = \mu_{32}\frac{|\vec{c}_2|^2}{|\vec{c}_3|^2 + \mu_{32}^2\,|\vec{c}_2|^2} = 0, \quad \nu_{21} = \mu_{31} = \frac{3}{4}, \quad \nu_{31} = \mu_{21} = -\frac{1}{2}\,.$$

Damit können wir nun auch den dritten Vektor der neuen Orthogonalbasis berechnen als

$$\vec{d}_3 = \vec{c}_2 - \nu_{32}\vec{d}_2 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}\,.$$

Wir ersetzen $\vec{c}_2, \vec{c}_3$ durch $\vec{d}_2, \vec{d}_3$ 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

$$\vec{b}_2 \leftarrow \vec{b}_2 - \vec{b}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \mu_{21} \leftarrow \mu_{21} - 1 = -\frac{1}{4}\,.$$

Um zu sehen, in welchem Fall wir sind, müssen wir das Längenquadrat zwei von

$$\vec{c}_2 + \mu_{21}\vec{c}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$$

mit $\frac{3}{4}\left|\vec{c}_1\right|^2 = 6$ vergleichen: Wir sind wieder in Fall 1 und müssen jetzt $\vec{b}_1$ und $\vec{b}_2$ miteinander vertauschen. Neuer erster Vektor der Orthogonalbasis wird $\vec{d}_1 = \vec{c}_2 + \mu_{21}\vec{c}_1$; nach GRAM-SCHMIDT muß das natürlich der neue Vektor $\vec{b}_1$ sein. Weiter ist

$$\nu_{21} = \mu_{21}\frac{\left|\vec{c}_1\right|^2}{\left|\vec{c}_2\right|^2 + \mu_{21}^2\left|\vec{c}_1\right|^2} = -1 \quad \text{und} \quad \vec{d}_2 = \vec{c}_1 - \nu_{21}\vec{d}_1 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}\,.$$

Die restlichen ν_{ij} sind

$$\nu_{31} = \mu_{31}\nu_{21} + \mu_{23}\frac{\left|\vec{c}_2\right|^2}{\left|\vec{d}_1\right|^2} = \frac{1}{2} \quad \text{und} \quad \nu_{32} = \mu_{31} - \mu_{32}\mu_{21} = -\frac{1}{2}\,.$$

Wir ersetzen $\vec{c}_1, \vec{c}_2$ durch $\vec{d}_1, \vec{d}_2$ 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:

$$\vec{b}_2 \leftarrow \vec{b}_2 + \vec{b}_1 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \quad \text{und} \quad \mu_{21} \leftarrow \mu_{21} + 1 = 0\,.$$

Da μ_{21} verschwindet, ist $\vec{c}_2 + \mu_{21}\vec{c}_1 = \vec{c}_2 = \vec{b}_2$, das alte $\vec{b}_1$; sein Längenquadrat ist sechs und damit größer als $\frac{3}{4}\left|\vec{c}_1\right|^2 = \frac{3}{2}$. 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

$$\vec{c}_3 + \mu_{32}\vec{c}_2 = \frac{1}{2}\begin{pmatrix} 3 \\ 3 \\ 6 \end{pmatrix}$$

hat Längenquadrat $\frac{27}{2}$, was größer ist als $\frac{3}{4}\left|\vec{c}_2\right|^2 = \frac{9}{2}$. 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

$$\vec{b}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad \vec{b}_2 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \quad \text{und} \quad \vec{b}_3 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Die zugehörige Orthogonalbasis des $\mathbb{R}^3$ besteht aus $\vec{c}_1 = \vec{b}_1$ und $\vec{c}_2 = \vec{b}_2$ sowie

$$\vec{c}_3 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} = \vec{b}_3 - \frac{1}{2}\vec{c}_1 + \frac{1}{2}\vec{c}_2 .$$

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.

$(\vec{b}_1, \ldots, \vec{b}_n)$ sei eine LLL-reduzierte Basis eines Gitters $\Gamma \subset \mathbb{R}^n$ und $(\vec{c}_1, \ldots, \vec{c}_n)$ sei die dazu nach GRAM-SCHMIDT berechnete Orthogonalbasis. Dann ist

$$\vec{b}_i = \vec{c}_i + \sum_{j=1}^{i-1} \mu_{ij}\vec{c}_j \Longrightarrow \left|\vec{b}_i\right|^2 = \left|\vec{c}_i\right|^2 + \sum_{j=1}^{i-1} \mu_{ij}^2 \left|\vec{c}_j\right|^2 \leq \left|\vec{c}_i\right|^2 + \frac{1}{4}\sum_{j=1}^{i-1} \left|\vec{c}_j\right|^2 ,$$

denn für eine LLL-reduzierte Basis sind alle $\left|\mu_{ij}\right| \leq \frac{1}{2}$. Außerdem ist wegen der LOVÁSZ-Bedingung $\left|\vec{c}_{j-1}\right|^2 \leq 2\left|\vec{c}_j\right|^2$; mehrfache Anwendung dieser Ungleichung führt auf $\left|\vec{c}_j\right|^2 \leq 2^{i-j}\left|\vec{c}_i\right|^2$ für alle $j < i$ und

$$\left|\vec{b}_i\right|^2 \leq \left|\vec{c}_i\right|^2 + \frac{1}{4}\sum_{j=1}^{i-1} 2^{j-i}\left|\vec{c}_i\right|^2 = \left(1 + \frac{1}{4}\sum_{j=1}^{i-1} 2^i\right)\left|\vec{c}_i\right|^2 = \frac{1+2^{i-1}}{2}\left|\vec{c}_i\right|^2$$
$$\leq 2^{i-1}\left|\vec{c}_i\right|^2 .$$

Ebenfalls wegen der LOVÁSZ-Bedingung gilt für $j < i$ die Ungleichung $\left|\vec{c}_j\right|^2 \leq 2^{i-j}\left|\vec{c}_i\right|^2$ und damit auch

$$\left|\vec{b}_j\right|^2 \leq 2^{j-1}\left|\vec{c}_j\right|^2 \leq 2^{j-1}2^{i-j}\left|\vec{c}_i\right|^2 = 2^{i-1}\left|\vec{c}_i\right|^2 .$$

Nun seien $\vec{v}_1, \ldots, \vec{v}_m$ irgendwelche linear unabhängige Vektoren aus dem Gitter Γ und k sei die kleinste Zahl mit der Eigenschaft, daß alle $\vec{v}_i$ im von $\vec{b}_1$ bis $\vec{b}_k$ aufgespannten Teilgitter liegen. Dann gibt es Koeffizienten λ_{ij}, ν_{ij} derart, daß

$$\vec{v}_i = \sum_{j=1}^{k} \lambda_{ij}\vec{b}_j = \sum_{j=1}^{k} \nu_{ij}\vec{c}_j$$

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 $\vec{b}_j$ als Linearkombinationen der $\vec{c}_j$ einsetzen, muß aber $\nu_{ik} = \lambda_{ik}$ und somit ganzzahlig sein, denn außer $\vec{b}_k$ liefert kein anderes $\vec{b}_j$ einen Beitrag mit $\vec{c}_k$.

Wir wählen ein i, für das $\lambda_{ik} = \nu_{ik}$ nicht verschwindet; wegen der Minimalität von k muß es das geben. Dann ist

$$\left|\vec{b}_k\right|^2 \leq 2^{k-1}\left|\vec{c}_k\right|^2 \leq 2^{k-1}\lambda_{ik}^2\left|\vec{c}_k\right|^2 \leq 2^{k-1}\left|\vec{v}_i\right|^2$$

und für $j < k$ ist

$$\left|\vec{b}_j\right|^2 \leq 2^{k-1}\left|\vec{c}_k\right|^2 \leq 2^{k-1}\left|\vec{v}_i\right|^2 .$$

Über die Zahlen i und k wissen wir nur, daß k zwischen m und n liegen muß (sonst wären die $\vec{v}_j$ linear abhängig) und $i \leq m$. Daher haben wir für alle $j \leq m$ die Abschätzung

$$\left|\vec{b}_j\right|^2 \leq 2^{n-1} \max\left\{\left|\vec{v}_1\right|^2, \ldots, \left|\vec{v}_m\right|^2\right\}$$

oder

$$\left|\vec{b}_j\right| \leq 2^{(n-1)/2} \max\left\{\left|\vec{v}_1\right|, \ldots, \left|\vec{v}_m\right|\right\}.$$

Speziell für $m = 1$ erhalten wir die Abschätzung

$$\left|\vec{b}_1\right| \leq 2^{(n-1)/2} \left|\vec{v}_1\right|$$

für jeden vom Nullvektor verschiedenen Gittervektor $\vec{v}_1$. Dies gilt insbesondere für den kürzesten solchen Vektor; die Länge von $\vec{b}_1$ übersteigt dessen Länge also höchstens um den Faktor $2^{(n-1)/2}$.

§11: Anwendung auf Faktorisierungsprobleme

Wie in §9 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 §9 hat f einen bis aufs Vorzeichen eindeutig bestimmten irreduziblen 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 §9 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 sind; 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 $\mathcal{M}$ von $2e$ 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 §9 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 §9 sagt uns, daß h_0 jedes b_j mit $j \in J$ teilt, also auch

$$h_1 = \mathrm{ggT}\big(\{b_j \mid j \in J\}\big) .$$

Da jedes dieser b_j durch h_1 teilbar ist und höchstens den Grad m hat, liegt es im Gitter

$$\mathbb{Z} \cdot h_1 \oplus \mathbb{Z} \cdot xh_1 \oplus \cdots \oplus \mathbb{Z} \cdot x^{m-\deg h_1} h_1 .$$

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 §10

$$\big\|b_j\big\|_2 \le 2^{m/2} \sqrt{\binom{2m}{m}}\, \|f\|_2 \quad \text{für } j = 0, \ldots, m - \deg h_0 .$$

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: Wir berechnen zunächst die Resultante von f und f' und wählen einen Primzahl p, die diese nicht teilt. Dann faktorisieren wir f mod p nach dem Algorithmus von BERLEKAMP; die Faktoren positiven Grades seien so normiert, daß ihre höchsten Koeffizienten alle eins sind. Außerdem berechnen wir die LANDAU-MIGNOTTE-Schranke für die Faktoren von f und wählen eine Zahl M, die größer ist, als deren Doppeltes.

