Wolfgang K. Seiler

Zahlentheorie

Vorlesung an der Universität Mannheim
im Frühjahrssemester 2011

Dieses Skriptum entstand parallel zur Vorlesung und wurde im wesentlichen nur von mir selbst systematisch korrekturgelesen. Glücklicherweise haben mir auch eine ganze Reihe von Studenten gefundene Fehler gemeldet; trotzdem kann man mit Sicherheit davon ausgehen, daß noch bei weitem nicht alle gefunden sind. Dabei handelt es sich wohl leider nicht immer nur um harmlose Tippfehler, sondern auch um Fehler bei den mathematischen Aussagen.

Das Skriptum ist daher in seiner Qualität auf keinen Fall mit einem Lehrbuch zu vergleichen; es sollte 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.

Inhalt

KAPITEL I: GANZE ZAHLEN UND IHRE PRIMZERLEGUNG 1
§1: Rationale und irrationale Zahlen 1
§2: Der Euklidische Algorithmus 5
§3: Der erweiterte EUKLIDische Algorithmus 8
§4: Der Aufwand des EUKLIDischen Algorithmus 13
§5: Die multiplikative Struktur der ganzen Zahlen 18
§6: Kongruenzenrechnung 20
§7: Der chinesische Restesatz 24
§8: Prime Restklassen 29

KAPITEL II: ANWENDUNGEN IN DER KRYPTOGRAPHIE 35
§1: New directions in cryptography 35
§2: Das RSA-Verfahren 40
§3: Weitere Anwendungen des RSA-Verfahrens 46
a) Identitätsnachweis 46
b) Elektronische Unterschriften 47
c) Blinde Unterschriften und elektronisches Bargeld 49
d) Bankkarten mit Chip 52
§4: Wie groß sollten die Primzahlen sein? 54
§5: Verfahren mit diskreten Logarithmen 56
§6: DSA 58
§7: Anwendungen bei SSL/TLS 60
§8: Ausblick 61

KAPITEL III: PRIMZAHLEN 63
§1: Die Verteilung der Primzahlen 63
§2: Das Sieb des ERATOSTHENES 78
§3: FERMAT-Test und FERMAT-Zahlen 81
§4: Der Test von MILLER und RABIN 90
§5: Der Test von AGRAWAL, KAYAL und SAXENA 92

KAPITEL IV: FAKTORISIERUNGSVERFAHREN 104
§1: Die ersten Schritte 106
a) Test auf Primzahl 106
b) Abdividieren kleiner Primteiler 107
§2: Die Verfahren von POLLARD und ihre Varianten 109
a) Die Monte-Carlo-Methode 109
b) Die $(p-1)$-Methode 115
c) Varianten 116
§3: Das Verfahren von FERMAT und seine Varianten 119

KAPITEL V: KETTENBRÜCHE 130
§1: Der Kettenbruchalgorithmus 130
§2: Geometrische Formulierung 133
§3: Optimale Approximation 137
§4: Kettenbrüche und Kalender 146
§5: Eine kryptographische Anwendung 158

KAPITEL VI: QUADRATISCHE ZAHLKÖRPER 161
§1: Grundbegriffe der Ringtheorie 161
§2: Die Elemente quadratischer Zahlkörper 164
§3: Die Hauptordnung eines Zahlkörpers 166
§4: Normen und Spuren in quadratischen Zahlkörpern 169
§5: EUKLIDische Ringe 171
§6: Einheiten in quadratischen Zahlkörpern 180
§7: Quaternionen 182

KAPITEL VII: QUADRATISCHE FORMEN 185
§1: Summen zweier Quadrate 185
§2: Anwendung auf die Berechnung von π 191
§3: Der Satz von LAGRANGE 197
§4: Quadratische Formen und Matrizen 200
§5: Kettenbruchentwicklung quadratischer Irrationalitäten 205
§6: Die PELLsche Gleichung 210

KAPITEL VIII: QUADRATISCHE RESTE 215
§1: Das LEGENDRE-Symbol 215
§2: Das quadratische Reziprozitätsgesetz 217
§3: Das JACOBI-Symbol 221
§4: Berechnung der modularen Quadratwurzel 225
§5: Anwendungen quadratischer Reste 231
a) Münzwurf per Telephon 232
a) Quadratische Formen und quadratische Reste 232
b) Münzwurf per Telephon 233
c) Akustik von Konzerthallen 235

KAPITEL IX: DIE FERMAT-VERMUTUNG FÜR ZAHLEN
UND FÜR POLYNOME 241
§1: Zahlen und Funktionen 241
§2: Pythagoräische Tripel 243
§3: Der Satz von MASON 246
§4: Die abc-Vermutung 249
§5: Die FREY-Kurve 253

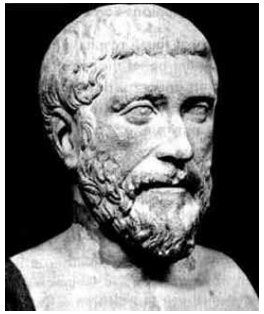
Kapitel 1
Ganze Zahlen und ihre Primzerlegung

§1: Rationale und irrationale Zahlen

Die Zahlentheorie beschäftigt sich, wie schon ihr Name sagt, mit Zahlen. Nun würde allerdings eine Umfrage wohl ergeben, daß sich nach Ansicht eines Großteils der Bevölkerung die gesamte Mathematik mit Zahlen beschäftigt – auch wenn beispielsweise im neunbändigen Analysislehrbuch von JEAN DIEUDONNÉ abgesehen von den dreiteiligen Abschnittsnummern praktisch keine Zahlen außer 0, 1 und 2 vorkommen.

Die Zahlentheorie unterscheidet sich dadurch von anderen Teilen der Mathematik, daß es dort vor allem um *ganze* Zahlen geht, oft sogar einfach um die natürlichen Zahlen $1, 2, 3, \ldots$. Die frühe griechische Philosophie der Pythagoräer beispielsweise stand unter dem Motto *Alles ist Zahl.* Sie konnten die musikalischen Harmonien auf einfache Verhältnisse natürlicher Zahlen zurückführen und waren überzeugt, daß dies auch für alle anderen Proportionen galt.

Umso größer war der Schock, als um 450 v.Chr. einer von ihnen, wahrscheinlich HIPPASSOS VON METAPONT, herausfand, daß es in der Geometrie Längenverhältnisse gibt, die sich *nicht* so beschreiben lassen. Schlimmer noch: Ein Beispiel dafür bietet ausgerechnet das Wahrzeichen der Pythagoräer, das Pentagramm. HIPPASSOS VON METAPONT nahm deshalb auch ein schlimmes Ende: Nach einigen Überlieferungen wurde er von den erzürnten Pythagoräern ertränkt, nach anderen ließen ihn die Götter als Strafe für seine Schandtat bei einem Schiffsuntergang ertrinken.



PYTHAGORAS VON SAMOS lebte etwa von 569 bis 475. Als ungefähr 18-Jähriger besuchte er Thales in Milot und ging auf dessen Rat 535 nach Ägypten, um mehr über Mathematik und Astronomie zu lernen. Im Tempel von Diospolis wurde er nach der dafür vorgesehen Ausbildung in die Priesterschaft aufgenommen. 525, bei der persischen Invasion Ägyptens, geriet er in Gefangenschaft und wurde nach Babylon gebracht, was er nutzte er, um die dortige Mathematik zu erlernen. 520 kam er frei und kehrte zurück nach Samos, zog aber schon bald weiter nach Croton in Süditalien, wo er eine religiöse und philosophische Schule gründete, die Pythagoräer.

Wir wollen uns hier zunächst mit einem einfacheren Fall als dem Pentagramm beschäftigen, beschäftigen, dem Verhältnis zwischen der Diagonale und der Seite eines Quadrats. In einem Quadrat der Seitenlänge a kann die Diagonale aufgefaßt werden als Hypothenuse eines gleichschenkligen rechtwinkligen Dreiecks mit Katheten der Länge a; sie hat daher nach dem Satz des PYTHAGORAS (der tatsächlich wohl einige hundert Jahre älter als PYTHAGORAS sein dürfte) die Länge $a\sqrt{2}$, und das gesuchte Verhältnis ist $\sqrt{2}$.

Wäre $\sqrt{2}$ als Verhältnis a/b zweier natürlicher Zahlen darstellbar, so könnten wir ohne Beschränkung der Allgemeinheit annehmen, daß mindestens eine der beiden Zahlen a und b ungerade ist: Andernfalls müßten wir einfach so lange durch zwei kürzen, bis dies der Fall ist.

Quadrieren wir beide Seiten der Gleichung $a/b = \sqrt{2}$, so erhalten wir die neue Gleichung $a^2/b^2 = 2$; demnach müßte $a^2 = 2b^2$ eine gerade Zahl sein. Damit müßte aber auch a gerade sein, denn das Quadrat einer ungeraden Zahl ist ungerade. Wenn aber a durch zwei teilbar ist, ist $a^2 = 2b^2$ durch vier teilbar, also wäre auch b^2 und damit auch b gerade, ein Widerspruch. Somit ist $\sqrt{2}$ keine rationale Zahl.

Auch das Verhältnis zwischen Umfang und Durchmesser eines Kreises, für das wir heute die Bezeichnung π verwenden, ist nicht rational, allerdings erfordert der Beweis dafür etwas mehr Arbeit. Ich möchte ihn trotzdem hier vorstellen, denn er zeigt sehr schön, wie zahlentheoretische Aussagen teilweise nur auf Umwegen über andere Gebiete der Mathematik bewiesen werden können. Beim folgenden Beweis führt

dieser Umweg über die reelle Analysis:

Wir gehen zunächst aus von einem beliebigen Polynom $P(x)$ mit reellen Koeffizienten von einem geraden Grad $2n$. Dazu definieren wir das Polynom

$$Q(x) = P(x) - P''(x) + P^{(4)}(x) - \cdots + (-1)^n P^{(2n)}(x)$$

als die alternierende Summe der Ableitungen gerader Ordnung von P. Weiter betrachten wir die Funktion $S(x) = Q'(x)\sin x - Q(x)\cos x$; ihre Ableitung ist

$$\begin{aligned} S'(x) &= Q''(x)\sin x + Q'(x)\cos x - Q'(x)\cos x + Q(x)\sin x \\ &= \big(Q''(x) + Q(x)\big)\sin x\,. \end{aligned}$$

In $Q''(x) = P''(x) - P^{(4)}(x) + \cdots + (-1)^{n-1}P^{(2n)}(x)$ kommen bis auf $P(x)$ genau dieselben Terme vor wie in $Q(x)$, allerdings mit dem jeweils anderen Vorzeichen. Daher ist $Q''(x) + Q(x) = P(x)$ und $S'(x) = P(x)\sin x$.

Somit ist $S(x) = Q'(x)\sin x - Q(x)\cos x$ eine Stammfunktion von $P(x)\sin x$, und wir erhalten die

Formel von Hermite:

$$\int_0^\pi P(x)\sin x\,dx = S(\pi) - S(0) = Q(\pi) + Q(0)\,,$$

denn $\sin 0 = \sin\pi = 0$, $\cos 0 = 1$ und $\cos\pi = -1$.

