Wolfgang K. Seiler

Computeralgebra

Vorlesung im Herbstsemester 2017
an der Universität Mannheim

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

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

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

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

Kapitel 0
Einführung

§1: Was ist Computeralgebra

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.

§2: Numerisches, exaktes und symbolisches Rechnen

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

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

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

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

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

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

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

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

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

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,

$$1{,}0000, \quad 2{,}0000, \quad 3{,}0000, \quad 4{,}0000, \quad 5{,}0000,$$

$$6{,}0000,\quad 7{,}0000,\quad 8{,}0001,\quad 8{,}9992,\quad 10{,}008,$$
$$10{,}957,\quad 12{,}383 \pm 0{,}10867i,\quad 14{,}374 \pm 0{,}77316i,$$
$$16{,}572 \pm 0{,}88332i,\quad 18{,}670 \pm 0{,}35064i,\quad 20{,}039\,.$$

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

§3: Unentscheidbarkeitsprobleme

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

$$\frac{10}{15} = \frac{2}{3}\,,\quad \sqrt{8} = 2\sqrt{2}\,,\quad \sqrt{4+2\sqrt{3}} = 1+\sqrt{3}$$
$$(a+b)^2 = a^2 + 2ab + b^2\,,\quad \frac{x^5-1}{x-1} = 1 + x + x^2 + x^3 + x^4\,,$$
$$X^5 - 15X^4 + 85X^3 - 225X^2 + 274X - 120$$
$$= (X-1)(X-2)(X-3)(X-4)(X-5)\,,$$
$$\sin x \cos x = \frac{\sin 2x}{2}\,,\quad 1 + \tan^2 x = \frac{1}{\cos^2 x}$$

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

Andererseits gehört es zu den Grundaufgaben jeglicher Art des Rechnens, daß man entscheiden muß, ob zwei Ausdrücke gleich sind. Dies

ist dann am einfachsten, wenn jeder Ausdruck intern durch eine eindeutig bestimmte kanonische Form dargestellt wird. In einem System, daß alle Ergebnisse auf eine solche kanonische Form bringt, lassen sich zwei Ausdrücke einfach dadurch auf Gleichheit testen, daß man ihre Differenz berechnet; die Ausdrücke sind genau dann gleich, wenn das Ergebnis die kanonische Darstellung der Null ist.

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

In älteren Computeralgebrasystemen wie REDUCE war es üblich, alles auszumultiplizieren; in den heute gebräuchlichen Systemen wie MAPLE und MATHEMATICA werden Umformungen nur noch durchgeführt, wenn es entweder für die jeweilige Rechnung notwendig ist (Zur Berechnung der Stammfunktion eines Polynoms muß dieses in ausmultiplizierter Form vorliegen) oder wenn es der Anwender explizit verlangt. Lediglich in einigen offensichtlichen Fällen bemühen sich auch diese Systeme um Normalisierung: Beispielsweise werden Brüche stets in gekürzter Form dargestellt und bei Summen werden gleichartige Terme zusammengefaßt.

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

Dieses negative Ergebnis kam hat seinen Ausgangspunkt in einem po-

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

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

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

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

Satz von Richardson: Es gibt kein Verfahren, das in endlich vielen Schritten entscheidet, ob ein beliebig vorgegebener Ausdruck bestehend aus rationalen Zahlen, π, einer Variablen x sowie den Funktionen +, $\cdot$, Sinus und Betrag gleich Null ist.

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

Mehr zum zehnten HILBERTschen Problem und seinen Konsequenzen findet man bei

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

Kapitel 1
Gröbner-Basen

Die klassische Aufgabe der Algebra besteht in der Lösung von Gleichungen und Gleichungssystemen. Im Falle eines Systems von Polynomgleichungen in mehreren Veränderlichen kann die Lösungsmenge sehr kompliziert sein und, sofern sie unendlich ist, möglicherweise nicht einmal explizit angebbar: Im Gegensatz zum Fall linearer Gleichungen können wir hier im allgemeinen keine endliche Menge von Lösungen finden, durch die sich alle anderen Lösungen ausdrücken lassen. Trotzdem gibt es Algorithmen, mit denen sich nichtlineare Gleichungssysteme deutlich vereinfachen lassen, und zumindest bei endlichen Lösungsmengen lassen sich diese auch konkret angeben – sofern wir die Nullstellen von Polynomen einer Veränderlichen explizit angeben können.

§ 1: Algebraische Vorbereitungen

Wenn wir lineare Gleichungssysteme mit dem GAUSS-Algorithmus lösen, verändern wir das Gleichungssystem sukzessive, indem wir Gleichungen so durch Linearkombinationen mit anderen Gleichungen ersetzen, daß sich an der Lösungsmenge nichts ändert. Indem wir eine lineare Gleichung

$$a_1 X_1 + \cdots + a_n X_n = b$$

über einem Körper k mit dem $(n+1)$-Tupel $(a_1, \ldots, a_n, b) \in k^{n+1}$ identifizieren, sehen wir leicht, daß die sämtlichen linearen Gleichungen in n Unbekannten über einem Körper k einen $(n+1)$-dimensionalen Vektorraum bilden; die Gleichungen eines konkreten linearen Gleichungssystems erzeugen darin einen Untervektorraum. Dieser besteht aus allen

Linearkombinationen der gegebenen Gleichungen, und das sind gleichzeitig alle linearen Gleichungen, die auf der Lösungsmenge des linearen Gleichungssystems verschwinden. Zwei lineare Gleichungssysteme haben somit genau dann die gleiche Lösungsmenge, wenn sie den gleichen Untervektorraum erzeugen.

Wenn wir Systeme nichtlinearer Gleichungen betrachten, ist es sinnvoll, die Menge aller möglicher Gleichungen nicht mehr nur als Vektorraum zu betrachten, sondern auch die Multiplikation mit Polynomen zuzulassen: Zur Lösung des Gleichungssystems

$$X^2Y^2 + 2X^3 - 3X^2 - X = 0 \quad \text{und} \quad Y^2 + X - 3 = 0$$

bietet sich etwa an, die zweite Gleichung mit X^2 zu multiplizieren und das Produkt $X^2Y^2 + X^3 - 3X^2 = 0$ von der ersten Gleichung zu subtrahieren; die Differenz $X^3 - X$ hängt nur noch von X ab und verschwindet bei 0 und ± 1. Setzen wir dies in die zweite Gleichung ein, erhalten wir die Lösungsmenge

$$\left\{ \left(0, \sqrt{3}\right), \left(0, -\sqrt{3}\right), \left(1, \sqrt{2}\right), \left(1, -\sqrt{2}\right), (-1, 2), (-1, -2) \right\}.$$

Wir sollten die Menge aller möglicher Gleichungen daher nicht mehr nur als einen Vektorraum betrachten, sondern als einen *Ring* im Sinne der folgenden Definition:

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.

Für die Betrachtung nichtlinearer Gleichungssysteme interessieren uns allerdings vor allem Polynomringe. Da auch diese kommutativ sind, vereinbaren wir:

Wenn nicht explizit etwas anderes gesagt wird, soll *Ring* im folgenden stets für einen *kommutativen Ring* stehe.

Definition: R sei ein Ring, und $X_1, \ldots, X_n$ seien n Symbole, die nicht in R liegen.
a) Ein *Monom* ist ein Produkt $X_1^{\alpha_1} \cdots X_n^{\alpha_n}$ mit nichtnegativen ganzen Zahlen $\alpha_1, \ldots, \alpha_n$. Die Summe der α_i bezeichnen wir als den *Grad* des Monoms.
b) Ein *Polynom* über R in den Variablen $X_1, \ldots, X_n$ ist eine endliche Linearkombination f von Monomen mit Koeffizienten aus R. Falls diese nicht Null ist, bezeichnen wir den größten Grad eines in f vorkommenden Monoms als den *Grad* deg f von f. Für das Polynom $f = 0$ definieren wir keinen Grad. *c)* Die Menge aller Polynome über R in den Variablen $X_1, \ldots, X_n$ bezeichnen wir als den *Polynomring* $R[X_1, \ldots, X_n]$ über R in den Variablen $X_1, \ldots, X_n$.

Es ist klar, daß $R[X_1, \ldots, X_n]$ mit der offensichtlichen Addition und Multiplikation ein Ring ist. Wir nehmen dabei natürlich an, daß die X_i untereinander kommutieren.

Wir interessieren uns vor allem für Polynomringe über Körpern; für Induktionsbeweise ist es aber oft nützlich, beispielsweise den Polynomring $k[X, Y]$ aufzufassen als den Polynomring in Y über dem Ring $R = k[X]$; daher die allgemeinere Definition.

Wie wir beim obigen Beispiel eines nichtlinearen Gleichungssystems gesehen haben, kann es bei der Lösung nützlich sein, nicht nur skalare Linearkombinationen der Gleichungen zu betrachten, sondern auch solche mit beliebigen Polynomen als Koeffizienten. Anstelle von Untervektorräumen des Polynomrings $k[X_1, \ldots, X_n]$ sollten wir daher Strukturen betrachten, in denen man Linearkombinationen mit beliebigen Ringelementen als Koeffizienten bilden kann, die sogenannten Ideale:

Definition: Eine nichtleere Teilmenge I eines Rings R heißt *Ideal*, in Zeichen $I \lhd R$, wenn gilt:
1.) Für je zwei Elemente $f, g \in I$ ist auch $f + g \in I$
2.) Für jedes $f \in I$ und jedes $r \in R$ liegt auch rf in I.

Bei den Produkten verlangen wir also, daß sie bereits dann in I liegen, wenn nur *ein* Faktor in I liegt.

Die Bedingung, daß ein Ideal mindestens ein Element enthalten muß, können wir auch ersetzen durch die Bedingung, daß es die Null von R enthalten muß, denn wenn es irgendein Element $f \in R$ enthält, muß es gemäß der zweiten Bedingung auch $0 \cdot f = 0$ enthalten.

Um mit dem Idealbegriff vertraut zu werden, betrachten wir zunächst Ideale im Ring der ganzen Zahlen:

Lemma: Zu jedem Ideal $I \lhd \mathbb{Z}$ gibt es eine ganze Zahl $n \in \mathbb{Z}$, so daß $I = \{nq | q \in \mathbb{Z}\}$.

Beweis: I ist nach Definition nicht leer, enthält also mindestens ein Element. Falls I nur aus der Null besteht, können wir $n = 0$ setzen und sind fertig. Wenn es ein Element $m \neq 0$ gibt, enthält das Ideal auch dessen sämtliche ganzzahlige Vielfachen, insbesondere also gibt es in I dann positive Zahlen. Die kleinste dieser Zahlen sei n. Wir wollen uns überlegen, daß I genau aus den ganzzahligen Vielfachen von n besteht.

Dazu sei $m \in I$ ein beliebiges Element von I. Wir dividieren m mit Rest durch q; das Ergebnis sei

$$m : n = q \quad \text{Rest } r \quad \text{mit} \quad 0 \leq r < n \,.$$

Dann liegt mit m und n auch $r = m - qn$ in I und ist echt kleiner als n. Da n die kleinste positive Zahl in I ist, muß daher $r = 0$ sein, d.h. $m = qn$ ist ein ganzzahliges Vielfaches von n.

■

Definition: *a)* Ist R ein Ring und $f \in R$ so bezeichnen wir

$$(f) \underset{\text{def}}{=} \{rf \mid r \in R\}$$

als das von f erzeugte *Hauptideal.*
b) R heißt *Hauptidealring,* wenn jedes Ideal von R ein Hauptideal ist.

Das gerade bewiesene Lemma zeigt also, daß $\mathbb{Z}$ ein Hauptidealring ist.

Allgemeiner definieren wir

Definition: Ist R ein Ring und ist $M \subset R$ eine Teilmenge von R, so ist das *von M erzeugte Ideal* (M) das kleinste Ideal von R, das M enthält, d.h. den Durchschnitt aller Ideale, die M enthalten. Für eine endliche Menge $M = \{f_1, \ldots, f_m\}$ schreiben wir (M) kurz als $(f_1, \ldots, f_m)$. Die Menge M bezeichnen wir als ein *Erzeugendensystem* des Ideals I.

Diese Definition macht nicht wirklich klar, wie das von M erzeugte Ideal aussieht. Da uns in der Computeralgebra nur endlich erzeugte Ideale interessieren, möchte ich mich auf diesen Fall beschränken; die Verallgemeinerung auf beliebige Mengen M sollte für jeden, der den folgenden Beweis verstanden hat, offensichtlich sein.

Lemma: $\left(f_1, \ldots, f_m\right) = \left\{ \sum_{i=1}^{m} r_i f_i \;\middle|\; r_i \in R \right\}$

Beweis: Da jedes Ideal, das $f_1, \ldots, f_m$ enthält, auch für $r_1, \ldots, r_m \in R$ die Elemente $r_i f_i$ enthält und damit auch deren Summe, ist klar, daß die rechte Seite in jedem Ideal enthalten ist, das die f_i enthält. Außerdem ist die rechtsstehende Menge selbst ein Ideal: Da sie die f_i enthält, ist sie nicht leer; die Summe zweier Elemente ist offensichtlich wieder ein Element, da wir einfach die Koeffizienten addieren müssen, und wenn wir ein Element mit einem beliebigen Element $r \in R$ multiplizieren,

werden einfach alle Koeffizienten mit r multipliziert. Somit ist die rechte Seite in der Tat das kleinste Ideal, das alle f_i enthält. ∎

Sei nun $R = k[X_1, \ldots, X_n]$ der Polynomring in n Variablen über einem Körper k,und seien $f_1, \ldots, f_m \in R$ Polynome. Wir interessieren uns für die Lösungsmenge des durch die f_i gegebenen Gleichungssystems, also die Menge aller $(x_1, \ldots, x_n) \in k^n$, für die alle f_i verschwinden. Wir definieren gleich allgemein

Definition: Die Nullstellenmenge einer Teilmenge $M \subseteq k[X_1, \ldots, X_n]$ ist

$$V(M) \underset{\text{def}}{=} \left\{(x_1, \ldots, x_n) \in k^n \;\middle|\; f(x_1, \ldots, x_n) = 0 \text{ für alle } f \in M\right\}.$$

Im Falle eine endlichen Menge $M = \{f_1, \ldots, f_m\}$ schreiben wir kurz $V(f_1, \ldots, f_m)$.

(In der algebraischen Geometrie bezeichnet man Mengen dieser Art als Varietäten; daher der Buchstabe V.)

Lemma: Ist $I = (f_1, \ldots, f_m)$ das von den f_i erzeugte Ideal, so ist

$$V(I) = V(f_1, \ldots, f_m).$$

Beweis: Da alle f_i in I liegen, ist natürlich $V(I) \subseteq V(f_1, \ldots, f_m)$. Umgekehrt sei $(x_1, \ldots, x_n)$ ein Element von $V(f_1, \ldots, f_m)$ und g irgendein Element von I. Nach dem vorigen Lemma gibt es Polynome $r_i \in R$; so daß $g = \sum_{i=1}^m r_i f_i$ ist. Damit ist auch

$$g(x_1, \ldots, x_n) = \sum_{i=1}^{m} r_i(x_1, \ldots, x_m) f_i(x_1, \ldots, x_m) = 0,$$

so daß $(x_1, \ldots, x_n)$ in $V(I)$ liegt. Damit ist das Lemma bewiesen. ∎

Dieses Lemma zeigt, daß zwei Gleichungssysteme

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

und

$$g_1(x_1, \ldots, x_n) = 0, \quad \ldots, \quad g_r(x_1, \ldots, x_n) = 0$$

die gleiche Lösungsmenge haben, wenn die Ideale $(f_1, \ldots, f_m)$ und $(g_1, \ldots, g_r)$ übereinstimmen.

Die Umkehrung dieser Aussage ist allerdings falsch. Ein einfaches Gegenbeispiel haben wir bereits bei nur einer Gleichung in einer Variablen: Die Gleichungen

$$x = 0, \qquad x^2 = 0, \qquad x^3 = 0, \qquad \ldots$$

haben allesamt nur die Null als Lösung, aber natürlich sind die Ideale $\left(x^d\right) \lhd k[X]$ für verschiedene Werte von d verschieden. Später werden wir diese Frage, wann so etwas vorkommt, genauer untersuchen.

Zum Abschluß dieses Paragraphen soll nur noch kurz festgehalten werden, wie sich Ideale und Nullstellenmengen zueinander verhalten. Dazu müssen wir zunächst die Summe und das Produkt zweier Ideale definieren:

Definition: *a)* Die Summe $I + J$ zweier Ideale I, J eines Rings R ist das kleinste Ideal, das sowohl I als auch J enthält.
b) Das Produkt IJ dieser Ideale ist das kleinste Ideal, das alle Produkte fg mit $f \in I$ und $g \in J$ enthält.

Man überlegt sich leicht (mit dem gleichen Argument, mit dem wir das Ideal $(f_1, \ldots, f_m)$ oben explizit bestimmt haben), daß $I + J$ gerade die Menge aller $f + g$ mit $f \in I$ und $g \in J$ ist; IJ dagegen enthält im allgemeinen auch Elemente, die sich *nicht* in der Form fg mit $f \in I$ und $g \in J$ darstellen lassen: Ist etwa $I = J = (X, Y) \lhd \mathbb{R}[X, Y]$, so enthält IJ mit $X^2 = X \cdot X$ und $Y^2 = Y \cdot Y$ auch deren Summe $X^2 + Y^2$, die sich nicht als Produkt zweier Polynome aus $\mathbb{R}[X, Y]$ schreiben läßt. Wenn wir $\mathbb{R}$ durch $\mathbb{C}$ ersetzen, läßt sich $X^2 + Y^2$ zwar zerlegen als $(X + iY)(X - iY)$, aber auch in $\mathbb{C}[X, Y]$ gibt es irreduzible Polynome in $(X, Y) \cdot (X, Y)$, die sich somit nicht als Produkt darstellen lassen. In IJ liegen daher auch alle (endlichen) Summen der Form $\sum f_i g_i$ mit $f_i \in I$ und $g_i \in J$; da diese (analog zum obigen Argument) ein Ideal bilden, besteht IJ genau aus diesen Summen.

Satz: Für zwei Ideale I, J im Polynomring $R = k[X_1, \ldots, X_n]$ gilt
a) Ist $I \subseteq J$, so ist $V(J) \subseteq V(I)$

b) $V(I + J) = V(I) \cap V(J)$
c) $V(IJ) = V(I) \cup V(J)$

Beweis: a) Sei $(x_1, \ldots, x_n) \in V(J)$. Dann verschwindet $f(x_1, \ldots, x_n)$ für alle $f \in J$, erst recht also für alle $f \in I$, d.h. $(x_1, \ldots, x_n) \in V(I)$.

b) Da $I + J$ das kleinste Ideal ist, das sowohl I als auch J enthält, liegt $V(I + J)$ nach *a)* sowohl in $V(I)$ als auch in $V(J)$, also auch in deren Durchschnitt. Liegt umgekehrt ein Punkt $(x_1, \ldots, x_n)$ sowohl in $V(I)$ als auch in $V(J)$, so liegt er auch in $V(I + J)$, denn wie wir gerade gesehen haben, läßt sich jedes Element von $I + J$ schreiben als $f + g$ mit $f \in I$ und $g \in J$, und sowohl f als auch g verschwinden im Punkt $(x_1, \ldots, x_n)$.

c) Da IJ erzeugt wird von den Produkten fg mit $f \in I$ und $g \in J$ und jedes dieser Produkte sowohl in I als auch in J liegt, ist IJ eine Teilmenge sowohl von I als auch von J; somit liegt $V(I) \cup V(J)$ nach *a)* in $V(IJ)$. Umgekehrt sei $(x_1, \ldots, x_n) \in V(IJ)$, liege aber nicht in $V(I)$. Dann gibt es ein $f \in I$ mit $f(x_1, \ldots, x_n) \neq 0$. Für jedes $g \in J$ liegt aber fg in IJ, so daß das Produkt $f(x_1, \ldots, x_n)g(x_1, \ldots, x_n)$ verschwinden muß. Da die Funktionswerte im Körper k liegen und der Faktor $f(x_1, \ldots, x_n)$ nicht verschwindet, muß $g(x_1, \ldots, x_n) = 0$ sein für alle $g \in J$; der Punkt liegt also in $V(J)$. Somit liegt er in jedem Fall in $V(I) \cup V(J)$.

■

§2: Gauß und Euklid

Zur (exakten) Lösung eines linearen Gleichungssystems in mehreren Veränderlichen verwenden wir üblicherweise den GAUSS-Algorithmus. Für die Lösung eines System von Polynomgleichungen höheren Grades in nur einer Veränderlichen können wir den EUKLIDischen Algorithmus verwenden, denn die gemeinsamen Nullstellen zweier Polynome in einer Veränderlichen sind gerade die Nullstellen ihres größten gemeinsamen Teilers, so daß wir das System durch mehrfache Anwendung des EUKLIDischen Algorithmus reduzieren können auf eine einzige Polynomgleichung.

Der um 1966 von BRUNO BUCHBERGER vorgestellte Ansatz zur Lösung nichtlinearer Gleichungssysteme in mehreren Veränderlichen kann als eine Kombination von Ideen hinter dem GAUSSschen Eliminationsverfahren und dem EUKLIDischen Algorithmus aufgefaßt werden; er hat Anwendungen, die weit über das Problem der Lösung nichtlinearer Gleichungssysteme hinausgehen. In der Tat wurde die Grundidee des Verfahrens bereits knapp vor BUCHBERGER, und ohne daß dieser davon wußte, von dem japanischen Mathematiker HEISUKE HIRONAKA entdeckt, der es für ein klassisches Problem der algebraischen Geometrie entwickelte: Für die damit bewiesene sogenannte Auflösung der Singularitäten einer algebraischen Varietät über einem Körper der Charakteristik Null erhielt HIRONAKA 1970 die Fields-Medaille, die höchste Auszeichnung der Mathematik.

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

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = u \quad \text{mit} \quad a_1 \neq 0 \tag{1}$$

$$b_1x_1 + b_2x_2 + \cdots + b_nx_n = v \tag{2}$$

die Variable X_1 mit Hilfe von (1) aus (2) eliminieren wollen, ersetzen wir die zweite Gleichung durch ihre Summe mit $-b_1/a_1$ mal der ersten. Die theoretische Rechtfertigung für diese Umformung besteht darin, daß das Gleichungssystem bestehend aus (1) und (2) sowie das neue Gleichungssystem dieselbe Lösungsmenge haben, und daran ändert sich auch dann nichts, wenn noch weitere Gleichungen dazukommen.

Ähnlich können wir vorgehen, wenn wir ein nichtlineares Gleichungssystem in nur einer Variablen betrachten: Am schwersten sind natürlich die Gleichungen vom höchsten Grad, also versuchen wir, die zu reduzieren auf Polynome niedrigeren Grades. Das kanonische Verfahren

dazu ist die Polynomdivision: Haben wir zwei Polynome

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

mit $e \leq d$, so dividieren wir f durch g, d.h. wir berechnen einen Quotienten q und einen Rest r derart, daß $f = qg + r$ ist und r entweder verschwindet oder kleineren Grad als g hat. Konkret: Bei jedem Divisionsschritt haben wir ein Polynom

$$f = c_\delta X^\delta + c_{\delta-1} X^{\delta-1} + \cdots + c_1 X + c_0 \quad \text{mit} \quad c_\delta \neq 0\,,$$

das wir für $\delta \geq e$ mit Hilfe des Divisors

$$g = b_e X^e + b_{e-1} X^{e-1} + \cdots + b_1 X + b_0$$

reduzieren, indem wir es ersetzen durch

$$f - \frac{b_e}{c_\delta} X^{\delta-e} g\,.$$

Das führen wir so lange fort, bis f auf Null oder ein Polynom von kleinerem Grad als e reduziert ist: Das ist dann der Divisionsrest r. Auch hier ist klar, daß sich nichts an der Lösungsmenge ändert, wenn man die beiden Gleichungen f, g ersetzt durch g, r, denn

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

d.h. f und g verschwinden genau dann für einen Wert x, wenn g und r an der Stelle x verschwinden.

In beiden Fällen ist die Vorgehensweise sehr ähnlich: Wir vereinfachen das Gleichungssystem schrittweise, indem wir eine Gleichung ersetzen durch ihre Summe mit einem geeigneter Vielfachen einer anderen Gleichung.

Dieselbe Strategie wollen wir auch anwenden Systeme von Polynomgleichungen in mehreren Veränderlichen. Erstes Problem dabei ist, daß wir nicht wissen, wie wir die Monome eines Polynoms anordnen sollen und damit, was der führende Term ist. Dazu gibt es eine ganze Reihe verschiedener Strategien, von denen je nach Anwendung mal die eine, mal die andere vorteilhaft ist.

§3: Monomordnungen und der Divisionsalgorithmus

Wir betrachten Polynome in n Variablen $X_1, \ldots, X_n$ über einem Körper k und setzen zur Abkürzung

$$X^\alpha = X_1^{\alpha_1} \cdots X_n^{\alpha_n} \quad \text{mit} \quad \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0 \,.$$

Terme der Form X^α haben wir in §1 als Monome bezeichnet und ihnen die Summe der α_i als Grad zugeordnet.

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 ein Polynom $f = \sum_{\alpha \in I} c_\alpha X^\alpha \in k[X_1, \ldots, X_n]$ mit $c_\alpha \neq 0$ für alle $\alpha \in I \subset \mathbb{N}_0$ sei γ das größte Element von I bezüglich einer fest gewählten Monomordnung. Dann bezeichnen wir bezüglich dieser Monomordnung

- γ = multideg f als Multigrad von f
- X^γ = FM(f) als führendes Monom von f
- c_γ = FK(f) als führenden Koeffizienten von f
- $c_\gamma X^\gamma$ = FT(f) als führenden Term von f

Der Grad deg f von f ist, wie in der Algebra üblich, der höchste Grad eines Monoms von f; je nach gewählter Monomordnung muß das nicht unbedingt der Grad des führenden Monoms sein.