Wir schreiben $f = f_1 f_2$ 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 mod 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

$$p^{ke} > 2^{md/2} \binom{2m}{m}^{d/2} \|f\|_2^{m+d}$$

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 $b_0, \ldots, b_t$.

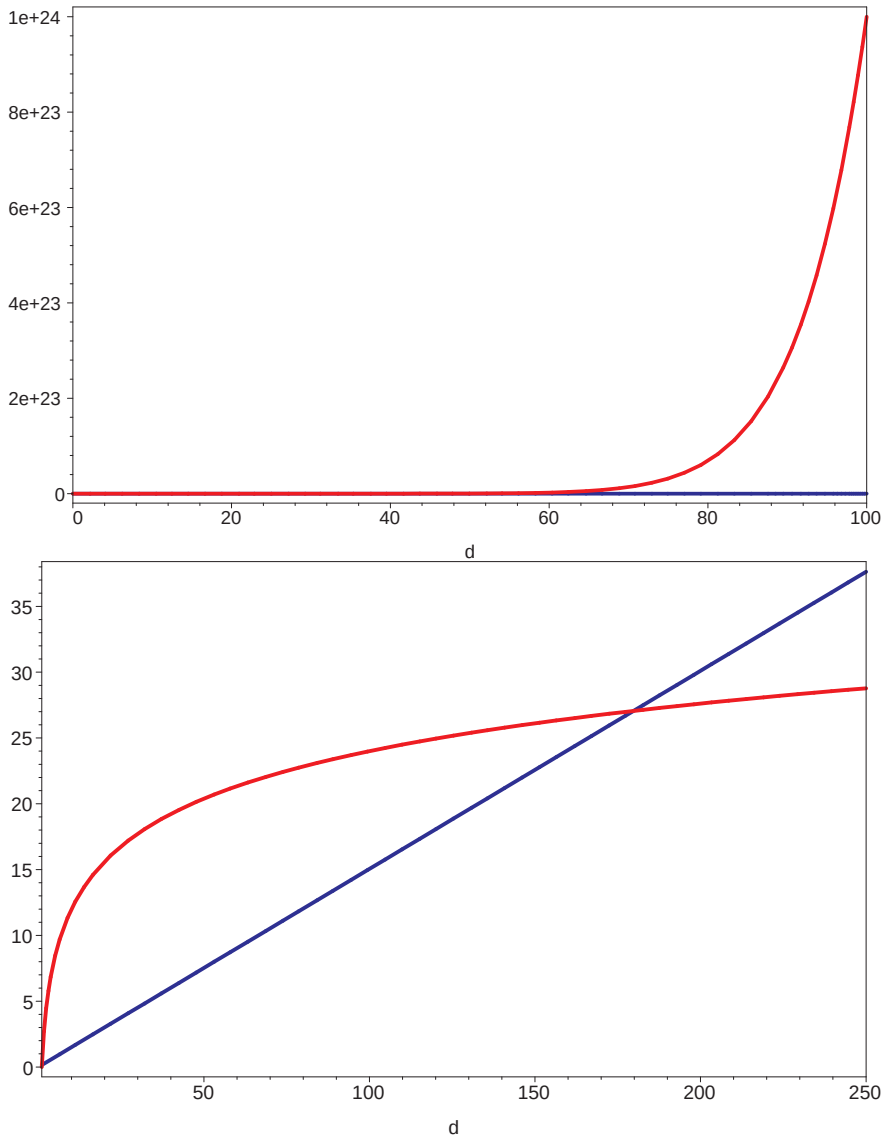
Wir ersetzen dann f_1 durch $f_1 h_0$ und f_2 durch f_2/h_0. Zur Faktorisierung des neuen f_2 modulo p brauchen wir natürlich keinen BERLEKAMP-Algorithmus, sondern können einfach testen, welche Faktoren des alten f_2 mod p in h_0 mod p (oder im neuen f_2 mod p) stecken.

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 LOVÁSZ 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 (erste der beiden folgenden Abbildungen), 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 zweiten 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 üblicherweise 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, 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,

§ 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_n über $k[x_1, \ldots, x_{n-1}]$ 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 4
Systeme von nichtlinearen Polynomgleichungen

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: Variablenelimination mit Resultanten

Wir wissen aus Kapitel 2, daß zwei Polynome $f, g \in R[x]$ über einem faktoriellen Ring R genau dann eine gemeinsame Nullstelle haben, wenn ihre Resultante verschwindet. Dies können wir anwenden, um aus einem System nichtlinearer Gleichungen eine Variable zu eliminieren und es so sukzessive auf Gleichungen in einer Veränderlichen zurückzuführen:

Im 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; dies wird wohl nur selten der Fall sein. Falls wir

die Polynome vorher faktorisieren und dann das eine Gleichungssystem ersetzen durch mehrere Systeme aus Polynomen kleineren Grades, 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

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

aus $k[x]$ 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.

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 für die Lösung eines nichtlinearen Gleichungssystems mit Resultanten betrachten wir die beiden Gleichungen

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

Ihre Resultante bezüglich x ist

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

Maple gibt deren 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 Maple ist das durchaus machbar.

Einfacher wird es aber, wenn wir y statt 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

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

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: Sie läßt sich nämlich als Linearkombination der beiden Polynome 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 nicht verschwinden. Die Gleichungen

$$x^i f = a_d x^{d+i} + \cdots + a_1 x^{1+i} + a_0 x^i \quad \text{und} \quad x^j g = b_e x^{e+j} + \cdots + b_1 x^{1+j} + b_0$$

für $i = 0, \ldots, e-1$ und $j = 0, \ldots, d-1$ können wir in Vektorschreibweise so zusammenfassen, daß wir den $(d+e)$-dimensionalen Vektor F mit Komponenten $x^{e-1}f, \ldots, xf, f, x^{d-1}g, \ldots, xg, g$ darstellen in der Form

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

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 LAGRANGEschen 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 LAGRANGE aufsummiert werden, genau eines der beiden Polynome. Nach der obigen Rechnung ist $pf + qg$ gleich der Determinante der r_k, also die Resultante. ∎

§2: Gauß und Euklid

Resultanten waren bereits im 19. Jahrhundert wohlbekannt. Erst 1966 entwickelte BRUNO BUCHBERGER einen alternatives Ansatz, dessen Bedeutung in der Computeralgebra – genau wie im Falle der Resultanten – weit über die Lösung nichtlinearer Gleichungssysteme hinausgeht. Mit diesem Verfahren wollen wir uns in den folgenden Paragraphen beschäftigen.

Ausgangspunkt sind der GAUSS-Algorithmus zur Lösung linearer Gleichungssysteme und der Algorithmus zur Polynomdivision, wie er im EUKLIDische Algorithmus zur Berechnung des ggT zweier Polynome verwendet wird:

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, uns 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

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = u \tag{1}$$

$$b_1x_1 + b_2x_2 + \cdots + b_nx_n = v \tag{2}$$

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

$$f = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \quad \text{und}$$
$$g = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0$$

mit $m \leq n$, 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 kleineren Grad als g hat. Konkret: Bei jedem Divisionsschritt haben wir ein Polynom

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

das wir mit Hilfe des Divisors

$$g = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0$$

reduzieren, indem wir es ersetzen durch

$$f - \frac{b_m}{a_n} x^{n-m} g \,,$$

und das führen wir so lange fort, bis f auf ein Polynom von kleinerem Grad als g 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. Wir wollen uns daher zunächst damit beschäftigen.

§3: Der Divisionsalgorithmus

Wir betrachten Polynome in n Variablen $x_1, \ldots, x_n$ und setzen zur Abkürzung

$$x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n} \quad \text{mit} \quad \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0 \,.$$

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 mit der 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 $f = \sum_{\alpha \in I} c_\alpha x^\alpha \in k[x_1, \ldots, x_n]$ 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 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 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; siehe Übungsblatt.

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 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 mir 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 auch mehrere Divisoren gleichzeitig behandeln kann.

Wir gehen also aus von einem Polynom $R = f \in k[x_1, \ldots, x_n]$, wobei k irgendein Körper ist, in dem wir rechnen können, meistens also $k = \mathbb{Q}$ oder $k = \mathbb{F}_p$. 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 irgendeinem 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 $f = a_1 f_1 + \cdots + a_m f_m + r$

1. Schritt (Initialisierung): Setze $a_1 = \cdots = a_m = r = 0$. Falls $f = 0$ endet der Algorithmus damit; andernfalls setzen wir $p = f$.

2. Schritt: Falls keiner der führenden Terme FT f_i den führenden Term FT p teilt, wird p ersetzt durch $p -$ FT p und r durch $r +$ FT p.

3. Schritt (Divisionsschritt): Andernfalls sei i der kleinste Index, für den FT f_i Teiler von FT r 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 zweiten oder dritten Schritts 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 folgt daraus, daß der Algorithmus nach endlich vielen Schritten abbrechen muß.

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

$$f = a_1 f_1 + a_2 f_2 + r \quad \text{mit} \quad a_1 = x + y, \quad a_2 = 1 \quad \text{und} \quad r = x + y + 1\,.$$

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:

$$f = a_1 f_1 + a_2 f_2 + r \quad \text{mit} \quad a_1 = x + 1, \quad a_2 = x \quad \text{und} \quad r = x + 1\,.$$

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 $f = xf_2$ 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

$$f_1(x_1, \ldots, x_n) = \cdots = f_r(x_1, \ldots, x_n) = 0 \;\text{ mit } f_i \in R = k[x_1, \ldots, x_n]\,;$$

seine Lösungsmenge sei $\mathcal{L} \subseteq k^n$.