Wir nehmen nun an, $\pi = a/b$ sei eine rationale Zahl und wenden die gerade bewiesene Formel an auf das spezielle Polynom

$$P(x) = \frac{x^n(a-bx)^n}{n!}\,;$$

wir erhalten

$$I_n \underset{\text{def}}{=} \int_0^\pi P(x)\sin x\,dx = Q(\pi) + Q(0)\,.$$

$P(x)$ ist im Intervall $(0,\,\pi)$ genau wie die Sinusfunktion überall positiv; daher ist auch $I_n > 0$. Außerdem ist $P(x)$ symmetrisch zu $\pi/2$, denn

$$P(\pi - x) = (\pi - x)^n\big(a - b(\pi - x)\big)^n = \Big(\frac{a}{b} - x\Big)^n (bx)^n = x^n(a-bx)^n\,.$$

Das Maximum von P in $(0,\,\pi)$ wird an derselben Stelle angenommen wie das der Funktion $f(x) = x(a-bx)$; wegen $f'(x) = a - 2bx$ handelt es sich dabei um die Intervallmitte $a/2b = \pi/2$, und der Funktionswert dort ist

$$P\Big(\frac{a}{2b}\Big) = \frac{1}{n!}\Big(\frac{a}{2b}\Big)^n\Big(\frac{ab}{2b}\Big)^n = \frac{1}{n!}\frac{a^{2n}}{(4b)^n}\,.$$

Schätzen wir das Integral ab durch Intervallänge mal Maximum des Integranden, erhalten wir daher die Ungleichung

$$I_n \le \pi\cdot\frac{1}{n!}\frac{a^{2n}}{(4b)^n}$$

und sehen, daß I_n für $n\to\infty$ gegen Null geht; denn $n!$ wächst stärker als jede Potenz einer reellen Zahl. Somit ist

$$\lim_{n\to\infty} I_n = 0\,.$$

Andererseits ist aber $I_n = Q(0) + Q(\pi)$, was in diesem Fall wegen der Symmetrie von P einfach $2Q(0)$ ist. Wir wollen uns überlegen, daß $Q(0)$ eine ganze Zahl ist. Dazu reicht es zu zeigen, daß alle Ableitungen von $P(x)$ an der Stelle Null ganzzahlige Werte annehmen. Nach der binomischen Formel ist

$$P(x) = \frac{x^n(a-bx)^n}{n!} = \frac{1}{n!}\sum_{i=0}^{n}\binom{n}{i}a^{n-i}(-b)^i x^{n+i}\,;$$

die k-te Ableitung verschwindet also für $k < n$ an der Stelle Null. Für $k = n + i \ge n$ ist

$$P^{(k)}(0) = \frac{1}{n!}\binom{n}{i}a^{n-i}(-b)^i(n+i)!$$

ebenfalls eine ganze Zahl, da der Nenner $n!$ Teiler von $(n+i)!$ ist und ansonsten nur ganze Zahlen dastehen.

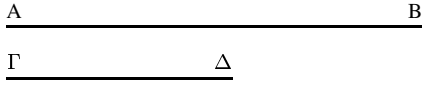
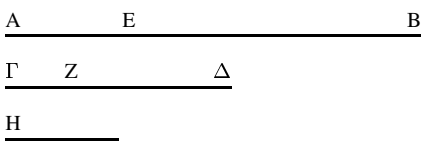
Somit ist also I_n für jedes $n\in\mathbb{N}$ eine positive ganze Zahl. Der Limes einer Folge positiver ganzer Zahlen kann aber unmöglich Null sein,

also führt die Annahme, $\pi = a/b$ sei eine rationale Zahl, zu einem Widerspruch, d.h. π ist irrational.

Wenn man schon dabei ist, kann man leicht auch noch viele andere wichtige Zahlen als irrational erkennen; auf dem ersten Übungsblatt ist ein Beweis für die Irrationalität der EULERschen Zahl e skizziert, und mit nur wenig mehr Aufwand als im Fall der Quadratwurzel aus zwei läßt sich auch leicht zeigen, daß *jede* Quadratwurzel einer natürlichen Zahl entweder ganzzahlig oder irrational ist. Dasselbe gibt für höhere Wurzeln und sogar für Nullstellen gewisser Polynome mit ganzzahligen Koeffizienten, allerdings brauchen wir für allgemeine Beweis einen Satz, den zwar fast jeder kennt, dessen Beweis man aber nur selten sieht: Die eindeutige Primzerlegung der natürlichen Zahlen. Der Beweis dieses Satzes wiederum verwendet eine Konstruktion, die wahrscheinlich bereits den Pythagoräern bekannt war und die wir heute als EUKLIDischen Algorithmus bezeichnen. Wie sich zeigen wird, ist er zusammen mit einer ganzen Reihe von Varianten ein nicht nur in der Zahlentheorie allgegenwärtiges Werkzeug; es lohnt sich also, ihn gleich jetzt zum Beginn der Vorlesung etwas ausführlicher zu betrachten.

§2: Der Euklidische Algorithmus

Bei EUKLID, in Proposition 2 des siebten Buchs seiner *Elemente,* wird er (in der Übersetzung von CLEMENS THAER in Oswalds Klassikern der exakten Wissenschaft) so beschrieben:

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

Aus heutiger Sicht erscheint hier die Voraussetzung, daß die betrachteten Größen nicht teilerfremd sein dürfen, seltsam. Sie erklärt sich daraus, daß in der griechischen Philosophie und Mathematik die Einheit eine Sonderrolle einnahm und nicht als Zahl angesehen wurde: Die Zahlen begannen erst mit der Zwei. Dementsprechend führt EUKLID in Proposition 1 des siebten Buchs fast wörtlich dieselbe Konstruktion durch für den Fall von teilerfremden Größen. Schon wenig später wurde die Eins auch in Griechenland als Zahl anerkannt, und für uns heute ist die Unterscheidung ohnehin bedeutungslos. Wir können die Bedingung, daß der ggT ungleich eins sein soll, also einfach ignorieren.

Das dem EUKLIDischen Algorithmus zugrunde liegende Prinzip der *Wechselwegnahme* oder wechselseitigen Subtraktion war in der griechischen Mathematik spätestens gegen Ende des fünften vorchristlichen Jahrhunderts bereits wohlbekannt unter dem Namen Antanairesis

(ἀνταναιρεσις) oder auch Anthyphairesis (ἀνθυφαιρεσις), und auch der Algorithmus selbst geht mit ziemlicher Sicherheit, wie so vieles in den Elementen, *nicht* erst auf EUKLID zurück: Seine *Elemente* waren das wohl mindestens vierte Buchprojekt dieses Namens, und alles spricht dafür, daß er vieles von seinen Vorgängern übernommen hat. Seine Elemente waren dann aber mit Abstand die erfolgreichsten, so daß die anderen in Vergessenheit gerieten und verloren gingen; EUKLID wurde schließlich als *der* Stoichist bekannt nach dem griechischen Titel στοιχεῖα der Elemente.

Es ist nicht ganz sicher, ob EUKLID wirklich gelebt hat; es ist möglich, wenn auch sehr unwahrscheinlich, daß EUKLID wie BOURBAKI einfach ein Pseudonym für eine Autorengruppe ist. (Das nebenstehende Bild aus dem 18. Jahrhundert ist reine Phantasie.) EUKLID ist vor allem bekannt als Autor der *Elemente,* in denen er die Geometrie seiner Zeit systematisch darstellte und (in gewisser Weise) auf wenige Definitionen sowie die berühmten fünf Postulate zurückführte; sie entstanden um 300 v. Chr. EUKLID arbeitete wohl am Museion in Alexandrien; außer den Elementen schrieb er noch ein Buch über Optik und weitere, teilweise verschollene Bücher.

Wenn wir nicht mit Zirkel und Lineal arbeiten, sondern rechnen, können wir die mehrfache „Wegnahme“ einer Strecke von einer anderen einfacher beschreiben durch eine Division mit Rest: Sind a und b die (als natürliche Zahlen vorausgesetzten) Längen der beiden Strecken und ist $a : b = q$ Rest r, so kann man q mal die Strecke b von a wegnehmen, und übrig bleibt eine Strecke der Länge r.

EUKLIDs Konstruktion wird dann zu folgendem Algorithmus:

Gegeben seien zwei natürliche Zahlen a, b.

Schritt 0: Setze $r_0 = a$ und $r_1 = b$.

Schritt i, $i \geq 1$**:** Falls r_i verschwindet, endet der Algorithmus mit $\mathrm{ggT}(a, b) = r_{i-1}$; andernfalls sei r_{i+1} der Rest bei der Division von r_{i-1} durch r_i.

EUKLID behauptet, daß dieser Algorithmus stets endet und daß das Ergebnis der größte gemeinsame Teiler der Ausgangszahlen a, b ist, d.h.

die größte natürliche Zahl, die sowohl a als auch b teilt.

Da der Divisionsrest r_{i+1} stets echt kleiner ist als sein Vorgänger r_i und eine Folge immer kleiner werdender nichtnegativer ganzer Zahlen notwendigerweise nach endlich vielen Schritten die Null erreicht, muß der Algorithmus in der Tat stets enden. Daß er mit dem richtigen Ergebnis endet, ist ebenfalls leicht zu sehen, denn im i-ten Schritt ist

$$r_{i-1} = q_i r_i + r_{i+1} \quad \text{oder} \quad r_{i+1} = r_{i-1} - q_i r_i \,,$$

so daß jeder gemeinsame Teiler von r_i und r_{i+1} auch ein Teiler von r_{i-1} ist und umgekehrt jeder gemeinsame Teiler von r_{i-1} und r_i auch r_{i+1} teilt. Somit haben r_i und r_{i-1} dieselben gemeinsamen Teiler wie r_i und r_{i+1}, insbesondere haben sie denselben größten gemeinsamen Teiler. Durch Induktion folgt, daß in jedem Schritt $\mathrm{ggT}(r_i, r_{i-1}) = \mathrm{ggT}(a, b)$ ist. Im letzten Schritt ist $r_i = 0$; da jede natürliche Zahl Teiler der Null ist, ist dann $r_{i-1} = \mathrm{ggT}(r_i, r_{i-1}) = \mathrm{ggT}(a, b)$, wie behauptet.

§3: Der erweiterte Euklidische Algorithmus

Mehr als zwei Tausend Jahre nach der Entdeckung von Anthyphairesis und EUKLIDischem Algorithmus, 1624 in Bourg-en-Bresse, stellte BACHET DE MÉZIRIAC in der zweiten Auflage seines Buchs *Problèmes plaisants et délectables qui se fonts par les nombres* Aufgaben wie die folgende:

Il y a 41 personnes en un banquet tant hommes que femmes et enfants qui en tout dépensent 40 sous, mais chaque homme paye 4 sous, chaque femme 3 sous, chaque enfant 4 deniers. Je demande combien il y a d'hommes, combien de femmes, combien d'enfants.

(Bei einem Bankett sind 41 Personen, Männer, Frauen und Kinder, die zusammen vierzig Sous ausgeben, aber jeder Mann zahlt vier Sous, jede Frau drei Sous und jedes Kind 4 Deniers. Ich frage, wie viele Männer, wie viele Frauen und wie viele Kinder es sind.)

CLAUDE GASPAR BACHET SIEUR DE MÉZIRIAC (1581-1638) verbrachte den größten Teil seines Lebens in seinem Geburtsort Bourg-en-Bresse. Er studierte bei den Jesuiten in Lyon und Milano und trat 1601 in den Orden ein, trat aber bereits 1602 wegen Krankheit wieder aus und kehrte nach Bourg zurück. Sein Buch erschien 1612; 1959 brachte der Verlag Blanchard eine vereinfachte Ausgabe heraus. Am bekanntesten ist BACHET für seine lateinische Übersetzung der *Arithmetika* von DIOPHANTOS. In einem Exemplar davon schrieb FERMAT seine Vermutung an den Rand. Auch Gedichte von BACHET sind erhalten. 1635 wurde er Mitglied der französischen Akademie der Wissenschaften.

Sobald man weiß, daß zwölf Deniers ein Sou sind (und zwanzig Sous ein Pfund), kann man dies in ein lineares Gleichungssystem übersetzen: Ist x die Zahl der Männer, y die der Frauen und z die der Kinder, so muß gelten $x + y + z = 41$ und $4x + 3y + \frac{1}{3}z = 40$.

Zur Lösung kann man zunächst die erste Gleichung nach z auflösen und in die zweite Gleichung einsetzen; dies führt auf die Gleichung

$$\frac{11}{3}x + \frac{8}{3}y = \frac{79}{3} \quad \text{oder} \quad 11x + 8y = 79\,.$$

Bei einer solchen Gleichung ist *a priori* nicht klar, ob es überhaupt Lösungen gibt: Die Gleichung $10x + 8y = 79$ beispielsweise kann keine haben, denn für ganze Zahlen x, y ist $10x + 8y$ stets gerade. Allgemein kann $ax + by = c$ höchstens dann ganzzahlige Lösungen haben, wenn der ggT von a und b Teiler von c ist.