Beispiele von Monomordnungen sind

a) Die lexikographische Ordnung: Hier ist $\alpha < \beta$ genau dann, wenn für den ersten Index i, in dem sich α und β unterscheiden, $\alpha_i < \beta_i$ ist. Betrachtet man Monome X^α als Worte über dem (geordneten) Alphabet $\{X_1, \ldots, X_n\}$, kommt hier ein Monom X^α genau dann vor X^β, wenn die entsprechenden Worte im Lexikon in dieser Reihenfolge gelistet werden. Die ersten beiden Forderungen an eine Monomordnung sind klar, und auch die Wohlordnung macht keine großen Probleme: Man betrachtet zunächst die Teilmenge aller Exponenten $\alpha \in I$ mit kleinstmöglichem α_1, unter diesen die Teilmenge mit kleinstmöglichem α_2, *usw.*, bis man bei α_n angelangt ist. Spätestens hier ist die verbleibende Teilmenge einelementig, und ihr einziges Element ist das gesuchte kleinste Element von I.

b) Die graduierte lexikographische Ordnung: Hier ist der Grad eines Monoms erstes Ordnungskriterium: Ist $\deg X^\alpha < \deg X^\beta$, so definieren wir $\alpha < \beta$. Falls beide Monome gleichen Grad haben, soll $\alpha < \beta$ genau dann gelten, wenn α im lexikographischen Sinne kleiner als β ist. Auch hier sind offensichtlich alle drei Forderungen erfüllt.

c) Die inverse lexikographische Ordnung: Hier ist $\alpha < \beta$ genau dann, wenn $\alpha_i < \beta_i$ für den *letzten* Index i, in dem sich α und β unterscheiden. Das entspricht offensichtlich gerade der lexikographischen Anordnung bezüglich des rückwärts gelesenen Alphabets $X_n, \ldots, X_1$. Entsprechend läßt sich natürlich auch bezüglich jeder anderen Permutation des Alphabets eine Monomordnung definieren, so daß diese Ordnung nicht sonderlich interessant ist – außer als Bestandteil der im folgenden definierten Monomordnung:

d) Die graduierte inverse lexikographische Ordnung: Wie bei der graduierten lexikographischen Ordnung ist hier der Grad eines Monoms erstes Ordnungskriterium: Falls $\deg X^\alpha < \deg X^\beta$, ist $\alpha < \beta$, und nur falls beide Monome gleichen Grad haben, soll $\alpha < \beta$ genau dann gelten, wenn α im Sinne der inversen lexikographischen Ordnung

größer ist als β. Man beachte, daß wir hier also nicht nur die Reihenfolge der Variablen invertieren, sondern auch die Ordnungsrelation im Fall gleicher Grade. Es ist nicht schwer zu sehen, daß auch damit eine Monomordnung definiert wird: Mit den ersten beiden Forderungen gibt es wie üblich keine Probleme, und wenn wir eine Menge M von Monomen haben, gibt es darin eine Teilmenge bestehend aus den Monomen kleinsten Grades. Da es für jeden Grad nur endlich viele Monome gibt, ist diese Menge endlich, hat also bezüglich der inversen lexikographischen Ordnung nicht nur ein kleinstes, sondern auch ein größtes Element. Dieses ist das kleinste Element von M bezüglich der graduierten invers lexikographischen Ordnung.

Für das folgende werden wir noch einige Eigenschaften einer Monomordnung benötigen, die in der Definition nicht erwähnt sind.

Als erstes wollen wir uns überlegen, daß bezüglich jeder Monomordnung auf $\mathbb{N}_0^n$ kein Element kleiner sein kann als $(0, \ldots, 0)$: Wäre nämlich $\alpha < (0, \ldots, 0)$, so wäre wegen der zweiten Eigenschaft auch

$$2\alpha = \alpha + \alpha < \alpha + (0, \ldots, 0) = \alpha$$

und so weiter, so daß wir eine unendliche Folge

$$\alpha > 2\alpha > 3\alpha > \cdots$$

hätten, im Widerspruch zur dritten Forderung.

Daraus folgt nun sofort, daß das Produkt zweier Monome größer ist als jeder der beiden Faktoren und damit auch, daß ein echter Teiler eines Monoms immer kleiner ist als dieses. Außerdem folgt, daß für ein Produkt von Polynomen stets $\mathrm{FM}(fg) = \mathrm{FM}(f) \cdot \mathrm{FM}(g)$ ist.

Die Eliminationsschritte beim GAUSS-Algorithmus können auch als Divisionen mit Rest verstanden werden, und beim EUKLIDischen Algorithmus ist ohnehin alles Division mit Rest. Für ein Verallgemeinerung der beiden Algorithmen auf Systeme nichtlinearer Gleichungssysteme brauchen wir also auch einen Divisionsalgorithmus für Polynome in mehreren Veränderlichen, der die eindimensionale Polynomdivision mit Rest und die Eliminationsschritte beim GAUSS-Algorithmus verallgemeinert.

Beim GAUSS-Algorithmus brauchen wir im allgemeinen mehr als nur einen Eliminationsschritt, bis wir eine Gleichung auf eine Variable reduziert haben; entsprechend wollen wir auch hier einen Divisionsalgorithmus betrachten, der gegebenenfalls 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 $k = \mathbb{F}_p$ oder eine endliche Erweiterung davon. Dieses Polynom wollen wir dividieren durch die Polynome $f_1, \ldots, f_m \in R$, d.h. wir suchen Polynome $a_1, \ldots, a_m, r \in R$, so daß

$$f = a_1 f_1 + \cdots + a_m f_m + r$$

ist, wobei r in 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 $f = a_1 f_1 + \cdots + a_m f_m + r$, wobei r kein Monom enthält, das durch das führende Monom eines der f_i teilbar ist.

1. Schritt (Initialisierung): Setze $a_1 = \cdots = a_m = r = 0$ und $p = f$.

2. Schritt (Endebedingung): Im Falle $p = 0$ endet der Algorithmus.

3. Schritt (Divisionsschritt): Falls keiner der führenden Terme FT f_i den führenden Term FT p teilt, wird p ersetzt durch $p -$ FT p und r durch $r +$ FT p. Andernfalls sei i der kleinste Index, für den FT f_i Teiler von FT p ist; der Quotient sei q. Dann wird a_i ersetzt durch $a_i + q$ und p durch $p - qf_i$. Weiter geht es mit dem 2. Schritt.

Offensichtlich ist die Bedingung $f - p = a_1 f_1 + \cdots + a_m f_m + r$ nach der Initialisierung im ersten Schritt erfüllt, und sie bleibt auch bei jeder

Anwendung des Divisionsschritts erfüllt. Außerdem endet der Algorithmus nach endlich vielen Schritten: Bei jedem Divisionsschritt wird der führende Term von p eliminiert, und alle Monome, die eventuell neu dazukommen, sind kleiner oder gleich dem führenden Monom von f_i. Da letzteres das (alte) führende Monom von p teilt, kann es nicht größer sein als dieses, d.h. der führende Term des neuen p ist kleiner als der des alten. Wegen der Wohlordnungseigenschaft einer Monomordnung kann es keine unendliche absteigende Kette von Monomen geben; daher muß der Algorithmus nach endlich vielen Schritten abbrechen.

Bei der klassischen Polynomdivision für Polynome in einer Variablen über einem Körper wissen wir, daß der Rest kleineren Grad hat als der Divisor. Das muß hier nicht der Fall sein; wir können nur sagen, daß der Rest keine Monome enthält, die durch den führenden Term eines der Divisoren f_i teilbar sind.

Um den Algorithmus besser zu verstehen, betrachten wir zunächst zwei Beispiele:

Als erstes dividieren wir $f = X^2Y + XY^2 + Y^2$ durch $f_1 = XY - 1$ und $f_2 = Y^2 - 1$.

Zur Initialisierung setzen wir $a_1 = a_2 = r = 0$ und $p = f$. Wir verwenden die lexikographische Ordnung; bezüglich derer ist der führende Term von p gleich X^2Y und der von f_1 gleich XY. Letzteres teilt X^2Y, wir setzen also

$$p \leftarrow p - Xf_1 = XY^2 + X + Y^2 \quad \text{und} \quad a_1 \leftarrow a_1 + X = X\,.$$

Neuer führender Term von p ist XY^2; auch das ist ein Vielfaches von XY, also setzen wir

$$p \leftarrow p - Yf_1 = X + Y^2 + Y \quad \text{und} \quad a_1 \leftarrow a_1 + Y = X + Y\,.$$

Nun ist X der führende Term von p, und der ist weder durch XY noch durch Y^2 teilbar, also kommt er in den Rest:

$$p \leftarrow p - X = Y^2 + Y \quad \text{und} \quad r \leftarrow r + X = X\,.$$

Der nun führende Term Y^2 von p ist gleichzeitig der führende Term von f_2 und nicht teilbar durch XY, also wird

$$p \leftarrow p - f_2 = Y + 1 \quad \text{und} \quad a_2 \leftarrow a_2 + 1 = 1\,.$$

Die verbleibenden Terme von p sind weder durch XY noch durch Y^2 teilbar, kommen also in den Rest, so daß wir als Ergebnis erhalten

$$f = a_1 f_1 + a_2 f_2 + r \quad \text{mit} \quad a_1 = X + Y, \quad a_2 = 1 \quad \text{und} \quad r = X + Y + 1 .$$

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 $f = X f_2$ ist, geht die Division hier ohne Rest auf. Der Divisionsalgorithmus erlaubt uns also nicht einmal die sichere Feststellung, ob f als Linearkombination der f_i darstellbar ist oder nicht; als alleiniges Hilfsmittel zur Lösung nichtlinearer Gleichungssysteme reicht er offenbar nicht aus. Daher müssen wir in den folgenden Paragraphen noch weitere Werkzeuge betrachten.

§4: Der Hilbertsche Basissatz

Die Grundidee des Algorithmus von BUCHBERGER besteht darin, das Gleichungssystem so abzuändern, daß möglichst viele seiner Eigenschaften bereits an den führenden Termen der Gleichungen ablesbar sind.

Angenommen, wir haben ein nichtlineares Gleichungssystem

$$f_1(X_1, \ldots, X_n) = \cdots = f_m(X_1, \ldots, X_n) = 0$$

mit $f_i \in R = k[X_1, \ldots, X_n]$; seine Lösungsmenge sei $\mathcal{L} \subseteq k^n$.

Wie wir aus §1 wissen, hängt $\mathcal{L}$ nur ab von dem Ideal $I = (f_1, \ldots, f_m)$; zur Lösung des Systems sollten wir daher versuchen, ein möglichst „einfaches“ Erzeugendensystem für dieses Ideal zu finden.

Ganz besonders einfach (wenn auch selten ausreichend) sind Ideale, die von Monomen erzeugt werden:

Definition: Ein Ideal $I \lhd R = k[X_1, \ldots, X_n]$ heißt *monomial*, wenn es 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 $X^{\gamma+\alpha}$ und fassen alle Summanden mit gleichem Monom zusammen, so entsteht eine k-Linearkombination verschiedener Monome, die insgesamt gleich X^β ist. Das ist aber nur möglich, wenn diese Summe aus dem einen Summanden X^β besteht, d.h. β läßt sich schreiben in der Form $\beta = \alpha + \gamma$ mit einem $\alpha \in A$ und einem $\gamma \in \mathbb{N}_0^n$.

Dies zeigt, daß ein Monom X^β genau dann in I liegt, wenn $\beta = \alpha + \gamma$ ist mit einem $\alpha \in A$ und einem $\gamma \in \mathbb{N}_0^n$, d.h. X^β ist das Produkt eines der erzeugenden Monome mit *irgendeinem* Monom. Das Ideal I besteht genau aus den Polynomen f, die sich als k-Linearkombinationen solcher Monome schreiben lassen.

Damit folgt insbesondere, daß ein Polynom f genau dann in einem monomialen Ideal I liegt, wenn jedes seiner Monome dort liegt.

Lemma von Dickson: Jedes monomiale Ideal in $R = k[X_1, \ldots, X_n]$ kann von endlich vielen Monomen erzeugt werden.

Der *Beweis* wird durch vollständige Induktion nach n geführt. Im Fall $n = 1$ ist alles klar, denn da sind die Monome gerade die Potenzen der einzigen Variable, und natürlich erzeugt jede Menge von Potenzen genau dasselbe Ideal wie die Potenz mit dem kleinsten Exponenten aus dieser Menge. Hier kommt man also sogar mit einem einzigen Monom aus.

Im Fall $n > 1$ und $\alpha \in \mathbb{N}_0^n$ setzen wir $X'^{\alpha} = X_1^{\alpha_1} \cdots X_{n-1}^{\alpha_{n-1}}$ und betrachten das Ideal

$$J = \left(X'^{\alpha} \;\middle|\; X^{\alpha} \in I \right) \lhd k[X_1, \ldots, X_{n-1}] .$$

Nach Induktionsvoraussetzung wird J erzeugt von endlich vielen Monomen X'^{α}

Jedes Monom aus aus dem endlichen Erzeugendensystem von J läßt sich in der Form X'^{α} schreiben mit einem $\alpha \in \mathbb{N}_0^n$, für das X^{α} in I liegt. Unter den Indizes α_n, die wir dabei jeweils an das $(n-1)$-Tupel $(\alpha_1, \ldots, \alpha_{n-1})$ anhängen, sei r der größte. Dann liegt $X'^{\alpha'} X_n^r$ für jedes Monom aus dem Erzeugendensystem von J in I und damit für jedes Monom aus J. Die endlich vielen Monome $X^{\alpha'} X_n^r$ erzeugen also zumindest ein Teilideal von I.

Es 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 Monom 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 Wien ein Mathematikstudium. Nach seiner Promotion ging er zu EMMY NOETHER nach Göttingen, um dort Algebra zu lernen. Aus materiellen Gründen mußte er schon bald nach Österreich zurück, konnte aber auch dort zunächst keine Anstellung finden, so daß er Kleinkraftwerke baute und im Hotel seines Vaters aushalf. Ein italienischen Mathematiker, der dort seinen Urlaub verbrachte, vermittelte ihm eine Stelle an der Universität Rom, die er 1939 wieder verlassen mußte, nachdem er sich beim Anschluß Südtirols an Italien für die deutsche Staatsbürgerschaft entschieden hatte. Während des zweiten Weltkriegs arbeitete er größtenteils an einem Forschungsinstitut der Luftwaffe, nach Kriegsende als Extraordinarius in Wien, dann als Ordinarius in Innsbruck, wo er 1980 starb. Seine Arbeiten beschäftigen sich mit der Algebra und algebraischen Geometrie sowie mit Methoden der Computeralgebra zur Lösung von Differentialgleichungen.

Die Theorie der GRÖBNER-Basen wurde von seinem Studenten BRUNO BUCHBERGER in dessen Dissertation entwickelt. BUCHBERGER wurde 1942 in Innsbruck geboren, wo er auch Mathematik studierte und 1966 bei GRÖBNER promovierte mit der Arbeit *Ein Algorithmus zum Auffinden der Basiselemente des Restklassenrings nach einem nulldimensionalen Polynomideal*. Er arbeitete zunächst als Assistent, nach seiner Habilitation als Dozent an der Universität Innsbruck, bis er 1974 einen Ruf auf den Lehrstuhl für Computermathematik an der Universität Linz erhielt. Dort gründete er 1987 das Research Institute for Symbolic Computation (RISC), dessen Direktor er bis 1999 war. 1989 initiierte er in Hagenberg (etwa 20 km nordöstlich von Linz) die Gründung eines Softwareparks mit angeschlossener Fachhochschule; er hat mittlerweile fast Tausend Mitarbeiter. Außer mit Computeralgebra beschäftigt er sich auch im Rahmen des Theorema-Projekts mit dem automatischen Beweisen mathematischer Aussagen.

Wie der obige Beweis des HILBERTschen Basissatzes zeigt, erzeugt eine GRÖBNER-Basis das Ideals. Außerdem hat jedes Ideal I im Polynomring eine GRÖBNER-Basis, denn nach dem Lemma von DICKSON hat das Ideal

der führenden Monome ein endliches Erzeugendensystem, und jedes Monom aus diesem Erzeugendensystem ist führendes Monom eines Polynoms $f_i \in I$. Die Menge der Polynome f_i ist offensichtlich eine GRÖBNER-Basis im Sinne der obigen Definition.

Bevor wir uns damit beschäftigen, wie man diese berechnen kann, wollen wir zunächst eine wichtige Eigenschaft betrachten.

$\{g_1, \ldots, g_m\}$ sei eine GRÖBNER-Basis eines Ideals $I \lhd R$. Wir wollen ein beliebiges Element $f \in R$ durch $g_1, \ldots, g_m$ dividieren. Dies liefert als Ergebnis

$$f = a_1 g_1 + \cdots + a_m g_m + r\,,$$

wobei kein Monom von r durch eines der Monome FM(g_i) teilbar ist. Wie wir wissen, sind allerdings bei der Polynomdivision im allgemeinen weder der Divisionsrest r noch die Koeffizienten a_i auch nur im entferntesten eindeutig. Wir wollen untersuchen, wie sich das hier verhält.

Angenommen. wir haben zwei Darstellungen

$$f = a_1 g_1 + \cdots + a_m g_m + r = b_1 g_1 + \cdots + b_m g_m + s$$

der obigen Form. Dann ist

$$(a_1 - b_1) g_1 + \cdots + (a_m - b_m) g_m = s - r\,.$$

Links steht ein Element von I, also auch rechts. Andererseits enthält aber weder r noch s ein Monom, das durch eines der Monome FM(g_i) teilbar ist, d.h. $r - s = 0$, da die FM(g_i) ja das Ideal FM(I) erzeugen. Somit ist bei der Division durch die Elemente einer GRÖBNER-Basis der Divisionsrest eindeutig bestimmt. Insbesondere ist f genau dann ein Element von I, wenn der Divisionsrest verschwindet. Wenn wir eine GRÖBNER-Basis haben, können wir also leicht entscheiden, ob ein gegebenes Element $f \in R$ im Ideal I liegt.

Nachdem im Fall einer GRÖBNER-Basis der Divisionsrest nicht von der Reihenfolge der Basiselemente abhängt, können wir ihn durch ein Symbol bezeichnen, das nur von der *Menge* $G = \{g_1, \ldots, g_m\}$ abhängt; wir schreiben $\overline{f}^G$.

Als nächstes wollen wir uns mit der Frage beschäftigen, wie wir für ein vorgegebenes Ideal I eine GRÖBNER-Basis bestimmen können.

Dazu müssen wir uns als erstes überlegen, *wie* das Ideal vorgegeben sein soll. Wenn wir damit rechnen wollen, müssen wir irgendeine Art von endlicher Information haben; was sich anbietet ist natürlich ein endliches Erzeugendensystem.

Wir gehen also aus von einem Ideal $I = (f_1, \ldots, f_m)$ und suchen eine GRÖBNER-Basis. Das Problem ist, daß die Monome 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 $i = 1, \ldots, n$. Das S-Polynom von f und g ist

$$S(f,g) = \frac{X^\gamma}{\mathrm{FT}(f)} \cdot f - \frac{X^\gamma}{\mathrm{FT}(g)} \cdot g\,.$$

Da $\frac{X^\gamma}{\mathrm{FT}(f)} \cdot f$ und $\frac{X^\gamma}{\mathrm{FT}(g)} \cdot g$ beide nicht nur dasselbe führende Monom X^γ haben, sondern es wegen der Division durch den führenden *Term* statt nur das führende Monom auch beide mit Koeffizient eins enthalten, fällt es bei der Bildung von $S(f,g)$ weg. Daher ist das führende Monom von $S(f,g)$ kleiner als X^γ. Das folgende Lemma ist der Kern des Beweises, daß S-Polynome alles sind, was wir brauchen, um GRÖBNER-Basen zu berechnen.

Lemma: Für die Polynome $f_1, \ldots, f_m \in R$ sei

$$S = \sum_{i=1}^{m} \lambda_i X^{\alpha_i} f_i \quad \text{mit} \quad \lambda_i \in k \quad \text{und} \quad \alpha_i \in \mathbb{N}_0^n$$

eine Linearkombination, zu der es ein $\delta \in \mathbb{N}_0^n$ gebe, so daß alle Summanden X^δ als führendes Monom haben, d.h. α_i + multideg $f_i = \delta_i$ für $i = 1, \ldots, m$. Falls multideg $S < \delta$ ist, gibt es Elemente $\lambda_{ij} \in k$, so daß

$$S = \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_{ij} X^{\gamma_{ij}} S(f_i, f_j)$$

ist mit $X^{\gamma_{ij}} = \mathrm{kgV}\big(\mathrm{FM}(f_i), \mathrm{FM}(f_j)\big)$.

Beweis: Der führende Koeffizient von f_i sei μ_i; dann ist $\lambda_i\mu_i$ der führende Koeffizient von $\lambda_i X^{\alpha_i} f_i$. Somit ist multideg S genau dann kleiner als δ, wenn $\sum_{i=1}^{m} \lambda_i\mu_i$ verschwindet. Wir normieren alle $X^{\alpha_i} f_i$ auf führenden Koeffizienten eins, indem wir $p_i = X^{\alpha_i} f_i/\mu_i$ betrachten; dann ist

$$\begin{aligned} S = \sum_{i=1}^{m} \lambda_i\mu_i p_i = {} & \lambda_1\mu_1(p_1 - p_2) + (\lambda_1\mu_1 + \lambda_2\mu_2)(p_2 - p_3) + \cdots \\ & + (\lambda_1\mu_1 + \cdots + \lambda_{m-1}\mu_{m-1})(p_{m-1} - p_m) \\ & + (\lambda_1\mu_1 + \cdots + \lambda_m\mu_m)p_m \,, \end{aligned}$$

wobei der Summand in der letzten Zeile genau dann verschwindet, wenn multideg $S < \delta$.

Da alle p_i denselben Multigrad δ und denselben führenden Koeffizienten eins haben, kürzen sich in den Differenzen $p_i - p_j$ die führenden Terme weg, genau wie in den S-Polynomen. In der Tat: Bezeichnen wir den Multigrad von $\mathrm{kgV}\big(\mathrm{FM}(f_i), \mathrm{FM}(f_j)\big)$ mit γ_{ij}, so ist

$$p_i - p_j = X^{\delta-\gamma_{ij}} S(f_i, f_j)\,.$$

Damit hat die obige Summendarstellung von S die gewünschte Form. ∎

Daraus folgt ziemlich unmittelbar

Satz: Ein Erzeugendensystem $f_1, \ldots, f_m$ eines Ideals I im Polynomring $R = k[X_1, \ldots, X_n]$ ist genau dann eine GRÖBNER-Basis, wenn jedes S-Polynom $S(f_i, f_j)$ bei der Division durch $f_1, \ldots, f_m$ Rest Null hat.

Beweis: Als R-Linearkombination von f_i und f_j liegt das S-Polynom $S(f_i, f_j)$ im Ideal I; falls $f_1, \ldots, f_m$ eine GRÖBNER-Basis von I ist, hat es also Rest Null bei der Division durch $f_1, \ldots, f_m$.

Umgekehrt sei $f_1, \ldots, f_m$ ein Erzeugendensystem von $I \lhd R$ mit der Eigenschaft, daß alle $S(f_i, f_j)$ bei der Division durch $f_1, \ldots, f_m$ (in irgendeiner Reihenfolge) Divisionsrest Null haben. Wir wollen zeigen, daß $f_1, \ldots, f_m$ dann eine GRÖBNER-Basis ist, daß also die führenden Monome $\mathrm{FM}(f_1), \ldots, \mathrm{FM}(f_m)$ das Ideal $\mathrm{FM}(I)$ erzeugen.

Sei also $f \in I$ ein beliebiges Element; wir müssen zeigen, daß FM(f) im von den 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, erhielten wir daher eine unendliche aufsteigende Folge von monomialen Idealen J_i, von denen jedes echt größer ist als sein Vorgänger:

$$J_1 < J_2 < \cdots J_i < J_{i+1} < \cdots .$$

Natürlich ist auch die Vereinigung J aller J_i ein monomiales Ideal, hat also nach dem Lemma von DICKSON ein endliches Erzeugendensystem $\{M_1, \ldots, M_q\}$. Da jedes M_j in einem J_i und damit auch in allen folgenden liegen muß, gibt es ein m, so daß alle M_j in J_m liegen. Damit ist $J = (M_1, \ldots, M_q) \subseteq J_m$, im Widerspruch zur Annahme, daß J_{m+1} und damit auch J echt größer als M_j ist.

Der Algorithmus kann natürlich auf mehrere offensichtliche Weisen optimiert werden: Beispielsweise stößt man beim wiederholten Durchlaufen des zweiten Schritts immer wieder auf dieselben S-Polynome, die daher nicht jedes Mal neu berechnet werden müssen, und wenn eines dieser Polynome einmal Divisionsrest Null hatte, hat es auch bei jedem weiteren Durchgang Divisionsrest Null, denn dann wird ja wieder durch dieselben Polynome (plus einiger neuer) dividiert. Es gibt inzwischen auch zahlreiche nicht offensichtliche Verbesserungen und Optimierungen; wir wollen uns aber mit dem Prinzip begnügen und stattdessen später noch einige andere Themen behandeln.

Der BUCHBERGER-Algorithmus hat den Nachteil, daß er das vorgegebene Erzeugendensystem in jedem Schritt größer macht ohne je ein Element zu streichen. Dies ist weder beim GAUSS-Algorithmus noch beim EUKLIDischen Algorithmus der Fall, bei denen jeweils eine Gleichung durch eine andere *ersetzt* wird. Obwohl wir sowohl die Eliminationsschritte des GAUSS-Algorithmus als auch die einzelnen Schritte der Polynomdivisionen beim EUKLIDischen Algorithmus durch S-Polynome ausdrücken können, *müssen* wir im allgemeinen Fall zusätzlich zu g und $S(f,g)$ auch noch das Polynom f beibehalten; andernfalls kann sich die Lösungsmenge ändern:

Als Beispiel können wir das Gleichungssystem

$$f(X,Y) = X^2Y + XY^2 + 1 = 0 \quad \text{und} \quad g(X,Y) = X^3 - XY - Y = 0$$

betrachten. Wenn wir mit der lexikographischen Ordnung arbeiten, sind hier die einzelnen Monome bereits der Größe nach geordnet, insbesondere stehen also die führenden Monome an erster Stelle und

$$S(f,g) = Xf(X,Y) - Yg(X,Y) = X^2Y^2 + XY^2 + X + Y^2 .$$

Der führende Term X^2Y^2 ist durch den führenden Term X^2Y von f teilbar; subtrahieren wir Yf vom S-Polynom, erhalten wir das nicht weiter reduzierbare Polynom

$$h(X,Y) = -XY^3 + XY^2 + X + Y^2 - Y .$$

Sowohl $g(X,Y)$ als auch $h(X,Y)$ verschwinden im Punkt $(0,0)$; dieser ist aber keine Lösung des Ausgangssystems, da $f(0,0) = 1$ nicht verschwindet.