Ist $g = g_1 f_1 + \cdots + g_r f_r$ mit $g_i \in R$ ein beliebiges Element des von $f_1, \ldots, f_r$ erzeugten Ideals $I \lhd R$, so ist für jede Lösung $(x_1, \ldots, x_r)$ aus $\mathcal{L}$ offensichtlich auch $g(x_1, \ldots, x_r) = 0$. Ist $I = (h_1, \ldots, h_s)$ eine andere Erzeugung von I, so hat das obige Gleichungssystem daher dieselbe Lösungsmenge wie das System

$$h_1(x_1, \ldots, x_n) = \cdots = h_s(x_1, \ldots, x_n) = 0\,.$$

Zur Lösung des Systems sollten wir daher versuchen, ein möglichst „einfaches“ Erzeugendensystem für das Ideal I 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 von (nicht notwendigerweise endlich vielen) Monomen erzeugt wird.

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 $\alpha \in A, \beta \in \mathbb{N}_0^n$ und $c \in k$. Sortieren wir diese Summanden nach den resultierenden Monomen $x^{\gamma+\alpha}$, 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 $\gamma \in \mathbb{N}_0^n$; das Ideal I selbst besteht

also 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.

Für $n > 1$ bezeichnen wir für $\alpha \in \mathbb{N}_0^n$ mit x'^{α} das Monom $x_1^{\alpha_1} \cdots x_{n-1}^{\alpha_{n-1}}$ 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 gibt aber natürlich auch noch Monome in I, 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.

■

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 Universität von 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.

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, im Widerspruch zu der oben gezeigten Eigenschaft eines monomialen Ideals, zu jedem seiner Elemente auch dessen sämtliche Monome zu enthalten.

Um monomiale Ideale auch für die Untersuchung solcher Ideale nützlich zu machen, wählen wir eine Monomordnung auf R und definieren für ein beliebiges Ideal $I \lhd R \underset{\text{def}}{=} k[x_1, \ldots, x_n]$ das monomiale Ideal

$$\mathrm{FM}(I) = \Big(\mathrm{FM}(f) \;\Big|\; f \in I \setminus \{0\}\Big)\,,$$

das von den führenden Monomen *aller* Elemente von I erzeugt wird – außer natürlich dem nicht existierenden führenden Term der Null.

Nach dem Lemma von DICKSON ist $\mathrm{FM}(I)$ 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 $\mathrm{FM}(x^\gamma f) = x^\gamma \,\mathrm{FM}(f)$, da für jede Monomordnung gilt $\alpha < \beta \Longrightarrow \alpha + \beta < \alpha + \gamma$. Somit wird $\mathrm{FM}(I)$ erzeugt von endlich vielen Monomen der Form $\mathrm{FM}(f_i)$, 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

$$f = a_1 f_1 + \cdots + a_m f_m + r\,.$$

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.

§5: Gröbner-Basen und der Buchberger-Algorithmus

Angesichts der Rolle der führenden Monome im obigen Beweis bietet sich folgende Definition an für eine Idealbasis, bezüglich derer möglichst viele Eigenschaften bereits an den führenden Monomen abgelesen werden können:

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, 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 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 war 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 dann zunächst als Assustent. 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 obige Beweis des HILBERTschen Basissatzes zeigt, erzeugt eine GRÖBNER-Basis das Ideal, und jedes Ideal im Polynomring hat eine GRÖBNER-Basis. Bevor wir uns damit beschäftigen, wie man diese

berechnen kann, wollen wir zunächst eine wichtige Eigenschaften betrachten.

Sei $g_1, \ldots, g_m$ 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 weder der Divisionsrest r noch die Koeffizienten a_i auch nur im entferntesten eindeutig. Sei etwa $f = a_1 g_1 + \cdots + a_m g_m + r = b_1 g_1 + \cdots + b_m g_m + s$; 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. $r - s = 0$. 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 überlegen, wie sich eine GRÖBNER-Basis eines vorgegebenen Ideals I finden läßt.

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 FM(f_i) im allgemeinen nicht ausreichen, um das monomiale Ideal 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; FM(f) = x^α und FM(g) = x^β 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 alle $i = 1, \ldots, n$. 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, d.h. $S(f,g)$ hat ein kleineres führendes Monom. 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 führen-

den 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, d.h. daß $\mathrm{FM}(f_1), \ldots, \mathrm{FM}(f_m)$ 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 FM(f) haben. Wegen FM($h_i f_i$) = FM(h_i) FM(f_i) liegt FM(f) daher im von den FM(f_i) erzeugten Ideal.

Falls sich die maximalen unter den führenden Termen 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 FM(f) als Element des von den 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, könnte das Ideal FM(I) kein endliches Erzeugendensystem haben, im Widerspruch zum Lemma von DICKSON. Also kann der zweite Schritt nur endlich oft durchlaufen werden.

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 für den Rest des Semesters lieber einige Anwendungen betrachten.

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

$$h(x,y) = -xy^3 + xy^2 + x + y^2 - y\boldsymbol{z}.$$

Sowohl $g(x,y)$ als auch das $h(x,y)$-Polynom 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. Dann ist

$$S(f_1,f_2) = yf_1 - xf_2 = -x^2$$

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

$$f_5 = S(f_2, f_3) = f_2 + yf_3 = -2y^2 + x\,.$$

Im so erweiterten 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:

$$S(f_1, f_4) = yf_1 - \frac{x}{2}f_4 = -2xy^2 = yf_4$$

$$S(f_1, f_5) = y^2f_1 + \frac{x^3}{2}f_5 = -2xy^3 + \frac{x^4}{2} = \frac{x}{2}f_1 + f_2 + y^2f_4 - f_5$$

$$S(f_2, f_4) = f_2 + \frac{x}{2}f_4 = -2y^2 + x = f_5$$

$$S(f_2, f_5) = yf_2 + \frac{x^2}{2}f_5 = \frac{x^3}{2} + xy - 2y^3 = \frac{1}{2}f_1 - \frac{1}{2}f_4 + yf_5$$

$$S(f_3, f_4) = -yf_3 - \frac{x}{2}f_4 = 0$$

$$S(f_3, f_5) = -y^2f_3 - \frac{x^2}{2}f_5 = \frac{1}{2}f_1 - \frac{1}{2}f_4$$

$$S(f_4, f_5) = -\frac{y}{2}f_4 - \frac{x}{2}f_5 = \frac{x^2}{2} = -\frac{1}{2}f_3$$

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 FT(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. ∎

Im obigen Beispiel haben wir die führenden Monome

$$\mathrm{FM}(f_1) = x^3, \quad \mathrm{FM}(f_2) = x^2y, \quad \mathrm{FM}(f_3) = x^2,$$
$$\mathrm{FM}(f_4) = xy \quad \text{und} \quad \mathrm{FM}(f_5) = y^2;$$

offensichtlich sind FM(f_1) und FM(f_2) durch FM(f_3) teilbar, so daß wir auf f_1 und f_2 verzichten können: Auch die Polynome f_3, f_4 und f_5 bilden eine GRÖBNER-Basis des von f_1 und f_2 erzeugten Ideals. 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 FT(g) im von den führenden Termen der übrigen Elemente erzeugten Ideal.

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 Termen 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 Terme der anderen Polynome teilbar sein.

Reduzierte GRÖBNER-Basis haben eine für das praktische Rechnen mit Idealen sehr wesentliche zusätzliche Eigenschaft:

Satz: Jedes Ideal $I \lhd k[x_1, \ldots, x_n]$ hat 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 FT(f) = FT(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 $f - g = 0$.

■

§6: Anwendungen von Gröbner-Basen

Die Eindeutigkeit der reduzierten GRÖBNER-Basis bedeutet, daß wir Ideale des Polynomrings durch endlich viele Daten eindeutig beschreiben können; insbesondere können wir entscheiden, ob zwei Mengen von Polynomen dasselbe Ideal erzeugen. Dies ist der Ausgangspunkt für zahlreiche Anwendungen von GRÖBNER-Basen in der kommutativen Algebra, algebraischen Geometrie, Kontrolltheorie und so weiter.

Wir wollen uns hier mit zwei einfacheren Anwendungen begnügen, zunächst dem Hauptproblem dieser Vorlesung, der Lösung algebraischer Gleichungen und Gleichungssysteme.

Wir gehen also 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

$$V(I) = \left\{ (x_1, \ldots, x_n) \in k^n \;\middle|\; f_i(x_1, \ldots, x_n) = 0 \text{ für } i = 1, \ldots, m \right\}.$$

Wir verwenden die lexikographische Ordnung mit $x_1 > \cdots > x_n$ und betrachten das von den f_i erzeugte Ideal $I \lhd k[x_1, \ldots, x_n]$.

Zur Lösung des Gleichungssystems wollen wir, wie von linearen Gleichungssystemen gewohnt, nacheinander die Variablen eliminieren; dazu definieren wir

Definition: Das k-te Eliminationsideal eines Ideal $I \lhd k[x_1, \ldots, x_n]$ ist $I_k = I \cap k[x_{k+1}, \ldots, x_n]$.

Satz: Ist G eine GRÖBNER-Basis von I bezüglich der lexikographischen Ordnung, so ist $G \cap I_k$ eine GRÖBNER-Basis von I_k.

Beweis: Die Elemente von $G = \{g_1, \ldots, g_m\}$ seien so angeordnet, daß $G \cap I_k = \{g_1, \ldots, g_r\}$ ist. Wir müssen zeigen, daß sich jedes $f \in I_k$ als Linearkombination von $g_1, \ldots, g_r$ darstellen läßt.

Der Divisionsalgorithmus bezüglich der lexikographischen Ordnung gibt uns eine Darstellung $f = h_1 g_1 + \cdots + h_n g_n$ von f als Element von I. Dabei mußten alle h_i mit $i > r$ verschwinden, denn da f in I_k liegt, kann bei der Division kein führender Term eines $g_i \notin I_k$ je den führenden Term des Dividenden teilen. Somit ist $G \cap I_k$ eine Basis von I_k. 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 $G \cap I_k$ 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. ∎

Ist I das von den Gleichungen eines nichtlinearen Gleichungssystems erzeugte Ideal, so ist jede Lösung $(x_1, \ldots, x_n)$ des Systems Nullstelle aller Polynome aus I, insbesondere also auch aller Polynome aus I_k. Für jede Lösung ist daher das Tupel $(x_{k+1}, \ldots, x_n)$ Nullstelle der Polynome aus I_k.