BACHET DE MÉZIRIAC hat bewiesen, daß sie in diesem Fall auch stets Lösungen hat; das Kernstück dazu ist seine Proposition XVIII, wo er zu zwei teilerfremden Zahlen a, b ganze Zahlen x, y konstruiert, für die $ax - by = 1$ ist: *Deux nombres premiers entre eux estant donnéz, treuuer le moindre multiple de chascun d'iceux, surpassant de l'unité un multiple de l'autre.* Die Methode ist eine einfache Erweiterung des EUKLIDischen Algorithmus, und genau wie letzterer nach EUKLID benannt ist, da ihn dieser rund 150 Jahre nach seiner Entdeckung in seinem Lehrbuch darstellte, heißt auch BACHETs Satz heute *Identität von* BÉZOUT, weil dieser ihn 142 Jahre später, im Jahre 1766, in seinem Lehrbuch beschrieb (und auf Polynome verallgemeinerte).

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

Zur Lösung von Problemen wie dem von BACHET wollen wir gleich allgemein den größten gemeinsamen Teiler zweier Zahlen als Linearkombination dieser Zahlen darstellen. Dazu ist nur eine kleine Erweiterung des EUKLIDischen Algorithmus notwendig, so daß man oft auch einfach vom erweiterten EUKLIDischen Algorithmus spricht.

Die Gleichung

$$r_{i-1} = q_i r_i + r_{i+1}$$

läßt sich umschreiben als

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

so daß r_{i+1} eine ganzzahlige Linearkombination von r_i und r_{i-1} ist. Da entsprechend auch r_i Linearkombination von r_{i-1} und r_{i-2} ist, folgt induktiv, daß der ggT von a und b als ganzzahlige Linearkombination von a und b dargestellt werden kann.

Algorithmisch sieht dies folgendermaßen aus:

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

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

Diese Relationen werden in jedem der folgenden Schritte erhalten:

Schritt i, $i \geq 1$**:** Falls r_i verschwindet, endet der Algorithmus mit

$$\mathrm{ggT}(a, b) = r_{i-1} = \alpha_{i-1}a + \beta_{i-1}b\,.$$

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

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

Dann ist

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

man setze also

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

Genau wie oben folgt, daß der Algorithmus für alle natürlichen Zahlen a und b endet und daß am Ende der richtige ggT berechnet wird; außerdem sind die α_i und β_i so definiert, daß in jedem Schritt $r_i = \alpha_i a + \beta_i b$ ist, insbesondere wird also im letzten Schritt der ggT als Linearkombination der Ausgangszahlen dargestellt.

Als Beispiel wollen wir den ggT von 200 und 148 als Linearkombination darstellen. Im nullten Schritt haben wir 200 und 148 als die trivialen Linearkombinationen

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

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

$$200 = 1 \cdot 148 + 52 \quad \text{und} \quad 52 = 1 \cdot 200 - 1 \cdot 148 \, .$$

Da auch $52 \neq 0$, dividieren wir im zweiten Schritt 148 durch 52:

$148 = 2 \cdot 52 + 44$ und $44 = 148 - 2 \cdot (1 \cdot 200 - 1 \cdot 148) = 3 \cdot 148 - 2 \cdot 200$.

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

$8 = 52 - 44 = (1 \cdot 200 - 1 \cdot 148) - (3 \cdot 148 - 2 \cdot 200) = 3 \cdot 200 - 4 \cdot 148$.

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

$4 = 44 - 5 \cdot 8 = (3 \cdot 148 - 2 \cdot 200) - 5 \cdot (3 \cdot 200 - 4 \cdot 148) = 23 \cdot 148 - 17 \cdot 200$.

Bei der Division von acht durch vier schließlich ist der Divisionsrest null; damit ist vier der ggT von 148 und 200 und kann in der angegebenen Weise linear kombiniert werden.

Zur Lösung des Problems von BACHET müssen wir die Gleichung $11x + 8y = 79$ betrachten. Dazu stellen wir zunächst den ggT von 11 und 8 als Linearkombination dieser Zahlen dar.

Elf durch acht ist eins Rest drei, also ist $3 = 1 \cdot 11 - 1 \cdot 8$.

Im nächsten Schritt dividieren wir acht durch drei mit dem Ergebnis zwei Rest zwei, also ist $2 = 1 \cdot 8 - 2 \cdot 3 = 1 \cdot 8 - 2 \cdot (1 \cdot 11 - 1 \cdot 8) = -2 \cdot 11 + 3 \cdot 8$.

Im letzten Schritt wird daher drei durch zwei dividiert und wir sehen erstens, daß der ggT gleich eins ist (was hier keine Überraschung ist), und zweitens, daß gilt $1 = 3 - 2 = (1 \cdot 11 - 1 \cdot 8) - (-2 \cdot 11 + 3 \cdot 8) = 3 \cdot 11 - 4 \cdot 8$.

Damit haben wir auch eine Darstellung von 79 als Linearkombination von elf und acht:

$$79 = 79 \cdot (3 \cdot 11 - 4 \cdot 8) = 237 \cdot 11 - 316 \cdot 8 \, .$$

Dies ist allerdings nicht die gesuchte Lösung: BACHET dachte sicherlich nicht an 237 Männer, -316 Frauen und 119 Kinder.

Nun ist aber die obige Gleichung $1 = 3 \cdot 11 - 4 \cdot 8$ nicht die einzige Möglichkeit zur Darstellung der Eins als Linearkombination von acht und elf: Da $8 \cdot 11 - 11 \cdot 8$ verschwindet, können wir ein beliebiges Vielfaches dieser Gleichung dazuaddieren und bekommen die allgemeinere Lösung

$$(3 + 8k) \cdot 11 - (4 + 11k) \cdot 8 = 1 \, .$$

Entsprechend können wir auch ein beliebiges Vielfaches dieser Gleichung zur Darstellung von 79 addieren:

$$79 = (237 + 8k) \cdot 11 - (316 + 11k) \cdot 8 \, .$$

Wir müssen k so wählen, daß sowohl die Anzahl $237 + 8k$ der Männer als auch die Anzahl $-(316 + 11k)$ der Frauen positiv oder zumindest nicht negativ wird, d.h. $-\frac{237}{8} \leq k \leq -\frac{316}{11}$. Da k ganzzahlig sein muß, kommt nur $k = -29$ in Frage; es waren also fünf Männer, drei Frauen und dazu noch $41 - 5 - 3 = 33$ Kinder. Ihre Gesamtausgaben belaufen sich in der Tat auf $5 \cdot 4 + 3 \cdot 3 + 33 \cdot \frac{1}{3} = 40$ Sous.

Entsprechend kann der erweiterte EUKLIDische Algorithmus zur Lösung anderer diophantischer Gleichungen verwendet werden, von Gleichungen also, bei denen nur ganzzahlige Lösungen interessieren. Wir betrachten nur die einfache lineare Gleichung

$$ax + by = c \quad \text{mit} \quad a, b, c \in \mathbb{Z}$$

für zwei Unbekannte $x, y \in \mathbb{Z}$.

Der größte gemeinsame Teiler $d = \text{ggT}(a, b)$ von a und b teilt offensichtlich jeden Ausdruck der Form $ax + by$ mit $x, y \in \mathbb{Z}$; falls d kein Teiler von c ist, kann es also keine ganzzahlige Lösung geben.

Ist aber $c = rd$ ein Vielfaches von d und ist $d = \alpha a + \beta b$ die lineare Darstellung des ggT nach dem erweiterten EUKLIDischen Algorithmus, so haben wir mit $x = r\alpha$ und $y = r\beta$ offensichtlich eine Lösung gefunden.

Ist (x', y') eine weitere Lösung, so ist

$$a(x - x') + b(y - y') = c - c = 0 \quad \text{oder} \quad a(x - x') = b(y' - y)\,.$$

$v = a(x - x') = b(y' - y)$ ist also ein gemeinsames Vielfaches von a und b und damit auch ein Vielfaches des kleinsten gemeinsamen Vielfachen von a und b. Dieses kleinste gemeinsame Vielfache ist ab/d, es muß also eine ganze Zahl m geben mit

$$x - x' = m \cdot \frac{b}{d} \quad \text{und} \quad y' - y = m \cdot \frac{a}{d}\,.$$

Die allgemeine Lösung der obigen Gleichung ist somit

$$x = r\alpha - m \cdot \frac{b}{d} \quad \text{und} \quad y = r\beta + m \cdot \frac{a}{d} \quad \text{mit} \quad m \in \mathbb{Z}\,.$$

§4: Der Aufwand des Euklidischen Algorithmus

Im Beweis, daß der EUKLIDische Algorithmus stets nach endlich vielen Schritten abbricht, hatten wir argumentiert, daß der Divisionsrest stets kleiner ist als der Divisor, so daß er irgendwann einmal null werden muß; dann endet der Algorithmus.

Damit haben wir auch eine obere Schranke für den Rechenaufwand zur Berechnung von $\text{ggT}(a, b)$: Wir müssen höchstens b Divisionen durchführen.

Das erscheint zwar auf den ersten Blick als ein recht gutes Ergebnis; wenn man aber bedenkt, daß der EUKLIDische Algorithmus heute in der Kryptographie auf über 600-stellige Zahlen angewendet wird, verliert diese Schranke schnell ihre Nützlichkeit: Da unser Universum ein geschätztes Alter von zehn Milliarden Jahren, also ungefähr $3 \cdot 10^{18}$ Sekunden hat, ist klar, daß auch der schnellste heutige Computer, der zu Beginn des Universum zu Rechnen begann, bis heute nur einen verschwindend kleinen Bruchteil von 10^{600} Divisionen ausgeführt hätte; Wäre 10^{600} eine realistische Aufwandsabschätzung, könnten wir an eine Anwendung des EUKLIDischen Algorithmus auf 600-stellige Zahlen nicht einmal denken.

Tatsächlich ist 10^{600} aber natürlich nur eine obere Schranke, von der wir bislang noch nicht wissen, wie realistisch sie ist. Um dies zu entscheiden, suchen wir die kleinsten natürlichen Zahlen a, b, für die n Divisionen notwendig sind; dies wird uns zurückführen auf ein Problem aus dem 13. Jahrhundert.

Im Falle $n = 1$ sind offensichtlich $a = b = 1$ die kleinstmöglichen Zahlen; wenn $a = b$ ist, kommt man immer mit genau einer Division aus.

Dies ist allerdings ein eher untypischer Fall, der sich insbesondere nicht rekursiv verallgemeinern läßt, denn ab dem zweiten Schritt des EUKLIDischen Algorithmus ist der Divisor stets kleiner als der Dividend: Ersterer ist schließlich der Rest bei der vorangegangenen Division und letzterer der Divisor. Die kleinsten natürlichen Zahlen $a \neq b$, für die man mit nur einer Division auskommt, sind offensichtlich $a = 2$ und $b = 1$.

Als nächstes suchen wir die kleinsten Zahlen a, b für die zwei Divisionen notwendig sind. Ist r der Rest bei der ersten Division, so ist $b : r$ die zweite Division. Für diese muß $r \geq 1$ und $b \geq 2$ sein, und $a = qb + r$, wobei q der Quotient bei der ersten Division ist. Dieser ist mindestens eins, die kleinstmöglichen Werte sind damit

$$r = 1, \quad b = 2 \quad \text{und} \quad a = b + r = 3\,.$$

Allgemeiner seien a_n und b_n die kleinsten Zahlen, für die n Divisionen notwendig sind, und r sei der Rest bei der ersten Division. Für die zweite

Division $b : r$ ist dann $b_n \geq a_{n-1}$ und $r \geq b_{n-1}$; die kleinstmöglichen Werte sind damit

$$r = b_{n-1}, \quad b_n = a_{n-1} \text{ und } a_n = b_n + r = a_{n-1} + b_{n-1} = a_{n-1} + a_{n-2}\,.$$

Da wir $a_1 = 2$ und $b_1 = 1$ kennen, können wir somit alle a_n und b_n berechnen; was wir erhalten, sind die sogenannten FIBONACCI-Zahlen.