Aus diesem Grund werden die nach dem BUCHBERGER-Algorithmus berechneten GRÖBNER-Basen oft sehr groß und unhandlich. Betrachten wir dazu als Beispiel das System aus den beiden Gleichungen

$$f_1 = X^3 - 2XY \quad \text{und} \quad f_2 = X^2Y - 2Y^2 + X$$

und berechnen eine GRÖBNER-Basis bezüglich der graduiert lexikographischen Ordnung.

$$S(f_1, f_2) = Y f_1 - X f_2 = -X^2$$

ist weder durch den führenden Term von f_1 noch den von f_2 teilbar, muß also als neues Element f_3 in die Basis aufgenommen werden.

$$S(f_1, f_3) = f_1 + X f_3 = -2XY$$

kann wieder mit keinem der f_i reduziert werden, muß also als neues Element f_4 in die Basis. Genauso ist es mit

$$f_5 = S(f_2, f_3) = f_2 + Y f_3 = -2Y^2 + X\,.$$

Für das so erweiterte Erzeugendensystem, bestehend aus den Polynomen

$$f_1 = X^3 - 2XY, \quad f_2 = X^2Y - 2Y^2 + X, \quad f_3 = -X^2,$$
$$f_4 = -2XY \quad \text{und} \quad f_5 = -2Y^2 + X\,,$$

sind die S-Polynome

$$S(f_1, f_2) = f_3, \quad S(f_1, f_3) = f_4 \quad \text{und} \quad S(f_2, f_3) = f_5$$

trivialerweise auf Null reduzierbar, die anderen Kombinationen müssen wir nachrechnen:

$$S(f_1, f_4) = Y f_1 + \frac{X^2}{2} f_4 = -2XY^2 = Y f_4$$

$$S(f_1, f_5) = Y^2 f_1 + \frac{X^3}{2} f_5 = -2XY^3 + \frac{X^4}{2} = \frac{X}{2} f_1 + f_2 + Y^2 f_4 - f_5$$

$$S(f_2, f_4) = f_2 + \frac{X}{2} f_4 = -2Y^2 + X = f_5$$

$$S(f_2, f_5) = Y f_2 + \frac{X^2}{2} f_5 = \frac{X^3}{2} + XY - 2Y^3 = \frac{1}{2} f_1 - \frac{1}{2} f_4 + Y f_5$$

$$S(f_3, f_4) = -Yf_3 - \frac{X}{2}f_4 = 0$$

$$S(f_3, f_5) = -Y^2 f_3 - \frac{X^2}{2}f_5 = \frac{1}{2}f_1 - \frac{1}{2}f_4$$

$$S(f_4, f_5) = -\frac{Y}{2}f_4 - \frac{X}{2}f_5 = \frac{X^2}{2} = -\frac{1}{2}f_3$$

Somit bilden diese fünf Polynome eine GRÖBNER-Basis des von f_1 und f_2 erzeugten Ideals.

Zum Glück brauchen wir aber nicht alle fünf Polynome. Das folgende Lemma gibt ein Kriterium, wann man auf ein Erzeugendes verzichten kann, und illustriert gleichzeitig das allgemeine Prinzip, wonach bei einer GRÖBNER-Basis alle wichtigen Eigenschaften anhand der führenden Termen ablesbar sein sollten:

Lemma: G sei eine GRÖBNER-Basis des Ideals $I \lhd k[X_1, \ldots, X_n]$, und $g \in G$ sei ein Polynom, dessen führendes Monom im von den führenden Monomen der restlichen Basiselemente erzeugten monomialen Ideal liegt. Dann ist auch $G \setminus \{g\}$ eine GRÖBNER-Basis von I.

Beweis: $G \setminus \{g\}$ ist nach Definition genau dann eine GRÖBNER-Basis von I, wenn die führenden Terme der Basiselemente das Ideal FM(I) erzeugen. Da G eine GRÖBNER-Basis von I ist und die führenden Terme egal ob mit oder ohne FT(g) dasselbe monomiale Ideal erzeugen, ist das klar. ∎

Man beachte, daß sich dieses Lemma nur anwenden läßt, wenn G eine GRÖBNER-Basis von I ist; wir können nicht schon während des Rechengangs im BUCHBERGER-Algorithmus Elemente streichen. Im obigen Beispiel etwa wird das Ideal $I = (f_1, f_2)$ natürlich auch erzeugt von f_1, f_2 und f_3; dabei ist FM(f_1) = X^3, FM(f_2) = X^2Y, und FM(f_3) = X^2 teilt beide dieser Monome. Wenn das Lemma auf die Basis f_1, f_2, f_3 anwendbar wäre, könnten wir also f_1 und f_2 streichen und f_3 wäre für sich allein eine GRÖBNER-Basis von I. Natürlich ist aber $I \neq (-X^2)$, denn weder f_1 noch f_2 sind Vielfache von X^2.

Von der Menge $\{f_1, f_2, f_3, f_4, f_5\}$ haben wir mit Hilfe des Kriteriums von BUCHBERGER verifiziert, daß sie eine GRÖBNER-Basis von I ist; deshalb können wir das Lemma darauf anwenden und f_1, f_2 streichen. Wir können das aber erst jetzt tun, denn im Verlauf der Berechnungen wurden f_1 und f_2 noch gebraucht um $f_4 = S(f_1, f_3)$ und $f_5 = S(f_2, f_3)$ zu konstruieren. Somit ist $I = (f_3, f_4, f_5)$, und darauf können wir das Lemma nicht weiter anwenden, denn

$$\mathrm{FM}(f_3) = X^2, \quad \mathrm{FM}(f_4) = XY \quad \text{und} \quad \mathrm{FM}(f_5) = Y^2,$$

und keines dieser drei Monome ist Vielfaches eines der anderen.

Zur weiteren Normierung können wir noch durch die führenden Koeffizienten teilen und erhalten dann die *minimale* GRÖBNER-Basis

$$\widetilde{f}_3 = X^2, \quad \widetilde{f}_4 = XY \quad \text{und} \quad \widetilde{f}_5 = Y^2 - \frac{X}{2}.$$

Definition: Eine *minimale* GRÖBNER-Basis von I ist eine GRÖBNER-Basis von I mit folgenden Eigenschaften:

1.) Alle $g \in G$ haben den führenden Koeffizienten eins

2.) Für kein $g \in G$ liegt FM(g) im von den führenden Monomen der übrigen Elemente erzeugten Ideal.

Da ein Monom X^α genau dann im von einer Menge M von Monomen erzeugten Ideal liegt, wenn es durch eines dieser Monome teilbar ist, können wir die zweite Bedingung auch so ausdrücken, daß es keine zwei Elemente $g \neq g'$ in G geben darf, für die FM(g) ein Teiler von FM(g') ist.

Es ist klar, daß jede GRÖBNER-Basis zu einer minimalen GRÖBNER-Basis verkleinert werden kann: Durch Division können wir alle führenden Koeffizienten zu eins machen ohne etwas an der Erzeugung zu ändern, und nach obigem Lemma können wir nacheinander alle Elemente eliminieren, die die zweite Bedingung verletzen.

Wir können aber noch mehr erreichen: Wenn nicht das führende, sondern einfach *irgendein* Monom eines Polynoms $g \in G$ im von den führenden Termen der übrigen Elemente erzeugten Ideal liegt, ist dieses Monom teilbar durch das führende Monom eines anderen Polynoms $h \in G$. Wir

können den Term mit diesem Monom daher zum Verschwinden bringen, indem wir g ersetzen durch g minus ein Vielfaches von h. Da sich dabei nichts an den führenden Termen der Elemente von G ändert, bleibt G eine GRÖBNER-Basis. Wir können somit aus den Elementen einer minimalen GRÖBNER-Basis Terme eliminieren, die durch den führenden Term eines anderen Elements teilbar sind. Was dabei schließlich entstehen sollte, ist eine *reduzierte* GRÖBNER-Basis:

Definition: Eine reduzierte GRÖBNER-Basis von I ist eine GRÖBNER-Basis von I mit folgenden Eigenschaften:
1.) Alle $g \in G$ haben den führenden Koeffizienten eins
2.) Für kein $g \in G$ liegt ein Monom von g im von den führenden Monomen der übrigen Elemente erzeugten Ideal.

Die minimale Basis im obigen Beispiel ist offenbar schon reduziert, denn außer $\widetilde{f}_5$ bestehen alle Basispolynome nur aus dem führendem Term, und bei $\widetilde{f}_5$ ist der zusätzliche Term linear, kann also nicht durch die quadratischen führenden Monome der anderen Polynome teilbar sein.

Reduzierte GRÖBNER-Basis haben eine für das praktische Rechnen mit Idealen sehr wichtige zusätzliche Eigenschaft:

Satz: Jedes Ideal $I \lhd k[X_1, \ldots, X_n]$ hat (bei vorgegebener Monomordnung) eine eindeutig bestimmte reduzierte GRÖBNER-Basis.

Beweis: Wir gehen aus von einer minimalen GRÖBNER-Basis G und ersetzen nacheinander jedes Element $g \in G$ durch seinen Rest bei der Polynomdivision durch $G \setminus \{g\}$. Da bei einer minimalen GRÖBNER-Basis kein führendes Monom eines Element das führende Monom eines anderen teilen kann, ändert sich dabei nichts an den führenden Termen, G ist also auch nach der Ersetzung eine minimale GRÖBNER-Basis. In der schließlich entstehenden Basis hat kein $g \in G$ mehr einen Term, der durch den führenden Term eines Elements von $G \setminus \{g\}$ teilbar wäre, denn auch wenn wir bei der Reduktion der einzelnen Elemente durch eine eventuell andere Menge geteilt haben, hat sich doch an den führenden Termen der Basiselemente nichts geändert. Also gibt es eine reduzierte GRÖBNER-Basis.

Nun seien G und G' zwei reduzierte GRÖBNER-Basen von I. Jedes Element $f \in G'$ liegt insbesondere in I, also ist $\overline{f}^{G} = 0$. Insbesondere muß der führende Term von f durch den führenden Term eines $g \in G$ teilbar sein. Umgekehrt ist aber auch $\overline{g}^{G'} = 0$, d.h. der führende Term von g muß durch den führenden Term eines Elements von $f' \in G'$ teilbar sein. Dieser führende Term teilt dann insbesondere den führenden Term von f, und da G' als reduzierte GRÖBNER-Basis minimal ist, muß $f' = f$ sein. Somit gibt es zu jedem $g \in G$ genau ein $f \in G'$ mit FM(f) = FM(g); insbesondere haben G und G' dieselbe Elementanzahl. Tatsächlich muß sogar $f = g$ sein, denn $f - g$ liegt in I, enthält aber keine Term, der durch den führenden Term irgendeines Elements von G teilbar wäre. Also ist $f - g = 0$.

■

Bemerkung: Die Forderung in den Definitionen von minimalen und reduzierten GRÖBNER-Basen, daß alle führenden Koeffizienten eins sein müssen, ist zwar nützlich für theoretische Diskussionen, führt aber im Falle von Polynomen mit rationalen Koeffizienten oft dazu, daß die Koeffizienten Nenner haben. Computeralgebrasysteme können zwar mit rationalen Zahlen rechnen, indem sie diese durch Paare teilerfremder ganzer Zahlen darstellen, aber diese Rechnungen sind erheblich aufwendiger als solche mit ganzen Zahlen. Daher liefern einige Computeralgebrasysteme beim Kommando zur Berechnung einer reduzierten GRÖBNER-Basis anstelle von Polynomen mit führendem Koeffizienten eins solche mit teilerfremden ganzzahligen Koeffizienten.

Kapitel 2
Systeme von nichtlinearen Polynomgleichungen

GRÖBNER-Basen haben eine Vielzahl von Anwendungen in der Algebra; wir wollen uns hier vor allem damit beschäftigen, wie sie direkt oder im Zusammenspiel mit anderen Methoden zur expliziten Lösung nichtlinearer Gleichungssysteme führen können. Explizit angebbar sind die Lösungen meist nur, wenn die Lösungsmenge endlich ist; daher werden wir uns meist auf solche Systeme beschränken und interessieren uns daher auch für Kriterien, wie wir einem Gleichungssystem die Endlichkeit seiner Lösungsmenge ansehen können.

§ 1: Gröbner-Basen für nichtlineare Gleichungssysteme

Wir gehen aus von m Polynomgleichungen

$$f_i(x_1, \ldots, x_n) = 0 \quad \text{mit} \quad f_i \in k[X_1, \ldots, X_n] \quad \text{für} \quad i = 1, \ldots, m$$

und suchen die Lösungsmenge

$$\left\{(x_1, \ldots, x_n) \in k^n \;\middle|\; f_i(x_1, \ldots, x_n) = 0 \text{ für } i = 1, \ldots, m\right\}.$$

Diese wird allerdings oft leer sein; für $f_1 = X^2 - 2$ und $f_2 = Y^2 - 3$ aus $\mathbb{Q}[X]$ etwa ist diese Menge leer, da die Lösungen $(\pm\sqrt{2}, \pm\sqrt{3})$ nicht in $\mathbb{Q}^2$ liegen. Wir betrachten daher meist noch einen zweiten Körper K, der k enthält, und interessieren uns allgemeiner für die Lösungsmenge in K^n:

Definition: *a)* Ist I ein Ideal in $k[X_1, \ldots, X_n]$, und ist K ein Körper, der k enthält, setzen wir

$$V_K(I) = \left\{(x_1, \ldots, x_n) \in K^n \;\middle|\; f(x_1, \ldots, x_n) = 0 \text{ für alle } f \in I\right\}.$$

b) Für $I = (f_1, \cdots, f_m)$ schreiben wir auch kurz $V_K(f_1, \ldots, f_m)$ an Stelle von $V_K(I)$.

Der Körper k sollte dabei möglichst klein sein, denn mit den Elementen dieses Körpers müssen wir rechnen, und je größer der Körper, desto aufwendiger sind seine Rechenoperationen. In konkreten Beispielen werden wir uns meist auf $k = \mathbb{Q}$ beschränken und – soweit möglich – sogar versuchen, unsere Konstruktionen in $\mathbb{Z}[X]$ durchzuführen.

Der Körper K hingegen sollte so groß sein, daß er für ein Gleichungssystem, daß in irgendeinem Körper eine nichtleere endliche Lösungsmenge hat, diese Lösungsmenge enthält. Wir werden meist $K = \mathbb{C}$ betrachten.

Wie wir bereits aus §1 des vorigen Kapitels wissen, hängt die Lösungsmenge des Gleichungssystems nur ab vom Ideal $I = (f_1, \ldots, f_m)$; wir suchen ein Erzeugendensystem $\{g_1, \ldots, g_r\}$ dieses Ideals, aus dem wir mehr über die Mengen

$$V_K(I) = V_K(f_1, \ldots, f_m) = V_K(g_1, \ldots, g_r)$$

ablesen können. Wir erwarten natürlich, daß wir hier vor allem im Falle einer geeigneten GRÖBNER-Basis $\{g_1, \ldots, g_r\}$ eventuell Erfolg haben.

Viele Lösungsansätze für Gleichungssysteme in mehreren Veränderlichen beruhen auf der Elimination von Variablen: Im ℓ-ten Schritt suchen wir nach Bedingungen, die ein $(n - \ell)$-Tupel $(x_{\ell+1}, \ldots, x_n)$ erfüllen muß, wenn es ein ℓ-Tupel $(x_1, \ldots, x_\ell)$ gibt, so daß $(x_1, \ldots, x_n)$ in $V(I)$ liegt. Eine solche Bedingung ist trivial: Für jedes Polynom $f \in I$, in dem die Variablen $X_1, \ldots, X_\ell$ nicht vorkommen, muß $f(x_{\ell+1}, \ldots, x_n) = 0$ sein.

Definition: *a)* Das ℓ-te *Eliminationsideal* eines Ideal $I \lhd k[X_1, \ldots, X_n]$ ist $I_\ell = I \cap k[X_{\ell+1}, \ldots, X_n]$.
b) Eine Monomordnung $<$ heißt *Eliminationsordnung* für $X_1, \ldots, X_\ell$, wenn jedes Monom, das mindestens eine der Variablen $X_1, \ldots, X_\ell$ enthält, größer ist als alle Monome, die nur $X_{\ell+1}, \ldots, X_n$ enthalten.

Die lexikographische Ordnung mit $X_1 > X_2 > \cdots > X_{n-1} > X_n$ ist offensichtlich für jedes ℓ eine Eliminationsordnung für $X_1, \ldots, X_\ell$, die

graduiert lexikographische aber nicht, da bezüglich dieser beispielsweise $X_1 < X_n^2$ ist.

Satz: Ist G eine GRÖBNER-Basis von I bezüglich einer Eliminationsordnung für $X_1, \ldots, X_\ell$, so ist $G \cap I_\ell$ eine GRÖBNER-Basis von I_ℓ.

Beweis: Die Elemente von $G = \{g_1, \ldots, g_m\}$ seien so angeordnet, daß $G \cap I_\ell = \{g_1, \ldots, g_r\}$ ist. Wir müssen zeigen, daß sich jedes $f \in I_\ell$ als Linearkombination von $g_1, \ldots, g_r$ mit Koeffizienten aus $k[X_{\ell+1}, \ldots, X_n]$ darstellen läßt.

Der Divisionsalgorithmus bezüglich der lexikographischen Ordnung gibt uns eine Darstellung $f = h_1 g_1 + \cdots + h_m g_m$ von f als Element von I. Die Polynome $g_{r+1}, \ldots, g_m$ enthalten jeweils mindestens eine der Variablen $X_1, \ldots, X_\ell$, und da wir eine Eliminationsordnung verwenden, muß auch das führende Monom eine dieser Variablen enthalten. Da kein Monom von f eine dieser Variablen enthält, kann im Divisionsalgorithmus das führende Monom eines dieser Polynome nie Teiler des führenden Monoms des jeweils betrachteten Polynoms p sein, Somit ist $h_{r+1} = \cdots = h_m = 0$, und in keinem der Polynome $h_1, \ldots, h_r$ kann eine der Variablen $X_1, \ldots, X_\ell$ auftreten. Dies zeigt, daß f im von $g_1, \ldots, g_r$ erzeugten Ideal von $k[X_{\ell+1}, \ldots, X_n]$ liegt, d.h. dieses Ideal wird von $g_1, \ldots, g_r$ erzeugt.

Um zu zeigen, daß es sich dabei sogar um eine GRÖBNER-Basis handelt, können wir zum Beispiel zeigen, daß alle $S(g_i, g_j)$ mit $i, j \leq r$ ohne Rest durch $g_1, \ldots, g_r$ teilbar sind. Da G nach Voraussetzung eine GRÖBNER-Basis ist, sind sie auf jeden Fall ohne Rest durch G teilbar, und wieder kann bei der Division nie der führende Term eines Dividenden durch den eines g_i mit $i > r$ teilbar sein, d.h. $S(g_i, g_j)$ ist als Linearkombination von $g_1, \ldots, g_r$ mit Koeffizienten aus $k[g_1, \ldots, g_r]$ darstellbar. ∎

Daraus ergibt sich eine Strategie zur Lösung nichtlinearer Gleichungssysteme nach Art des GAUSS-Algorithmus: Wir gehen aus von der lexikographischen Ordnung, die ja für jedes ℓ eine Eliminationsordnung für $X_1, \ldots, X_\ell$ ist, und bestimmen eine (reduzierte) GRÖBNER-Basis für das von den Gleichungen erzeugte Ideal des Polynomrings $k[X_1, \ldots, X_n]$.

Dann betrachten als erstes das Eliminationsideal I_{n-1}. Dieses besteht nur aus Polynomen in X_n; falls wir mit einer reduzierten GRÖBNER-Basis arbeiten, gibt es darin höchstens ein solches Polynom.

Falls es ein solches Polynom gibt, muß jede Lösung des Gleichungssystem als letzte Komponente eine von dessen Nullstellen haben. Wir bestimmen daher diese Nullstellen (in K) und setzen sie nacheinander in das restliche Gleichungssystem ein. Dadurch erhalten wir Gleichungssysteme in $n-1$ Unbekannten, wo wir nach Gleichungen nur in X_{n-1} suchen können. Diese erhalten wir, indem wir bei allen Erzeugenden des Eliminationsideals I_{n-2} für X_n nacheinander die Werte aus $V_K(I_{n-1}) \subset k$ einsetzen. Nachdem wir so $V_K(I_{n-2}) \subset K^2$ bestimmt haben, können wir analog die Mengen $V_K(I_{n-3}) \subset K^3$ und so weiter bis $V_K(I) \subset K^n$ bestimmen.

Betrachten wir noch einmal das Beispiel gegen Ende von §5 des vorigen Kapitels mit

$$f_1 = X^3 - 2XY \quad \text{und} \quad f_2 = X^2Y - 2Y^2 + X\,.$$

Dort hatten wir die reduzierte GRÖBNER-Basis bezüglich der graduiert lexikographischen Ordnung berechnet; sie besteht aus

$$g_1 = X^2, \quad g_2 = XY \quad \text{und} \quad g_3 = Y^2 - \frac{X}{2}\,.$$

Da die graduiert lexikographische Ordnung keine Eliminationsordnung für X ist, können wir nicht erwarten, daß $\{g_1, g_2, g_3\} \cap k[Y]$ ein Erzeugendensystem des Eliminationsideals $(f_1, f_2) \cap k[Y]$ liefert, und in der Tat liegt keines der g_i in $k[Y]$. Zufälligerweise liegt aber $g_1 = X^2$ in $k[X]$, wir wissen also, daß für jede Lösung (x, y) des Gleichungssystems $x = 0$ sein muß. $g_2 = XY$ verschwindet für alle solche Punkte automatisch, und $g_3 = Y^2 - X/2$ verschwindet genau dann, wenn auch $y = 0$ ist. Somit ist $V(f_1, f_2) = \{(0, 0)\}$.

Wenn wir das Gleichungssystem mit dem hier vorgestellten Verfahren lösen wollen. müssen wir mit der lexikographischen Ordnung arbeiten. Da die führenden Terme von f_1 und f_2 bei beiden Ordnungen gleich sind und viele der zu berechnenden S-Polynome nur aus einem Term

bestehen, ändert sich zunächst nichts: Wie bei der graduiert lexikographischen Ordnung kommen wir auf

$$f_3 = S(f_1, f_2) = -X^2, \quad f_4 = S(f_1, f_3) = -2XY \quad \text{und}$$
$$f_5 = S(f_2, f_3) = X - 2Y^2 .$$

Auch $S(f_1, f_4) = -2XY^2 = Y f_4$ kann wie dort auf Null reduziert werden, bei der Berechnung von $S(f_1, f_5)$ ist jetzt aber nicht mehr Y^2, sondern X das führende Monom. Somit ist

$$S(f_1, f_5) = f_1 - X^2 f_5 = 2X^2Y^2 - 2XY = 2Y f_2 + 2f_4 + 4Y^3 ,$$

das S-Polynom läßt sich also modulo $\{f_1, f_2, f_3, f_4, f_5\}$ nicht auf Null reduzieren und wir müssen $f_6 = 4Y^3$ als neues Element in die Basis aufnehmen. Erst jetzt zeigt eine mühsame Rechnung, die man am besten seinem Computer überläßt, daß $S(f_i, f_j)$ für alle $1 \leq i < j \leq 6$ modulo $\{f_1, f_2, f_3, f_4, f_5, f_6\}$ auf Null reduziert werden kann, womit wir eine GRÖBNER-Basis gefunden haben.

Die führenden Monome der sechs Basiselemente bezüglich der lexikographischen Ordnung sind

$$\begin{array}{lll} \text{FM}(f_1) = X^3 , & \text{FM}(f_2) = X^2Y , & \text{FM}(f_3) = -X^2 , \\ \text{FM}(f_4) = -2XY , & \text{FM}(f_5) = X, , & \text{FM}(f_6) = 4Y^3 ; \end{array}$$

wir können also f_1 bis f_4 eliminieren. Die reduzierte GRÖBNER-Basis bedeutet besteht somit aus $g_1 = X - 2Y^2$ und $g_2 = Y^3$.

Das Eliminationsideal I_1 wird daher erzeugt von $g_2 = Y^3$, d.h. für jede Lösung (x, y) muß y verschwinden. Setzen wir $y = 0$ in g_1 ein, so sehen wir, daß auch x verschwinden muß, der Nullpunkt ist also die einzige Lösung.

Es war ein Zufall, daß wir dieses Ergebnis auch der GRÖBNER-Basis bezüglich der graduiert lexikographischen Ordnung ansehen konnten; bei komplizierteren Systemen wird dort oft jedes Basiselement alle Variablen enthalten, so daß wir nichts sehen können. Trotzdem kann die graduiert lexikographische Ordnung zur Lösung nichtlinearer Gleichungssysteme nützlich sein: 1993 publizierten J.C. FAUGÈRE, P. GIANNI, D. LAZARD und T. MORA einen heute nach ihren Anfangsbuchstaben

als FGLM benannten Algorithmus, der für ein Ideal I mit endlicher Nullstellenmenge $V(I)$ effizient eine GRÖBNER-Basis bezüglich der lexikographischen Ordnung bestimmt auf dem Umweg über die graduiert lexikographische Ordnung. Wir werden später sehen, daß wir im Falle einer endlichen Lösungsmenge diese auch ausgehend von einer beliebigen GRÖBNER-Basis mit alternativen Techniken bestimmen können.

Nun kann es beim obigen Verfahren für nichtlineare Gleichungssysteme natürlich vorkommen, daß I_{n-1} das Nullideal ist; falls unter den Lösungen des Systems unendlich viele Werte für die letzte Variable vorkommen, muß das sogar so sein. Es kann sogar vorkommen, daß *alle* Eliminationsideale außer $I_0 = I$ das Nullideal sind. In diesem Fall führt die gerade skizzierte Vorgehensweise zu nichts.

Bevor wir uns darüber wundern, sollten wir uns überlegen, was wir überhaupt unter der Lösung eines nichtlinearen Gleichungssystems verstehen wollen. Im Falle einer endlichen Lösungsmenge ist das klar: Dann wollen wir eine Auflistung der sämtlichen Lösungstupel. Bei einer unendlichen Lösungsmenge ist das aber nicht mehr möglich. Im Falle eines linearen Gleichungssystems wissen wir, daß die Lösungsmenge ein affiner Raum ist; wir können sie daher auch wenn sie unendlich sein sollte durch endlich viele Daten eindeutig beschreiben, zum Beispiel durch eine spezielle Lösung und eine Basis des Lösungsraums des zugehörigen homogenen Gleichungssystems.