Daraus ergibt sich eine Strategie zur Lösung nichtlinearer Gleichungssysteme nach Art des GAUSS-Algorithmus: Wir bestimmen zunächst eine (reduzierte) GRÖBNER-Basis für das von den Gleichungen erzeugte Ideal

des Polynomrings $k[x_1, \ldots, x_n]$ und 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 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, und so weiter.

Im obigen Beispiel etwa besteht die reduzierte GRÖBNER-Basis bezüglich der lexikographischen Ordnung aus den beiden Polynomen

$$g_1 = x - 2y^2 \quad \text{und} \quad g_2 = y^3 .$$

Das Eliminationsideal I_1 wird also erzeugt von $g_2 = y^3$, d.h. für jede Lösung (x, y) muß y verschwinden. Setzen wir $y = 0$ in g_1, so sehen wir, daß auch x verschwinden muß, der Nullpunkt ist also die einzige Lösung.

Nun kann es 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 Kurven; 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 zweier elliptischer Zylinder. 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.

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 haben wir also 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 suchen Polynome $f_1, \ldots, f_r$, die genau auf der Menge aller $(x_1, \ldots, x_n)$ verschwinden, für die es eine solche Darstellung gibt.

Dazu wählen wir eine lexikographische Ordnung, 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.

§7: Der Hilbertsche Nullstellensatz

In diesem Paragraphen wollen wir Kriterien für die Lösbarkeit eines nichtlinearen Gleichungssystems sowie für die Endlichkeit der Lösungsmenge herleiten. Außerdem überlegen wir uns, wann ein Polynom g auf der Lösungsmenge eines nichtlinearen Gleichungssystems verschwindet. Natürlich muß es dann verschwinden, wenn es im von den Gleichungen erzeugten Ideal liegt, aber die Umkehrung dazu gilt nicht: Die Nullstellenmenge des System mit der einzigen Gleichung $(x - y)^3 = 0$ etwa besteht genau aus den Punkten (x, y) mit $x = y$, und dort verschwindet auch das lineare Polynom $x - y$, das schon aus Gradgründen nicht im von $(x - y)^3$ erzeugten Ideal liegen kann.

Wenn wir über einem endlichen Körper arbeiten, beispielsweise dem Körper $\mathbb{F}_p$, haben wir das zusätzliche Problem, daß es Polynome gibt, die auf ganz $\mathbb{F}_p^n$ verschwinden: Nach dem kleinen Satz von FERMAT ist beispielsweise $x^p - x = 0$ für alle $x \in \mathbb{F}_p$, und daraus lassen sich leicht Polynome in n Variablen konstruieren, die auf ganz $\mathbb{F}_p^n$ verschwinden. Wir wollen uns als erstes überlegen, daß dieses Phänomen bei unendlichen Körpern nicht auftreten kann:

Lemma: k sei ein unendlicher Körper und $f \in k[x_1, \ldots, x_n]$ sei nicht das Nullpolynom. Dann gibt es $a_1, \ldots, a_n \in k$ derart, daß $f(a_1, \ldots, a_n)$ nicht verschwindet.

Wir führen den *Beweis* durch Induktion nach n: Für Polynome einer Veränderlichen folgt dies aus der Tatsache, daß ein vom Nullpolynom verschiedenes Polynom höchstens so viele Nullstellen haben kann, wie sein Grad angibt, also auch jeden Fall endlich viele. In einem unendlichen Körper muß es daher Elemente geben, für die das Polynom nicht verschwindet.

Für $n > 1$ schreiben wir $f = f_d x_n^d + f_{d-1} x^{d-1} + \cdots + f_1 x + f_0$ als Polynom in x_n mit Koeffizienten $f_i \in k[x_1, \ldots, x_{n-1}]$, wobei wir annehmen können, daß der führende Koeffizient f_d nicht das Nullpolynom ist. Nach Induktionsannahme gibt es dann $a_1, \ldots, a_{n-1} \in k$, für die $f_d(a_1, \ldots, a_{n-1})$ nicht verschwindet. Setzen wir $x_1, \ldots, x_{n-1}$ auf diese Werte, ist daher $f(a_1, \ldots, a_{n-1}, x_n) \in k[x_n]$ nicht das Nullpolynom, hat also nur endlich viele Nullstellen. Wählen wir für $a_n \in k$ irgendein Element, das keine Nullstelle ist, muß $f(a_1, \ldots, a_{n-1}, a_n)$ von Null verschieden sein.

■

Korollar: Das Polynom $f \in k[x_1, \ldots, x_n]$ über dem unendlichen Körper k habe den Gesamtgrad d. Dann gibt es Elemente $\lambda_i \in k$, für die das Polynom $f(y_1 + \lambda_1 y_n, \ldots, y_{n-1} + \lambda_{n-1} y_n, y)$ aus dem Polynomring $k[y_1, \ldots, y_n] = k[y_1, \ldots, y_{n-1}][y_n]$ als Polynom in y_n den führenden Term $c y_n^d$ hat mit einem $c \neq 0$ aus k.

Den *Beweis* führen wir wieder durch Induktion nach n: Für $n = 1$ ist $y_1 = x_1$, und die Behauptung trivial; sei also $n > 1$. Wir schreiben

$$\begin{aligned} g(y_1, \ldots, y_n) &\underset{\text{def}}{=} f(y_1 + \lambda_1 y_n, \ldots, y_{n-1} + \lambda_{n-1} y_n, y) \\ &= \sum_e a_e(\lambda_1, \ldots, \lambda_{n-1}) y^e \end{aligned}$$

als Polynom in den y_i mit Koeffizienten aus $k[\lambda_1, \ldots, \lambda_{n-1}]$; die Summe Läuft also über gewisse n-Tupel $e \in \mathbb{N}_0^n$ vom Grad höchstens d. Da wir in f für jedes x_i einen in y_n linearen Ausdruck eingesetzt haben, führt jedes Monom von f nach Einsetzen und Ausmultiplizieren zu einer Summe, in der ein Term mit y_n^d vorkommt; der Koeffizient von y_n^d ist also nicht das Nullpolynom aus $k[\lambda_1, \ldots, \lambda_{n-1}]$. Nach dem gerade bewiesenen Lemma gibt es daher $\lambda_i \in k$, für die dieser Koeffizient von Null verschieden ist, und mit diesen λ_i gilt die Behauptung.

■

Dieses eher technische Korollar sagt also, daß wir durch eine lineare Koordinatentransformation immer erzwingen können, daß eine der Variablen mit dem Gesamtgrad als Exponenten auftritt. Dies benutzen wir, um zu untersuchen, wann ein Gleichungssystem unlösbar ist.

Dabei wollen wir die *Unlösbarkeit* in einem starken Sinn interpretieren: Die Gleichung $x^2 + 1 = 0$ hat beispielsweise zwar keine reelle Lösung, aber sie hat die beiden komplexen Lösungen $x = \pm i$. So eine Gleichung wollen wir nicht als unlösbar betrachten. Wir definieren

Definition: *a)* Ein Körper k heißt *algebraisch abgeschlossen*, wenn jedes nichtkonstante Polynom aus $k[x]$ mindestens eine Nullstelle hat.
b) Ist I ein Ideal in $k[x_1, \ldots, x_n]$ und ist k' ein Körper, der k enthält, setzen wir

$$V_{k'}(I) = \left\{(z_1, \ldots, z_n) \in k'^n \;\middle|\; f(z_1, \ldots, z_n) = 0 \text{ für alle } f \in I\right\}.$$

c) Erzeugen die Polynome $f_1, \ldots, f_m \in k[x_1, \ldots, x_n]$ das Ideal I, so sei $V_{k'}(f_1, \ldots, f_m) = V_{k'}(I)$.

Schwache Form des Hilbertschen Nullstellensatzes: k sei ein Körper, K ein algebraisch abgeschlossener Körper, der k enthält, und I sei ein Ideal im Polynomring $k[x_1, \ldots, x_n]$ über k. Dann ist $V_K(I) = \emptyset$ genau dann, wenn das Ideal I die Eins enthält.

In Gleichungen ausgedrückt heißt dies, daß das Gleichungssystem

$$f_i(x_1, \ldots, x_n) = 0 \quad \text{für } i = 1, \ldots, m$$

genau dann selbst in K keine Lösung hat, wenn die Eins im Ideal $(f_1, \ldots, f_m)$ liegt, wenn es also Polynome $g_1, \ldots, g_m \in k[x_1, \ldots, x_n]$ gibt, für die $g_1 f_1 + \cdots + g_m f_m = 1$ ist.

Der *Beweis* erfolgt auch hier wieder durch vollständige Induktion nach der Anzahl der Variablen:

Für $n = 1$ ist jedes Ideal $I \lhd k[x]$ ein Hauptideal; es sei erzeugt von $f \in k[x]$. Nach Definition eines algebraisch abgeschlossenen Körpers hat f genau dann keine Nullstelle in K, wenn f konstant ist, und das ist äquivalent dazu, daß I die Eins enthält.

Für $n > 1$ betrachten wir ein Erzeugendensystem $f_1, \ldots, f_m$ von I. Nach dem obigen Korollar können wir annehmen, daß f_1 den Term x_n^d enthält, wobei d den Grad von f_1 bezeichnet: Eine lineare Koordinatentransformation ändert schließlich nichts daran, ob $V_K(I)$ leer ist oder nicht und auch nichts daran, ob I die Eins enthält oder nicht.

Wir führen eine neue Variable u ein und betrachten das Polynom

$$h = f_2 + uf_3 + \cdots + u^{m-2} f_m \in k[x_1, \ldots, x_n, u]$$

sowie die Resultante $\mathrm{Res}_{x_n}(f_1, h) \in k[x_1, \ldots, x_{n-1}, u]$. Wir schreiben sie als Polynom

$$\mathrm{Res}_{x_n}(f_1, h) = a_\ell(x_1, \ldots, x_{n-1})u^\ell + \cdots + a_0(x_1, \ldots, x_{n-1})$$

in u mit Koeffizienten aus $k[x_1, \ldots, x_{n-1}]$.