Sie sind durch folgende Rekursionsformel definiert:

$$F_0 = 0, \qquad F_1 = 1 \qquad \text{und} \qquad F_n = F_{n-1} + F_{n-2} \quad \text{für } n \geq 2\,.$$

FIBONACCI führte sie ein, um die Vermehrung einer Karnickelpopulation durch ein einfaches Modell zu berechnen. In seinem 1202 erschienenen Buch *Liber abaci* schreibt er:

Ein Mann bringt ein Paar Karnickel auf einen Platz, der von allen Seiten durch eine Mauer umgeben ist. Wie viele Paare können von diesem Paar innerhalb eines Jahres produziert werden, wenn man annimmt, daß jedes Paar jeden Monat ein neues Paar liefert, das vom zweiten Monat nach seiner Geburt an produktiv ist?

LEONARDO PISANO (1170–1250) ist heute vor allem unter seinem Spitznamen FIBONACCI bekannt; gelegentlich nannte er sich auch BIGOLLO, auf Deutsch *Tunichtgut* oder *Reisender.* Er ging in Nordafrika zur Schule, kam aber 1202 zurück nach Pisa. Seine Bücher waren mit die ersten, die die indisch-arabischen Ziffern in Europa einführten. Er behandelt darin nicht nur Rechenaufgaben für Kaufleute, sondern auch zahlentheoretische Fragen, beispielsweise daß man die Quadratzahlen durch Aufaddieren der ungeraden Zahlen erhält. Auch betrachtet er Beispiele nichtlinearer Gleichungen, die er approximativ löst, und erinnert an viele in Vergessenheit geratene Ergebnisse der antiken Mathematik.

Wie wir gerade gesehen haben, kann man mit den FIBONACCI-Zahlen nicht nur Karnickelpopulationen beschreiben, sondern – wie GABRIEL LAMÉ 1844 entdeckte – auch eine Obergrenze für den Aufwand beim EUKLIDischen Algorithmus angeben:

Satz von Lamé (1844)**:** Die kleinsten natürlichen Zahlen a, b, für die beim EUKLIDischen Algorithmus $n \geq 2$ Divisionen benötigt werden, sind $a = F_{n+2}$ und $b = F_{n+1}$. ∎

GABRIEL LAMÉ (1795–1870) studierte von 1813 bis 1817 Mathematik an der Ecole Polytechnique, danach bis 1820 Ingenieurwissenschaften an der Ecole des Mines. Auf Einladung Alexanders I. kam er 1820 nach Rußland, wo er in St. Petersburg als Professor und Ingenieur unter anderem Vorlesungen über Analysis, Physik, Chemie und Ingenieurwissenschaften hielt. 1832 erhielt er einen Lehrstuhl für Physik an der Ecole Polytechnique in Paris, 1852 einen für mathematische Physik und Wahrscheinlichkeitstheorie an der Sorbonne. 1836/37 war er wesentlich am Bau der Eisenbahnlinien Paris-Versailles und Paris–St. Germain beteiligt.

(Für $n = 1$ gilt der Satz nur, wenn wir zusätzlich voraussetzen, daß $a \neq b$ ist; für $n \geq 2$ ist dies automatisch erfüllt.)

Um die Zahlen F_n durch eine geschlossene Formel darzustellen, können wir (genau wie man es auch für die rekursive Berechnung per Computer tun würde) die Definitionsgleichung der FIBONACCI-Zahlen als

$$\begin{pmatrix} F_{n+1} \\ F_n \end{pmatrix} = A \begin{pmatrix} F_n \\ F_{n-1} \end{pmatrix} \quad \text{mit} \quad A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

schreiben; dann ist

$$\begin{pmatrix} F_n \\ F_{n-1} \end{pmatrix} = A^{n-1} \begin{pmatrix} F_1 \\ F_0 \end{pmatrix} = A^{n-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}\,.$$

Das charakteristische Polynom von A ist

$$\det(A - \lambda E) = (1 - \lambda)(-\lambda) - 1 = \lambda^2 - \lambda - 1\,;$$

die Eigenwerte von A sind daher $\lambda_{1/2} = \frac{1}{2} \pm \frac{1}{2}\sqrt{5}$. Bezeichnet B die Matrix, deren Spalten aus den zugehörigen Eigenvektoren besteht, so ist also $A = B^{-1} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} B$ und

$$\begin{pmatrix} F_n \\ F_{n-1} \end{pmatrix} = A^{n-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = B^{-1} \begin{pmatrix} \lambda_1^{n-1} & 0 \\ 0 & \lambda_2^{n-1} \end{pmatrix} B \begin{pmatrix} 1 \\ 0 \end{pmatrix}\,.$$

Auch ohne die Matrix B zu berechnen, wissen wir somit, daß sich F_n in der Form $F_n = a\lambda_1^{n-1} + b\lambda_2^{n-1}$ darstellen läßt. Für $n = 1$ und $n = 2$ erhalten wir die beiden Bedingungen

$$1 = a\lambda_1^0 + b\lambda_2^0 = a + b \quad \text{und} \quad 1 = a\lambda_1 + b\lambda_2 \,.$$

Damit ist $b = 1 - a$, und die zweite Gleichung wird zu

$$a(\lambda_1 - \lambda_2) + \lambda_2 = a\sqrt{5} + \lambda_2 = 1 \Longrightarrow a = \frac{1-\lambda_2}{\sqrt{5}} = \frac{\lambda_1}{\sqrt{5}} \,.$$

Also ist $b = 1 - a = -\lambda_2/\sqrt{5}$ und

$$F_n = \frac{\lambda_1^n - \lambda_2^n}{\sqrt{5}} \,.$$

Numerisch ist

$$\lambda_1 = \frac{1+\sqrt{5}}{2} \approx 1{,}618034, \quad \lambda_2 = 1 - \lambda_1 = \frac{1-\sqrt{5}}{2} \approx -0{,}618034$$

und $\sqrt{5} \approx 2{,}236068$; der Quotient $\lambda_2^n/\sqrt{5}$ ist also für jedes n kleiner als $1/2$. Daher können wir F_n auch einfacher berechnen als nächste ganze Zahl zu $\lambda_1^n/\sqrt{5}$. Insbesondere folgt, daß F_n exponentiell mit n wächst.

Die Gleichung $\lambda^2 - \lambda - 1 = 0$ läßt sich umschreiben als $\lambda(\lambda - 1) = 1$ oder $\lambda : 1 = 1 : (\lambda - 1)$. Diese Gleichung charakterisiert den *goldenen Schnitt:* Stehen zwei Strecken a und b in diesem Verhältnis, so auch die beiden Strecken b und $a - b$. Die positive Lösung λ_1 wird traditionell mit dem Buchstaben ϕ bezeichnet; F_n ist also der zur nächsten ganzen Zahl gerundete Wert von $\phi^n/\sqrt{5}$.

Die beiden kleinsten Zahlen, für die wir n Divisionen brauchen, sind nach LAMÉ $a = F_{n+2}$ und $b = F_{n+1}$. Aus der geschlossenen Formel für die FIBONACCI-Zahlen folgt

$$\begin{aligned} n &\approx \log_\phi \sqrt{5}\, b - 1 = \log_\phi b + \log_\phi \sqrt{5} - 1 = \frac{\ln b}{\ln \phi} + \frac{\ln \sqrt{5}}{\ln \phi} - 1 \\ &\approx 2{,}078 \ln b + 0{,}672 \,. \end{aligned}$$

Für beliebige Zahlen $a > b$ können nicht mehr Divisionen notwendig sein als für die auf b folgenden nächstgrößeren FIBONACCI-Zahlen, also gibt obige Formel für jedes b eine obere Grenze. Die Anzahl der Divisionen wächst daher nicht (wie oben bei der naiven Abschätzung) wie b, sondern höchstens wie $\log b$. Für sechshundertstellige Zahlen a, b müssen wir daher nicht mit 10^{600} Divisionen rechnen, sondern mit weniger als drei Tausend, was auch für weniger leistungsfähige Computer problemlos und schnell möglich ist.

Tatsächlich gibt natürlich auch die hier berechnete Schranke nur selten den tatsächlichen Aufwand wieder; fast immer werden wir mit erheblich weniger auskommen. Im übrigen ist auch alles andere als klar, ob wir den ggT auf andere Weise nicht möglicherweise schneller berechnen können. Da wir aber für Zahlen der Größenordnung, die in heutigen Anwendungen interessieren, selbst mit der Schranke für den schlimmsten Fall ganz gut leben können, sei hier auf diese Fragen nicht weiter eingegangen. Interessenten finden mehr dazu z.B. in den Abschnitten 4.5.2+3 des Buchs

DONALD E. KNUTH: The Art of Computer Programming, *vol. 2:* Seminumerical Algorithms, *Addison-Wesley,* ^{2}1981

§5: Die multiplikative Struktur der ganzen Zahlen

Eine Primzahl ist bekanntlich eine natürliche Zahl p, die genau zwei Teiler hat, nämlich die Eins und sich selbst. Der erweiterte EUKLIDische Algorithmus liefert eine wichtige Folgerung aus dieser Definition:

Lemma: Wenn eine Primzahl das Produkt ab zweier natürlicher Zahlen teilt, teilt sie mindestens einen der Faktoren.

Beweis: Angenommen, die Primzahl p sei kein Teiler von a, teile aber ab. Da der ggT von a und p Teiler von p und ungleich p ist, muß er notgedrungen gleich eins sein; es gibt also eine Darstellung

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

Dann ist $b = \alpha ab + \beta pb$ durch p teilbar, denn sowohl ab also auch pb sind Vielfache von p. ∎

Daraus folgt induktiv

Satz: Jede natürliche Zahl läßt sich bis auf Reihenfolge eindeutig als ein Produkt von Primzahlpotenzen schreiben.

Beweis: Wir zeigen zunächst, daß sich jede natürliche Zahl überhaupt als Produkt von Primzahlpotenzen schreiben läßt. Falls dies nicht der Fall wäre, gäbe es ein minimales Gegenbeispiel M. Dies kann nicht die Eins sein, denn die ist ja das leere Produkt, und es kann auch keine Primzahl sein, denn die ist ja das Produkt mit sich selbst als einzigem Faktor. Somit hat M einen echten Teiler N, d.h. $1 < N < M$. Da M das minimale Gegenbeispiel war, lassen sich N und M/N als Produkte von Primzahlpotenzen schreiben, also auch $M = N \times M/N$.

Bleibt noch zu zeigen, daß die Produktdarstellung bis auf die Reihenfolge der Faktoren eindeutig ist. Auch hier gäbe es andernfalls wieder ein minimales Gegenbeispiel M, das somit mindestens zwei verschiedene Darstellungen

$$M = \prod_{i=1}^{r} p_i^{e_i} = \prod_{j=1}^{s} q_j^{f_j}$$

hätte. Da die Eins durch kein Produkt dargestellt werden kann, in dem wirklich eine Primzahl vorkommt, ist $M > 1$ und somit steht in jedem der beiden Produkte mindestens eine Primzahl.

Da p_1 Teiler von M ist, teilt es auch das rechtsstehende Produkt, also nach dem gerade bewiesenen Lemma mindestens einen der Faktoren, d.h. mindestens ein q_j. Da q_j eine Primzahl ist, muß dann $p_1 = q_j$ sein. Da M als minimales Gegenbeispiel vorausgesetzt war, unterscheiden sich die beiden Produkte, aus denen dieser gemeinsame Faktor gestrichen wurde, höchstens durch die Reihenfolge der Faktoren, und damit gilt dasselbe für die beiden Darstellungen von M. ∎

Als erste Anwendung dieses Satzes können wir zeigen

Satz: Die reelle Zahl x erfülle die Gleichung

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0 \quad \text{mit} \quad a_i \in \mathbb{Z}\,.$$

Dann ist x entweder ganzzahlig oder irrational.