Bei nichtlinearen Gleichungssystemen gibt es im allgemeinen keine solche Beschreibung unendlicher Lösungsmengen: Die Lösungsmenge des Gleichungssystems

$$X^2 + 2Y^2 + 3Z^2 = 100 \quad \text{und} \quad 2X^2 + 3Y^2 - Z^2 = 0$$

etwa ist die Schnittmenge eines Ellipsoids mit einem elliptischen Kegel; sie besteht aus zwei ovalen Kurven höherer Ordnung. Die GRÖBNER-Basis besteht in diesem Fall aus den beiden Polynomen

$$X^2 - 11Z^2 + 300 \quad \text{und} \quad Y^2 + 7Z^2 - 200\,,$$

stellt uns dieselbe Menge also dar als Schnitt eines hyperbolischen und eines elliptischen Zylinders. Eine explizitere Beschreibung der Lösungsmenge ist schwer vorstellbar.

Auf der Basis von STURMschen Ketten, dem Lemma von THOM und Verallgemeinerungen davon hat die semialgebraische Geometrie Methoden entwickelt, wie man auch allgemeinere Lösungsmengen nichtlinearer Gleichungssysteme durch eine sogenannte zylindrische Zerlegung qualitativ beschreiben kann; dazu wird der $\mathbb{R}^n$ in Teilmengen zerlegt, in denen die Lösungsmenge entweder ein einfaches qualitatives Verhalten hat oder aber leeren Durchschnitt mit der Teilmenge. Dadurch kann man insbesondere feststellen, in welchen Regionen des $\mathbb{R}^n$ Lösungen zu finden sind; diese Methoden sind Gegenstand der reell-algebraischen Geometrie.

In manchen Fällen lassen sich Lösungsmengen parametrisieren; wie man mit Methoden der algebraischen Geometrie zeigen kann, ist das aber im allgemeinen nur bei Gleichungen kleinen Grades der Fall und kommt daher für allgemeine Lösungsalgorithmen nicht in Frage.

Stets möglich ist das umgekehrte Problem, d.h. die Beschreibung einer parametrisch gegebenen Menge in impliziter Form. Hier gehen wir aus von Gleichungen der Form

$$x_1 = \varphi_1(t_1, \ldots, t_m), \quad \ldots, \quad x_n = \varphi_n(t_1, \ldots, t_m)\,,$$

und wir suchen Polynome $f_1, \ldots, f_r$ aus $k[X_1, \ldots, X_n]$, die auf der Menge aller jener $(x_1, \ldots, x_n)$ verschwinden, für die es eine solche Darstellung gibt (und eventuell noch auf Grenzwerten davon).

Dazu wählen wir eine lexikographische Ordnung auf dem Polynomring $k[T_1, \ldots, T_m, X_1, \ldots, X_n]$, bei der alle T_i größer sind als die X_j, und bestimmen eine GRÖBNER-Basis für das von den Polynomen $X_i - \varphi_i(T_1, \ldots, T_m)$ erzeugte Ideal. Dessen Schnitt mit $k[X_1, \ldots, X_n]$ ist ein Eliminationsideal, hat also als Basis genau die Polynome aus der GRÖBNER-Basis, in denen keine T_i vorkommen.

Fast genauso können wir auch zu einer vorgegebenen endlichen Menge von Punkten ein Gleichungssystem konstruieren, das genau diese Menge als Lösungsmenge hat; dies spielt beispielsweise in der algebraischen Statistik eine Rolle, wenn zu einem vorgegebenen Design die damit schätzbaren Modelle identifiziert werden sollen.

Wir gehen aus von r Punkten

$$P_i = \left(x_1^{(i)}, \ldots, x_n^{(i)}\right) \in k^n, \quad i = 1, \ldots, r,$$

und suchen ein Ideal $I \lhd k[X_1, \ldots, X_n]$, dessen Elemente genau in den Punkten P_i verschwinden. Im Falle nur eines Punktes P_i können wir einfach das Ideal

$$I_i = \left(X_1 - x_1^{(i)}, \ldots, X_n - x_n^{(i)}\right)$$

nehmen; bei mehreren Punkten brauchen wir den Durchschnitt der Ideale I_1 bis I_r, für den wir kein offensichtliches Erzeugendensystem haben.

Betrachten wir stattdessen die Punkte

$$Q_i = \left(t_1^{(i)}, \ldots, t_r^{(i)}, x_1^{(i)}, \ldots, x_n^{(i)}\right) \in k^{r+n} \quad \text{mit} \quad t_j^{(i)} = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{sonst} \end{cases},$$

so erzeugen die Polynome

$$\left(X_j - x_j^{(i)}\right) T_i \in k[T_1, \ldots, T_r, X_1, \ldots, X_n]$$

für $i = 1, \ldots, n$ und $j = 1, \ldots, r$ zusammen mit dem Polynom $T_1 + \cdots + T_r - 1$ ein Ideal, das alle Punkte Q_i als Nullstellen hat: Die Polynome $\left(X_j - x_j^{(i)}\right) T_i$ verschwinden in Q_i, da $x_j^{(i)}$ die j-te Koordinate von Q_i ist, und für $\ell \neq i$ verschwindet $\left(X_j - x_j^{(i)}\right) T_\ell$, da $t_\ell^{(i)}$ verschwindet.

Ist umgekehrt $Q = (t_1, \ldots, t_r, x_1, \ldots, x_n) \in k^{r+n}$ keiner der Punkte Q_i, so gibt es für jedes i mindestens eine Koordinate, in der sich Q von Q_i unterscheidet. Ist dies etwa die j-te Koordinate, so ist $X_j - x_j^{(i)}$ in Q von Null verschieden; $\left(X_j - x_j^{(i)}\right) T_i$ kann daher nur verschwinden, wenn $t_i = 0$ ist. Dies kann aber nicht für alle i der Fall sein, denn die Summe der t_i ist eins, da $T_1 + \cdots + T_r - 1$ verschwindet. Somit liegt Q nicht in $V(J)$.

Damit haben wir ein Ideal $J \lhd k[T_1, \ldots, T_r, X_1, \ldots, X_n]$ gefunden, dessen Nullstellen genau die Punkte $Q_1, \ldots, Q_r \in k^{r+n}$ sind. Die Punkte $P_1, \ldots, P_r$ sind die Projektionen der Q_i von k^{n+r} nach k^n; deshalb ist klar, daß alle Polynome aus

$$I \underset{\text{def}}{=} J \cap k[X_1, \ldots, X_n]$$

in den Punkten P_i verschwinden. Wir erhalten ein Erzeugendensystem dieses Ideals, indem wir bezüglich einer Eliminationsordnung für $T_1, \ldots, T_r$ eine GRÖBNER-Basis von J berechnen und davon nur die Polynome betrachten, die keine der Variablen T_i enthalten.

§2: Der Hilbertsche Nullstellensatz

Wie wir wissen, stimmen die Lösungsmengen zweier Gleichungssysteme

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

und

$$g_1(x_1, \ldots, x_n) = \cdots = g_p(x_1, \ldots, x_n) = 0$$

überein, wenn die Ideale $(f_1, \ldots, f_m)$ und $(g_1, \ldots, g_p)$ übereinstimmen. Umgekehrt folgt aber nicht aus der Gleichheit der Lösungsmengen, daß auch die Ideale gleich sein müssen. In diesem Paragraphen wollen wir genauer untersuchen, was hier gilt.

Als erstes müssen wir uns überlegen, *wo* wir nach Lösungen suchen: Wie wir bereits in §1 gesehen haben, wird die Lösungsmenge über dem kleinsten Körper, der alle Koeffizienten der Polynome enthält, oft leer sein, obwohl es Lösungen in größeren Körpern gibt. In den meisten Beispielen betrachten wir $k = \mathbb{Q}$ und $K = \mathbb{C}$ sein; wie wir wissen hat in $\mathbb{C}$ zumindest jedes nichtkonstante Polynom in einer Veränderlichen eine Nullstelle. Körper mit dieser Eigenschaft bezeichnen wir als *algebraisch abgeschlossen:*

Definition: Ein Körper k heißt *algebraisch abgeschlossen*, wenn jedes nichtkonstante Polynom $f \in k[X]$ mindestens eine Nullstelle in k hat.

Durch Polynomdivision folgt leicht induktiv:

Lemma: Ist k algebraisch abgeschlossen, so läßt sich jedes Polynom vom Grad d aus $k[X]$ schreiben als

$$f = c(X - x_1) \cdots (X - x_d) \quad \text{mit} \quad c \in k \setminus \{0\} \ \text{ und } \ x_1, \ldots, x_d \in k\,.$$

Die x_i müssen dabei nicht notwendigerweise verschieden sein.

■

Der Körper K soll im folgenden stets algebraisch abgeschlossen sein; zur Vereinfachung der Beweise wollen wir zusätzlich annehmen, daß er überabzählbar viele Elemente enthält. Die Sätze aus diesem Paragraphen gelten zwar auch ohne diese Zusatzvoraussetzung, jedoch erfordern die Beweise dann einen größeren Aufwand.

Sei also für den Rest dieses Paragraphen k irgendein Körper, und K sei ein algebraisch abgeschlossener Körper mit überabzählbar vielen Elementen, der k enthält.

Als erstes wollen wir uns mit der Frage beschäftigen, für welche Ideale $I \lhd k[X_1, \ldots, X_n]$ die Lösungsmenge $V_K(I)$ in K^n leer ist. Ein Beispiel ist offensichtlich: Natürlich ist $I = k[X_1, \ldots, X_n]$ ein Ideal, und da es insbesondere die Konstante eins enthält, ist $V_K(I) = \emptyset$. Eine (schwache) Form des HILBERTschen Nullstellensatzes besagt, daß dies das einzige Beispiel ist. Zur Vorbereitung des Beweises definieren wir

Definition: R sei ein Ring.
a) $I \lhd R$ ist ein *echtes* Ideal, falls $I \neq R$.
b) Ein echtes Ideal $\mathfrak{m} \lhd R$ heißt *maximales* Ideal, wenn R das einzige Ideal ist, das $\mathfrak{m}$ als echte Teilmenge enthält.
c) Ein echtes Ideal $\mathfrak{p} \lhd R$ heißt *Primideal*, wenn gilt: Liegt für zwei Elemente $f, g \in R$ das Produkt fg in $\mathfrak{p}$, so liegt mindestens einer der Faktoren f, g in $\mathfrak{p}$.

Wie aus der Zahlentheorie bekannt, teilt eine Primzahl p genau dann das Produkt zweier Zahlen a, b, wenn sie mindestens einen der beiden Faktoren teilt; in $\mathbb{Z}$ sind also die von den Primzahlen erzeugten Hauptideale Primideale. Dazu kommt wegen der Nullteilerfreiheit auch noch das Nullideal.

Durch vollständige Induktion beweist man leicht

Lemma: Ist $\mathfrak{p}$ ein Primideal und liegt ein Produkt $f_1 \cdots f_n$ von Elementen $f_i \in R$ in $\mathfrak{p}$, so liegt mindestens einer der Faktoren f_i in $\mathfrak{p}$. ∎

Lemma: Jedes maximale Ideal $\mathfrak{m} \lhd R$ ist ein Primideal.

Beweis: Das Produkt fg zweier Elemente $f, g \in R$ liege in $\mathfrak{m}$. Falls $f \in \mathfrak{m}$ sind wir fertig; andernfalls ist $\mathfrak{m}+(f) = R$ wegen der Maximalität von $\mathfrak{m}$; es gibt also Elemente $m \in \mathfrak{m}$ und $h \in R$, so daß $m + hf = 1$ ist. Damit ist $g = mg + hfg \in \mathfrak{m}$, denn $m \in \mathfrak{m}$ und $fg \in \mathfrak{m}$. ∎

Lemma: Jedes echte Ideal $I \lhd k[X_1, \ldots, X_n]$ liegt in einem maximalen Ideal $\mathfrak{m} \lhd k[X_1, \ldots, X_n]$.

Beweis: Falls I selbst maximal ist, sind wir fertig; andernfalls gibt es ein echtes Ideal I_1, das I als echte Teilmenge enthält. Auch wenn I_2 ein maximales Ideal ist, sind wir fertig; andernfalls gibt es ein echtes Ideal I_3, das I_2 als echte Teilmenge enthält, und so weiter. Wenn dieses Verfahren nach endlich vielen Schritten abbricht, haben wir ein maximales Ideal gefunden, das I enthält; andernfalls gibt es eine unendliche aufsteigende Folge von Idealen $I \subset I_1 \subset I_2 \subset \cdots$. Die Vereinigung aller I_j ist selbst ein Ideal in $k[X_1, \ldots, X_n]$ und hat damit nach dem HILBERTschen Basissatz ein endliches Erzeugendensystem $\{f_1, \ldots, f_m\}$. Jedes f_i liegt in einem der Ideale I_j und damit auch in allen I_ℓ mit $\ell > j$. Wegen der Endlichkeit des Erzeugendensystems gibt es daher einen Index r derart, daß alle f_i in I_r liegen. Dann ist aber $I = I_r = I_{r+1} = \cdots$, im Widerspruch zu der Annahme, daß jedes I_j echte Teilmenge von I_{j+1} ist. Somit bricht das Verfahren nach endlich vielen Schritten ab und liefert ein maximales Ideal $\mathfrak{m}$, in dem I enthalten ist. ∎

(Tatsächlich gilt auch dieses Lemma für beliebige Ringe; da dort der HILBERTsche Basissatz nicht gelten muß, beweist man es im allgemeinen Fall mit Hilfe des ZORNschen Lemmas.)

Schwache Form des Hilbertschen Nullstellensatzes: Für ein echtes Ideal $I \lhd k[X_1, \ldots, X_n]$ ist $V_K(I) \neq \emptyset$.

Beweis: Nach dem HILBERTschen Basissatz hat jedes Ideal I ein endliches Erzeugendensystem $\{f_1, \ldots, f_m\}$. Wir betrachten das von den f_i erzeugte Ideal $\bar{I}$ in $K[X_1, \ldots, X_n]$. Da eine Basis des k-Vektorraums $k[X_1, \ldots, X_n]/I$ auch Basis des K-Vektorraums $K[X_1, \ldots, X_n]/\bar{I}$ ist, muß auch $\bar{I}$ ein echtes Ideal von $K[X_1, \ldots, X_n]$ sein und liegt somit in

einem maximalen Ideal $\mathfrak{m} \lhd K[X_1, \ldots, K_n]$. Der Satz folgt somit aus der folgenden alternativen Version des HILBERTschen Nullstellensatzes:

Satz: Die maximalen Ideale $\mathfrak{m} \lhd K[X_1, \ldots, X_n]$ sind genau die Ideale

$$\mathfrak{m} = (X_1 - x_1, \ldots, X_n - x_n) \quad \text{mit} \quad (x_1, \ldots, x_n) \in K^n \, .$$

Beweis: $\mathfrak{m} = (X_1 - x_1, \ldots, X_n - x_n)$ ist der Kern der Abbildung

$$\begin{cases} K[X_1, \ldots, X_n] \to K \\ \qquad\qquad f \mapsto f(x_1, \ldots, x_n) \end{cases} .$$

Ist daher I ein Ideal, das $\mathfrak{m}$ echt enthält, so muß der Vektorraum $K[X_1, \ldots, X_n]/I$ ein echter Untervektorraum von $K[X_1, \ldots, X_n]/\mathfrak{m}$ sein. Da letzterer nach dem Homomorphiesatz isomorph zum eindimensionalen Vektorraum K ist, muß dies der Nullraum sein. Somit ist $I = K[X_1, \ldots, X_m]$, d.h. $\mathfrak{m}$ ist ein maximales Ideal.

Umgekehrt sei $\mathfrak{m}$ ein maximales Ideal. Wenn wir zeigen können, daß es Elemente $x_1, \ldots, x_n$ gibt, für die $X_i - x_i$ in $\mathfrak{m}$ liegt, ist $(X_1 - x_1, \ldots, X_n - x_n) \subseteq \mathfrak{m}$, und da links ein maximales Ideal steht, müssen beide Seiten gleich sein.

Angenommen, es gibt ein $i \in \{1, \ldots, n\}$, für das $X_i - x$ für kein $x \in K$ im Ideal $\mathfrak{m}$ liegt. Wegen der Maximalität von $\mathfrak{m}$ ist dann

$$\mathfrak{m} + (X - x) = K[X_1, \ldots, X_n] \quad \text{für alle } x \in K \, .$$

Somit gibt es für jedes $x \in K$ ein Polynom $f_x \in \mathfrak{m}$ sowie ein Polynom $h_x \in K[X_1, \ldots, X_n]$ derart, daß

$$f_x + h_x \cdot (X_i - x) = 1$$

ist. Da $1 \notin \mathfrak{m}$, ist dabei $h_x \neq 0$. Wir wählen für jedes $x \in K$ ein festes Polynom h_x (und damit auch f_x), das obige Gleichung erfüllt, und setzen $K_d = \{x \in K \mid \deg h_x = d\}$ für jedes $d \in \mathbb{N}_0$. Da K nach Voraussetzung überabzählbar viele Elemente enthält und K die Vereinigung der K_d ist, muß mindestens eine der Mengen K_d unendlich viele Elemente enthalten. (Nur an dieser Stelle geht die Voraussetzung der Überabzählbarkeit ein.)

Wir wählen eine solche Menge K_d und betrachten den Vektorraum $K[X_1, \ldots, K_n]_d$ aller Polynome vom Grad höchstens d. Da es nur endlich viele Monome vom Grad höchstens d gibt, ist dies ein endlichdimensionaler K-Vektorraum. Wir wählen eine natürliche Zahl r, die größer ist als seine Dimension, und dazu r Elemente $x^{(1)}, \ldots, x^{(r)} \in K$ mit $h_{x^{(i)}} \in k[X_1, \ldots, X_n]_d$. Dann muß es Elemente $\lambda_1, \ldots, \lambda_r \in K$ geben, die nicht allesamt verschwinden, derart, daß

$$\lambda_1 h_{x^{(1)}} + \cdots + \lambda_r h_{x^{(r)}} = 0$$

ist.

Dazu definieren wir

$$g = \sum_{j=1}^{r} \lambda_j \prod_{\ell \neq j} \left(X_i - x^{(\ell)}\right) \in K[X_i]\,.$$

Dieses Polynom liegt auch in $\mathfrak{m}$, denn wegen

$$1 = f_{x^{(j)}} + h_{x^{(j)}}\left(X_i - x^{(j)}\right) \quad \text{für } j = 1, \ldots, r$$

ist

$$\begin{aligned} g &= \sum_{j=1}^{r} \lambda_j \left(f_{x^{(j)}} + h_{x^{(j)}}\left(X_i - x^{(j)}\right)\right) \prod_{\ell \neq j} \left(X_i - x^{(\ell)}\right) \in K[X_i] \\ &= \sum_{j=1}^{r} \lambda_j f_{x^{(j)}} \prod_{\ell \neq j} \left(X_i - x^{(\ell)}\right) + \left(\sum_{j=1}^{r} \lambda_j h_{x^{(j)}}\right) \prod_{\ell=1}^{n} \left(X_i - x^{(\ell)}\right) \\ &= \sum_{j=1}^{r} \lambda_j \prod_{\ell \neq j} \left(X_i - x^{(\ell)}\right) f_{x^{(j)}} \in \mathfrak{m}\,, \end{aligned}$$

da $\sum_{j=1}^{r} \lambda_j h_{x^{(j)}}$ verschwindet und alle $f_{x^{(j)}}$ in $\mathfrak{m}$ liegen.

g ist nicht das Nullpolynom, denn für jeden Index ν ist

$$g\left(x^{(\nu)}\right) = \sum_{j=1}^{r} \lambda_j \prod_{\ell \neq j} \left(x^{(\nu)} - x^{(\ell)}\right) = \lambda_\nu \prod_{\ell \neq \nu} \left(x^{(\nu)} - x^{(\ell)}\right)\,.$$

Da die $x^{(\ell)}$ paarweise verschieden sind und mindestens ein λ_ν nicht verschwindet, muß mindestens einer dieser Werte von Null verschieden sein.

Da g in $\mathfrak{m}$ liegt, kann g auch keine von Null verschiedene Konstante sein, hat also einen positiven Grad e. Über dem algebraisch abgeschlossenen Körper K zerfällt g daher in Linearfaktoren:

$$g = c(X_i - z_1)\dots(X_i - z_e) \quad \text{mit} \quad c \in K \setminus \{0\},\ z_1, \dots z_e \in k\,.$$

g liegt in $\mathfrak{m}$, aber nach Voraussetzung liegt keiner der Faktoren $X_i - z_j$ in $\mathfrak{m}$, und die Konstante $c \neq 0$ natürlich auch nicht. Dies ist ein Widerspruch, denn als maximales Ideal ist $\mathfrak{m}$ insbesondere ein Primideal. ∎

Somit hat also jedes echte Ideal $I \lhd k[X_1, \dots, X_n]$ zumindest in einem Erweiterungskörper K von k mindestens eine Nullstelle. Damit folgt umgekehrt

Satz: Das Gleichungssystem

$$f_1(x_1, \dots, x_n) = \dots = f_m(x_1, \dots, x_n) = 0$$

mit $f_1, \dots, f_m \in k[X_1, \dots, X_n]$ ist genau dann in jedem Erweiterungskörper K von k unlösbar, wenn es Polynome $h_1, \dots, h_m$ in $X_1, \dots, X_n$ gibt, so daß $h_1 f_1 + \dots + h_m f_m = 1$ ist.

Beweis: Im Falle der Unlösbarkeit ist das von $f_1, \dots, f_m$ erzeugte Ideal der ganze Polynomring, enthält also insbesondere die Eins. Da

$$(f_1, \dots, f_m) = \left\{ h_1 f_1 + \dots + h_m f_m \;\middle|\; h_1, \dots, h_m \in k[X_1, \dots, X_n] \right\},$$

hat auch die Eins eine Darstellung der verlangten Form.

Ist umgekehrt $h_1 f_1 + \dots + h_m f_m = 1$ für irgendwelche Polynome $h_1, \dots, h_m$, so ist für jeden Erweiterungskörper K von k und jedes n-Tupel $(x_1, \dots, x_n) \in K^n$

$$h_1(x_1, \dots, x_n) f_1(x_1, \dots, x_n) + \dots + h_m(x_1, \dots, x_n) f_m(x_1, \dots, x_n) = 1\,,$$

so daß nicht alle $f_j(x_1, \dots, x_n)$ verschwinden können. ∎

Wenn wir eine GRÖBNER-Basis eines Ideals I kennen, ist es einfach zu entscheiden, ob $I = k[X_1, \dots, X_n]$ ist (oder äquivalent, ob $1 \in I$): Da

der führende Term eines jeden Polynoms aus I 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.

Aus dem gerade bewiesenen Satz folgt mit einem 1929 von J.L. RABINOWITSCH gefundenen Trick die

Starke Form des Hilbertschen Nullstellensatzes: k sei ein beliebiger Körper und K ein überabzählbarer algebraisch abgeschlossener Erweiterungskörper von k. Falls für ein Ideal $I \lhd k[X_1, \ldots, X_n]$ ein Polynom $f \in k[X_1, \ldots, X_n]$ auf ganz $V_K(I)$ verschwindet, gibt es ein $q \in \mathbb{N}$, so daß f^q in I liegt.

Beweis: Wir erweitern den Polynomring $k[X_1, \ldots, X_n]$ mit einer neuen Variablen X_{n+1} zu $k[X_1, \ldots, X_{n+1}]$ und betrachten dort für ein Erzeugendensystem $\{f_1, \ldots, f_m\}$ von I das Gleichungssystem

$$f_1(x_1, \ldots, x_n) = \cdots = f_m(x_1, \ldots, x_n) = 1 - x_{n+1} f(x_1, \ldots, x_n) = 0 \,.$$

Für jeden Punkt $(x_1, \ldots, x_n, x_{n+1}) \in K^{n+1}$, für den die $f_j(x_1, \ldots, x_n)$ verschwinden, verschwindet auch $f(x_1, \ldots, x_n)$, d.h.

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

Somit haben diese $n + 1$ Gleichungen keine gemeinsame Nullstelle; es gibt also Polynome $h_1, \ldots, h_{m+1} \in k[X_1, \ldots, X_{n+1}]$ derart, daß

$$h_1 f_1 + \cdots + h_m f_m + h_{m+1}(1 - X_{n+1} f) = 1$$

ist. Diese Gleichung bleibt gültig, wenn wir überall für X_{n+1} ein Polynom oder eine rationale Funktion in $X_1, \ldots, X_n$ einsetzen; wir setzen $X_{n+1} = 1/f$. Die h_j werden dann zu rationalen Funktionen in $X_1, \ldots, X_n$, wobei alle Nenner Potenzen von f sind. Ist f^q die höchste dieser Potenzen, so erhalten wir nach Multiplikation mit f^q eine Gleichung der Form

$$\widetilde{h}_1 f_1 + \cdots + \widetilde{h}_m f_m = f^q$$

mit $\widetilde{h}_j = f^q h_j\left(X_1, \ldots, X_n, 1/f\right) \in k[X_1, \ldots, X_n]$. Dies zeigt, daß f^q in $I = (f_1, \ldots, f_m)$ liegt.

■

Definition: R sei ein Ring und $I \lhd R$ ein Ideal von R. Das *Radikal* von I ist die Menge

$$\sqrt{I} \underset{\text{def}}{=} \left\{ f \in R \;\middle|\; \exists q \in \mathbb{N} : f^q \in I \right\}.$$

Das Radikal besteht also aus allen Ringelementen, die eine Potenz in I haben. Es ist selbst ein Ideal, denn sind $f, g \in \sqrt{I}$ zwei Elemente mit $f^p \in I$ und $g^q \in I$, so sind in

$$(f+g)^{p+q} = \sum_{\ell=0}^{p+q} \binom{p+q}{\ell} f^{p+q-\ell} g^{\ell}$$

die ersten q Summanden Vielfache von f^p, und die restlichen p sind Vielfache von g^q. Somit liegt jeder Summand in I, also auch die Summe. Für ein beliebiges $r \in R$ liegt natürlich auch rf in $\sqrt{I}$, denn seine q-te Potenz $(rf)^q = r^q f^q$ liegt in I, sobald f^q in I liegt.