Wie wir am Ende von §1 gesehen haben, läßt sich die Resultante zweier Polynome als Linearkombination dieser Polynome darstellen; es gibt daher Polynome $p, q \in k[x_1, \ldots, x_n, u]$, so daß gilt

$$\mathrm{Res}_{x_n}(f_1, h) = pf_1 + qh = pf_1 + qf_2 + quf_3 + \cdots + qu^{m-2} f_m \,.$$

Vergleichen wir dies mit obiger Darstellung der Resultante als Polynom in u, sehen wir, daß die Koeffizientenpolynome $a_i(x_1, \ldots, x_{n-1})$ im Ideal $I = (f_1, \ldots, f_m)$ liegen müssen.

Wenn wir zeigen können, daß diese Polynome keine gemeinsame Nullstelle in K^{n-1} haben, wissen wir nach Induktionsannahme, daß sich die Eins in $k[x_1, \ldots, x_{n-1}]$ als Linearkombination der a_i darstellen läßt; da diese Polynome in I liegen, liegt die Eins somit erst recht in I, und wir sind fertig.

Angenommen, die a_i hätten eine gemeinsame Nullstelle $(z_1, \ldots, z_{n-1})$ in K^{n-1}. Dann wäre $\mathrm{Res}_{x_n}(f_1, h)(z_1, \ldots, z_{n-1}, u) \in k[u]$ das Nullpolynom. Somit hätten die beiden Polynome

$$f_1(z_1, \ldots, z_{n-1}, x_n) \in k[x_n] \;\text{ und }\; h(z_1, \ldots, z_{n-1}, x_n, u] \in k[x_n, u]$$

einen nichtkonstanten gemeinsamen Faktor. In K gäbe es dann eine Nullstelle z_n von $f_1(z_1, \ldots, z_{n-1}, x_n)$, für die $h(z_1, \ldots, z_{n-1}, z_n, u)$ das Nullpolynom wäre. Nach Definition von h verschwänden dann nicht nur $f_1(z_1, \ldots, z_n)$, sondern auch alle $f_j(z_1, \ldots, z_n)$ für $j = 2, \ldots, m$. Damit läge $(z_1, \ldots, z_n)$ in $V_K(I)$, was wir aber als leer vorausgesetzt haben. Damit ist klar, daß die a_i keine gemeinsame Nullstelle haben können, und der Satz ist bewiesen.

∎

Ob ein Ideal die Eins enthält oder nicht, kann man seiner GRÖBNER-Basis leicht ansehen: Da der führende Term eines jeden Polynoms aus dem Ideal durch den führenden Term eines Elements der GRÖBNER-Basis teilbar sein muß, enthält diese im Falle eines Ideals, das die Eins enthält, ein Polynom, dessen führendes Monom die Eins ist. Da diese bezüglich jeder Monomordnung das kleinste Monom ist, muß somit die GRÖBNER-Basis eine Konstante enthalten. Die zugehörige minimale und erst recht die reduzierte GRÖBNER-Basis besteht in diesem Fall nur aus der Eins.

Kriterium: Ein nichtlineares Gleichungssystem ist genau dann unlösbar selbst über einem algebraisch abgeschlossenen Körper, der k enthält, wenn seine reduzierte GRÖBNER-Basis nur aus der Eins besteht. ∎

Die starke Form des HILBERTschen Nullstellensatzes sagt uns allgemein, welche Polynome auf der Nullstellenmenge eines Ideals verschwinden:

Hilbertscher Nullstellensatz: $I \lhd k[x_1, \ldots, x_n]$ sei ein Ideal, und das Polynom $g \in k[x_1, \ldots, x_n]$ verschwinde in jedem Punkt von $V_K(I)$, wobei K ein algebraisch abgeschlossener Körper sei, der k enthält. Dann gibt es eine natürliche Zahl r, so daß g^r in I liegt.

Beweis: $f_1, \ldots, f_m$ sei ein Erzeugendensystem von I und

$$J = (f_1, \ldots, f_m, Tg - 1) \lhd k[x_1, \ldots, x_n, T] .$$

Für einen Punkt $(z_1, \ldots, z_n, t) \in V_K(J)$ müßte einerseits gelten

$$f_j(z_1, \ldots, z_n) = 0 \quad \text{für alle } j = 1, \ldots, m ,$$

so daß $(z_1, \ldots, z_n)$ in $V_K(I)$ läge; andererseits wäre auch

$$tg(z_1, \ldots, z_n) - 1 = 0 .$$

Da g in allen Punkten aus $V_K(I)$ verschwindet, ist das nicht möglich, also ist $V_K(J) = \emptyset$. Nach der schwachen Form des HILBERTschen Nullstellensatzes muß somit die Eins in J liegen; es gibt also Polynome $a_0, \ldots, a_m \in k[x_1, \ldots, x_n, T]$, so daß

$$a_1 f_1 + \cdots + a_n f_n + a_0(Tg - 1) = 1$$

ist. Im Quotientenkörper von $k[x_1, \ldots, x_n]$ können wir in dieser Identität $T = 1/g$ einsetzen. Dadurch können die Summanden Potenzen von g in ihre Nenner bekommen; durch Multiplikation mit der höchsten auftretenden Potenz g^r erhalten wir eine Polynomgleichung der Form

$$b_1 f_1 + \cdots + b_m f_m = g^r \quad \text{mit} \quad b_j \in k[x_1, \ldots, x_n] \,.$$

Dies beweist die Behauptung.

■

Definition: R sei ein Ring und $I \lhd R$ ein Ideal von R. Das *Radikal* von I ist die Menge

$$\sqrt{I} \underset{\text{def}}{=} \left\{ a \in R \;\middle|\; \exists n \in \mathbb{N} : a^n \in I \right\} .$$

Das Radikal besteht also aus allen Ringelementen, die eine Potenz in I haben. Es ist selbst ein Ideal, denn sind $a, b \in \sqrt{I}$ zwei Elemente mit $a^n \in I$ und $b^m \in I$, so sind in

$$(a + b)^{n+m} = \sum_{k=0}^{n+m} \binom{n+m}{k} a^{n+m-k} b^k$$

die ersten m Summanden Vielfache von a^n, und die restlichen n sind Vielfache von b^m. Somit liegt jeder Summand in I, also auch die Summe. Für ein beliebiges $r \in R$ liegt natürlich auch ra in $\sqrt{I}$, denn seine n-te Potenz $(ra)^n = r^n a^n$ liegt in I.

Falls ein Ideal mit seinem Radikal übereinstimmt, enthält es *alle* Polynome, die auf $V_K(I)$ verschwinden; zwei Polynome nehmen genau dann in jedem Punkt von $V_K(I)$ denselben Wert an, wenn ihre Differenz in I liegt, wenn sie also modulo I dieselbe Restklasse definieren.

Wenn das Ideal I nicht mit seinem Radikal übereinstimmt, gilt zwar nicht mehr *genau dann*, aber wir können trotzdem die Elemente des Faktorvektorraums $A = k[x_1, \ldots, x_n]/I$ auffassen als Funktionen von $V_K(I)$ nach K: Für jede Restklasse und jeden Punkt aus $V_K(I)$ nehmen wir einfach irgendein Polynom aus der Restklasse und setzen die Koordinaten des Punkts ein. Da die Differenz zweier Polynome aus derselben Restklasse in I liegt, wird sie nach Einsetzen des Punktes zu Null, der Wert hängt also nicht ab von der Wahl des Polynoms.

Auch Polynome aus $K[x_1, \ldots, x_n]$ definieren in dieser Weise Funktionen $V_K(I) \to K$; hinreichend (aber nicht notwendig) dafür, daß zwei Polynome dieselbe Funktion definieren ist, daß ihre Differenz im von I erzeugten Ideal $\bar{I} \lhd K[x_1, \ldots, x_n]$ liegt.

Im Falle von Polynomen einer Veränderlichen ist jedes Ideal von $k[x]$ ein Hauptideal; ist $I = (f)$ mit einem Polynom $f \neq 0$ vom Grad d, so können wir die Restklassen repräsentieren durch die Polynome vom Grad höchstens $d - 1$, denn jedes Polynom $g \in k[x]$ hat dieselbe Restklasse wie sein Divisionsrest bei der Polynomdivision durch f. Somit ist $A = k[x]/I$ in diesem Fall ein d-dimensionaler Vektorraum. Da $V_K(I)$ gerade aus den Nullstellen von f in K besteht, von denen es höchstens d verschiedene gibt, liefert die Dimension von A eine obere Schranke für die Elementanzahl von $V_K(I)$; wenn wir die Nullstellen mit ihrer Vielfachheit zählen, ist die Dimension von A sogar *gleich* der Gesamtzahl der Nullstellen.

Dies gilt auch für Polynome mehrerer Veränderlicher, ist allerdings schwerer zu beweisen. Vielfachheiten werden wir erst im nächsten Paragraphen betrachten; hier begnügen wir uns mit dem folgenden

Satz: I sei ein Ideal im Polynomring $k[x_1, \ldots, x_n]$ über dem Körper k, und K sei ein algebraisch abgeschlossener Körper, in dem k enthalten sei. Dann gilt: $V_K(I)$ ist genau dann endlich, wenn $A = k[x_1, \ldots, x_n]/I$ ein endlichdimensionaler Vektorraum ist. In diesem Fall ist die Dimension von A eine obere Schranke für die Elementanzahl von $V_K(I)$.

Den recht umfangreichen *Beweis* führen wir in mehreren Schritten:

1. Schritt: Wenn der Vektorraum A endliche Dimension hat, ist $V_K(I)$ endlich.

Bezeichnet nämlich d die Dimension von A, so sind für jedes i die Potenzen $1, x_i, \ldots, x_i^d$ linear abhängig; es gibt also ein Polynom aus $k[x_i]$, das modulo I zur Null wird und somit in I liegt. Für jeden Punkt aus $V_K(I)$ muß die i-te Koordinate eine Nullstelle dieses Polynoms sein, so daß diese nur endlich viele Werte annehmen kann. Da dies für alle i gilt, ist $V_K(I)$ endlich.