Beweis: Jede rationale Zahl x kann als Quotient $x = p/q$ zweier zueinander teilerfremder ganzer Zahlen p und q geschrieben werden. Multiplizieren wir die Gleichung

$$\left(\frac{p}{q}\right)^n + a_{n-1}\left(\frac{p}{q}\right)^{n-1} + \cdots + a_1\left(\frac{p}{q}\right) + a_0 = 0$$

mit q^n, erhalten wir die nennerlose Gleichung

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

Auflösen nach p^n führt auf

$$\begin{aligned} p^n &= -a_{n-1}p^{n-1}q - \cdots - a_1 pq^{n-1} - a_0 q^n \\ &= q(-a_{n-1}p^{n-1} - \cdots - a_1 pq^{n-2} - a_0 q^{n-1})\,, \end{aligned}$$

d.h. q muß ein Teiler von p^n sein, was wegen der Eindeutigkeit der Primfaktorzerlegung von p^n sowie der vorausgesetzten Teilerfremdheit von p und q nur für $q = \pm 1$ der Fall sein kann. Somit ist x eine ganze Zahl, wie behauptet. ∎

§6: Kongruenzenrechnung

Zwei ganze Zahlen lassen sich im allgemeinen nicht durcheinander dividieren. Trotzdem – oder gerade deshalb – spielen Teilbarkeitsfragen in der Zahlentheorie eine große Rolle. Das technische Werkzeug zu ihrer Behandlung ist die Kongruenzenrechnung.

Definition: Wir sagen, zwei ganze Zahlen $x, y \in \mathbb{Z}$ seien kongruent modulo m für eine natürliche Zahl m, in Zeichen

$$x \equiv y \bmod m\,,$$

wenn $x - y$ durch m teilbar ist.

Die Kongruenz modulo m definiert offensichtlich eine Äquivalenzrelation auf $\mathbb{Z}$: Jede ganze Zahl ist kongruent zu sich selbst, denn $x - x = 0$ ist durch jede natürliche Zahl teilbar; wenn $x - y$ durch m teilbar ist, so auch $y - x = -(x - y)$, und ist schließlich $x \equiv y \bmod m$ und

$y \equiv z \bmod m$, so sind $x - y$ und $y - z$ durch m teilbar, also auch ihre Summe $x - z$, und damit ist auch $x \equiv z \bmod m$.

Zwei Zahlen $x, y \in \mathbb{Z}$ liegen genau dann in derselben Äquivalenzklasse, wenn sie bei der Division durch m denselben Divisionsrest haben; es gibt somit m Äquivalenzklassen, die den m möglichen Divisionsresten $0, 1, \ldots, m-1$ entsprechen.

Lemma: Ist $x \equiv x' \bmod m$ und $y \equiv y' \bmod m$, so ist auch

$$x \pm y \equiv x' \pm y' \bmod m \quad \text{und} \quad x'y' \equiv xy \bmod m\,.$$

Beweis: Sind $x - x'$ und $y - y'$ durch m teilbar, so auch

$$(x \pm y) - (x' \pm y') = (x - x') \pm (y - y') \qquad \text{und}$$
$$xy - x'y' = x(y - y') + y'(x - x')$$

∎

Im folgenden wollen wir das Symbol „mod“ nicht nur in Kongruenzen wie $x \equiv y \bmod m$ benutzen, sondern auch – wie in vielen Programmiersprachen üblich – als Rechenoperation:

Definition: Für eine ganze Zahl x und eine natürliche Zahl m bezeichnet $x \bmod m$ jene ganze Zahl $0 \le r < m$ mit $x \equiv r \bmod m$.

$x \bmod m$ ist also einfach der Divisionsrest bei der Division von x durch m.

Da nach dem gerade bewiesenen Lemma die Addition, Subtraktion und Multiplikation mit Kongrenzen vertauschbar sind, können wir auf der Menge aller Äquivalenzklassen Rechenoperationen einführen. Übersichtlicher wird das, wenn wir statt dessen die Menge

$$\mathbb{Z}/m \underset{\text{def}}{=} \{0, 1, \ldots, m-1\}$$

betrachten. Wir definieren eine Addition durch

$$x \oplus y = (x + y) \bmod m = \begin{cases} x + y & \text{falls } x + y < m \\ x + y - m & \text{sonst} \end{cases}$$

und entsprechend eine Multiplikation gemäß

$$x \odot y = (xy) \bmod m\,.$$

Für $m = 4$ haben wir also folgende Operationen:

$\oplus$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

und

$\odot$	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	0	2
3	0	3	2	1

Um diese Tabellen zu interpretieren, sollten wir uns an einige Grundbegriffe aus der Algebra erinnern:

Definition: *a)* Eine Gruppe ist eine Menge G zusammen mit einer Verknüpfung $\times\colon G \times G \to G$, für die gilt
1.) $(x \times y) \times z = x \times (y \times z)$ für alle $x, y, z \in G$.
2.) Es gibt ein Element $e \in G$, so daß $e \times x = x \times e = x$ für alle $x \in G$.
3.) Zu jedem $x \in G$ gibt es ein $x' \in G$, so daß $x \times x' = x' \times x = e$ ist.

Die Gruppe heißt kommutativ oder abelsch, wenn zusätzlich noch gilt

4.) $x \times y = y \times x$ für alle $x, y \in G$.

b) Eine Abbildung $\varphi\colon G \to H$ zwischen zwei Gruppen G und H mit Verknüpfungen $\times$ und $\otimes$ heißt (Gruppen-)Homomorphismus, falls für alle $x, y \in G$ gilt: $\varphi(x \times y) = \varphi(x) \otimes \varphi(y)$. Ist φ zusätzlich bijektiv, reden wir von einem Isomorphismus. Die Gruppen G und H heißen isomorph, in Zeichen $G \cong H$, wenn es einen Isomorphismus $\varphi\colon G \to H$ gibt.

c) Ein Ring ist eine Menge R zusammen mit zwei Verknüpfungen $+, \cdot\colon R \times R \to R$, so daß gilt
1.) Bezüglich + ist R eine abelsche Gruppe.
2.) $(x \cdot y) \cdot z = x \cdot (y \cdot z)$ für alle $x, y, z \in R$.
3.) Es gibt ein Element $1 \in R$, so daß $1 \cdot x = x \cdot 1 = x$ für alle $x \in R$.
4.) $x \cdot (y + z) = x \cdot y + x \cdot z$ und $(x + y) \cdot z = x \cdot z + y \cdot z$ für alle $x, y, z \in R$.
Der Ring heißt kommutativ, wenn zusätzlich noch gilt
5.) $x \cdot y = y \cdot x$ für alle $x, y \in R$.

d) Eine Abbildung $\varphi\colon R \to S$ zwischen zwei Ringen heißt (Ring-) Homomorphismus, wenn für alle $x, y \in R$ gilt

$$\varphi(r+s) = \varphi(r) + \varphi(s) \quad \text{und} \quad \varphi(r \cdot s) = \varphi(r) \cdot \varphi(s)\,,$$

wobei + und $\cdot$ auf der linken Seite jeweils die Operationen von R bezeichnen und rechts die von S. Auch hier reden wir von einem Isomorphismus, wenn φ bijektiv ist, und bezeichnen $R \cong S$ als isomorph, wenn es einen Isomorphismus $\varphi\colon R \to S$ gibt.

Lemma: Für jedes $m \in \mathbb{N}$ ist $\mathbb{Z}/m$ mit den Operationen $\oplus$ und $\odot$ ein Ring.

Beweis: Wir betrachten die Abbildung

$$\varphi\colon \begin{cases} \mathbb{Z} \to \mathbb{Z}/m \\ x \mapsto x \bmod m \end{cases}.$$

Nach dem obigen Lemma ist

$$\varphi(x+y) = \varphi(x) \oplus \varphi(y) \quad \text{und} \quad \varphi(xy) = \varphi(x) \odot \varphi(y)\,.$$

Da $\mathbb{Z}$ bezüglich + eine abelsche Gruppe ist, gilt somit dasselbe für $\mathbb{Z}/m$ bezüglich $\oplus$: Wenn zwei ganze Zahlen gleich sind, sind schließlich auch ihre Divisionsreste modulo m gleich. Das Neutralelement ist $\varphi(0) = 0$, und das additive Inverse ist $\varphi(-x) = m - \varphi(x)$. Auch die Eigenschaften von $\odot$ folgen sofort aus den entsprechenden Eigenschaften der Multiplikation ganzer Zahlen; das Neutralelement ist $\varphi(1) = 1$. ∎

Man beachte, daß $\mathbb{Z}/m$ im allgemeinen kein Körper ist: In $\mathbb{Z}/4$ beispielsweise ist $2 \odot 2 = 0$, und in einem Körper kann ein Produkt nur verschwinden, wenn mindestens einer der beiden Faktoren verschwindet.

Im folgenden werden wir die Rechenoperationen in $\mathbb{Z}/m$ einfach mit + und $\cdot$ bezeichnen, wobei jedesmal aus dem Zusammenhang klar sein sollte, ob wir von der Addition und Multiplikation in $\mathbb{Z}/m$ oder der in $\mathbb{Z}$ reden. Der Malpunkt wird dabei, wie üblich, oft weggelassen.

§7: Der chinesische Restesatz

Der Legende nach zählten chinesische Generäle ihre Truppen, indem sie diese mehrfach antreten ließen in Reihen verschiedener Breiten $m_1, \ldots, m_r$ und jedesmal nur die Anzahl a_r der Soldaten in der letzten Reihe zählten. Aus den r Relationen

$$x \equiv a_1 \bmod m_1, \quad \ldots, \quad x \equiv a_r \bmod m_r$$

bestimmten sie dann die Gesamtzahl x der Soldaten.

Ob es im alten China wirklich Generäle gab, die soviel Mathematik konnten, sei dahingestellt; Beispiele zu diesem Satz finden sich jedenfalls bereits 1247 in den chinesischen *Mathematischen Abhandlungen in neun Bänden* von CH'IN CHIU-SHAO (1202–1261), allerdings geht es dort nicht um Soldaten, sondern um Reis.

CH'IN CHIU-SHAO oder QIN JIUSHAO wurde 1202 in der Provinz Sichuan geboren. Auf eine wilde Jugend mit vielen Affairen folgte ein wildes und alles andere als gesetztreues Berufsleben in Armee, öffentlicher Verwaltung und illegalem Salzhandel. Als Jugendlicher studierte er an der Akademie von Hang-chou Astronomie, später brachte ihm ein unbekannter Lehrer Mathematik bei. Insbesondere studierte er die in vorchristlicher Zeit entstandenen *Neun Bücher der Rechenkunst,* nach deren Vorbild er 1247 seine deutlich anspruchsvolleren *Mathematischen Abhandlungen in neun Bänden* publizierte, die ihn als einen der bedeutendsten Mathematiker nicht nur Chinas ausweisen. Zum chinesischen Restesatz schreibt er, daß er ihn von den Kalendermachern gelernt habe, diese ihn jedoch nur rein mechanisch anwendeten ohne ihn zu verstehen. CH'IN CHIU-SHAO starb 1261 in Meixian, wohin er nach einer seiner vielen Entlassungen wegen krimineller Machenschaften geschickt worden war.

Wir wollen uns zunächst überlegen, unter welchen Bedingungen ein solches Verfahren überhaupt funktionieren kann. Offensichtlich können die obigen r Relationen eine natürliche Zahl nicht eindeutig festlegen, denn ist x eine Lösung und M irgendein gemeinsames Vielfaches der sämtlichen m_i, so ist $x + M$ auch eine – M ist schließlich modulo aller m_i kongruent zur Null.