Mit diesem neuen Begriff können wir den obigen Satz umformulieren:

Satz: Ein Polynom $f \in k[X_1, \ldots, X_n]$ verschwindet genau dann auf $V_K(I)$, wenn $f \in \sqrt{I}$.

■

Anders ausgedrückt heißt dies

Satz: Für zwei Ideale $I, J \lhd k[X_1, \ldots, X_n]$ ist $V_K(I) = V_K(J)$ genau dann, wenn $\sqrt{I} = \sqrt{J}$ ist.

■

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, denn nach dem HILBERTschen Basissatz hat es ein endlich Erzeugendensystem $\{f, \ldots, f_m\}$ und wird daher offensichtlich von $f = \text{ggT}(f_1, \ldots, f_m)$ erzeugt. 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. Im nächsten Paragraphen wollen wir uns überlegen, wie man ähnliche Ergebnisse auch für Systeme von Polynomgleichungen in mehreren Veränderlichen finden kann.

§3: Gleichungssysteme mit endlicher Lösungsmenge

Auch hier gehen wir wieder aus von einem beliebigen Körper k sowie einem algebraisch abgeschlossenen Erweiterungskörper K mit überabzählbar vielen Elementen. Letztere Bedingung ist nur notwendig, weil wir sie im Beweis des HILBERTschen Nullstellensatzes verwendet haben; wie bereits dort erwähnt, gibt es auch Beweise für den Fall, daß K ein beliebiger algebraisch abgeschlossener Körper ist, so daß alle Sätze dieses Paragraphen tatsächlich auch ohne die Voraussetzung der Überabzählbarkeit von K gelten.

Satz: I sei ein Ideal im Polynomring $k[X_1, \ldots, X_n]$ über dem Körper k, und K sei ein überabzählbarer algebraisch abgeschlossener Körper, in dem k enthalten sei. Dann gilt: $V_K(I)$ ist genau dann endlich, wenn der Faktorring $A = k[X_1, \ldots, X_n]/I$ ein endlichdimensionaler k-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 daher die i-te Koordinate eine Nullstelle dieses Polynoms sein. Damit kann die i-te Koordinate nur endlich viele Werte annehmen, und da dies für alle i gilt, ist $V_K(I)$ endlich.

2. Schritt: $\bar{I}$ sei das von I in $K[X_1, \ldots, X_n]$ erzeugte Ideal. Wenn $V_K(I)$ endlich ist, hat der K-Vektorraum $\bar{A} = K[X_1, \ldots, X_n]/\bar{I}$ 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 $\{b_1, \ldots, b_r\}$ und zu jedem Basiselement b_i ein Polynom $B_i \in k[X_1, \ldots, X_n]$, das modulo I gleich b_i ist. Zusammen mit einer Basis von I als k-Vektorraum bilden die B_i dann eine k-Vektorraumbasis von $k[X_1, \ldots, X_n]$. Über K wird die Basis von I zu einer K-Vektorraumbasis von $\bar{I}$, da sich jedes Element von $\bar{I}$ als eine K-Linearkombination von Elementen aus I schreiben läßt. Zusammen mit den B_i, die wir auch als Elemente von $K[X_1, \ldots, X_n]$ aufâã|fassen können, erhalten wir sowohl über k als auch über K eine Basis des ganzen jeweiligen Polynomrings, und damit ist klar, daß die Restklassen der B_i modulo $\bar{I}$ den Faktorring $\bar{A}$ erzeugen. Somit ist dieser als K-Vektorraum endlichdimensional.

Die Gleichheit von $\dim_k A$ und $\dim_K \bar{A}$ folgt, falls wir zeigen können, daß die Restklassen der B_i modulo I linear unabhängig sind.

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. $\bar{\lambda}_1 B_1 + \cdots + \bar{\lambda}_r B_r$ liegt daher in $\bar{I}$, also erst recht in $\bar{I}$, so daß auch $\lambda_1 \bar{b}_1 + \cdots + \lambda_r \bar{b}_r$ 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

$$\lambda_1 B_1 + \cdots + \lambda_r B_r = g_1 f_1 + \cdots + g_m f_m \,.$$

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

$$\lambda_1 B_1 + \cdots + \lambda_r B_r - \sum_{j=1}^{m} \sum_{\ell=1}^{r_j} \mu_{j\ell} M_{j\ell} f_j = 0$$

mit Elementen $\mu_{j\ell} \in K$, 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. Seine Lösungsmenge über k ist ein k-Vektorraum, für den uns der GAUSS-Algorithmus eine Basis liefert. Da der GAUSS-Algorithmus nirgends aus dem Körper hinausführt, in dem die Koeffizienten liegen, ist dies auch eine Basis des Lösungsraums über K; die beiden Vektorräume haben also dieselbe Dimension. Da wir wissen, daß es über K eine nichttriviale Lösung gibt, muß es daher auch über k eine geben,

Es gibt somit 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 über k, also ist A endlichdimensional.

Im nächsten Schritt wollen wir das Zählen der Lösungen zurückführen auf das Zählen von Nullstellen eines Polynoms einer Veränderlichen.

Definition: Ein Polynom $u \in K[X_1, \ldots, X_n]$ heißt *separierend*, wenn es für keine zwei Elemente von $V_K(I)$ denselben Wert annimmt.

4. Schritt: Falls $V_K(I)$ endlich ist, gibt es ein separierendes homogenes lineares Polynom $u = c_1X_1 + \cdots c_nX_n$. Wir können dabei für u eines der speziellen Polynome

$$u_a = X_1 + aX_2 + a^2X_3 + \cdots + a^{n-1}X_n$$

wählen, wobei a in einer beliebig vorgebbaren Teilmenge von K mit mehr als $(n-1)\binom{s}{2} = \frac{1}{2}s(s-1)(n-1)$ Elementen liegt.

Für je zwei verschiedene Punkte $z, w \in V_K(I)$ ist $u_a(z) = u_a(w)$ genau dann, wenn

$$(z_1 - w_1) + (z_2 - w_2)a + (z_3 - w_3)a^2 + \cdots + (z_n - w_n)a^{n-1}$$

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. Ist $s = \#V_K(I)$ endlich, so gibt es $\binom{s}{2}$ Paare aus voneinander verschiedenen Elementen; somit gibt es höchstens $(n-1)\binom{s}{2}$ Elemente $a \in K$, für die $u_a(z) = u_a(w)$ für *irgendwelche* voneinander verschiedene Elemente von $V_K(I)$.

(Hier haben wir benutzt, daß jeder algebraisch abgeschlossene Körper unendlich ist. Falls bereits k unendlich ist, etwa $k = \mathbb{Q}$, können wir sogar ein $a \in k$ finden 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. Im hier meistens betrachteten Fall $k = \mathbb{Q}$ können wir etwa eine ganze Zahl a mit $0 \leq a \leq (n-1)\binom{s}{2}$ wählen.

5. Schritt: Die Elementanzahl s 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 4. 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 $\bar{I}$ in $\bar{A}$ und wollen uns überlegen, daß die Elemente $1, \tilde{u}, \ldots, \tilde{u}^{s-1} \in \bar{A}$ linear unabhängig sind: Angenommen, es gibt eine Relation der Form $\sum_{\ell=0}^{s-1} \lambda_\ell \tilde{u}^\ell = 0$ mit $\lambda_\ell \in K$. Das Polynom $\sum_{\ell=0}^{s-1} \lambda_\ell u^\ell \in K[X_1, \ldots, X_n]$ liegt dann in $\bar{I}$, verschwindet also für jedes der s Elemente von $V_K(I)$. Da u für jedes dieser Elemente einen anderen Wert annimmt, hat das Polynom $\sum_{\ell=0}^{r-1} \lambda_\ell U^\ell \in k[U]$ einerseits mindestens s verschiedene Nullstellen in K, andererseits ist sein Grad kleiner als s. Das ist nur für das Nullpolynom möglich; somit verschwinden alle Koeffizienten λ_ℓ, was die behauptete lineare Unabhängigkeit beweist. Damit enthält $\bar{A}$ mindestens s linear unabhängige Elemente, d.h. $r = \dim_K \bar{A} \geq s = \#V_K(I)$. Damit ist die Behauptung und auch der gesamte Satz bewiesen.

■

Betrachten wir als Beispiel das von $f = X^2 + Y^2 - 1$ und $g = X - Y$ erzeugte Ideal $I \lhd \mathbb{Q}[X, Y]$. Seine Lösungsmenge ist, geometrisch gesehen, der Schnitt des Einheitskreises mit der ersten Winkelhalbierenden, besteht also aus den beiden Punkten $\left(\frac{1}{2}\sqrt{2}, \frac{1}{2}\sqrt{2}\right)$ und $\left(-\frac{1}{2}\sqrt{2}, -\frac{1}{2}\sqrt{2}\right)$.

Der Polynomring $\mathbb{Q}[X, Y]$ hat als $\mathbb{Q}$-Vektorraum eine Basis bestehend aus allen Monomen $X^a Y^b$ mit $a, b \in \mathbb{N}_0$. Modulo I sind X und Y äquivalent, und damit ist $X^a Y^b \sim X^{a+b}$. Außerdem ist $2X^2$ äquivalent zu $X^2 + Y^2$, und das wiederum ist wegen f äquivalent zu 1, d.h. $X^2 \sim \frac{1}{2}$. Daher ist jedes Monom äquivalent entweder zu einer Konstanten (falls $a + b$ gerade) oder einem skalaren Vielfachen von X. Da I kein Polynom der Form $\lambda X + \mu$ enthält, sind X und 1 modulo I linear unabhängig; somit bilden ihre Restklassen eine Basis des Vektorraums $\mathbb{Q}[X, Y]/I$.

Ersetzen wir in diesem Beispiel g durch $X^2 - Y^2 = (X + Y)(X - Y)$, so schneiden wir den Kreis mit beiden Winkelhalbierenden und haben nun eine vierelementige Lösungsmenge

$$V_{\mathbb{C}}(I) = \left\{ \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), \left(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right), \left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right) \right\}.$$

Modulo dem neuen Ideal I sind X und Y nicht mehr äquivalent, sondern nur noch X^2 und Y^2. Jedes Monom ist somit äquivalent entweder zu

einer X-Potenz oder zu einem Monom der Form X^aY. Da auch hier $X^2 \sim \frac{1}{2}$, ist es somit äquivalent zu einem skalaren Vielfachen eines der Monome $1, X, Y$ oder XY. Da keine Linearkombination dieser vier Monome in I liegt, bilden ihre Restklassen eine Basis von $\mathbb{Q}[X,Y]/I$.

In diesen beiden Beispielen waren sowohl die Lösungsmengen als auch Basen der Faktorringe einfach zu finden; im Allgemeinen ist das eher nicht der Fall. Wenn wir eine GRÖBNER-Basis des Ideals I kennen, können wir leicht eine Vektorraumbasis des Faktorrings konstruieren:

Definition: $I \lhd k[X_1, \ldots, X_n]$ sei ein Ideal und G sei eine GRÖBNER-Basis bezüglich irgendeiner Monomordnung auf $k[X_1, \ldots, X_n]$. Ein Monom in $X_1, \ldots, X_n$ heißt *Standardmonom* (bezüglich G), wenn es für kein $g \in G$ durch das führende Monom von g teilbar ist.

Satz: Für jede GRÖBNER-Basis G eines Ideals $I \lhd k[X_1, \ldots, X_n]$ bilden die Restklassen der Standardmonome eine Vektorraumbasis von $k[X_1, \ldots, X_n]/I$.

Beweis: Zunächst sind diese Restklassen linear unabhängig, denn jede nichttriviale Linearkombination der Null entspräche einem Polynom h aus I, dessen sämtliche Monome Standardmonome sind. Da die führenden Monome der Elemente von G das Ideal FM(I) erzeugen, müßte daher FM(h) Vielfaches eines FM(g) mit $g \in G$ sein, was der Definition eines Standardmonoms widerspricht.

Für ein beliebiges $f \in k[X_1, \ldots, X_n]$ liefert uns der Divisionsalgorithmus eine Darstellung

$$f = \sum_{g \in G} a_g g + r \quad \text{mit} \quad a_g, r \in k[X_1, \ldots, X_n]\,,$$

wobei r eine k-Linearkombination von Standardmonomen ist. Da die Summe der $a_g g$ in I liegt, ist f also äquivalent zu einer k-Linearkombination von Standardmonomen, so daß seine Restklasse die entsprechende Linearkombination von deren Restklassen ist. ∎

Dieser Satz gilt unabhängig davon, ob $k[X_1, \ldots, X_n]/I$ als Vektorraum endlichdimensional ist; er liefert uns auch ein einfaches Kriterium dafür,

wann er endliche Dimension hat und wann somit die Lösungsmenge $V_K(I)$ endlich ist:

Lemma: G sei eine GRÖBNER-Basis eines Ideals $I \lhd k[X_1, \ldots, X_n]$ bezüglich irgendeiner Monomordnung. $V_K(I)$ ist genau dann endlich, wenn G für jedes i ein Polynom enthält, dessen führendes Monom eine X_i-Potenz ist.

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 Standardmonome. Somit hat A endliche Dimension, und $V_K(I)$ ist endlich.

Ist umgekehrt $V_K(I)$ endlich, so enthält $\overline{I}$ 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 $\overline{I}$ 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. ∎

§4: Multiplizitäten

Eine Teilmenge eines Ideals I eines Polynomrings ist nach Definition genau dann eine GRÖBNER-Basis, wenn die führenden Monome ihrer Elemente das Ideal FM(I) erzeugen. Im Falle eines Polynomrings in nur einer Veränderlichen X ist das führende Monom die höchste vorkommende X-Potenz; damit gilt:

Satz: Für ein Ideal $I \lhd k[X]$ bildet jedes Element $f \in I$ mit minimalem Grad eine GRÖBNER-Basis von I. Insbesondere ist jedes Ideal ein Hauptideal.

∎

Ist d der minimale Grad eines Polynoms aus I, so sind die Standardmonome gerade die X-Potenzen $1, X, \ldots, X^{d-1}$, d.h. $k[X]/I$ hat die

Dimension d. Die Nullstellen von I in einem algebraisch abgeschlossenen Erweiterungskörper von k sind genau die Nullstellen des erzeugenden Polynoms, und wie wir wissen, ist deren Anzahl, mit Vielfachheiten gezählt, gleich d.

Um auch im mehrdimensionalen Fall zu einer entsprechenden Aussage zu kommen, müssen wir auch hier eine Vielfachheit oder, wie man auch sagt, Multiplizitäten definieren. Im Falle von Polynomen einer Veränderlichen können wir hier mit Ableitungen arbeiten; für $k = \mathbb{R}$ 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 im Punkt z eine r-fache Nullstelle habe. Dann ist $f = (X - z)^r g$ mit einem Polynom $g \in k[X]$, das an der Stelle z nicht verschwindet. Der im vorigen Paragraphen eingeführte Faktorraum $A = K[X]/(f)$ hat als Basis die Potenzen X^ℓ mit $0 \leq \ell < \deg f$; 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 ist aber das Gegenteil der Fall: Wenn wir die üblichen Regeln der Bruchrechnung anwenden, 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 $\ell \geq r$, 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 A, und die Dimensionen der so definierten Unterräume zu den verschiedenen Nullstellen von f ergänzen sich zur Dimension von 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:

$$(f, g) \sim (r, s) \iff \exists h \in R \setminus \{0\} : h(fs - rg) = 0\,.$$

Die Gleichung $h(fs-rg) = 0$ ist natürlich äquivalent zu $h \cdot fs = h \cdot rg$; bis auf den Faktor h entspricht sie also dem aus der Bruchrechnung bekannten Überkreuzmultiplizieren. Fall R nullteilerfrei ist, können wir auf den Faktor h verzichten, denn dann folgt aus $h(fs - rg) = 0$, daß $fs-rg = 0$ sein muß. Wie wir oben gesehen haben, ist $k[X_1, \ldots, X_n]/I$ allerdings im Allgemeinen kein Integritätsbereich.

Die Äquivalenzklasse des Paars (f, g) wird mit $\frac{f}{g}$ bezeichnet, die Menge aller Äquivalenzklassen mit $S^{-1}R$. Addition und Multiplikation definieren wird nach den Regeln der Bruchrechnung als

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

und wir müssen uns überlegen, daß diese Verknüpfungen wohldefiniert sind, daß das Ergebnis also nicht von der Wahl spezieller Repräsentanten (f, g) und (r, s) abhängt:

Sind $(f,g) \sim (\tilde{f},\tilde{g})$ und $(r,s) \sim (\tilde{r},\tilde{s})$, so ist

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

Nach Definition gibt es Elemente $h,t \in R \setminus \{0\}$, so daß $h \cdot f\tilde{g} = h \cdot \tilde{f}g$ und $t \cdot r\tilde{s} = t \cdot \tilde{r}s$ ist. Dann ist

$$\begin{aligned} ht \cdot (\tilde{f}\tilde{s} + \tilde{r}\tilde{g}) \cdot gs &= t \cdot (h \cdot \tilde{f}g)s\tilde{s} + h \cdot (t \cdot \tilde{r}s)g\tilde{g} \\ &= t \cdot (h \cdot f\tilde{g})s\tilde{s} + h \cdot (t \cdot r\tilde{s})g\tilde{g} \\ &= ht \cdot (fs + rg) \cdot \tilde{q}\tilde{s} , \end{aligned}$$

d.h.

$$\left(\tilde{f}\tilde{s} + \tilde{r}\tilde{g}, \tilde{g}\tilde{s}\right) \sim (fs + rg, gs) .$$

Entsprechend ist

$$ht \cdot \tilde{f}\tilde{r}gs = (h \cdot \tilde{f}g)(t \cdot \tilde{r}s) = (h \cdot f\tilde{g})(t \cdot r\tilde{s}) = ht \cdot fr\tilde{g}\tilde{s} ,$$

das heißt

$$\left(\tilde{f}\tilde{r}, \tilde{g}\tilde{s}\right) \sim (fr, gs) .$$

Die größte multiplikativ abgeschlossene Teilmenge eines Integritätsbereichs R ist $S = R \setminus \{0\}$; in diesem Fall ist $S^{-1}R$ ein Körper, den wir als den *Quotientenkörper* Quot R von R kennen.

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 $\frac{h}{1} = \frac{hg}{g} = \frac{0}{g} = 0$.

Weitere typische Beispiele multiplikativ abgeschlossener Teilmengen eines Rings sind die Potenzen eines Nichtnullteilers oder auch das Komplement eines Primideals: Ein Ideal $\mathfrak{p} \lhd R$ heißt bekanntlich *Primideal* wenn für je zwei Elemente $f,g \in R$ mit $fg \in \mathfrak{p}$ mindestens einer der

beiden Faktoren f, g in $\mathfrak{p}$ liegt. Dies ist offensichtlich äquivalent dazu, daß die Menge $R \setminus \mathfrak{p}$ 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 addieren und 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 einen 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 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 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 $\bar{A} \cong \bigoplus_{x \in V_K(I)} \bar{A}_x$

Beweis: Wie wir aus dem vorigem Paragraphen wissen (4. Schritt im Beweis des großen 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 $x \in V_K(I)$ mit x_1 und betrachten die LAGRANGE-Polynome

$$s_x = \frac{\prod_{y \in V_K(I) \setminus \{x\}} (X_1 - y)}{\prod_{y \in V_K(I) \setminus \{x\}} (x_1 - y)} \in K[X_1]\,;$$

offensichtlich ist $s_x(x) = 1$ und $s_x(y) = 0$ für alle $y \neq x$ aus $V_K(I)$. Somit verschwindet das Produkt $s_x s_y$ zweier solcher Funktionen in jedem Punkt von $V_K(I)$; nach dem HILBERTschen Nullstellensatz liegt daher eine Potenz von $s_x s_y$ im Ideal $\bar{I}$. Bezeichnet r den größten Exponenten, den wir für eines der Produkte $s_x s_y$ brauchen, haben daher die Polynome $t_x = s_x^r$ die Eigenschaft, daß $t_x t_y$ für $x \neq y$ in $\bar{I}$ liegt, und $t_x(x) = 1$.

Wir betrachten nun das Ideal $J \lhd K[X_1, \ldots, X_n]$, das von I und den sämtlichen t_x erzeugt wird. Es hat offensichtlich keine gemeinsame Nullstelle, denn die gemeinsamen Nullstellen von $\bar{I}$ sind die $x \in V_K(I)$,

und für jedes dieser x ist $t_x(x) = 1$. Nach der schwachen Form des HILBERTschen Nullstellensatzes enthält J daher die Eins; es gibt also Polynome $p_x \in K[X_1, \ldots, X_n]$ und ein Polynom $p \in \bar{I}$, so daß

$$\sum_{x \in V_K(I)} p_x t_x + p = 1$$

ist. Die Restklassen $e_x \in \bar{A}$ von $p_x t_x$ modulo $\bar{I}$ erfüllen die Gleichungen

1.) $\sum_{x \in V_K(I)} e_x = 1$
2.) $e_x e_y = 0$ für $x \neq y$ aus $V_K(I)$
3.) $e_x^2 = e_x$
4.) $e_x(x) = 1$

Die erste Gleichung ist klar, denn gehen wir in der Gleichung

$$\sum_{x \in V_K(I)} p_x t_t + p = 1$$

zu Restklassen modulo $\bar{I}$ über, wird p zur Klasse der Null und $p_x t_x$ zu e_x. Für $x \neq y$ ist $t_x t_y \in \bar{I}$; modulo $\bar{I}$ verschwindet das Produkt und damit auch $e_x e_y$, was die zweite Gleichung beweist.

Die dritte Gleichung folgt aus den ersten beiden: Nach der ersten ist $1 - e_x$ gleich der Summe der übrigen e_y, also ist

$$e_x - e_x^2 = e_x(1 - e_x) = e_x \sum_{y \neq x} e_y = \sum_{y \neq x} e_x e_y = 0 .$$

Für die vierte Gleichung schließlich beachten wir, daß für $x \neq y$ mit $s_y(x)$ auch $t_y(x)$ verschwindet, d.h.

$$\sum_{y \in V_K(I)} p_y(x) t_y(x) + p(x) = p_x(x) t_x(x) = e_x(x) = 1 .$$

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 $\bar{A}e_x$ isomorph ist zur Lokalisierung von $\bar{A}$ bei x; der Isomorphismus ist gegeben durch

$$\begin{cases} \bar{A}e_x \to \bar{A}_x \\ fe_x \mapsto \dfrac{f}{1} \end{cases} .$$

Zum Nachweis der Bijektivität konstruieren wir eine Umkehrabbildung $\bar{A}_x \to \bar{A}e_x$ wie folgt: Zu jedem $g \in \bar{A}$ mit $g(x) \neq 0$ setzen wir

$$\tilde{g} \underset{\text{def}}{=} \frac{g}{g(x)} - 1 \in \bar{A}_x, \quad \text{d.h.} \quad g = g(x)(1 + \tilde{g}) .$$

Da $\tilde{g}(x)$ verschwindet und $e_x(w) = 0$ für alle $w \neq x$, verschwindet $\tilde{g}e_x$ 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 $\left(\tilde{g}e_x\right)^N = \tilde{g}^N e_x$ die Null von $\bar{A}$ ist. Dann ist

$$(1 + \tilde{g})e_x \cdot \left(1 - \tilde{g} + \tilde{g}^2 - \cdots + (-1)^{N-1}\tilde{g}^{N-1}\right)e_x = (1 - \tilde{g}^N)e_x = e_x ;$$

im Ring $\bar{A}_x$ hat also $1+\tilde{g}$ ein Inverses und damit auch $ge_x = g(x)(1+\tilde{g})e_x$. Wir bilden daher den Bruch $f/g \in \bar{A}_x$ ab auf

$$f \cdot \frac{1}{g(x)} \cdot \left(1 - \tilde{g} + \tilde{g}^2 - \cdots + (-1)^{N-1}\tilde{g}^{N-1}\right)e_x \in \bar{A}_x ,$$

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 $\bar{A}e_x$ ist; das ist klar, da die Summe der e_x gleich eins ist und $e_x e_y = 0$ für $x \neq y$.

■

Im Falle, daß alle Nullstellen einfach sind, läßt sich dieser Satz einfacher formulieren: Die Elemente von $V_K(I)$ seien die Punkte

$$x^{(j)} = (x_1(j), \ldots, x_n^{(j)}) \quad \text{für } j = 1, \ldots, r ,$$

und für jedes j betrachten wir das maximale Ideal

$$\mathfrak{m}_j = (X_1 - x_1^{(j)}, \ldots, X_n - x_n^{(j)})$$

von $\bar{R} = K[X_1, \ldots, X_n]$. Für jedes j ist $\mathfrak{m}_j$ der Kern der Abbildung

$$\begin{cases} \bar{R} \to K \\ f \mapsto f(x^{(j)}) \end{cases} .$$

Da $\mathfrak{m}_j$ als maximales Ideal insbesondere prim ist, ist $S_j = \bar{R} \setminus \mathfrak{m}_j$ eine multiplikativ abgeschlossene Menge und wir können die entsprechend definierte Abbildung für $S_j^{-1}\bar{R}$ betrachten; ihr Kern ist das von $\mathfrak{m}_j$ in $S_j^{-1}\bar{R}$ erzeugte Ideal $\mathfrak{m}_j S_j^{(-1)}\bar{R}$, und nach dem Homomorphiesatz ist

$$\bar{R}/\mathfrak{m}_j = S_j^{-1}\bar{R}/\mathfrak{m}_j S_j^{-1}\bar{R} \cong K .$$

Da die Nullstelle $x^{(j)}$ einfach ist, muß auch $S_j^{-1}\bar{R}/IS_j^{-1}\bar{R}$ ein eindimensionaler Vektorraum, also isomorph zu K sein, und da $IS_j^{-1}\bar{R}$ in $\mathfrak{m}S_j^{-1}\bar{R}$ enthalten ist, müssen die beiden Ideale übereinstimmen. Die haben somit nach dem vorigen Satz einen Isomorphismus

$$\bar{A} = \bar{R}/\bar{I} \cong \bigoplus_{j=1}^{r} \bar{R}/\mathfrak{m}_j \cong K^r ,$$

der durch die Abbildung $f \mapsto \big(f(x^{(1)}), \ldots, f(x^{(r)})\big)$ gegeben ist. Um das Ideal $\bar{I}$ mit den $\mathfrak{m}_j$ in Verbindung zu bringen, brauchen wir die ringtheoretische Version des chinesischen Restesatzes:

Satz: $I_1, \ldots, I_r$ seien Ideale eines Rings R derart, daß $I_j + \bigcap_{\ell \neq j} I_\ell = R$ ist für alle j. Dann ist $R/\bigcap_{j=1}^{r} I_j \cong \bigoplus_{j=1}^{r} R/I_j$.