2. Schritt: Wenn $V_K(I)$ endlich ist, hat der K-Vektorraum $\bar{A}$ endliche Dimension.

Besteht $V_K(I)$ nur aus endlich vielen Punkten, so nimmt jede der Koordinatenfunktionen $x_1, \ldots, x_n$ auf $V_K(I)$ nur endlich viele Werte an; es gibt also für jedes i ein Polynom aus $K[x_i]$, das auf ganz $V_K(I)$ verschwindet. Nach dem HILBERTschen Nullstellensatz muß eine Potenz dieses Polynoms in $\bar{I}$ liegen, es gibt also auch in $\bar{I}$ für jedes i ein Polynom nur in x_i. Somit gibt es einen Grad d_i derart, daß sich x_i^e für $e \geq d_i$ modulo $\bar{I}$ durch die endlich vielen x_i-Potenzen $1, x_i, \ldots, x_i^{d_i-1}$ ausdrücken läßt. Damit läßt sich auch jedes Monom aus $K[x_1, \ldots, x_n]$ modulo $\bar{I}$ durch jene Monome ausdrücken, bei denen jede Variable x_i höchstens mit Exponent $d_i - 1$ auftritt. Da es nur endlich viele solche Monome gibt, ist $K[x_1, \ldots, x_n]/\bar{I}$ ein endlichdimensionaler K-Vektorraum.

3. Schritt: A ist genau dann endlichdimensional, wenn $\bar{A}$ endlichdimensional ist; in diesem Fall haben beide dieselbe Dimension.

Ist A endlichdimensional, so wählen wir eine Basis und zu jedem Basiselement ein Polynom aus $k[x_1, \ldots, x_n]$, das modulo I gleich diesem Element ist. Diese Polynome liegen erst recht in $K[x_1, \ldots, x_n]$, und es ist klar, daß ihre Restklassen modulo $\bar{I}$ den Vektorraum $\bar{A}$ erzeugen. Somit ist auch $\bar{A}$ endlichdimensional. Die Gleichheit von $\dim_k A$ und $\dim_K \bar{A}$ folgt, falls wir zeigen können, daß dieses Erzeugendensystem linear unabhängig ist.

Dazu zeigen wir die folgende, etwas allgemeinere Aussage: Sind $B_1, \ldots, B_r$ Polynome aus $k[x_1, \ldots, x_n]$ mit Restklassen $b_1, \ldots, b_r$ modulo I und Restklassen $\bar{b}_1, \ldots, \bar{b}_r$ modulo $\bar{I}$, so sind die b_i genau dann linear abhängig, wenn es die $\bar{b}_i$ sind.

Die eine Richtung ist einfach: Falls die b_i linear abhängig sind, gibt es Skalare $\lambda_i \in k$, die nicht alle verschwinden, so daß $\lambda_1 b_1 + \cdots + \lambda_r b_r$ der Nullvektor aus A ist. $\lambda_1 B_1 + \cdots + \lambda_r B_r$ liegt daher in I, also erst recht in $\bar{I}$, so daß auch $\lambda_1 b_1 + \cdots + \lambda_r b_r$ der Nullvektor aus $\bar{A}$ ist.

Wenn die $\bar{b}_i$ linear abhängig sind, gibt es $\lambda_i \in K$, so daß $\lambda_1 \bar{b}_1 + \cdots + \lambda_r \bar{b}_r$ der Nullvektor aus $\bar{A}$ ist, d.h. $\lambda_1 B_1 + \cdots + \lambda_r B_r$ liegt in $\bar{I}$. Da die λ_i

nicht in k liegen müssen, nützt und das noch nichts, um etwas über die b_i auszusagen.

Um trotzdem deren lineare Abhängigkeit zu beweisen, wählen wir ein endliches Erzeugendensystem $f_1, \ldots, f_m$ des Ideals I; wir wissen dann, daß es Polynome $g_1, \ldots, g_m$ aus $K[x_1, \ldots, x_n]$ gibt mit

$$\lambda_1 B_1 + \cdots + \lambda_r B_r = g_1 f_1 + \cdots + g_m f_m \, .$$

ist. Die Polynome g_j sind K-Linearkombinationen von Monomen $M_{j\ell}$ in den Variablen X_i. Die obige Gleichung ist also äquivalent zu einer Gleichung der Form

$$\lambda_1 B_1 + \cdots + \lambda_r B_r - \sum_{j=1}^{m} \sum_{\ell=1}^{r_j} \nu_{j\ell} M_{j\ell} f_j = 0$$

mit Elementen $\nu_{j\ell} \in K$, die von den g_j abhängen. Sortieren wir diese Gleichung nach Monomen, können wir dies so interpretieren, daß ein (recht großes) lineares Gleichungssystem in den Variablen λ_i und $\mu_{j\ell}$ eine nichttriviale Lösung hat. Da die B_i und die f_j Polynome mit Koeffizienten aus k sind, ist dies ein homogenes lineares Gleichungssystem mit Koeffizienten aus k. Es hat genau dann nichttriviale Lösungen über k, wenn der Rang seiner Matrix kleiner ist als die Anzahl der Variablen, was man durch das Verschwinden gewisser Determinanten charakterisieren kann.

Da k in K enthalten ist, können wir dieses Gleichungssystem auch über K betrachten; an den Bedingungen für die Lösbarkeit ändert sich dadurch nichts, denn eine Determinante mit Einträgen aus k verschwindet natürlich in K genau dann, wenn sie in k verschwindet.

Somit muß das Gleichungssystem auch eine nichttriviale Lösung über k haben, es gibt also bereits Elemente $\lambda'_i \in k$ und $\mu_{j\ell} \in k$, die das Gleichungssystem lösen. Damit ist dann

$$\lambda'_1 B_1 + \cdots + \lambda'_r B_r = g'_1 f_1 + \cdots + g'_m f_m$$

mit Polynomen $g'_j \in k[x_1, \ldots, x_n]$, die linke Seite liegt also im Ideal I. Somit ist $\lambda'_1 b_1 + \cdots + \lambda'_r b_r$ der Nullvektor in A. Die λ'_i können nicht allesamt verschwinden, denn ansonsten müßte mindestens ein $\mu_{j\ell} \neq 0$

sein, Null wäre also gleich einer nichttrivialen Linearkombination von Monomen, was absurd ist. Also sind auch die b_i linear abhängig.

Bleibt noch zu zeigen, daß A endlichdimensional ist, wenn $\bar{A}$ endlichdimensional ist. Das folgt sofort aus der gerade gezeigten Äquivalenz der linearen Abhängigkeit über k und über K: Hat $\bar{A}$ die endliche Dimension d, so ist jede Teilmenge von $\bar{A}$ mit mehr als d Elementen linear abhängig. Damit ist, wie wir gerade gesehen haben, auch jede Teilmenge von mehr als d Elementen aus A linear abhängig, also A endlichdimensional.

4. Schritt: Falls $V_K(I)$ endlich ist, gibt es ein homogenes lineares Polynom $u = c_1x_1 + \cdots c_nx_n$ aus $K[x_1, \ldots, x_n]$, das für jeden Punkt aus $V_K(I)$ einen anderen Wert annimmt.

Wir betrachten die Polynome $u_a = x_1 + ax_2 + \cdots + a^{n-1}x_n$ zu den verschiedenen Elementen $a \in K$. Für je zwei verschiedene Nullstellen $z, w \in V_K(I)$ ist $u_a(z) = u_a(w)$ genau dann, wenn

$$(z_1 - w_1) + (z_2 - w_2)a + \cdots + (z_n - w_n)a^{n-1}$$

verschwindet. Die Koordinaten z_i, w_i von z und w sind Elemente von K; die $a \in K$, für die $u_a(z) = u_a(w)$ ist, sind also die Nullstellen eines Polynoms in einer Veränderlichen über K vom Grad höchstens $n - 1$. Daher gibt es höchstens $n - 1$ Werte $a \in K$, für die $u_a(z) = u_a(w)$ ist. Wenn $V_K(I)$ endlich ist, gibt es auch nur endlich viele verschiedene Paare aus voneinander verschiedenen Elementen; somit gibt es nur endlich viele $a \in K$, für die $u_a(z) = u_a(w)$ sein kann für *irgendwelche* voneinander verschiedene Elemente von $V_K(I)$. Da K als algebraisch abgeschlossener Körper unendlich sein muß, gibt es somit Polynome u_a, die für je zwei verschiedene Elemente von $V_K(I)$ verschiedene Werte annehmen. Falls bereits k ein unendlicher Körper ist, können wir sogar entsprechende $a \in k$ finden; in diesem Fall gibt es also schon in $k[x_1, \ldots, x_n]$ solche Polynome.

6. Schritt: Die Elementanzahl r von $V_K(I)$ ist höchstens gleich der Dimension von A.

Da wir im 3. Schritt gesehen haben, daß $\dim_k A = \dim_K \bar{A}$ ist, können wir auch mit dieser Dimension argumentieren. Aus dem 5. Schritt wissen

wir, daß es ein Polynom $u \in K[x_1, \ldots, x_n]$ gibt, das für jedes Element von $V_K(I)$ einen anderen Wert annimmt. Wir ersetzen u durch seine Restklasse $\tilde{u}$ modulo I in $\bar{A}$. Wir wollen uns überlegen, daß die Elemente $1, \tilde{u}, \ldots, \tilde{u}^{r-1} \in \bar{A}$ linear unabhängig sind, wenn $V_K(I)$ mindestens r Elemente enthält: Falls es eine Relation der Form $\sum_{\ell=0}^{r-1} \lambda_\ell \tilde{u}^\ell = 0$ gäbe mit $\lambda_\ell \in k$, so läge das Polynom $\sum_{\ell=0}^{r-1} \lambda_\ell u^\ell$ in $\bar{I}$, würde also für jedes der r Elemente von $V_K(I)$ verschwinden. Da u für jedes dieser Elemente einen anderen Wert annimmt, ist dies bei einem Polynom vom Grad r nur möglich, wenn alle Koeffizienten λ_ℓ verschwinden, was die behauptete lineare Unabhängigkeit beweist. Somit enthält $\bar{A}$ mindestens r linear unabhängige Elemente, d.h. $\dim_K \bar{A} \geq r$. Damit ist die Behauptung und auch der gesamte Satz bewiesen.