Außerdem gibt es Relationen obiger Form, die unlösbar sind, beispielsweise das System

$$x \equiv 2 \bmod 4 \quad \text{und} \quad x \equiv 3 \bmod 6\,,$$

dessen erste Gleichung nur gerade Lösungen hat, während die zweite nur ungerade hat. Das Problem hier besteht darin, daß zwei ein gemeinsamer Teiler von vier und sechs ist, so daß jede der beiden Kongruenzen auch etwas über x mod 2 aussagt, wobei diese beiden Aussagen hier einander widersprechen.

Dieses Problem können wir dadurch umgehen, daß wir nur Moduln m_i zulassen, die paarweise teilerfremd sind. Dies hat auch den Vorteil, daß jedes gemeinsame Vielfache der m_i Vielfaches des Produkts aller m_i sein muß, so daß wir x modulo einer vergleichsweise großen Zahl kennen.

Chinesischer Restesatz: Das System von Kongruenzen

$$x \equiv a_1 \bmod m_1, \quad \ldots, \quad x \equiv a_r \bmod m_r$$

hat für paarweise teilerfremde Moduln m_i genau eine Lösung x mit $0 \leq x < m_1 \cdots m_r$. Jede andere Lösung $y \in \mathbb{Z}$ läst sich in der Form $x + km_1 \cdots m_r$ schreiben mit $k \in \mathbb{Z}$.

Mit den Begriffen aus dem vorigen Paragraphen läßt sich dies auch anders formulieren: Die Zahl x mod m_i können wir auffassen als Element von $\mathbb{Z}/m_i$, das r-Tupel aus allen diesen Zahlen also als Element von $\mathbb{Z}/m_1 \times \cdots \times \mathbb{Z}/m_r$. Man überlegt sich leicht, daß das kartesische Produkt von zwei oder mehr Gruppen wieder eine Gruppe ist: Die Verknüpfung wird einfach komponentenweise definiert, und das Neutralelement ist dasjenige Tupel, dessen sämtliche Komponenten Neutralelemente der jeweiligen Faktoren sind. Genauso folgt, daß das kartesische Produkt von zwei oder mehr Ringen wieder ein Ring ist.

In algebraischer Formulierung haben wir dann die folgende Verschärfung des obigen Satzes:

Chinesischer Restesatz *(Algebraische Form)***:** Die Abbildung

$$\varphi\colon \begin{cases} \mathbb{Z}/m_1 \cdots m_r \to Z/m_1 \times \cdots \times \mathbb{Z}/m_r \\ \qquad\qquad x \mapsto (x \bmod m_1, \ldots, x \bmod m_r) \end{cases}$$

ist ein Isomorphismus von Ringen.

Wir *beweisen* den Satz in dieser algebraischen Form:

Zunächst ist

$$\psi\colon \begin{cases} \mathbb{Z} \to \mathbb{Z}/m_1 \times \cdots \times \mathbb{Z}/m_r \\ x \mapsto (x \bmod m_1, \ldots, x \bmod m_r) \end{cases}$$

ein Ringhomomorphismus, denn nach dem Lemma aus dem vorigen Paragraphen ist der Übergang zu den Restklassen modulo jeder der Zahlen m_i mit Addition und Multiplikation vertauschbar. Da $\psi(x)$ nur von x mod $m_1 \cdots m_r$ abhängt, folgt daraus, daß auch φ ein Ringhomomorphismus ist.

φ ist injektiv, denn ist $\varphi(x) = \varphi(y)$, so ist $x \bmod m_i = y \bmod m_i$ für alle i; da die m_i paarweise teilerfremd sind, ist $x - y$ somit durch das Produkt der m_i teilbar, was für $x, y \in \mathbb{Z}/m_1 \cdots m_r$ nur im Fall $x = y$ möglich ist.

Nun ist φ aber eine Abbildung zwischen endlichen Mengen, die beide aus je $m_1 \cdots m_r$ Elementen bestehen. Jede injektive Abbildung zwischen zwei gleichmächtigen endlichen Mengen ist zwangsläufig auch surjektiv, also bijektiv, und somit ist φ ein Isomorphismus. ∎

Aus Sicht der chinesischen Generäle ist dieser Beweis enttäuschend: Angenommen, ein General weiß, daß höchstens Tausend Soldaten vor ihm stehen. Er läßt sie in Zehnerreihen antreten; in der letzten Reihe stehen fünf Soldaten. Bei Elferreihen sind es neun, bei Dreizehnerreihen sechs. Da $10 \cdot 11 \cdot 13 = 1430$ größer ist als Tausend, weiß er, daß dies die Anzahl eindeutig festlegt. Er weiß aber nicht, wie viele Soldaten nun tatsächlich vor ihm stehen. Als General hat er natürlich die Möglichkeit, einige Soldaten abzukommandieren, die für jede Zahl bis Tausend die Divisionsreste modulo 9, 10 und 13 berechnen müssen, bis sie auf das Tripel $(5, 9, 6)$ stoßen. Wir als Mathematiker sollten jedoch eine effizientere Methode finden.

Dazu verhilft uns der erweiterte EUKLIDische Algorithmus:

Wir beginnen mit dem Fall zweier Kongruenzen

$$x \equiv a \bmod m \quad \text{und} \quad y \equiv b \bmod n$$

mit zueinander teilerfremden Zahlen m und n. Ihr ggT eins läßt sich nach dem erweiterten EUKLIDischen Algorithmus als $1 = \alpha m + \beta n$ schreiben. Somit ist

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

also löst

$$x = \beta n a + \alpha m b \equiv \begin{cases} a & \mod m \\ b & \mod n \end{cases}$$

das Problem.

x ist natürlich nicht die einzige Lösung; wenn wir ein gemeinsames Vielfaches von m und n addieren, ändert sich nichts an den Kongruenzen. Da wir von teilerfremden Zahlen ausgegangen sind, ist das Produkt das kleinste gemeinsame Vielfache; die allgemeine Lösung ist daher

$$x + (\beta n + \lambda b)a + (\alpha m - \lambda a)b\,.$$

Insbesondere ist die Lösung eindeutig modulo nm.

Als Beispiel betrachten wir die beiden Kongruenzen

$$x \equiv 1 \mod 17 \quad \text{und} \quad x \equiv 5 \mod 19\,.$$

Wir müssen als erstes den erweiterten EUKLIDischen Algorithmus auf die beiden Moduln 17 und 19 anwenden:

$$19 : 17 = 1 \text{ Rest } 2 \Longrightarrow 2 = 19 - 17$$

$$17 : 2 = 8 \text{ Rest } 1 \Longrightarrow 1 = 17 - 8 \cdot 2 = 9 \cdot 17 - 8 \cdot 19$$

Also ist $9 \cdot 17 = 153 \equiv 0 \mod 17$ und $\equiv 1 \mod 19$; außerdem ist $-8 \cdot 19 = -152$ durch 19 teilbar und $\equiv 1 \mod 17$. Die Zahl

$$x = -152 \cdot 1 + 153 \cdot 5 = 613$$

löst somit das Problem. Da 613 größer ist als $17 \cdot 19 = 323$, ist allerdings nicht 613 die kleinste positive Lösung, sondern $613 - 323 = 290$.

Bei mehr als zwei Kongruenzen gehen wir rekursiv vor: Wir lösen die ersten beiden Kongruenzen $x \equiv a_1 \mod m_1$ und $x \equiv a_2 \mod m_2$ wie gerade besprochen; das Ergebnis ist eindeutig modulo $m_1 m_2$. Ist c_2 eine feste Lösung, so läßt sich die Lösung schreiben als Kongruenz

$$x \equiv c_2 \mod m_1 m_2\,,$$

und da die m_i paarweise teilerfremd sind, ist auch $m_1 m_2$ teilerfremd zu m_3. Mit EUKLID können wir daher das System

$$x \equiv c_2 \mod m_1 m_2 \quad \text{und} \quad x \equiv a_3 \mod m_3$$

lösen und die Lösung schreiben als

$$x \equiv c_3 \mod m_1 m_2 m_3$$

und so weiter, bis wir schließlich x modulo dem Produkt aller m_i kennen und somit das Problem gelöst haben.

Im Beispiel des oben angesprochenen Systems

$$x \equiv 5 \mod 10, \quad x \equiv 9 \mod 11, \quad x \equiv 6 \mod 13$$

lösen wir also zunächst nur das System

$$x \equiv 5 \mod 10 \quad \text{und} \quad x \equiv 9 \mod 11\,.$$

Da $1 = 11 - 10$, ist $11 \equiv 0 \mod 11$ und $11 \equiv 1 \mod 10$; entsprechend ist $-10 \equiv 0 \mod 10$ und $-10 \equiv 1 \mod 11$. Also ist

$$x = 5 \cdot 11 - 9 \cdot 10 = -35$$

eine Lösung; die allgemeine Lösung ist $-35 + 110k$ mit $k \in \mathbb{Z}$. Die kleinste positive Lösung ist $-35 + 110 = 75$.

Unser Ausgangssystem ist somit äquivalent zu den beiden Kongruenzen

$$x \equiv 75 \mod 110 \quad \text{und} \quad x \equiv 6 \mod 13\,.$$

Um es zu lösen, müssen wir zunächst die Eins als Linearkombination von 110 und 13 darstellen. Hier bietet sich keine offensichtliche Lösung an, also verwenden wir den erweiterten EUKLIDischen Algorithmus:

$$110 : 13 = 8 \text{ Rest } 6 \Longrightarrow 6 = 110 - 8 \cdot 13$$

$$13 : 6 = 2 \text{ Rest } 1 \Longrightarrow 1 = 13 - 2 \cdot 6 = 17 \cdot 13 - 2 \cdot 110$$

Also ist $17 \cdot 13 = 221 \equiv 1 \mod 110$ und $\equiv 0 \mod 13$; genauso ist $-2 \cdot 110 = 220 \equiv 1 \mod 13$ und $\equiv 9 \mod 110$. Eine ganzzahlige Lösung unseres Problems ist somit

$$75 \cdot 221 - 6 \cdot 220 = 15\,255\,.$$

Die allgemeine Lösung ist

$$15\,255 + k \cdot 110 \cdot 13 = 15\,255 + 1\,430k \quad \text{mit} \quad k \in \mathbb{Z}\,.$$

Da $15\,255 : 1\,430 = 10$ Rest 955 ist, hatte der General also 955 Soldaten vor sich stehen.

Alternativ läßt sich die Lösung eines Systems aus r Kongruenzen auch in einer geschlossenen Form darstellen allerdings um den Preis einer n-maligen statt $(n-1)$-maligen Anwendung des EUKLIDischen Algorithmus und größeren Zahlen schon von Beginn an: Um das System

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

zu lösen, berechnen wir zunächst für jedes i das Produkt

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

der sämtlichen übrigen m_j und bestimmen dazu ganze Zahlen α_i, β_i, für die gilt $\alpha_i m_i + \beta_i \widehat{m}_i = 1$ Dann ist

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

Natürlich wird x hier – wie auch bei den obigen Formel – oft größer sein als das Produkt der m_i; um die kleinste Lösung zu finden, müssen wir also noch modulo diesem Produkt reduzieren.

Im obigen Beispiel wäre

$$\begin{array}{lll} m_1 = 10 & \widehat{m}_1 = 11 \cdot 13 = 143 & 1 = 43 \cdot 10 - 3 \cdot 143 \\ m_2 = 11 & \widehat{m}_2 = 10 \cdot 13 = 130 & 1 = -59 \cdot 11 + 5 \cdot 130 \\ m_3 = 13 & \widehat{m}_3 = 10 \cdot 11 = 110 & 1 = 17 \cdot 13 - 2 \cdot 110 \,, \end{array}$$

also

$$x = -3 \cdot 143 \cdot 5 + 5 \cdot 130 \cdot 9 - 2 \cdot 110 \cdot 6 = -2145 + 5850 - 1320 = 2385 \,.$$

Modulo $10 \cdot 11 \cdot 13$ erhalten wir natürlich auch hier wieder 955.