Beweis: Wir betrachten die Abbildung

$$\begin{cases} R \to \bigoplus_{j=1}^{r} R/I_j \\ f \mapsto (f \bmod I_1, \ldots, f \bmod I_r) \end{cases} .$$

Ihr Kern besteht aus allen Elementen von f, die modulo jedem I_j verschwinden, die also in allen I_j liegen. Somit ist der Kern der Durchschnitt der I_j und der Satz folgt aus dem Homomorphiesatz, wenn wir zeigen können, daß die Abbildung surjektiv ist.

Dies beweisen wir durch vollständige Induktion nach r: Für $r = 1$ ist das klar; sei also $r > 1$ und $(f_1, \ldots, f_r) \in R^r$. Nach Induktionsvoraussetzung gibt es ein Element $f^* \in R$, so daß $f^* \bmod I_j = f_j \bmod I_j$ für $1 \leq j < r$, und dieses Element f^* ist eindeutig modulo $I^* = \bigcap_{j=1}^{r-1} I_j$. Nach Voraussetzung ist $I^* + I_r = R$; es gibt also Elemente $g \in I^*$ und $h \in I_r$ mit $g + h = 1$. Modulo I^* ist $g = 0$ und $h = 1$, modulo I_r ist $g = 1$ und $h = 0$. Somit ist $f = hf^* + gf_r$ modulo I^* gleich f^* und modulo I_r gleich f_r, d.h. $f \bmod I_j = f_j \bmod I_j$ für alle j, so daß f ein Urbild von $(f_1, \ldots, f_r)$ ist.

∎

Da die Ideale $\mathfrak{m}_j$ allesamt maximal sind, erfüllen sie die Voraussetzung dieses Satzes, d.h.

$$\bar{R} \Big/ \bigcap_{j=1}^{r} \mathfrak{m}_j \cong \bigoplus_{j=1}^{r} \bar{R}/\mathfrak{m}_j \,.$$

Andererseits ist die rechte Seite auch isomorph zu $\bar{A} = \bar{R}/\bar{I}$, und natürlich liegt $\bar{I}$ im Durchschnitt der $\mathfrak{m}_j$, denn dieser Durchschnitt besteht gerade aus den sämtlichen Polynomen, die in den Punkten $x^{(1)}, \ldots, x^{(r)}$ verschwinden. Da die Faktorringe übereinstimmen, muß somit

$$\bar{I} = \bigcap_{j=1}^{r} \mathfrak{m}_j$$

sein. Damit ist $\bar{I}$ ein Radikalideal, denn die $\mathfrak{m}_j$ sind als maximale Ideale insbesondere Primideale. Liegt für ein $f \in \bar{R}$ eine Potenz $f^m \in \bar{I}$, so liegt sie in jedem $\mathfrak{m}_j$, und da ein Produkt nur dann in einem Primideal liegen kann, wenn mindestens einer der Faktoren dort liegt, folgt $f \in \mathfrak{m}_j$ für alle j, also $f \in \bar{I}$.

§5: Die explizite Bestimmung der Lösungsmenge

Wir gehen weiterhin aus von einem Gleichungssystem

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

mit Polynomen $f_1, \ldots, f_m$ in n Variablen $X_1, \ldots, X_n$ über einem Körper k; dazu betrachten wir einen algebraisch abgeschlossenen Erweiterungskörper K (mit überabzählbar vielen Elementen). Wir betrachten das Ideal $I = (f_1, \ldots, f_m)$ in $k[X_1, \ldots, X_n]$ sowie das von den gleichen Polynomen in $K[X_1, \ldots, X_n]$ erzeugte Ideal $\bar{I}$. Wir wollen annehmen, daß $V_K(I)$ endlich ist; im vorigen Paragraphen haben wir gesehen, daß wir das leicht nachprüfen können, sobald wir eine GRÖBNER-Basis von I bezüglich *irgendeiner* Monomordnung haben. Dieser GRÖBNER-Basis können wir allerdings nicht unbedingt ansehen, wie die Lösungen aussehen.

Wir wissen aber aus §3, daß diese Basis wegen der Endlichkeit von $V_K(I)$ für jede der Variablen X_i ein Polynom enthalten muß, dessen führender Term eine Potenz von X_i ist. Falls wir die GRÖBNER-Basis bezüglich der lexikographischen Ordnung bezüglich irgendeiner Anordnung der Variablen berechnet haben, kann dieses Basiselement kein Monom enthalten in dem eine Variable vorkommt, die in der gewählten Anordnung vor X_i steht. Für ein Polynom, dessen führender Term eine Potenz der letzten Variablen ist, kann also keine andere Variable vorkommen, und wir haben ein Polynom in nur einer Veränderlichen.

Bei einem Basiselement, dessen führender Term eine Potenz der vorletzten Variable ist, kann entsprechend außer dieser Variablen nur noch die letzte vorkommen, und so weiter. Wenn wir mit der Anordnung $X_1 > X_2 > \cdots > X_n$ arbeiten, haben wir daher zu jedem i ein Polynom in der Basis, das nur die Variablen $X_i, \ldots, X_n$ enthält. Für $i = n$ haben wir also ein Polynom nur in X_n, für $i = n - 1$ eines in X_{n-1} und X_n, und so weiter. Die GRÖBNER-Basis enthält somit ein Gleichungssystem in Treppengestalt, und durch sukzessives Lösen von Polynomgleichungen in einer Veränderlichen können wir, ähnlich wie beim GAUSS-Algorithmus, schrittweise die gesamte Lösungsmenge berechnen. Bei Polynomgleichungen höheren Grades kann freilich deren Lösung problematisch sein.

Am einfachsten wäre die Situation, wenn alle Polynome, die mehr als eine Variable enthalten, von der Form $X_i - g_i$ mit einem Polynom $g_i \in k[X_j]$ für eine in der Anordnung nach X_i kommende Variable X_j wären. In diesem Fall müßten wir nur noch eine Polynomgleichung

höheren Grades lösen, die für die letzte Variable, und ansonsten könnten wir uns mit der Auswertung der Polynome g_i an bekannten Werten begnügen.

Definition: Eine GRÖBNER-Basis hat die Form des *Shape*-Lemmas, wenn sie genau ein Polynom in nur einer Variablen X enthält und jedes andere Basiselement von der Form $Y - g_Y(X)$ ist mit einer Variablen $Y \neq X$ und einem Polynom g_Y in X.

Wir wollen uns überlegen, wann wir so eine GRÖBNER-Basis bekommen können.

Wir gehen der Einfachheit halber wieder aus von der Standardanordnung $X_1 > X_2 > \cdots > X_n$. Wenn wir eine GRÖBNER-Basis der gewünschten Art haben, ist offensichtlich jeder Lösungspunkt $(x_1, \ldots, x_n) \in V_K(I)$ durch seine X_n-Koordinate eindeutig bestimmt; wenn das nicht der Fall ist, haben wir keine Chance. Die lineare Funktion X_n muß daher separierend sein im Sinne der Definition aus §3. Wie wir dort gesehen haben, gibt es stets eine separierende Linearform gibt, und der Beweis zeigt auch, daß dies (einen unendlichen Grundkörper k vorausgesetzt) sogar für *fast alle* Linearformen gilt: Die Koeffizienten von denen, für die es nicht gilt, müssen Polynomgleichungen erfüllen und liegen daher in einer Menge niedrigerer Dimension. Daher haben wir durchaus Chancen, daß vielleicht eine der Variablen separierend ist – es sei denn, natürlich, das Koordinatensystem wäre speziell an die Lösungsmenge angepaßt. In diesem Fall würde uns aber eine zufällig gewählte lineare Koordinatentransformation mit großer Wahrscheinlichkeit mindestens eine separierende Koordinate liefern.

Nehmen wir also ein, die Variable X_n sei separierend. Da die Menge $V_K(I)$ endlich ist, gibt es auf jeden Fall ein Polynom $g \in K[X_n]$, das genau in den X_n-Koordinaten der Punkte aus $V_K(I)$ verschwindet. Fassen wir g auf als Polynom aus $K[X_1, \ldots, X_n]$, verschwindet g auf ganz $V_K(I) = V_K(\bar{I})$; nach der starken Form des HILBERTschen Nullstellensatzes muß eine Potenz von g in $\bar{I}$ liegen, und natürlich ist auch diese Potenz ein Polynom nur in X_n und hat genau die X_n-Koordinaten der Punkte aus $V_K(I)$ als Nullstellen.

Da wir die Variable X_n als separierend angenommen haben, sind die restlichen Komponenten der Lösungspunkte durch die X_n-Komponente eindeutig bestimmt. Da es nur endlich viele Lösungspunkte gibt, können wir daher für jedes $i < n-1$ ein Interpolationspolynom $g_i \in K[X_n]$ finden, so daß für jedes $(x_1, \ldots, x_n) \in V_K(I)$ gilt: $x_i = g_i(x_n)$. Damit verschwindet auch das Polynom $X_i - g_i$ auf ganz $\underline{V}_K(I)$. Leider folgt daraus wieder nur, daß eine Potenz von $X_i - g_i$ in $\overline{I}$ liegt, und die kann deutlich unangenehmer aussehen.

Wir müssen daher zusätzlich annehmen, daß das Ideal $I = (f_1, \ldots, f_m)$ sein eigenes Radikal ist, also ein sogenanntes Radikalideal. Damit haben wir alle notwendigen Voraussetzungen zusammen und können zeigen, daß wir eine GRÖBNER-Basis der gewünschten Form konstruieren können.

Satz: Ist I ein Radikalideal mit endlicher Lösungsmenge, und ist X_n separierend für $V_K(I)$, so gibt es Polynome $g_1, \ldots, g_n \in k[X_n]$ derart, daß

$$\{X_1 - g_1, \ldots, X_{n-1} - g_{n-1}, g_n\}$$

die reduzierte GRÖBNER-Basis von I ist bezüglich jeder lexikographischen Ordnung, die X_n an die letzte Stelle setzt.

Zum *Beweis* müssen wir uns nur überlegen, daß es eine reduzierte GRÖBNER-Basis dieser Gestalt gibt; da reduzierte GRÖBNER-Basen durch die Monomordnung eindeutig bestimmt sind, folgt dann die Behauptung. Wir überlegen uns zunächst, daß $\overline{I}$ eine GRÖBNER-Basis dieser Form hat mit $g_i \in K[X_n]$. Das Argument ist im wesentlichen das gleiche wie oben: $V_K(I)$ ist eine endliche Menge; sie enthalte die r Elemente $(x_1^{(\nu)}, \ldots, x_n^{(\nu)})$ für $j = 1, \ldots, r$. Da X_n separierend ist, sind die $x_n^{(\nu)}$ paarweise verschieden; das Polynom $g_n = (X_n - x_n^{(1)}) \cdots (X_n - x_n^{(r)})$ hat also lauter verschiedene Nullstellen und verschwindet, wenn wir es als Polynom in $X_1, \ldots, X_n$ betrachten, auf $V_K(I)$. Da $\overline{I}$ ein Radikalideal ist, liegt g_n somit in $\overline{I}$.

Für $i < n$ ist die X_i-Koordinate eines Punktes aus $V_K(I)$ durch die X_n-Koordinate eindeutig bestimmt; wir können daher, beispielsweise nach LAGRANGE oder NEWTON ein Interpolationspolynom $g_i \in K[X_n]$

für die r Punkte vom Grad höchstens $r-1$ für die r-Punkte $(x_n^{(\nu)}, x_i^{(\nu)})$ finden. Wir wollen uns überlegen, daß die Polynome

$$X_1 - g_1, X_2 - g_2, \ldots, X_{n-1} - g_{n-1}, g_n$$

eine GRÖBNER-Basis von $\bar{I}$ bilden. Dazu wenden wir das Kriterium von BUCHBERGER an: Für $i, j < n$ und $X_i > X_j$ bezüglich der gewählten Monomordnung ist

$$S(X_i - g_i, X_j - g_j) = X_j(X_i - g_i) - X_i(X_j - g_j) = X_i g_j - X_j g_i \,.$$

Das führende Monom X_i FM(g_j) davon ist durch X_i teilbar, aber durch kein X_ℓ mit $X_\ell > X_i$; bei der Anwendung des Divisionsalgorithmus subtrahieren wir also im ersten Schritt das Polynom FT(g_j)($X_i - g_j$). Falls die Differenz noch ein Monom der Form $X_i X_n^a$ enthält, ist sie von der Form $X_i\widetilde{g} - X_j g_i + \widetilde{h}$ mit $\widetilde{g}, \widetilde{h} \in K[X_n]$ und $\deg \widetilde{g} < \deg g_j$. Dann können wir mit diesem Rest genauso verfahren, und durch eventuell weitere Wiederholungen können wir alle Monome der Form $X_i X_n^a$ mit $a \geq 0$ eliminieren und kommen wir schließlich zu einem Ausdruck der Form $-X_j g_i + h^*$ mit $h^* \in K[X_n]$. Jetzt ist der führende Term durch X_j teilbar, wir addieren also im ersten Schritt $X_j FT(g_i)$ und eliminieren dann nacheinander alle Monome der Form $X_j X_n^a$. Was übrig bleibt, ist ein Polynom $h^\# \in K[X_n]$. Da $S(X_i - g_i, X_j - g_j)$ in $\bar{I}$ liegt und wir beim Divisionsalgorithmus bislang nur Vielfache der Polynome $X_i - g_i$ und $X_j - g_j$ subtrahiert haben, die ebenfalls in $\bar{I}$ liegen, ist auch $h^\# \in \bar{I}$. Somit verschwindet $h^\# \in K[X_n]$ in den X_n-Komponenten aller Punkte aus $V_K(I)$, ist also durch alle Linearfaktoren $X_n - x_n^{(\nu)}$ teilbar und, da diese paarweise verschieden sind, auch durch deren Produkt g_n. Falls $h^\#$ nicht verschwindet, hat es also mindestens den Grad r und damit ein führendes Monom X_n^s mit $s \geq r$; der Divisionsalgorithmus reduziert dieses mit g_n schließlich auf Null.

Bleiben noch die S-Polynome

$$S(X_i - g_i, g_n) = X_n^r(X_i - g_i) - X_i g_n \,.$$

Sie sind von der Form $X_i g + h$ mit $g, h \in K[X_n]$ und können daher mit dem Divisionsalgorithmus genau wie oben auf Null reduziert werden.

Dies zeigt, daß wir in der Tat eine GRÖBNER-Basis haben. Die führenden Monome sind $X_1, \ldots, X_{n-1}, X_n^r$, von denen keines ein anderes teilt. Da alle führenden Koeffizienten eins sind, ist diese GRÖBER-Basis minimal. Sie ist sogar reduziert, denn da die g_i für $i < n$ höchstens Grad $r-1$ haben, ist keines ihrer Monome durch FM(g_n) = X^r teilbar. Damit haben wir also eine GRÖBNER-Basis der gewünschten Form für I gefunden; nach allem was wir bisher wissen, können die Koeffizienten der Polynome daraus beliebige Elemente von K sein.

Nun berechnen wir nach dem BUCHBERGER-Algorithmus, ausgehend von den Polynomen $f_1, \ldots, f_m \in k[X_1, \ldots, X_n]$, eine GRÖBNER-Basis von I bezüglich der gleichen Monomordnung und machen daraus eine reduzierte GRÖBNER-Basis. Bei keinem der dabei durchgeführten Rechenschritte verlassen wir den Polynomring $k[X_1, \ldots, X_n]$, so daß auch alle Elemente der so berechneten reduzierten GRÖBNER-Basis dort liegen müssen. Wegen der Eindeutigkeit der reduzierten GRÖBNER-Basis bei gegebener Monomordnung muß das Ergebnis gleich dem obigen sein, die Polynome $X_i - g_i$ für $i < n$ und g_n liegen also in $k[X_1, \ldots, X_n]$, d.h. alle g_i liegen in $k[X_n]$. Da die Bestimmung der reduzierten GRÖBNER-Basis für das von $f_1, \ldots, f_m$ erzeugte Ideal nach dem BUCHBERGER-Algorithmus in $k[X_1, \ldots, X_n]$ genauso verläuft wie in $K[X_1, \ldots, X_n]$, ist die angegebene GRÖBNER-Basis auch eine von I, womit der Satz vollständig bewiesen ist.

■

§6: Berechnung von Vielfachheiten und Radikalen

Der Satz am Ende des vorigen Paragraphen läßt sich leider nur anwenden, wenn die gegebenen Gleichungen ein Radikalideal erzeugen und zusätzlich eine der Variablen separierend ist. Letzteres läßt sich durch eine generische Koordinatentransformation immer erreichen, ersteres aber wird oft nicht der Fall sein. In diesem Paragraphen wollen wir uns überlegen, wie man ein gegebenes Gleichungssystem so ergänzen kann, daß die Gleichungen ein Radikalideal erzeugen. In diesem wie auch im nächsten Paragraphen arbeiten wir vor allem mit Methoden der linearen Algebra

Wir betrachten weiterhin nur Gleichungssysteme mit endlicher Lösungsmenge; dann ist der Restklassenring $A = k[X_1, \ldots, X_n]/I$ ein endlichdimensionaler k-Vektorraum, und die Standardmonome bezüglich irgendeiner GRÖBNER-Basis von I bilden eine Basis dieses Vektorraums. Es ist wichtig, daß wir hier mit einer GRÖBNER zu *irgendeiner* Monomordnung arbeiten können und nicht wie im vorigen Paragraphen eine lexikographische Ordnung benötigen, denn wie bereits erwähnt ist der BUCHBERGER-Algorithmus für die lexikographische Ordnung im Allgemeinen besonders langsam. K sei weiterhin ein (überabzählbarer) algebraisch abgeschlossener Körper, der k enthält.

Für ein beliebiges Element $f \in A$ betrachten wir die k-lineare Abbildung

$$L_f: \begin{cases} A \to A \\ g \mapsto fg \end{cases}$$

Für diese gilt der

Satz von Stickelberger: L_f induziert für jedes $x \in V_K(I)$ eine lineare Abbildung $\bar{A}_x \to \bar{A}_x$. Diese hat $f(x)$ als ihren einzigen Eigenwert; dessen Vielfachheit ist also die Vielfachheit $\mu(x)$ der Nullstelle x.

Beweis: Wie wir aus §4 wissen, enthält $\bar{A}$ für jedes $x \in V_K(I)$ ein idempotentes Element e_x derart, daß $\bar{A}_x \cong Ae_x$. Für ein jedes Element $ge_x \in \bar{A}e_x$ ist daher $L_f(ge_x) = f(ge_x) = (fg)e_x = L_f(g)e_x$, so daß das Bild wieder in $\bar{A}e_x$ liegt.

$\big(f - f(x)\big)e_x$ verschwindet auf ganz $V_K(I)$, da e_x für alle $y \neq x$ aus der Lösungsmenge verschwindet. Nach der starken Form des HILBERTschen Nullstellensatzes liegt daher eine Potenz, etwa die m-te, eines Repräsentanten dieses Elements im Ideal $\bar{I}$. Im Restklassenring ist daher

$$\Big(\big(f - f(x)\big)e_x\Big)^m = \big(f - f(x)\big)^m e_x^m = \big(f - f(x)\big)^m e_x = 0\,,$$

d.h. $L^m_{f-f(x)}$ ist die Nullabbildung auf $\bar{A}e_x \cong \bar{A}_x$. ∎

Aus diesem Satz können wir sofort Aussagen über die Spur, die Determinante und das charakteristische Polynom von L_f ablesen:

Korollar: Für $f \in A$ gilt
a) Sp $L_f = \sum_{x \in V_K(I)} \mu(x) f(x)$
b) det $L_f = \prod_{x \in V_K(I)} f(X)^{\mu(x)}$
c) Das charakteristische Polynom von L_f ist

$$\det(L_f - T \cdot \mathrm{id}) = \prod_{x \in V_K(I)} \big(T - f(x)\big)^{\mu(x)}$$

■

Vor allem die Spuren der linearen Abbildungen L_f werden für uns im folgenden wichtig sein. Man kann sie einfach berechnen, sobald man *irgendeine* GRÖBNER-Basis G des Ideal I kennt: Die Standardmonome bezüglich G bilden eine Vektorraumbasis $\omega_1, \ldots, \omega_r$ des Restklassenrings A über k. Zur Berechnung der Spur von L_f müssen wir die Elemente $L_f(\omega_i)$ in dieser Basis darstellen. Dazu müssen wir ein Polynom $F \in k[X_1, \ldots, X_n]$ wählen mit Restklasse f; meist wird f ohnehin in dieser Weise gegeben sein. Wenn wir einfach die Monome ω_i mit F multiplizieren, erhalten wir im Allgemeinen kein Polynom, das sich als Linearkombination der Monome ω_j schreiben läßt, denn wir multiplizieren ja im Polynomring, nicht im Restklassenring. Daher müssen wir die Produkte $F\omega_i$ mit dem Divisionsalgorithmus modulo der GRÖBNER-Basis G reduzieren; der Divisionsrest ist eine Linearkombination der Standardmonome. In dieser müssen wir den Koeffizienten von ω_i bestimmen, und diese Koeffizienten müssen wir über alle i aufsummieren.

Bei älteren Computeralgebrasystemen kann es hier zu Problemen kommen: Für den BUCHBERGER-Algorithmus ist es egal, ob wir den exakten Divisionsrest des S-Polynoms verwenden oder ein skalares Vielfaches davon. Da der exakte Rest im Fall $k = \mathbb{Q}$ oft große Nenner hat, wird die weitere Berechnung effizienter, wenn man ihn mit einer ganzen Zahl multipliziert derart, daß das Ergebnis nur noch ganzzahlige Koeffizienten hat. Auf diese Weise arbeitet beispielsweise das Kommando `poly_normal_form` für den Divisionsalgorithmus in Maxima (und wahrscheinlich auch in vielen anderen Versionen von macsyma). Für Spurberechnungen ist ein solcher modifizierter Divisionsrest natürlich unbrauchbar; wir brauchen den exakten Rest.

Zu dessen Bestimmung, auch mit einem solchen System, gibt es mehrere Möglichkeiten. Am einfachsten ist die Situation, wenn man bereits ein Element von $V_K(I)$ kennt, für das das zu reduzierende Polynom P nicht verschwindet. Da die Differenz von P und dem (korrekten) Divisionsrest im Ideal I liegt, nehmen beide in diesem Punkt denselben Wert an. Wertet man daher sowohl P als auch das vom Computeralgebrasystem berechnete Vielfache des Rests an diesem Punkt aus, findet man den Multiplikator, mit dem auf den korrekten Rest kommt.

Leider kennt man nur selten ein solches Element von $V_K(I)$. Der folgende Ansatz funktioniert allgemein: Angenommen, wie kennen für ein Polynom P ein Vielfaches Q des korrekten Divisionsrests bezüglich der GRÖBNER-Basis G. Im Falle $Q = 0$ ist auch der korrekte Divisionsrest gleich Null. Andernfalls gibt es genau eine Zahl $\lambda \in k$, für die $P - \lambda Q$ in I liegt: Gäbe es nämlich ein $\mu \neq \lambda$ mit $P - \mu Q \in I$, so läge auch die Differenz $(\mu - \lambda)Q$ und damit auch Q in I, d.h. auch P müßte in I liegen, so daß wir bei der Division durch die Elemente einer GRÖBNER-Basis von I Rest Null erhalten müßten.

Wir betrachten nun λ als eine neue Variable und rechnen im Polynomring über $k(\lambda)$. Auch hier können wir den BUCHBERGER-Algorithmus durchführen und erhalten als Rest ein Polynom mit Koeffizienten aus $k(\lambda)$. Wir wir uns bereits überlegt haben, gibt es genau ein $\lambda \in k$, für das alle diese Koeffizienten verschwinden, und für dieses ist λQ der korrekte Divisionsrest.

Wir betrachten nun für jedes $h \in A$ die Bilinearform

$$\underset{h}{\text{SpB}}: \begin{cases} A \times A \to K \\ (f, g) \mapsto \text{Sp}\, L_{fgh} \end{cases}$$

und die dazugehörige quadratische Form, die sogenannte HERMITE-Form

$$Q_h: \begin{cases} A \to K \\ f \mapsto \text{Sp}\, L_{f^2 h} \end{cases}$$

Im Falle $h = 1$ werden wir den Index h in beiden Fällen meist weglassen.

Für die Beweise der folgenden Sätze benötigen wir die VANDERMONDEsche Determinante. Für Leser, die sie nicht aus der Linearen Algebra oder Numerik kennen, sei sie hier kurz definiert und berechnet.

Der Franzose ALEXANDRE THÉOPHILE VANDERMONDE (1735–1796) war zunächst Musiker; erst im Alter von 35 Jahren begann er sich für Mathematik zu interessieren und publizierte in den Jahren 1771 und 1772 vier Arbeiten über Gleichungen, Determinanten und über das Problem, einen Springer so über ein Schachbrett zu bewegen, daß er jedes Feld genau einmal betritt. Die VANDERMONDEsche Determinante ist nirgends in seinem publizierten Werk zu finden; sie wurde erst um 1935 von HENRI LEBESGUE (1875–1941) nach ihm benannt.