■

In der Computeralgebra interessieren wir uns nicht in erster Linie für abstrakte Sätze über Nullstellenmenge und Ideale; wir wollen die Lösungsmengen eines Systems von Polynomgleichungen möglichst explizit angeben. Die beste Chance dazu haben wir, wenn die Lösungsmenge endlich ist; daher interessieren wir uns für möglichst einfache Kriterien dafür, daß $V_K(I)$ eine endliche Menge ist. (Die eigentlich sehr viel interessantere Frage nach der Endlichkeit von $V_k(I)$ ist um soviel schwieriger zu beantworten, daß sie jenseits unserer Ambitionen bleiben muß; halbwegs allgemeine Resultate hierzu sind zumindest derzeit unbekannt.)

Wie wir gerade gesehen haben, ist diese Endlichkeit äquivalent zur Endlichdimensionalität des Vektorraums A; wir suchen daher Kriterien, die dies garantieren. Da wir uns im Kapitel über GRÖBNER-Basen befinden, sollten diese auch damit etwas zu tun haben.

Wir betrachten daher zwar weiterhin ein Gleichungssystem der Form

$$f_j(x_1, \ldots, x_n) = 0 \quad \text{für } j = 1 \ldots, m \quad \text{mit} \quad f_j \in k[_1, \ldots, x_n]\,,$$

nehmen aber an, daß wir eine GRÖBNER-Basis G des von den Polynomen $f_j \in k[x_1, \ldots, x_n]$ erzeugten Ideals I bezüglich irgendeiner Monomordnung kennen.

Wir müssen dann entscheiden, ob der Vektorraum $A = k[x_1, \ldots, x_n]/I$ endliche Dimension hat. Im eindimensionalen Fall ist das einfach: I ist

dann ein Hauptideal, eine reduzierte GRÖBNER-Basis besteht nur aus einem Element, und wenn dieses ein Polynom vom Grad d ist, hat $A = k[x]/I$ die Restklassen der Elemente $1, x, \ldots, x^{d-1}$ als Basis.

Ähnlich können wir auch im Falle mehrerer Veränderlicher argumentieren: Wenden wir den Divisionsalgorithmus an auf ein beliebiges Polynom und die GRÖBNER-Basis, erhalten wir eine Darstellung des Polynoms als Summe einer Linearkombination mit Koeffizienten aus $k[x_1, \ldots, x_n]$ von Elementen der GRÖBNER-Basis und einem Rest. Dieser ist eine k-Linearkombination von Monomen, die durch kein führendes Monom eines Elements der GRÖBNER-Basis teilbar sind. Somit bilden diese Monome eine Basis des Vektorraums A; falls es nur endlich viele davon gibt, ist A endlichdimensional.

Einfacher ist das folgende Kriterium:

Lemma: $V_K(I)$ ist genau dann endlich, wenn die GRÖBNER-Basis von I (bezüglich irgendeiner Monomordnung) für jedes i ein Polynom mit einer x_i-Potenz als führenden Term enthält.

Beweis: Falls die GRÖBNER-Basis für jedes i ein Polynom mit führendem Monom $x_i^{d_i}$ enthält, ist jedes Monom, in dem ein x_i mit einem Exponenten größer oder gleich d_i vorkommt, durch das führende Monom eines Elements der GRÖBNER-Basis teilbar. Die Monome, für die das nicht der Fall ist, haben für jedes i einen Exponenten echt kleiner d_i; es gibt also nur endlich viele solche Monome. Somit hat A endliche Dimension, und $V_K(I)$ ist endlich.

Ist umgekehrt $V_K(I)$ endlich, so enthält $\bar{I}$ für jedes i ein Polynom aus $K[x_i]$ – siehe Schritt 2 im Beweis des obigen Satzes. Da die GRÖBNER-Basis von I gleichzeitig eine GRÖBNER-Basis von $\bar{I}$ ist, muß das führende Monom eines ihrer Elemente die höchste x_i-Potenz in diesem Polynom teilen, muß also selbst eine Potenz von x_i sein. ∎

§8: Multiplizitäten

Um, wie im eindimensionalen Fall, statt einer Ungleichung eine Gleichung für die Anzahl der Nullstellen zu bekommen, müssen wir diesen

eine Vielfachheit oder, wie man auch sagt, Multiplizitäten zuordnen. Im Falle von Polynomen einer Veränderlichen können wir diese mit Hilfe von Ableitungen definieren; falls wir über den reellen Zahlen arbeiten, reicht zur Bestimmung der Multiplizität daher die Kenntnis einer beliebig kleinen ε-Umgebung.

In der Algebra haben wir keine ε-Umgebungen, aber wir können uns auch mit algebraischen Methoden auf die Umgebung eines Punktes konzentrieren: Wir betrachten einfach an Stelle von Polynomen beliebige rationale Funktionen, von denen wir nur verlangen, daß der Nenner im betrachteten Punkt nicht verschwindet.

Sei zunächst $f \in k[x]$ ein Polynom einer Veränderlichen, das bei $x = z$ eine r-fache Nullstelle habe. Dann ist $f(x) = (x-z)^r g(x)$ mit einem Polynom $g \in k[x]$, das bei $x = z$ nicht verschwindet. Der im vorigen Paragraphen eingeführte Faktorraum $A = K[x]/(f)$ hat als Basis die Potenzen x^ℓ mit $0 \le \ell < \deg f$; alternativ können wir natürlich auch die entsprechenden Potenzen $(x-z)^\ell$ nehmen. Dann verschwindet ein Element von A genau dann im Punkt z, wenn es im von den $(x-z)^\ell$ mit $\ell > 0$ aufgespannten Untervektorraum liegt.

Wenn wir alle anderen Elemente von A als Nenner zulassen, sollte man zunächst erwarten, daß A dadurch größer wird. Tatsächlich aber ist das Gegenteil der Fall: Wenn wir von den üblichen Regeln der Bruchrechnung ausgehen, ist beispielsweise

$$(x-z)^r = \frac{(x-z)^r}{1} = \frac{(x-z)^r g}{g} = \frac{f}{g} = \frac{0}{g} = 0\,,$$

denn wir rechnen ja modulo f, und g ist als Nenner zugelassen, da $g(z)$ nicht verschwindet. Entsprechendes gilt für alle $(x-z)^\ell$ mit $\ell \ge r$, nicht aber für die mit $\ell < r$, denn hier bräuchten wir ja noch mindestens einen Faktor $(x-z)$, um im Zähler auf f zu kommen, und Funktionen, die in z verschwinden, sind im Nenner nicht erlaubt. Durch das Einführen solcher Nenner verringert sich also die Dimension von A; der neue Vektorraum hat nur noch die Dimension r, was gleich der Vielfachheit der Nullstelle z ist. Wir können ihn über die Basis aus den $(x-z)^\ell$ mit $\ell < r$ identifizieren mit einem r-dimensionalen Untervektorraum

von $\bar{A}$, und die Dimensionen der so definierten Unterräume zu den verschiedenen Nullstellen von f ergänzen sich zur Dimension von $\bar{A}$.

Für Polynome einer Veränderlichen ist das sicherlich eine sehr umständliche Art der Betrachtung; sie hat aber den Vorteil, daß sie sich auf Polynome in mehreren Veränderlichen verallgemeinern läßt.

Als erstes müssen wir klar definieren, was oben kurz als die „Einführung von Nennern“ bezeichnet wurde:

Definition: R sei ein (kommutativer) Ring.
a) Eine Teilmenge $S \subseteq R \setminus \{0\}$ heißt *multiplikativ abgeschlossen,* wenn sie mit je zwei Elementen $f, g \in S$ auch deren Produkt enthält.
b) Die *Lokalisierung* von R nach der multiplikativ abgeschlossenen Menge S ist die Menge aller Paare $(f, g) \in R \times S$ modulo der folgenden Äquivalenzrelation:

$$(f, g) \sim (f', g') \Longleftrightarrow \exists h \in R \setminus \{0\} : h(fg' - f'g) = 0 \,.$$

Die Äquivalenzklasse des Paars (f, g) wird mit $\frac{f}{g}$ bezeichnet, die Menge aller Äquivalenzklassen mit $S^{-1}R$. Sie wird zum Ring durch die Verknüpfungsdefinitionen

$$\frac{f}{g} + \frac{f'}{g'} = \frac{fg' + f'g}{gg'} \quad \text{und} \quad \frac{f}{g} \cdot \frac{f'}{g'} = \frac{ff'}{gg'} \,.$$

Man sollte sich kurz überlegen, daß diese Verknüpfungen wohldefiniert sind, daß das Ergebnis also nicht von der Wahl spezieller Repräsentanten (f, g) und (f', g') abhängt; im wesentlichen ist dies die gleiche Rechnung wie bei der Einführung des Quotientenkörpers in Kapitel 2§6.

Falls R nullteilerfrei ist, können wir in der Definition der Äquivalenzrelation auf des Element h verzichten, denn wenn $h(fg' - f'g)$ für ein $h \neq 0$ verschwindet, muß der zweite Faktor Null sein. Da unsere Ringe A und $\bar{A}$ im allgemeinen nicht nullteilerfrei sind, ist – wie wir gerade am Beispiel der Polynome einer Veränderlichen gesehen haben – die Möglichkeit zur Erweiterung mit Nullteilern wesentlich.

Die größte multiplikativ abgeschlossene Teilmenge eines Integritätsbereichs R ist $R \setminus \{0\}$; in diesem Fall ist $S^{-1}R$ der Quotientkörper.

Falls R Nullteiler enthält, d.h. Elemente $g \neq 0$, zu denen es ein $h \neq 0$ gibt mit $gh = 0$, ist $R \setminus \{0\}$ nicht mehr multiplikativ abgeschlossen: Zwar liegen g und h in $R \setminus \{0\}$, nicht aber deren Produkt. In diesem Fall besteht die größte multiplikativ abgeschlossene Teilmenge $S \subset R$ aus allen $f \in R$, für die es kein $g \neq 0$ gibt mit $fg = 0$, wir müssen also außer der Null auch noch alle Nullteiler ausschließen. Die Menge $S^{-1}R$ wird in diesem Fall als *vollständiger Quotientenring* von R bezeichnet. Man beachte, daß sich R in so einem Fall nicht injektiv in $S^{-1}R$ einbetten läßt: Für einen Nullteiler h und ein $g \neq 0$ mit $hg = 0$ ist $h/1 = hg/g = 0/g = 0$.