Damit kennen wir nun auch zwei konstruktive Beweise des chinesischen Restesatzes und wissen, wie man Systeme von Kongruenzen mit Hilfe des erweiterten EUKLIDischen Algorithmus lösen kann.

§8: Prime Restklassen

Wie wir gesehen haben, können wir auch in $\mathbb{Z}/m$ im allgemeinen nicht dividieren. Allerdings ist Division doch sehr viel häufiger möglich als in den ganzen Zahlen. Dies wollen wir als nächstes genauer untersuchen:

Lemma: Zu zwei gegebenen natürlichen Zahlen a, m gibt es genau dann ein $x \in \mathbb{N}$, so daß $ax \equiv 1 \bmod m$, wenn ggT$(a, m) = 1$ ist.

Beweis: Wenn es ein solches x gibt, gibt es dazu ein $y \in \mathbb{N}$, so daß $ax = 1 + my$, d.h. $1 = ax - my$. Damit muß jeder gemeinsame Teiler von a und m Teiler der Eins sein, a und m sind also teilerfremd.

Sind umgekehrt a und m teilerfremd, so gibt es nach dem erweiterten EUKLIDischen Algorithmus $x, y \in \mathbb{Z}$ mit $ax + my = 1$. Durch (gegebenenfalls mehrfache) Addition der Gleichung $am - ma = 0$ läßt sich nötigenfalls erreichen, daß a positiv wird, und offensichtlich ist $ax \equiv 1 \bmod m$. ∎

Definition: Ein Element $a \in \mathbb{Z}/m$ heißt prime Restklasse, wenn ggT$(a, m) = 1$ ist.

Nach dem gerade bewiesenen Lemma gibt es somit zu jeder primen Restklasse a ein $x \in \mathbb{Z}/m$, so daß dort $ax = 1$ ist. Damit ist das folgende Lemma nicht verwunderlich:

Lemma: Die primen Restklassen aus $\mathbb{Z}/m$ bilden bezüglich der Multiplikation eine Gruppe.

Beweis: Wir müssen uns zunächst überlegen, daß das Produkt zweier primer Restklassen wieder eine prime Restklasse ist. Sind $a, b \in \mathbb{Z}/m$ beide teilerfremd zu m, so auch ab, denn wäre p ein gemeinsamer Primteiler von ab und m, so wäre p als Primzahl auch Teiler von a oder b, also gemeinsamer Teiler von a und m oder von b und m. Die Eins ist natürlich eine prime Restklasse, und auch die Existenz von Inversen ist kein Problem: Nach dem vorigen Lemma gibt es ein $x \in \mathbb{Z}$, so daß $ax \equiv 1 \bmod m$ ist, und die andere Richtung dieses Lemmas zeigt, daß auch x mod m eine prime Restklasse ist. Das Assoziativgesetz der Multiplikation gilt für alle Elemente von $\mathbb{Z}/m$, erst recht also für die primen Restklassen. ∎

Definition: Die Gruppe $(\mathbb{Z}/m)^\times$ der primen Restklassen heißt *prime Restklassengruppe,* ihre Ordnung wird mit $\varphi(m)$ bezeichnet. $\varphi : \mathbb{N} \to \mathbb{N}$ heißt EULERsche φ-Funktion.

LEONHARD EULER (1707–1783) wurde in Basel geboren und ging auch dort zur Schule und, im Alter von 14 Jahren, zur Universität. Dort legte er zwei Jahre später die Magisterprüfung in Philosophie ab und begann mit dem Studium der Theologie; daneben hatte er sich seit Beginn seines Studium unter Anleitung von JOHANN BERNOULLI mit Mathematik beschäftigt. 1726 beendete er sein Studium in Basel und bekam eine Stelle an der Petersburger Akademie der Wissenschaften, die er 1727 antrat. Auf Einladung FRIEDRICHS DES GROSSEN wechselte er 1741 an die preußische Akademie der Wissenschaften; nachdem sich das Verhältnis zwischen den beiden dramatisch verschlechtert hatte, kehrte er 1766 nach St. Petersburg zurück. Im gleichen Jahr erblindete er vollständig; trotzdem schrieb er rund die Hälfte seiner zahlreichen Arbeiten (Seine gesammelten Abhandlungen umfassen 73 Bände) danach. Sie enthalten bedeutende Beiträge zu zahlreichen Teilgebieten der Mathematik, Physik, Astronomie und Kartographie.

Lemma: *a)* Für zwei zueinander teilerfremde Zahlen $n, m \in \mathbb{N}$ ist $\varphi(nm) = \varphi(n)\varphi(m)$.
b) Für $m = \prod_{i=1}^{r} p_i^{e_i}$ ist $\varphi(m) = \prod_{i=1}^{r} \left(p_i^{e_i-1}(p_i - 1)\right)$.

Beweis: a) Eine Zahl a ist genau dann teilerfremd zum Produkt nm, wenn a mod n teilerfremd zu n und a mod m teilerfremd zu m ist. Da nach dem chinesischen Restesatz $\mathbb{Z}/nm \cong \mathbb{Z}/n \times \mathbb{Z}/m$ ist, ist daher auch $(\mathbb{Z}/nm)^\times \cong (\mathbb{Z}/n)^\times \times (\mathbb{Z}/m)^\times$.

b) Wegen *a)* genügt es, dies für Primzahlpotenzen p^e zu beweisen. Eine Zahl a ist genau dann teilerfremd zu p^e, wenn sie kein Vielfaches von p ist. Unter den Zahlen von 1 bis p^e gibt es genau p^{e-1} Vielfache von p, also ist $\varphi(p^e) = p^e - p^{e-1} = p^{e-1}(p-1)$. ∎

Korollar: $\mathbb{Z}/m$ ist genau dann ein Körper, wenn m eine Primzahl ist.

Beweis: Das einzige, was $\mathbb{Z}/m$ zu einem Körper eventuell fehlt, ist die Existenz von multiplikativen Inversen für alle von null verschiedenen Elemente. Dies ist offenbar äquivalent zur Formel $\varphi(m) = m - 1$, und die gilt nach dem Lemma genau dann, wenn m prim ist. ∎

Der Körper $\mathbb{Z}/p$ mit p Elementen wird üblicherweise mit $\mathbb{F}_p$ bezeichnet; die zugehörige prime Restklassengruppe $(\mathbb{Z}/p)^\times = \mathbb{F}_p \setminus \{0\}$ entsprechend als $\mathbb{F}_p^\times$. Dabei steht das „$\mathbb{F}$“ für *finit*. Im Englischen werden endliche Körper gelegentlich auch als *Galois fields* bezeichnet, so daß man hier auch die Abkürzung GF(p) sieht. *Field* ist das englische Wort für Körper; das gelegentlich in Informatikbüchern zu lesende Wort *Galoisfeld* ist also ein Übersetzungsfehler.

Wir wollen uns als nächstes überlegen, daß die multiplikative Gruppe dieses Körpers aus den Potenzen eines einzigen Elements besteht. Dazu brauchen wir zunächst noch ein Lemma aus der Gruppentheorie:

Definition: Die Ordnung eines Elements a einer (multiplikativ geschriebenen) Gruppe G ist die kleinste natürliche Zahl r, für die a^r gleich dem Einselement ist. Falls es keine solche Zahl gibt, sagen wir, a habe unendliche Ordnung.

Lemma (LAGRANGE): In einer endlichen Gruppe teilt die Ordnung eines jeden Elements die Gruppenordnung.

Beweis: Die Potenzen des Elements a bilden zusammen mit der Eins eine Untergruppe H von G, deren Elementanzahl gerade die Ordnung r von H ist. Wir führen auf G eine Äquivalenzrelation ein durch die Vorschrift $g \sim h$, falls gh^{-1} in H liegt. Offensichtlich besteht die Äquivalenzklasse eines jeden Elements $g \in G$ aus genau r Elementen, nämlich $g, ga, \ldots, ga^{r-1}$. Da G die Vereinigung aller Äquivalenzklassen ist, muß die Gruppenordnung somit ein Vielfaches von r sein. ∎

JOSEPH-LOUIS LAGRANGE (1736–1813) wurde als GIUSEPPE LODOVICO LAGRANGIA in Turin geboren und studierte dort zunächst Latein. Erst eine alte Arbeit von HALLEY über algebraische Methoden in der Optik weckte sein Interesse an der Mathematik, woraus ein ausgedehnter Briefwechsel mit EULER entstand. In einem Brief vom 12. August 1755 berichtete er diesem unter anderem über seine Methode zur Berechnung von Maxima und Minima; 1756 wurde er, auf EULERS Vorschlag, Mitglied der Berliner Akademie; zehn Jahre später zog er nach Berlin und wurde dort EULERS Nachfolger als mathematischer Direktor der

Akademie. 1787 wechselte er an die Pariser Académie des Sciences, wo er bis zu seinem Tod blieb und unter anderem an der Einführung des metrischen Systems beteiligt war. Seine Arbeiten umspannen weite Teile der Analysis, Algebra und Geometrie.

Korollar: Für zwei zueinander teilfremde Zahlen a, m ist

$$a^{\varphi(m)} \equiv 1 \bmod m\,.$$

Beweis: Klar, denn $\varphi(m)$ ist die Ordnung der primen Restklassengruppe modulo m. ∎

Für eine Primzahl $N = p$ bezeichnet man diese Aussage auch als den *kleinen Satz von* FERMAT:

Satz (FERMAT): Für jede nicht durch die Primzahl p teilbare natürliche Zahl a ist $a^{p-1} \equiv 1 \bmod p$. Für alle $a \in \mathbb{Z}$ ist $a^p \equiv a \bmod p$.

Beweis: Die erste Aussage ist klar, da $\varphi(p) = p - 1$ ist. Für die zweite müssen wir nur noch beachten, daß für durch p teilbare Zahlen a sowohl a^p als auch a kongruent null modulo p sind. ∎

Der französische Mathematiker PIERRE DE FERMAT (1601–1665) wurde in Beaumont-de-Lomagne im Departement Tarn et Garonne geboren. Bekannt ist er heutzutage vor allem für seine 1994 von ANDREW WILES bewiesene Vermutung, wonach die Gleichung $x^n + y^n = z^n$ für $n \geq 3$ keine ganzzahlige Lösung mit $xyz \neq 0$ hat. Dieser „große" Satzes von FERMAT, von dem FERMAT lediglich in einer Randnotitz behauptete, daß er ihn beweisen könne, erklärt den Namen der obigen Aussage. Obwohl FERMAT sich sein Leben lang sehr mit Mathematik beschäftigte und wesentliche Beiträge zur Zahlentheorie, Wahrscheinlichkeitstheorie und Analysis lieferte, war er hauptberuflich Jurist.

Satz: Die multiplikative Gruppe eines endlichen Körpers ist zyklisch.

Beweis: Da die multiplikative Gruppe eines Körpers mit q Elementen aus allen Körperelementen außer der Null besteht, hat sie die Ordnung $q-1$, d.h. nach LAGRANGE ist die Ordnung eines jeden Elements ein Teiler

von $q - 1$. Wir müssen zeigen, daß es mindestens ein Element gibt, dessen Ordnung *genau* $q - 1$ ist.

Für jeden Primteiler p_i von $q - 1$ hat die Polynomgleichung

$$x^{(q-1)/p_i} = 1$$

höchstens $(q - 1)/p_i$ Lösungen im Körper; es gibt also zu jedem p_i ein Körperelement a_i mit $a_i^{(q-1)/p_i} \neq 1$.

q_i sei die größte Potenz von p_i, die $q - 1$ teilt, und $g_i = a_i^{(q-1)/q_i}$ die $(q - 1)/q_i$-te Potenz von a_i. Dann ist

$$g_i^{q_i} = a_i^{q-1} = 1 \quad \text{und} \quad g_i^{\frac{q_i}{p_i}} = a_i^{\frac{q-1}{p_i}} \neq 1\,;$$

g_i hat also die Ordnung q_i. Da die verschiedenen q_i Potenzen verschiedener Primzahlen p_i sind, hat daher das Produkt g aller g_i das Produkt aller q_i als Ordnung, also $q - 1$. Damit ist die multiplikative Gruppe des Körpers zyklisch. ∎

Definition: Ein Element g eines endlichen Körpers k heißt *primitive Wurzel,* wenn es die zyklische Gruppe $k^\times$ erzeugt.