Definition: Für $a_1, \ldots, a_n \in K$ ist die VANDERMONDEsche Determinante

$$V(a_1, \ldots, a_n) = \begin{vmatrix} 1 & a_1 & a_1^2 & \ldots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \ldots & a_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_n & a_n^2 & \ldots & a_n^{n-1} \end{vmatrix}.$$

Zur Berechnung dieser Determinante nach dem LAPLACEschen Entwicklungssatz subtrahieren wir zunächst die erste Zeile von jeder der folgenden; da sich der Wert der Determinanten dadurch nicht ändert, ist

$$V(a_1, \ldots, a_n) = \begin{vmatrix} 1 & a_1 & a_1^2 & \ldots & a_1^{n-1} \\ 0 & a_2 - a_1 & a_2^2 - a_1^2 & \ldots & a_2^{n-1} - a_1^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_n - a_1 & a_n^2 - a_1^2 & \ldots & a_n^{n-1} - a_1^{n-1} \end{vmatrix}.$$

Wenn wir hier nach der ersten Spalte entwickeln, muß nur eine einzige $(n-1) \times (n-1)$-Determinante berücksichtigt werden, alle anderen haben den Vorfaktor Null. Also ist

$$V(a_1, \ldots, a_n) = \begin{vmatrix} a_2 - a_1 & a_2^2 - a_1^2 & \ldots & a_2^{n-1} - a_1^{n-1} \\ a_3 - a_1 & a_3^2 - a_1^2 & \ldots & a_3^{n-1} - a_1^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_n - a_1 & a_n^2 - a_1^2 & \ldots & a_n^{n-1} - a_1^{n-1} \end{vmatrix}$$

gleich jeder Determinanten, die durch Streichung der ersten Spalte und der ersten Zeile entsteht.

Hier können wir in jeder Zeile die jeweils vorne stehende Differenz ausklammern, denn genau wie

$$x^k - 1 = (x-1)(x^{k-1} + x^{k-2} + \cdots + x + 1)$$

durch $(x-1)$ teilbar ist, ist auch

$$a_i^k - a_1^k = (a_i - a_1)(a_i^{k-1} + a_i^{k-2}a_1 + a_i^{k-3}a_1^2 + \cdots + a_i a_1^{k-2} + a_1^{k-1})$$

durch $(a_i - a_1)$ teilbar; den Quotienten schreiben wir kurz als $q_{i,k-1}$:

$$q_{i,k-1} \underset{\text{def}}{=} a_i^{k-1} + a_i^{k-2}a_1 + a_i^{k-3}a_1^2 + \cdots + a_i a_1^{k-2} + a_1^{k-1}\,.$$

Wegen der Linearität der Determinante können wir jeden Faktor, den wir aus einer Zeile (oder Spalte) ausklammern, vor die Determinante ziehen und erhalten für $V(a_1, \ldots, a_n)$ somit den Wert

$$(a_2 - a_1)(a_3 - a_1)\cdots(a_n - a_1) \begin{vmatrix} 1 & q_{21} & q_{22} & \cdots & q_{2,n-2} \\ 1 & q_{31} & q_{32} & \cdots & q_{3,n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & q_{n1} & q_{n2} & \cdots & q_{n,n-2} \end{vmatrix}.$$

Die Nützlichkeit dieser Formel steht und fällt damit, daß wir die q_{ij} gut miteinander in Verbindung bringen können. Für verschiedene Indizes i haben die entsprechenden Ausdrücke offensichtlich wenig miteinander zu tun; sie enthalten nicht einmal dieselben Variablen. Schreiben wir allerdings

$$\begin{aligned} q_{ij} &= a_i^j + a_i^{j-1}a_1 + a_i^{j-2}a_1^2 + \cdots a_i a_1^{j-1} + a_1^j \\ &= a_i^j + a_1(a_i^{j-1} + a_i^{j-2}a_1 + \cdots a_i a_1^{j-2} + a_1^{j-1})\,, \end{aligned}$$

so sehen wir, daß

$$q_{ij} = a_i^j + a_1 q_{i,j-1} \quad \text{oder} \quad q_{ij} - a_1 q_{i,j-1} = a_i^j$$

ist. Subtrahieren wir also zuerst a_1 mal die vorletzte Spalte von der letzten, so werden die Einträge der letzten Spalte zu a_i^{n-2}. Entsprechend subtrahieren wir a_1 mal die $(n-2)$-te Zeile von der $(n-1)$-ten und erhalten lauter Einträge a_i^{n-3} und so weiter, bis schließlich die Subtraktion des a_1-fachen der ersten Spalte von der zweiten die Einträge der letzteren zu

$$q_{i1} - a_1 = (a_i + a_1) - a_1 = a_i$$

macht. Somit ist $V(a_1, \ldots . a_n)$ gleich

$$(a_2 - a_1)(a_3 - a_1) \cdots (a_n - a_1) \begin{vmatrix} 1 & a_2 & a_2^2 & \ldots & a_2^{n-2} \\ 1 & a_3 & a_3^2 & \ldots & a_3^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_n & a_n^2 & \ldots & a_n^{n-2} \end{vmatrix} .$$

Die Determinante rechts ist offensichtlich wieder eine VANDERMONDEsche Determinante, allerdings mit um eins verminderter Zeilen- und Spaltenzahl und mit einer Variablen weniger.

Damit haben wir die Rekursionsformel

$$V(a_1, \ldots, a_n) = (a_2 - a_1)(a_3 - a_1) \cdots (a_n - a_1)\, V(a_2, \ldots, a_n) ,$$

die es erlaubt, die Berechnung von $V(a_1, \ldots, a_n)$ auf eine einzige VANDERMONDEsche Determinante der Größe $(n-1) \times (n-1)$ zurückzuführen.

Zur vollständigen Berechnung von $V(a_1, \ldots, a_n)$ fehlt uns jetzt nur noch ein Induktionsanfang; direktes Nachrechnen zeigt sofort, daß

$$V(a_n) = \det(1) \quad \text{und} \quad V(a_{n-1}, a_n) = \begin{vmatrix} 1 & a_{n-1} \\ 1 & a_n \end{vmatrix} = a_n - a_{n-1}$$

ist, also folgt induktiv

$$V(a_1, \ldots, a_n) = \prod_{j<i} (a_i - a_j) .$$

Satz: Ein Polynom $f \in k[X_1, \ldots, X_n]$ liegt genau dann im Radikal von I, wenn f mod I in

$$\operatorname{Kern} Q \underset{\text{def}}{=} \left\{ f \in A \;\middle|\; \operatorname{SpB}(f, g) = 0 \quad \forall g \in A \right\}$$

liegt.

Beweis: Für $f \in \sqrt{I}$ verschwindet f für alle $x \in V_K(I)$; nach obigem Korollar ist daher für alle $g \in A$

$$\operatorname{SpB}(f, g) = \operatorname{Sp} L_{fg} = \sum_{x \in V_K(I)} \mu(x) f(x) g(x) = 0 .$$

Umgekehrt sei SpB$(f,g) = 0$ für alle $g \in A$, d.h.

$$\sum_{x \in V_K(I)} \mu(x) f(x) g(x) = 0 \quad \forall g \in A\,.$$

Sei $V_K(I) = \{x^{(1)}, \ldots, x^{(s)}\}$. Wie wir aus §3 wissen, gibt es eine separierende Linearform $u \in A$, d.h. eine Funktion, die für jedes $x^{(j)}$ einen anderen Wert annimmt. Wir betrachten dazu das Produkt der Matrix

$$M = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ u(x^{(1)}) & u(x^{(2)}) & \ldots & u(x^{(s)}) \\ u^2(x^{(1)}) & u^2(x^{(2)}) & \ldots & u^2(x^{(s)}) \\ \vdots & \vdots & \ddots & \vdots \\ u^{s-1}(x^{(1)}) & u^{s-q}(x^{(2)}) & \ldots & u^{s-1}(x^{(s)}) \end{pmatrix}$$

mit dem Vektor

$$v = \begin{pmatrix} \mu(x^{(1)}) f(x^{(1)}) \\ \mu(x^{(2)}) f(x^{(2)}) \\ \mu(x^{(s)}) f(x^{(s)}) \end{pmatrix}\,.$$

Der ℓ-te Eintrag von Mv ist

$$\sum_{j=1}^{s} u(x^{(j)})^{\ell-1} \cdot \mu(x^{(j)}) f(x^{(j)}) = \text{SpB}(u^{\ell-1}, f) = 0\,,$$

da $f \in \operatorname{Kern} Q$.

Die Determinante von M (oder genauer seiner Transponierten) ist eine VANDERMONDEsche Determinante. Da u separierend ist, sind alle $u(x^{(j)})$ verschieden, so daß det M nach obiger Formel nicht verschwindet. Das homogene lineare Gleichungssystem $Mv = 0$ hat daher nur die triviale Lösung $v = 0$. Damit verschwinden alle $\mu(x)f(x)$, also alle $f(x)$, d.h. f verschwindet auf $V_K(I)$ und muß daher im Radikal von I liegen. ∎

Satz: Für alle $h \in A$ ist $\operatorname{Rang} Q_h = \#\{x \in V_K(I) \mid h(x) \neq 0\}$. Insbesondere ist $\operatorname{Rang} Q$ gleich der Anzahl der Lösungen in $V_K(I)$.

Beweis: Wie im vorigen Beweis sei $u \in A$ eine separierende Linearform, und $V_K(I) = \{x^{(1)}, \ldots, x^{(s)}\}$. Dann sind $1, u, \ldots u^{s-1}$ linear unabhängig, denn ist

$$\lambda_0 + \lambda_1 u + \cdots + \lambda_s u^{s-1} = 0\,,$$

so ist insbesondere

$$\lambda_0 + \lambda_1 u(x^{(j)}) + \cdots + \lambda_s u^{s-1}(x^{(j)}) = 0$$

für $j = 1, \ldots, s$. Dies ist ein homogenes lineares Gleichungssystem für $\lambda_0, \ldots, \lambda_{s-1}$, dessen Matrix eine VANDERMONDEsche Determinante hat, die nicht verschwindet, da u separierend ist. Somit gibt es nur die triviale Lösung $\lambda_0 = \cdots = \lambda_{s-1} = 0$.

Falls $s < r = \dim_k A$ ist, können wir daher das System der Vektoren $\omega_1 = 1$, $\omega_2 = u, \ldots, \omega_s = u^{s-1}$ ergänzen zu einer Basis $\omega_1, \ldots, \omega_r$ von A.

Für jedes Element $g = g_1\omega_1 + \cdots + g_r\omega_r \in A$ (mit $g_\ell \in k$ ist nach obigem Korollar

$$Q_h(g) = \operatorname{Sp} L_{g^2 h} = \sum_{j=1}^{s} \mu(x^{(j)}) \left(\sum_{\ell=1}^{r} g_\ell \omega_\ell(x^{(j)}) \right)^2 \cdot h(x^{(j)}) \,.$$

$v \in K^s$ sei der Vektor mit Komponenten $v_j = \sum_{\ell=1}^{r} g_\ell \omega_\ell(x^{(j)})$, und Δ sei die Diagonalmatrix mit Einträgen $\mu(x^{(j)})h(x^{(j)})$. Dann ist

$$Q_h(g) = v^T \Delta v \,.$$

Mit

$$\Gamma = \begin{pmatrix} 1 & \omega_1(x^{(1)}) & \ldots & \omega_s(x^{(1)}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_1(x^{(s)}) & \ldots & \omega_s(x^{(s)}) \end{pmatrix} \in K^{s \times r} \quad \text{ist} \quad v = \Gamma \begin{pmatrix} g_1 \\ \vdots \\ g_r \end{pmatrix}$$

und damit

$$Q_h(q) = (g1, \ldots, g_r)\Gamma^T \Delta \Gamma \begin{pmatrix} g_1 \\ \vdots \\ g_r \end{pmatrix} \,.$$

Γ hat Rang s, da die ersten s Spalten eine VANDERMONDEsche Determinante haben, und der Rang von Δ ist gleich der Anzahl der $x \in V_K(I)$ mit $h(x) \neq 0$, also eventuell kleiner als s. Somit ist der Rang von $Q_h = \Gamma^T \Delta \Gamma$ gleich letzterer Anzahl, wie behauptet.

■

Ist $f = f_1\omega_1 + \cdots + f_r\omega_r$ mit $j_i \in k$ ein weiteres Element von A, so ist

$$fgh = \sum_{i=1}^{r}\sum_{j=1}^{r} hf_ig_j\omega_i\omega_j \,;$$

da für zwei Matrizen $A, B \in k^{r\times r}$ und zwei Skalare λ, μ in k gilt $\mathrm{Sp}(\lambda A + \mu B) = \lambda \,\mathrm{Sp}\, A + \mu \,\mathrm{Sp}\, B$, ist also

$$\mathrm{Sp}\, L_{fgh} = h\sum_{i=1}^{r}\sum_{j=1}^{r} f_ig_j \,\mathrm{Sp}(\omega_i\omega_j)\,.$$

Mit der Matrix

$$\mathrm{SpM} \underset{\text{def}}{=} \begin{pmatrix} \mathrm{Sp}(\omega_1\omega_1) & \ldots & \mathrm{Sp}(\omega_1\omega_r) \\ \vdots & \ddots & \vdots \\ \mathrm{Sp}(\omega_r\omega_1) & \ldots & \mathrm{Sp}(\omega_1\omega_r) \end{pmatrix}$$

ist also

$$\mathrm{Sp}(L_{fgh}) = (f_1, \ldots, f_r)\,\mathrm{SpM} \begin{pmatrix} g_1 \\ \vdots \\ g_r \end{pmatrix}\,.$$

Ein Polynom $f \in k[X_1, \ldots, X_n]$ liegt nach dem vorletzten Satz genau dann in $\sqrt{I}$, wenn seine Restklasse $f + I$ im Kern von Q liegt, wenn also für $f + I = f_1\omega_1 + \cdots + f_r\omega_r$ gilt

$$(f_1, \ldots, f_r)\,\mathrm{SpM}\, g = 0 \quad \forall g \in k^r\,.$$

Dies ist genau dann der Fall, wenn

$$(f_1, \ldots, f_r)\,\mathrm{SpM} = (0, \ldots, 0)$$

ist. Transposition auf beiden Seiten macht daraus das homogene lineare Gleichungssystem

$$\mathrm{SpM}(f_1, \ldots, f_r)^T = 0\,,$$

das wir explizit aufstellen und lösen können. Die Vektoren $p^{(1)}, \ldots, p^{(t)}$ aus k^r seien eine Basis des Lösungsraums. Dann sind die Polynome $P^{(j)} = \sum_{i=1}^{r} p_i^{(j)}\omega_i$ nach dem vorletzten Satz Elemente von $\sqrt{I}$, und $\sqrt{I}$ wird erzeugt von diesen Polynomen und den Erzeugenden von I.

§7: Univariate rationale Darstellungen

GRÖBNER-Basen in der Form des *Shape*-Lemmas sind ideal für die Lösung eines nichtlinearen Gleichungssystems, hängen aber von nicht unmittelbar verifizierbaren Voraussetzungen ab und erfordern die Berechnung einer GRÖBNER-Basis bezüglich einer lexikographischen Ordnung, was oft sehr aufwendig ist. Für die praktische Berechnung der Lösungen ist eine leichte Modifikation praktisch genauso gut:

Wir gehen wie üblich aus von einem Ideal $I \lhd k[X_1, \ldots, X_n]$ und suchen $V_K(I)$.

Definition: Eine Lösung durch univariate rationale Darstellungen besteht aus einem Polynom $\chi \in k[T]$ und n rationalen Funktionen $\varphi_i \in k(T)$, so daß gilt: Sind $t_1, \ldots, t_s$ die Nullstellen von χ in K, so ist $V_K(I) = \left\{ \big(\varphi_1(t_i), \ldots, \varphi_n(t)\big) \;\middle|\; i = 1, \ldots, s \right\}$, und die Vielfachheit einer jeden Lösung ist gleich der Vielfachheit der entsprechenden Nullstelle von χ.

Die neue Variable T erhalten wir dabei über eine separierende Linearform für $V_K(I)$.

Konkret sei zunächst $u \in A$ beliebig und $\chi_u \in k[T]$ sei das charakteristische Polynom von L_u. Nach dem Korollar zum Satz von STICKELBERGER ist

$$\chi_u = \prod_{x \in V_K(I)} \big(T - u(x)\big)^{\mu(x)} .$$

Definition: Die i-te NEWTON-Summe von χ_u ist

$$s_i \underset{\text{def}}{=} \sum_{x \in V_K(I)} \mu(x) u(x)^i .$$

Nach dem Satz von STICKELBERGER ist $s_i = \operatorname{Sp} u^i$, läßt sich also bei Kenntnis von u und einer Basis von A explizit berechnen. Wir wollen uns überlegen, wie wir daraus auch χ_u berechnen können. χ_u ist ein

Polynom vom Grad $r = \dim_k A$ mit höchstem Koeffizienten eins; wir schreiben es im Gegensatz zu unserer sonstigen Konvention als

$$\chi_u = \sum_{i=0}^{r} b_i T^{r-i} \quad \text{mit} \quad b_i \in k$$

und wollen die b_i bestimmen.

Die LEIBNIZ-Regel zur Berechnung der Ableitung eines Produkts ist $(uv)' = u'v + uv'$; für $u, v \neq 0$ ist also

$$\frac{(uv)'}{uv} = \frac{u'}{u} + \frac{v'}{v},$$

und entsprechend auch für Produkte von mehr als zwei Funktionen. Daher ist

$$\frac{\chi_u'}{\chi_u} = \sum_{x \in V_K(I)} \frac{\frac{d}{dT}\big(T - u(x)\big)^{\mu(x)}}{\big(T - u(x)\big)^{\mu(x)}} = \sum_{x \in V_K(I)} \frac{\mu(x)}{T - u(x)}.$$

Die Summanden können wir nach der Summenformel für die geometrische Reihe auch schreiben als

$$\frac{\mu(x)\,T^{-1}}{1 - \frac{u(x)}{T}} = \frac{\mu(x)}{T} \sum_{j \geq 0} \left(\frac{u(x)}{T}\right)^j;$$

daher ist

$$\frac{\chi_u'}{\chi_u} = \sum_{x \in V_K(I)} \frac{\mu(x)}{T} \sum_{j \geq 0} \frac{u(x)^j}{T^j} = \sum_{j \geq 0} \sum_{x \in V_K(I)} \mu(x) \frac{u(x)^j}{T^{j+1}}$$

$$= \sum_{j \geq 0} \frac{\sum\limits_{x \in V_K(I)} u(x)^j}{T^{j+1}} = \sum_{j \geq 0} \frac{\operatorname{Sp} u^j}{T^{j+1}}.$$

Durch Multiplikation mit $\chi_u = \sum b_i T^{r-i}$ erhalten wir

$$\chi_u' = \sum_{j \geq 0} \frac{\operatorname{Sp} u^j}{T^{j+1}} \cdot \sum_{i=0}^{r} b_i T^{r-i} = \sum_{j \geq 0} \sum_{i=0}^{r} b_i T^{r+j-i-1} \operatorname{Sp}(u^j).$$

Andererseits ist

$$\chi_u' = \sum_{\ell=0}^{r-1} (r - \ell) b_\ell T^{r-\ell-1}.$$

Durch Koeffizientenvergleich folgt

$$(r-\ell)b_\ell = \sum_{\nu=0}^{\ell} \operatorname{Sp} u^\nu b_{\ell-\nu}\,.$$

Für $\nu = 0$ tritt auch rechts der Koeffizient b_ℓ auf; sein Koeffizient ist $\operatorname{Sp} u^0$, also die Spur der Identität. Deren Abbildungsmatrix auf dem r-dimensionalen Vektorraum A ist die $r\times r$-Einheitsmatrix mit Spur r, also steht rechts rb_ℓ. Subtrahieren wir dies auf beiden Seiten und dividieren wir durch $-\ell$, erhalten wir

$$b_\ell = -\frac{1}{\ell}\sum_{\nu=1}^{\ell} \operatorname{Sp} u^\nu b_{\ell-\nu}\,,$$

und damit lassen sich ausgehend von $b_0 = 1$ die folgenden b_ℓ rekursiv berechnen. Damit können wir χ_u berechnen, ohne $V_K(I)$ zu kennen, und wenn wir die Nullstellen dieses Polynoms in einer Veränderlichen bestimmen können, kennen wir die Werte, die u auf $V_K(I)$ annimmt.

Falls u separierend ist, ist jedes $x \in V_K(I)$ eindeutig durch seinen Wert $u(x)$ bestimmt; um eine univariate rationale Darstellung der Lösungen zu bekommen, müssen wir uns dann überlegen, wie wir x aus $u(x)$ durch rationale Funktionen berechnen können.

Dazu betrachten wir für ein beliebiges Element $v \in A$ das Polynom

$$g_u(v,T) = \sum_{x\in V_K(I)} \mu(x)v(x) \prod_{\alpha\in u\left(V_K(I)\smallsetminus\{x\}\right)} (T-\alpha) \in K[T]\,.$$

Wenn u separierend ist und wir diese Polynome explizit kennen, können wir damit die Werte $v(x)$ berechnen, denn dann sind für alle $x \in V_K(I)$ die $u(x)$ verschieden, und für jedes $y \in V_K(I)$ ist

$$\frac{g_u\big(v,u(y)\big)}{g_u\big(1,u(y)\big)} = \frac{\displaystyle\sum_{x\in V_K(I)} \mu(x)v(x) \prod_{\alpha\in u\left(V_K(I)\smallsetminus\{x\}\right)} (u(x)-\alpha)}{\displaystyle\sum_{x\in V_K(I)} \mu(x) \prod_{\alpha\in u\left(V_K(I)\smallsetminus\{x\}\right)} (u(x)-\alpha)}\,.$$

Für $x \neq y$ nimmt α für einen der Faktoren auch den Wert $u(y)$ an, so daß das Produkt verschwindet; daher haben wir sowohl im Zähler als auch im Nenner tatsächlich nur den Summanden mit $x = y$ stehen, d.h.

$$\frac{g_u\big(v, u(y)\big)}{g_u\big(1, u(y)\big)} = \frac{\mu(y)v(y) \prod\limits_{\alpha \in u\left(V_K(I) \smallsetminus \{y\}\right)} (u(y) - \alpha)}{\mu(y) \prod\limits_{\alpha \in u\left(V_K(I) \smallsetminus \{y\}\right)} (u(y) - \alpha)} = v(y)\,.$$

Wenn wir für v die Koordinatenfunktionen einsetzen, können wir so aus $u(x)$ die Koordinaten von x berechnen.

Dazu müssen wir aber zunächst in der Lage sein, das Polynom $g_u(v, T)$ ohne Kenntnis von $V_K(I)$ zu berechnen. Dazu beginnen wir mit dem quadratfreien Teil von χ_u. Wir beschränken uns im folgenden der Einfachheit halber auf den Fall, daß der Körper k die Charakteristik Null hat.

Definition: Für ein Polynom $f \in k[T]$ bezeichnen wir

$$f^{\#} = \frac{f}{\mathrm{ggT}(f, f')}$$

als den *quadratfreien Anteil* von f.

Ist $t \in K$ eine e-fache Nullstelle von f, so können wir f schreiben als $f = (T - t)^e g$ für ein Polynom $g \in K[T]$ mit $g(t) \neq 0$. Dann hat $f' = e(T-t)^{e-1}g + (T-t)^e g'$ für $e \geq 2$ eine $(e-1)$-fache Nullstelle bei t und für $e = 1$ keine; dasselbe gilt auch für den größten gemeinsamen Teiler von f und f'. Dieser hat dann als Nullstellen genau die mehrfachen Nullstellen von f; der Quotient $f^{\#}$ hat also dieselben Nullstellen wie f, aber jeweils nur mit Vielfachheit eins. Speziell für χ_u ist also

$$\chi_u^{\#} = \prod_{\alpha \in u\left(V_K(I)\right)} (T - \alpha)^s a_i T^{s-1}\,.$$

(Um einen konstanten Faktor brauchen wir uns nicht zu kümmern, da der ggT zweier Polynome über K ohnehin nur bis auf einen konstanten Faktor definiert ist. Wir nehmen einfach irgendeinen ggT und normalisieren dann den Quotienten auf höchsten Koeffizienten eins.)

Um die Koeffizienten von $g_u(v,T)$ zu berechnen, betrachten wir den Quotienten

$$\frac{g_u(v,T)}{\chi_u^{\#}} = \frac{\displaystyle\sum_{x\in V_K(I)} \mu(x)v(x) \prod_{\alpha\in u\left(V_K(I)\smallsetminus\{x\}\right)} (T-\alpha)}{\displaystyle\prod_{\alpha\in u\left(V_K(I)\right)} (T-\alpha)}$$

$$= \sum_{x\in V_K(I)} \frac{\mu(x)v(x)}{T-u(x)} = \sum_{j\geq 0} \frac{\displaystyle\sum_{x\in V_K(I)} \mu(x)v(x)u(x)^j}{T^{j+1}} = \sum_{j\geq 0} \frac{\mathrm{Sp}(vu^j)}{T^{j+1}}\,,$$

wobei wir bei dieser Berechnung wieder den gleichen Weg über die Summenformel der geometrischen Reihe gegangen sind wie oben bei der Berechnung von χ_u/χ_u'. Multiplikation mit $\chi_u^{\#}$ führt auf die gewünschte Darstellung

$$g_u(v,T) = \sum_{\ell=0}^{s-1} \sum_{j=0}^{s-\ell-1} \mathrm{Sp}(vu^j) a_\ell T^{s-\ell-j-1}\,.$$

Insbesondere zeigt dies, daß $g_u(v,T) \in K[T]$ tatsächlich bereits in $k[T]$ liegt.

Zur Berechnung des Polynoms für konkrete Werte von T erinnern wir uns an das HORNER-Schema zur Berechnung eines Polynoms $f = \sum_{i=0}^{n} c_i T^{n-1}$: Definieren wir rekursiv

$$H_0(f) = c_0 \quad \text{und} \quad H_{j+1}(f) = H_j(f)T + c_j\,,$$

so ist $H_j(f) = \sum_{i=0}^{j} c_j T^{j-i}$ und somit ist

$$g_u(v,T) = \sum_{\ell=0}^{s-1} \mathrm{Sp}(vu^\ell) H_{s-\ell-1}(\chi^{\#})\,.$$

Für einen Algorithmus zur Berechnung einer univariaten rationalen Darstellung fehlt uns nun nur noch ein Kandidat für die separierende Linearform u. Wie wir aus §3 wissen, gibt es nur endlich viele $a \in k$, für die

$$u_a = X_1 + aX_2 + a^2X_3 + \cdots + a^{n-1}X_n$$

nicht separierend ist; genauer gesagt sind es höchstens $(n-1)\binom{s}{2}$ Stück. Da wir uns in diesem Paragraphen auf Charakteristik Null beschränken, ist $\mathbb{Z} \subset k$, und für mindestens eine ganze Zahlen a mit $0 \leq a \leq (n-1)\binom{s}{2}$ ist u_a separierend. Wir können daher wie folgt vorgehen:

1. Schritt: Berechne eine GRÖBNER-Basis G von I bezüglich irgendeiner Monomordnung, z.B. der graduiert lexikographischen. $\omega_1, \ldots, \omega_r$ sei die Menge der Standardmonome bezüglich G; sie bilden eine Basis des Restklassenrings A.