Weitere typische Beispiele multiplikativ abgeschlossener Teilmengen eines Rings sind die Potenzen eines Nichtnullteilers oder auch das Komplement eines Primideals: Ein Ideal $I \lhd R$ heißt *Primideal* wenn für je zwei Elemente $f, g \in R$ mit $fg \in I$ mindestens einer der beiden Faktoren f, g in I liegt. Dies ist offensichtlich äquivalent dazu, daß die Menge $R \setminus I$ multiplikativ abgeschlossen ist.

Wir interessieren uns für Ideale $I \lhd k[x_1, \ldots, x_n]$, für die $V_K(I)$ eine endliche Menge ist; dabei bezeichnet K wie üblich einen algebraisch abgeschlossenen Körper, der k enthält. Die Elemente der Vektorräume $A = k[x_1, \ldots, x_n]/I$ und $\bar{A} = K[x_1, \ldots, x_n]/\bar{I}$ können wir als Funktionen auf $V_K(I)$ mit Werten in K interpretieren. Da sich Funktionen miteinander multiplizieren lassen, sind auch A und $\bar{A}$ Ringe, deren Multiplikation offensichtlich mit der im Polynomring kompatibel ist. Für jedes $z \in V_K(I)$ ist die Menge

$$S_z = \left\{ f \in \bar{A} \;\middle|\; f(z) \neq 0 \right\}$$

multiplikativ abgeschlossen, denn die Funktionswerte liegen ja im (nullteilerfreien) Körper K. Diese Lokalisierungen wollen wir im folgenden genauer untersuchen.

Definition: *a)* $\bar{A}_z \underset{\text{def}}{=} S_z^{-1}\bar{A}$
b) Die *Vielfachheit* oder *Multiplizität* einer Nullstelle $z \in V_K(I)$ ist die Dimension von $\bar{A}_z$ als K-Vektorraum.

Wie wir oben gesehen haben, entspricht dies für Polynome einer Veränderlichen der gewohnten Vielfachheit; wir wollen uns überlegen,

daß sich die Vielfachheiten der verschiedenen Elemente von $V_K(I)$ auch im Falle von Polynomen mehrerer Veränderlichen zu $\dim_K \bar{A}$ addieren.

Dazu benötigen wir noch einen Begriff aus der Linearen Algebra:

Definition: $V_1, \ldots, V_r$ seien Vektorräume über dem Körper k. Die direkte Summe

$$\bigoplus_{i=1}^{r} V_i = V_1 \oplus \cdots \oplus V_r$$

ist als Menge gleich dem kartesischen Produkt $V_1 \times \cdots V_r$ der Vektorräume; die Vektorraumaddition ist definiert durch

$$(v_1, \ldots, v_r) + (w_1, \ldots, w_r) = (v_1 + w_1, \ldots, v_r + w_r)\,,$$

und für die Multiplikation mit einem Skalar $\lambda \in k$ setzen wir

$$\lambda(v_1, \ldots, v_n) = (\lambda v_1, \ldots, \lambda v_n)\,.$$

Die Vektorräume V_i können identifiziert werden mit jenen Untervektorräumen von $\oplus_{i=1}^{r} V_i$, in denen alle Komponenten außer eventuell der i-ten gleich dem Nullvektor sind.

Wenn alle Räume V_i endliche Dimensionen haben, ist die Dimension ihrer direkten Summe offensichtlich einfach die Summe dieser Dimensionen: Wählen wir in jedem der Vektorräume V_i eine Basis und fassen wir V_i auf als Untervektorraum der direkten Summe, so ist die Vereinigung der Basen der V_i offensichtlich eine Basis des Summenraums. Insbesondere ist jeder endlichdimensionale k-Vektorraum mit einer Basis $b_1, \ldots, b_n$ isomorph zur direkten Summe der eindimensionalen Untervektorräume kb_i.

Satz: Ist $V_K(I)$ endlich, so ist $\bar{A} \cong \bigoplus_{z \in V_K(I)} \bar{A}_z$

Beweis: Wie wir aus dem vorigem Paragraphen wissen (4. Schritt im Beweis des letzten Satzes), gibt es ein homogenes lineares Polynom über K, das für jeden Punkt aus $V_K(I)$ einen anderen Wert annimmt. Durch einen linearen Koordinatenwechsel können wir erreichen, daß x_1

diese Eigenschaft hat. Wir bezeichnen die x_1-Koordinate eines Punktes $z \in V_K(I)$ mit z_1 und betrachten die LAGRANGE-Polynome

$$s_z = \frac{\prod_{w \in V_K(I) \setminus \{z\}} (x_1 - w_1)}{\prod_{w \in V_K(I) \setminus \{z\}} (z_1 - w_1)} \in K[x_1] ;$$

offensichtlich ist $s_z(z) = 1$ und $s_z(w) = 0$ für alle $w \neq z$ aus $V_K(I)$. Somit verschwindet das Produkt $s_z s_w$ zweier solcher Funktionen in jedem Punkt von $V_K(I)$; nach dem HILBERTschen Nullstellensatz liegt daher eine Potenz von $s_z s_w$ im Ideal $\bar{I}$. Bezeichnet r den größten Exponenten, den wir für eines der Produkte $s_z s_w$ brauchen, haben daher die Polynome $t_z = s_z^r$ die Eigenschaft, daß $t_z t_w$ für $z \neq w$ in $\bar{I}$ liegt, und $t_z(z) = 1$.

Wir betrachten nun das Ideal J von $K[x_1, \ldots, x_n]$, das von I und den sämtlichen t_z erzeugt wird. Es hat offensichtlich keine gemeinsame Nullstelle, denn die gemeinsamen Nullstellen von $\bar{I}$ sind die $z \in V_K(I)$, und für jedes dieser z ist $t_z(z) = 1$. Nach der schwachen Form des HILBERTschen Nullstellensatzes enthält J daher die Eins; es gibt also Polynome $p_z \in K[x_1, \ldots, x_n]$ und ein Polynom $p \in \bar{I}$, so daß

$$\sum_{z \in V_K(I)} p_z t_z + p = 1$$

ist. Die Restklassen $e_z \in \bar{A}$ von $p_z t_z$ modulo $\bar{I}$ erfüllen die Gleichungen

1.) $\sum_{z \in V_K(I)} e_z = 1$
2.) $e_z^2 = e_z$
3.) $e_z(z) = 1$
4.) $e_z e_w = 0$ für $z \neq w$ aus $V_K(I)$

Die einzige noch nicht gezeigte Aussage ist 2.); sie folgt aus der Gleichung

$$e_z - e_z^2 = e_z(1 - e_z) = e_z \sum_{w \neq z} e_w = \sum_{w \neq z} e_z e_w = 0 .$$

Elemente e eines Rings R mit der Eigenschaft $e^2 = e$ bezeichnet man als *Idempotente;* sie haben die Eigenschaft, daß das Ideal $(e) = Re$ selbst ein Ring ist mit e als der Eins, denn $(ae)(be) = abe^2 = abe$ für alle $a, b \in R$.

Wir wollen uns als nächstes überlegen, daß der Ring $\bar{A}e_z$ isomorph ist zur Lokalisierung von $\bar{A}$ bei z; der Isomorphismus ist gegeben durch

$$\begin{cases} \bar{A}e_z \to \bar{A}_z \\ fe_z \mapsto \dfrac{f}{1} \end{cases} .$$

Zum Nachweis der Bijektivität konstruieren wir eine Umkehrabbildung $\bar{A}_z \to \bar{A}e_z$ wie folgt: Zu jedem $g \in \bar{A}$ mit $g(z) \neq 0$ setzen wir

$$\tilde{g} \underset{\text{def}}{=} \frac{g}{g(z)} - 1 \in \bar{A}_z, \quad \text{d.h.} \quad g = g(z)(1+\tilde{g}) .$$

Da $\tilde{g}(z)$ verschwindet und $e_z(w) = 0$ für alle $w \neq z$, verschwindet $\tilde{g}e_z$ auf ganz $V_K(I)$. Nach dem HILBERTschen Nullstellensatz gibt es somit eine Potenz eines Repräsentanten, die in $\bar{I}$ liegt, d.h. es gibt eine natürliche Zahl N, so daß $\big(\tilde{g}e_z\big)^N = \tilde{g}^N e_z$ die Null von $\bar{A}$ ist. Dann ist

$$(1+\tilde{g})e_z \cdot \big(1 - \tilde{g} + \tilde{g}^2 - \cdots + (-1)^{N-1}\tilde{g}^{N-1}\big)e_z = (1 - \tilde{g}^N)e_z = e_z ;$$

im Ring $\bar{A}_z$ hat also $1+\tilde{g}$ ein Inverses und damit auch $ge_z = g(z)(1+\tilde{g})e_z$. Wir bilden daher den Bruch $f/g \in \bar{A}_z$ ab auf

$$f \cdot \frac{1}{g(z)} \cdot \big(1 - \tilde{g} + \tilde{g}^2 - \cdots + (-1)^{N-1}\tilde{g}^{N-1}\big)e_z \in \bar{A}_z ,$$

und mit Hilfe der gerade durchgeführten Rechnung folgt leicht, daß die beiden Abbildungen zueinander invers, also Isomorphismen sind.

Zum Beweis des Satzes fehlt nun nur noch, daß $\bar{A}$ die direkte Summe der Ringe $\bar{A}e_z$ ist; das ist klar, da die Summe der e_z gleich eins ist und $e_z e_w = 0$ für $z \neq w$.

■

Weiteres über den Umgang mit Multiplizitäten und die Lösung nichtlinearer Gleichungssysteme mit endlicher Lösungsmenge findet man zum Beispiel in dem Übersichtsartikel

LAUREANO GONZALEZ-VEGA, FABRICE ROUILLIER, MARIE-FRANÇOISE ROY: Symbolic Recipes for Polynomial System Solving *in:* ARJEH M. COHEN, HANS CUYPERS, HANS STERK [EDS.]: Some Tapas of Computer Algebra, *Springer,* 1999,

dessen Anfangsteil ich in diesem und dem vorigen Paragraphen weitgehend folgte.