2. Schritt: Wähle irgendein a mit $0 \leq a \leq (n-1)\binom{s}{2}$, setze $u = u_a$ und berechne χ_u.

3. Schritt: Berechne das quadratfreie Polynom $\chi_u^{\#}$ zu χ_u. Falls es mit χ_u übereinstimmt, ist χ_u quadratfrei und damit u separierend. Andernfalls berechne man den Rang s der Matrix SpM aus dem letzten Paragraphen. Wegen $s = \#V_K(I)$ ist u auch separierend, wenn $\deg \chi_u = s$ ist. Ist $\deg \chi_u < s$, so ist u nicht separierend; dann muß man zurück zu Schritt 2 und dort ein neues, bislang noch nicht betrachtetes a auswählen.

4. Schritt: Berechne für die Koordinatenfunktionen $X_1, \ldots, X_n$ die Polynome $g_u(X_i, T)$ und $g_u(1, T)$.

5. Schritt: Bestimme die Nullstellen $t_1, \ldots, t_s$ von $\chi_u^{\#}$.

6. Schritt: Die Elemente von $V_K(I)$ sind die n-Tupel

$$\left(\frac{g(X_1, t_j)}{g(1, t_j)}, \ldots, \frac{g(X_n, t_j)}{g(1, t_j)} \right)$$

für $j = 1, \ldots, s$.

Der problematischste Schritt bei diesem Algorithmus wird im Allgemeinen der fünfte sein, denn allgemeine algebraische Formeln zur Bestimmung der Nullstellen eines Polynoms vom Grad d gibt es nur für $d \leq 4$. Die Computeralgebra kennt zwar Algorithmen zur Zerlegung eines Polynoms in seine irreduziblen Faktoren, aber die meisten Polynome aus $\mathbb{Q}[T]$ sind irreduzibel. Sätze aus der reellen Algebra erlauben es, die reellen Nullstellen eines Polynoms einzugrenzen, d.h. für ein Polynom mit s reellen Nullstellen lassen sich (beliebig kurze) Intervalle

$[a_i, b_i]$ für $i = 1, \ldots, s$ bestimmen deren jedes genau eine der Nullstellen enthält. Mit etwas mehr Aufwand lassen sich auch Intervalle für die Real- und Imaginärteile der komplexen Nullstellen finden. Mit den so bestimmten Nullstellen läßt sich auch rechnen, so daß im sechsten Schritt auch die Koordinaten der Lösungstupel in dieser Form angeben kann.

Das Nennerpolynom $g_u(1, T)$ verhindert, daß die obige Darstellung zu einer Basis des Ideals I führt. Da wir zur Bestimmung von $V_K(I)$ nur Nullstellen von χ_u einsetzen, können wir die Division eventuell verhindern: Wenn wir ein Polynom $g^* \in k[T]$ finden, so daß

$$\frac{1}{g_u(1,t)} = g^*(t) \qquad \text{für alle } t \in K \text{ mit } \chi_u(t) = 0\,,$$

können wir die Division durch $g_u(1, t)$ ersetzen durch eine Multiplikation mit $g^*(t)$, wobei wir das Ergebnis natürlich modulo χ_u reduzieren können.

Um zu sehen, wann dies der Fall ist, schauen wir uns

$$g_u(1,T) = \sum_{x \in V_K(I)} \mu(x) \prod_{\alpha \in u\left(V_K(I) \smallsetminus \{x\}\right)} (T - \alpha)$$

etwas genauer an: Die Struktur der rechten Seite erinnert an die Ableitung von

$$\chi_u = \prod_{x \in V_K(I)} \big(T - u(x)\big)^{\mu}(x)\,,$$

denn nach der LEIBNIZ-Regel ist

$$\chi_u' = \sum_{x \in V_K(I)} \mu(x)\big(T - u(x)\big)^{\mu(x)-1} \prod_{x \in V_K(I) \smallsetminus \{x\}} \big(T - u(x)\big)^{\mu(x)}\,.$$

Dividieren wir dies durch $\prod_{x \in V_K(I)} \big(T - u(x)\big)^{\mu(x)-1}$, erhalten wir

$$\begin{aligned} &\sum_{x \in V_K(I)} \mu(x) \prod_{x \in V_K(I) \smallsetminus \{x\}} \big(T - u(x)\big) \\ &= \sum_{x \in V_K(I)} \mu(x) \prod_{\alpha \in u\left(V_K(I) \smallsetminus \{x\}\right)} (T - \alpha)\,. \end{aligned}$$

Wie wir uns bereits bei der Berechnung von $\chi_u^{\#}$ überlegt haben, ist in Charakteristik Null eine e-fache Nullstelle eines Polynoms f eine $(e-1)$-fache Nullstelle der Ableitung; der ggT von f und f' ist also gerade das Produkt der $(T-\alpha)^{e_\alpha - 1}$, wobei α die Nullstellen von f durchläuft und e_α die Nullstellenordnung von α bezeichnet. Somit ist

$$g_u(1,T) = \frac{\chi_u'}{\mathrm{ggT}(\chi_u, \chi_u')} .$$

Falls χ_u keine mehrfachen Nullstellen hat, ist $\mathrm{ggT}(\chi_u, \chi_u') = 1$, also $g_u(1,T) = \chi_u'$, und nach dem erweiterten EUKLIDischen Algorithmus lassen sich Polynome $g^*, h^* \in k[T]$ bestimmen derart, daß

$$g^*\chi_u' + h^*\chi_u = g^* g_u(1,T) + h^*\chi_u = 1 \quad \text{und} \quad g^* g_u(1,T) \equiv 1 \bmod \chi_u$$

ist.

§8: Nullstellen und Eigenwerte

1992 stellten die Wiener Numeriker AUZINGER und STETTER auf einer Tagung in Singapur ein Verfahren vor, wie man Lösungen nichtlinearer Gleichungssysteme zurückführen kann auf die Bestimmung von Eigenwerten geeigneter Matrizen. Sie arbeiteten dabei natürlich mit numerischen Methoden, aber das Verfahren läßt sich auch für exaktes symbolisches Rechnen anwenden.

Die Matrizen, um die es geht, sind Abbildungsmatrizen für lineare Abbildungen des Restklassenrings A auf sich selbst; auch hier müssen wir also wieder annehmen, daß das gegebene Gleichungssystem eine endliche Lösungsmenge hat, so daß A als Vektorraum endlichdimensional ist.

Wir benötigen zunächst eine Vektorraumbasis von A; dazu nehmen wir wieder die Standardmonome bezüglich irgendeiner GRÖBNER-Basis des Ideals. Dann betrachten wir für jede Variable X_i die lineare Abbildung

$$L_{X_i}: \begin{cases} A \to A \\ f \mapsto X_i f \end{cases} ,$$

wobei mit $X_i f$ natürlich die Restklasse gemeint ist, deren Darstellung in der Basis aus Standardmonomen wir mit dem Divisionsalgorithmus bestimmen können. $C^{(i)}$ sei die Abbildungsmatrix von L_{X_i} bezüglich dieser Basis.

Lemma: Die Matrizen $C^{(i)}$ kommutieren, d.h. für alle i, j ist

$$C^{(i)}C^{(j)} = C^{(j)}C^{i)} .$$

Beweis: Die zugehörigen linearen Abbildungen L_{X_i} und L_{X_j} kommutieren, denn für alle $f \in A$ ist

$$L_{X_i}\big(L_{X_j}(f)\big) = X_i X_j f = X_j X_i f = L_{X_j}\big(L_{X_i}(f)\big) .$$

■

Diese Kommutativität erlaubt uns, eine Basis von A zu finden bezüglich derer alle diese Matrizen obere Dreiecksmatrizen werden, denn es gilt:

Satz: Ist $M \subset k^{n\times n}$ eine Menge miteinander kommutierender Matrizen, so gibt es eine Basis von K^n, bezüglich derer alle diese Matrizen obere Dreiecksgestalt haben. Falls die Matrizen aus einer Teilmenge $N \subseteq M$ diagonalisierbar sind, läßt sich diese Basis so bestimmen, daß alle Matrizen aus N Diagonalgestalt haben.

Die Formulierung *Eine Matrix C hat bezüglich einer Basis von K^n obere Dreiecksgestalt* soll dabei natürlich bedeuten, daß die Abbildungsmatrix der linearen Abbildung

$$\begin{cases} K^n \to K^n \\ \quad v \mapsto Cv \end{cases}$$

bezüglich dieser Basis eine obere Dreiecksmatrix ist; entsprechend ist auch die Diagonalgestalt zu interpretieren. Offensichtlich hat eine Matrix genau dann obere Dreiecksgestalt bezüglich einer Basis $b_1, \ldots, b_n$ von K^n, wenn für jedes $r \leq n$ der von $b_1, \ldots, b_r$ erzeugte Untervektorraum auf sich selbst abgebildet wird, wenn also Cb_r für jedes r eine Linearkombination der Vektoren $b_1, \ldots, b_r$ ist.

Den *Beweis* des Satzes führen wir durch vollständige Induktion nach n. Der Induktionsanfang $n = 1$ ist trivial, da jede 1×1-Matrix eine Diagonalmatrix ist.

Nun sei $n > 1$. Falls alle Matrizen aus M skalare Vielfache der Einheitsmatrix sind, gibt es nichts mehr zu beweisen; andernfalls gibt es mindestens eine Matrix $C \in M$, die kein Vielfaches einer Diagonalmatrix ist. Falls es in N eine solche Matrix gibt, nehmen wir diese. Da der Körper K algebraisch abgeschlossen ist, hat C mindestens einen Eigenwert $\lambda \in K$, und die Dimension des Eigenraums

$$U = \left\{ v \in K^n \;\middle|\; Cv = \lambda v \right\}$$

ist zwar echt größer als Null, aber kleiner als n. Für jede weitere Matrix $B \in M$ und jeden Vektor $v \in U$ ist

$$C(Bv) = (CB)v = (BC)v = B(Cv) = B(\lambda v) = \lambda Bv\,,$$

d.h. auch Bv liegt in U. Somit operieren die Matrizen $B \in M$ auch auf U über die linearen Abbildungen

$$\begin{cases} U \to U \\ v \mapsto Bv \end{cases}.$$

Die Abbildungsmatrix dieser linearen Abbildung bezüglich irgendeiner festen Basis von U sei $\overline{B}$. Natürlich kommutieren auch die Matrizen $\overline{B}$, und da $\dim U < n$ gibt es nach Induktionsannahme eine Basis von U, bezüglich derer alle $\overline{B}$ Diagonalgestalt haben.

Falls $C \in N$, ist K^n die direkte Summe aller Eigenräume von C, und das gleiche Argument zeigt, daß auch alle anderen Eigenräume von allen $B \in M$ auf sich selbst abgebildet werden. Nach Induktionsannahme hat jeder dieser Eigenräume eine Basis, bezüglich derer alle Matrizen aus M obere Dreiecksgestalt haben und die aus N sogar Diagonalgestalt. Setzen wir diese Basen zusammen, erhalten wir eine Basis von K^n mit derselben Eigenschaft.

Falls wir C nicht aus M wählen können, wissen wir immerhin, daß U eine entsprechende Basis hat. Daraus folgt insbesondere, daß der erste Basisvektor b_1 dieser Basis von allen Matrizen $B \in M$ auf ein skalares Vielfaches $\lambda_B b_1$ abgebildet wird.

Betrachten wir nun den Faktorraum $W = K^n / Kb_1$. Auch dort operiert M, denn sind $v, w \in K^n$ zwei Vektoren, die sich nur um ein

Vielfaches von b_1 unterscheiden, etwa $w = v + \lambda b_1$, so ist für jede Matrix $B \in M$

$$Bw = Bv + B\lambda w = Bv + \lambda Bw = Bv + \lambda\lambda_B b_1 ;$$

die Bilder unterscheiden sich also auch nur um ein Vielfaches von b_1 und haben somit die gleiche Restklasse modulo Kb_1.

Nach Induktionsannahme gibt es eine Basis von W derart, daß alle diese linearen Abbildungen bezüglich dieser Basis eine obere Dreiecksmatrix als Abbildungsmatrix haben. $b_2, \ldots, b_r \in K^n$ seien Repräsentanten dieser Basisvektoren in K^n. Dann wird der von $b_1, \ldots, b_r$ von jeder Matrix $B \in M$ auf sich selbst abgebildet, d.h. alle Matrizen aus M haben obere Dreiecksgestalt bezüglich dieser Basis, Da wir C nicht aus N wählen konnten, sind alle Matrizen aus N, so es überhaupt welche gibt, skalare Vielfache der Einheitsmatrix und haben damit ohnehin bezüglich jeder Basis Diagonalgestalt.

■

Für uns bedeutet dies, daß es eine Basis von $\bar{A}$ gibt, bezüglich derer alle Abbildungen L_{X_i} obere Dreiecksgestalt haben. $b_1, \ldots, b_r$ seinen Polynome aus $K[X_1, \ldots, X_n]$, deren Restklassen modulo I eine solche Basis bilden.

Nehmen wir zunächst der Einfachheit halber an, das Ideal I sei ein Radikalideal, und die Abbildungsmatrizen aller L_{X_i} seien diagonalisierbar. Dann sind alle b_i Eigenvektoren aller L_{X_i}; zu jedem i gibt es also ein λ_i, so daß $X_i b_i \equiv \lambda_i b_i \bmod I$, d.h. $(X_i - \lambda_i)b_i \in I$. Der Eigenvektor b_i kann nicht der Nullvektor sein, liegt also nicht in I. Da I ein Radikalideal ist, liegt jedes Polynom, das auf $V_K(I)$ verschwindet, in I; daher gibt es mindestens ein $x \in V_K(I)$ mit $b_i(x) \neq 0$. Für dieses x muß dann $X_i - \lambda_i$ verschwinden, d.h. die i-te Koordinate von x muß gleich dem Eigenwert λ_i sein.

Dies gilt auch im allgemeinen Fall, denn der Restklassenring $\bar{A}$ ist isomorph zur direktem Summe seiner Lokalisierungen zu den Lösungen aus $V_K(I)$, und nach dem Satz von STICKELBERGER hat die Mutiplikation mit einem Polynom f auf $\bar{A}_x$ genau einen Eigenwert, nämlich $f(x)$, Seine algebraische Vielfachheit ist gleich der Multiplizität $\mu(x)$, also gleich der Dimension von A_x.

§9: Resultanten

Fast alle Verfahren, die wir bislang betrachtet haben, stammen aus der Zeit nach 1965; nichtlineare Gleichungssysteme interessierten Mathematiker aber natürlich bereits lange vorher. und sie entwickelten auch Methoden zu ihrer Lösung. Eine auch noch heute oft wichtige Technik sind die im 19. Jahrhundert entwickelten Resultanten.

Wir beginnen mit zwei Polynomen

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

und

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

in einer Veränderlichen, lassen aber für die Koeffizienten nicht nur Elemente aus einem Körper zu, sondern aus einem beliebigen faktoriellen Ring:

Definition: *a)* Ein Element $f \in R$ eines Rings R heißt *Einheit,* falls es ein $f' \in R$ gibt, so daß $f f' = 1$ ist.

b) $f \neq 0$ heißt *irreduzibel,* wenn f keine Einheit ist und bei jeder Produktdarstellung $f = gh$ von f entweder g oder h eine Einheit ist.

c) Ein *faktorieller Ring* ist ein Integritätsbereich, in dem sich jedes Element $f \neq 0$ bis auf Reihenfolge und Einheiten eindeutig als Produkt irreduzibler Elemente schreiben läßt.

In diesem Sinne sind beispielsweise die ganzen Zahlen ein faktorieller Ring; die einzigen Einheiten sind ± 1, und die irreduziblen Elemente sind die Primzahlen und ihre negativen. *Bis auf Reihenfolge und Einheiten eindeutig* heißt, daß wir uns nicht daran stören, daß

$$10 = 2 \cdot 5 = 5 \cdot 2 = (-2) \cdot (-5) = (-5) \cdot (-2)$$

ist.

Auch jeder Körper ist ein faktorieller Ring; hier ist jedes von Null verschiedene Element eine Einheit, und es gibt gar keine irreduziblen Elemente.

Nach einem Satz von GAUSS ist auch der Polynomring über einem faktoriellen Ring wieder faktoriell; der Beweis ist in praktisch jedem Lehrbuch der Algebra zu finden und wird auch meist in der Vorlesung *Algebra* gegeben. Daraus folgt induktiv, daß jeder Polynomring in endlich vielen Variablen $X_1, \ldots, X_n$ über $\mathbb{Z}$ oder einem Körper faktoriell ist. Man überlegt sich leicht, daß die Einheiten von $R[X]$ genau die von R sind, In einem Polynomring über einem Körper ist ein Polynom genau dann irreduzibel, wenn es sich nicht als Produkt zweier Polynome positiven Grades schreiben läßt; in $\mathbb{Z}[X]$ muß noch zusätzlich der ggT der Koeffizienten gleich eins sein, da man sonst diesen als Faktor nehmen könnte. Außerdem sind alle Primzahlen und ihre negativen irreduzible Elemente von $\mathbb{Z}[X]$.

Die Resultante zweier Polynome f, g wie oben soll uns ein Kriterium dafür geben, daß f und g einen gemeinsamen Faktor positiven Grades haben. Ist h ein solcher Faktor, so ist

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

ein gemeinsames Vielfaches von f und g, dessen Grad

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

höchstens gleich $d + e - 1$ ist.

Haben umgekehrt f und g ein gemeinsames Vielfaches vom Grad höchstens $d + e - 1$, so hat auch ihr kleinstes gemeinsames Vielfaches S höchstens den Grad $d + e - 1$. (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 S 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 $d+e-\deg S$ ist mindestens gleich eins. Damit ist gezeigt:

Lemma: Zwei Polynome $f, g \in R[X]$ haben genau dann einen gemeinsamen Teiler positiven Grades, wenn es nichtverschwindende Polynome $u, v \in R[X]$ mit $uf = vg$ und Graden $\deg u \leq \deg g - 1$ und $\deg v \leq \deg f - 1$.

■

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

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

und

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

Die Koeffizienten von X^r in uf und vg sind

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

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

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

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

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

$$\begin{aligned}
a_d u_{e-1} - b_e v_{d-1} &= 0 \\
a_{d-1}u_{e-1} + a_d u_{e-2} - b_{e-1}v_{d-1} - b_e v_{d-2} &= 0 \\
a_{d-2}u_{e-1} + a_{d-1}u_{e-2} + a_d u_{e-3} & \\
-b_{e-2}v_{d^{-1}} - b_{e-1}v_{d-2}b_e v_{d-3} &= 0 \\
\cdots &
\end{aligned}$$

$$a_0u_3 + a_1u_2 + a_2u_1 + a_3u_0 - b_0v_3 - b_1v_2 - b_2v_1 - b_3v_0 = 0$$
$$a_0u_2 + a_1u_1 + a_2u_0 - b_0v_2 - b_1v_1 - b_2v_0 = 0$$
$$a_0u_1 + a_1u_0 - b_0v_1 - b_1v_0 = 0$$
$$a_0u_0 - b_0v_0 = 0$$

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

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

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

Definition: Die obige Matrix heißt SYLVESTER-Matrix: ihre Determinante ist die *Resultante*

$$\mathrm{Res}(f, g) = \mathrm{Res}_X(f, g)$$

der beiden Polynome f und g bezüglich der Variablen X.

Der Index X ist dann notwendig, wenn schon der Ring R ein Polynomring ist, so daß f und g Polynome in mehreren Veränderlichen sind und nicht klar ist, bezüglich welcher von ihnen die Resultante gebildet wird.

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

Damit haben wir gezeigt

Satz: Zwei Polynome $f, g \in R[X]$ über dem faktoriellen Ring R haben genau dann einen gemeinsamen Faktor positiven Grades, wenn ihre Resultante verschwindet. ∎

Angenommen, wir haben zwei Polynome $f, g \in k[X, Y]$ und suchen deren gemeinsame Nullstellenmenge über einem algebraisch abgeschlossenen Körper K, der k enthält. Dann können wir f und g auffassen als Polynome in Y über $k[X]$ und ihre Resultante $\mathrm{Res}_Y(f, g) \in k[X]$ betrachten. Für einen speziellen Wert $x \in K$ können wir auch die Resultante der beiden Polynome $f(x, Y)$ und $g(x, Y)$ aus $K[Y]$ betrachten. Offensichtlich ist diese gleich dem Wert den wir erhalten, wenn wir x in $\mathrm{Res}_Y(f, g) \in k[X]$ einsetzen.

Die Resultante von $f(x, Y)$ und $g(x, Y)$ aus $K[Y]$ verschwindet genau dann, wenn diese beiden Polynome einen gemeinsamen Faktor positiven Grades haben. Da K algebraisch abgeschlossen ist, ist dies äquivalent dazu, daß sie eine gemeinsame Nullstelle in K haben. Somit gibt es zu einem vorgegebenen $x \in K$ genau dann ein $y \in K$, so daß $(x, y) \in V_K(f, g)$, wenn x eine Nullstelle von

$\text{Res}_Y(f,g) \in k[X]$ ist. Eine Strategie zur Lösung des des Gleichungssystems $f(x,y) = g(x,y) = 0$ kann also darin bestehen, zunächst die Nullstellen von $\text{Res}_Y(f,g)$ zu bestimmen und für jede dieser Nullstellen x dann die beiden Polynome $f(x,Y)$ und $g(x,Y)$ zu betrachten. Dann kann man entweder die Nullstellen eines dieser Polynome bestimmen und durch Einsetzen überprüfen, welche davon auch Nullstellen des anderen sind, oder man berechnet den ggT der beiden Polynome. Seine Nullstellen sind die y-Koordinaten der Lösungen mit erster Koordinate x.

Für ein allgemeines Gleichungssystem

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

betrachten wir die $f_i \in k[X_1,\ldots,X_n]$ als Polynome in X_n mit Koeffizienten aus $k[X_1,\ldots,x_{N-1}]$. Falls die Resultante $\text{Res}_{X_n}(f_i,f_j)$ 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

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

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

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

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

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

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

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. Tatsächlich verwendet aber natürlich ohnehin niemand den Entwicklungssatz von LAGRANGE um eine große Determinante zu berechnen; dessen Nützlichkeit beschränkt sich definitiv auf kleineren Spielzeugdeterminanten, wie sie vor allem in Mathematikklausuren vorkommen. In realistischen Anwendungen wird man die Matrix durch Zeilen- und/oder Spaltenoperationen auf Dreiecksform bringen und dann die Determinante einfach als Produkt der Diagonaleinträge berechnen oder man tat dies über eine LR- oder QR-Zerlegung. Das dauert für die SYLVESTER-Matrix zweier Polynome der Grade dreißig und vierzig auf heutigen Computern weniger als eine halbe Minute.

Stellt man allerdings keine Matrix auf, sondern verlangt von einem Computeralgebrasystem einfach, daß es die Resultante der beiden Polynome berechnen soll, hat man das Ergebnis nach weniger als einem Zehntel der Zeit. Einer der Schlüssel dazu ist wieder einmal der EUKLIDische Algorithmus.

Angenommen, wir haben zwei Polynome f, g in einer Variablen X über

einem faktoriellen Ring R:

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

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

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

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

Das ist freilich nur dann möglich, wenn $R[X]$ ein EUKLIDischer Ring ist, also im wesentlichen nur dann, wenn R ein Körper ist. Das ist aber keine so große Einschränkung wie es scheint, denn wir können statt in $R[X]$ im Polynomring über dem Quotientenkörper von R rechnen; da die Resultante ein eindeutig bestimmtes Element von R ist, muß spätestens das Endergebnis unserer Rechnung in R liegen. Die Zwischenergebnisse können freilich recht große Nenner bekommen – ein wohlbekanntes Problem der Computeralgebra, das uns bereits beim EUKLIDischen Algorithmus für Polynome begegnet ist.

Der zentrale Punkt beim EUKLIDischen Algorithmus ist, daß die gemeinsamen Teiler von f und g genau dieselben sind wie die von f und h. Insbesondere haben also f und g genau dann einen gemeinsamen Teiler von positivem Grad, wenn f und h einen haben, d.h. $\mathrm{Res}_X(f, g)$ verschwindet genau dann, wenn $\mathrm{Res}_X(f, h)$ verschwindet. Damit sollte es also einen Zusammenhang zwischen den beiden Resultanten geben, und den können wir zur Berechnung von $\mathrm{Res}_X(f, g)$ ausnützen, denn natürlich ist $\mathrm{Res}_X(f, h)$ kleiner und einfacher als $\mathrm{Res}_X(f, g)$.

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

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

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

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

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

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

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

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

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

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

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

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

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

Dies läßt sich nun nach Art des EUKLIDischen Algorithmus iterieren: Berechnen wir wie dort die Folge der Reste $r_1 = h$ der Division von g durch f und dann (mit $r_0 = g$) weiter r_{i+1} gleich dem Rest bei der Division von r_i durch r_{i-1}, so können wie die Berechnung von $\mathrm{Res}_X(f,g)$ durch Multiplikation mit Potenzen der führenden Koeffizienten der Divisoren zurückführen auf die viel kleineren Resultanten $\mathrm{Res}_X(r_i, r_{i+1})$. Sobald r_{i+1} eine Konstante ist, egal ob Null oder nicht, haben wir eine explizite Formel und der Algorithmus endet. Für den Fall, daß f größeren Grad als g hat brauchen wir noch

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

Beweis: Wir müssen in der SYLVESTER-Matrix e Zeilen zu f mit den d Zeilen zu g vertauschen. Dies kann beispielsweise so realisiert werden, daß wir die unterste f-Zeile nacheinander mit jeder der g-Zeilen

vertauschen, bis sie nach d Vertauschungen schließlich unten steht. Dies müssen wir wiederholen, bis alle f-Zeilen unten stehen, wir haben also insgesamt de Zeilenvertauschungen. Somit ändert sich das Vorzeichen der Determinante um den Faktor $(-1)^{de}$.

∎

Zum Abschluß dieses Paragraphen wollen wir uns noch überlegen, daß die Resultante zweier Polynome noch aus einem anderen Grund für jede gemeinsame Nullstelle verschwinden muß: 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 beide nicht verschwinden. Die Gleichungen

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

und

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

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

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

darstellen in der Form

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

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

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

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

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.

■