Selbst im Fall der Körper $\mathbb{F}_p$ gibt es keine Formel, mit der man eine solche primitive Wurzel explizit in Abhängigkeit von p angeben kann. Üblicherweise wählt man zufällig ein Element aus und testet, ob es die Ordnung $p - 1$ hat. Die Wahrscheinlichkeit dafür ist offenbar $\varphi(p - 1) : (p - 1)$, was für die meisten Werte von p recht gut ist. Der Test, ob die Ordnung gleich $p - 1$ ist, läßt sich allerdings nur dann effizient durchführen, wenn die Primteiler p_i von $p - 1$ bekannt sind, denn dann kann man einfach testen, ob alle Potenzen mit den Exponenten $(p - 1)/p_i$ von eins verschieden sind. Für große Werte von p, wie sie in der Kryptographie benötigt werden, kann dies ein Problem sein, so daß man hier im allgemeinen von faktorisierten Zahlen r ausgeht und dann testet, ob $r + 1$ prim ist. Im Kapitel über Primzahlen werden wir geeignete Tests kennenlernen.

Kapitel 2
Anwendungen in der Kryptographie

§1: New directions in cryptography

In der klassischen Kryptographie verläuft die Entschlüsselung entweder genauso oder zumindest sehr ähnlich wie die Verschlüsselung; insbesondere kann jeder, der eine Nachricht verschlüsseln kann, jede andere entsprechend verschlüsselte Nachricht auch entschlüsseln. Man bezeichnet diese Verfahren daher als *symmetrisch.*

Der Nachteil eines solchen Verfahrens besteht darin, daß in einem Netzwerk jeder Teilnehmer mit jedem anderen einen Schlüssel vereinbaren muß. In militärischen Netzen war dies üblicherweise so geregelt, daß das gesamte Netz denselben Schlüssel benutzte, der in einem Codebuch für jeden Tag im voraus festgelegt war; in kommerziellen Netzen wie beispielsweise einem Mobilfunknetz ist so etwas natürlich unmöglich.

1976 publizierten MARTIN HELLMAN, damals Assistenzprofessor in Stanford, und sein Forschungsassistent WHITFIELD DIFFIE eine Arbeit mit dem Titel *New directions in cryptography* (IEEE Trans. Inform. Theory **22**, 644–654), in der sie vorschlugen, den Vorgang der *Ver*schlüsselung und den der *Ent*schlüsselung völlig voneinander zu trennen: Es sei schließlich nicht notwendig, daß der Sender einer verschlüsselten Nachricht auch in der Lage sei, diese zu *ent*schlüsseln.

Der Vorteil eines solchen *asymmetrischen* Verfahrens wäre, daß jeder potentielle Empfänger nur einen einzigen Schlüssel bräuchte und dennoch sicher sein könnte, daß nur er selbst seine Post entschlüsseln kann. Der Schlüssel müßte nicht einmal geheimgehalten werden, da es ja

(meistens) nichts schadet, wenn jedermann Nachrichten *ver*schlüsseln kann. In einem Netzwerk mit n Teilnehmern bräuchte man also nur n Schlüssel, um es jedem Teilnehmer zu erlauben, mit jedem anderen sicher zu kommunizieren. Die Schlüssel könnten sogar in einem öffentlichen Verzeichnis stehen. Bei einem symmetrischen Kryptosystem wäre der gleiche Zweck nur erreichbar mit $\frac{1}{2}n(n-1)$ Schlüsseln, die auf einem sicheren Weg wie etwa bei einem persönliches Treffen oder durch vertrauenswürdige Boten ausgetauscht werden müßten.

BAILEY WHITFIELD DIFFIE wurde 1944 geboren. Erst im Alter von zehn Jahren lernte er lesen; im gleichen Jahr hielt eine Lehrerin an seiner New Yorker Grundschule einen Vortrag über Chiffren. Er ließ sich von seinem Vater alle verfügbare Literatur darüber besorgen, entschied sich dann 1961 aber doch für ein Mathematikstudium am MIT. Um einer Einberufung zu entgehen, arbeitete er nach seinem Bachelor bei Mitre; später, nachdem sein Interesse an der Kryptographie wieder erwacht war, kam er zu Martin Hellman nach Stanford, der ihn als Forschungsassistent einstellte. Seit 1991 arbeitet er als *chief security officer* bei Sun Microsystems, seit 2010 ist er bei ICANN für Sicherheit zuständig. `labs.oracle.com/people/mybio.php?uid=18607` .

MARTIN HELLMAN wurde 1945 in New York geboren. Er studierte Elektrotechnik zunächst bis zum Bachelor an der dortigen Universität; für Master und Promotion studierte er in Stanford. Nach kurzen Zwischenaufenthalten am Watson Research Center der IBM und am MIT wurde er 1971 Professor an der Stanford University. Seit 1996 ist er emeritiert, gibt aber immer noch Kurse, mit denen er Schüler für mathematische Probleme interessieren will. Seine home page findet man unter `http://www-ee.stanford.edu/~hellman/` .

DIFFIE und HELLMAN machten nur sehr vage Andeutungen, wie so ein System mit öffentlichen Schlüsseln aussehen könnte. Es ist zunächst einmal klar, daß ein solches System keinerlei Sicherheit gegen einen Gegner mit unbeschränkter Rechenkraft (In der Kryptographie spricht man von einem BAYESschen Gegner) bieten kann, denn die Verschlüsselungsfunktion ist eine bijektive Abbildung zwischen endlichen Mengen,

und jeder, der die Funktion kennt, kann zumindest im Prinzip auch ihre Umkehrfunktion berechnen.

Wer im Gegensatz zum BAYESschen Gegner nur über begrenzte Ressourcen verfügt, kann diese Berechnung allerdings möglicherweise nicht mit realistischem Aufwand durchführen, und nur darauf beruht die Sicherheit eines Kryptosystems mit öffentlichen Schlüsseln. Da Computer immer leistungsfähiger werden und auch immer neue und bessere Algorithmen gefunden werden, sind solche Systeme insbesondere nicht für die Ewigkeit gedacht, sondern nur für die einigermaßen überschaubare Zukunft. Derzeit geht man in der Kryptologie davon aus, daß kein Gegner 2^{128} oder gar mehr Entschlüsselungsversuche durchführen kann und bezeichnet ein System daher als sicher, wenn trotz intensiver Untersuchung kein Angriff bekannt ist, der mit geringerem Aufwand und einer nicht vernachlässigbaren Erfolgsaussicht Entschlüsselungen liefert. Man spricht dann von 128-Bit-Sicherheit, wobei diese Sicherheit natürlich nicht bewiesen ist.

Einige Stellen, darunter auch das Bundesamt für Sicherheit in der Informationstechnik, halten derzeit noch eine 100-Bit-Sicherheit für ausreichend. Dieser Auffassung schließen sich unsere Banken an: Bei den derzeit in Deutschland üblichen Bankkarten für Geldautomaten ist die Geheimzahl mit dem sogenannten Triple-DES geschützt, dessen Sicherheit bei knapp 112 Bit liegt.

DIFFIE und HELLMAN bezeichnen eine Funktion, deren Umkehrfunktion nicht mit vertretbarem Aufwand (im obigen Sinne) berechnet werden kann, als *Einwegfunktion* und wollen solche Funktionen zur Verschlüsselung verwenden. Das allein führt allerdings noch nicht zu einem praktikablen Kryptosystem, denn bei einer echten Einwegfunktion ist es auch für den legitimen Empfänger nicht möglich, seinen Posteingang zu entschlüsseln. DIFFIE und HELLMAN schlagen deshalb eine Einwegfunktion mit *Falltür* vor, wobei der legitime Empfänger zusätzlich zu seinem öffentlichen Schlüssel noch über einen geheimen Schlüssel verfügt, mit dem er (und nur er) diese Falltür öffnen kann.

Natürlich hängt alles davon ab, ob es solche Einwegfunktionen mit Falltür wirklich gibt. DIFFIE und HELLMAN gaben keine an, und es gab

unter den Experten einige Skepsis bezüglich der Möglichkeit, solche Funktionen zu finden.

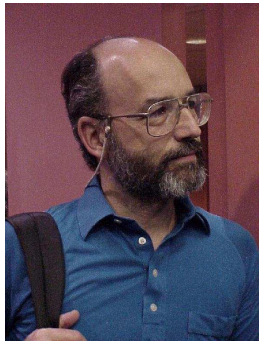
Tatsächlich aber gab es damals bereits Systeme, die auf solchen Funktionen beruhten, auch wenn sie nicht in der offenen Literatur dokumentiert waren: Die britische *Communications-Electronics Security Group* (CESG) hatte bereits Ende der sechziger Jahre damit begonnen, nach entsprechenden Verfahren zu suchen, um die Probleme des Militärs mit dem Schlüsselmanagement zu lösen, aufbauend auf (impraktikablen) Ansätzen von AT&T zur Sprachverschlüsselung während des zweiten Weltkriegs. Die CESG sprach nicht von Kryptographie mit öffentlichen Schlüsseln, sondern von *nichtgeheimer Verschlüsselung,* aber das Prinzip war das gleiche.

Erste Ideen dazu sind in einer auf Januar 1970 datierten Arbeit von JAMES H. ELLIS zu finden, ein praktikables System in einer auf den 20. November 1973 datierten Arbeit von CLIFF C. COCKS. Wie im Milieu üblich, gelangte nichts über diese Arbeiten an die Öffentlichkeit; erst 1997 veröffentlichten die *Government Communications Headquarters* (GCHQ), zu denen CESG gehört, einige Arbeiten aus der damaligen Zeit; eine Zeitlang waren sie auch auf dem Server `http://www.cesg.gov.uk/` zu finden, wo sie allerdings inzwischen anscheinend wieder verschwunden sind.

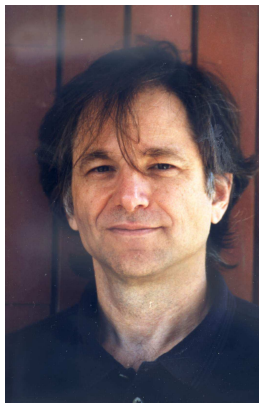
Im akademischen Bereich gab es ein Jahr nach Erscheinen der Arbeit von DIFFIE und HELLMAN das erste Kryptosystem mit öffentlichen Schlüsseln: Drei Wissenschaftler am Massachusetts Institute of Technology fanden nach rund vierzig erfolglosen Ansätzen 1977 schließlich jenes System, das heute nach ihren Anfangsbuchstaben mit RSA bezeichnet wird: RON RIVEST, ADI SHAMIR und LEN ADLEMAN.

RIVEST, SHAMIR und ADLEMAN gründeten eine Firma namens RSA Computer Security Inc., die 1983 das RSA-Verfahren patentieren ließ und auch nach Auslaufen dieses Patents im September 2000 noch erfolgreich im Kryptobereich tätig ist. 2002 erhielten RIVEST, SHAMIR und ADLEMAN für die Entdeckung des RSA-Systems den TURING-Preis der *Association for Computing Machinery* ACM, ein jährlich vergebener Preis, der als eine der höchsten Auszeichnungen der Informatik gilt.

RONALD LINN RIVEST wurde 1947 in Schenectady im US-Bundesstaat New York geboren. Er studierte zunächst Mathematik an der Yale University, wo er 1969 seinen Bachelor bekam; danach studierte er in Stanford Informatik. Nach seiner Promotion 1974 wurde er Assistenzprofessor am Massachusetts Institute of Technology, wo er heute einen Lehrstuhl hat. Er arbeitet immer noch auf dem Gebiet der Kryptographie und entwickelte eine ganze Reihe weiterer Verfahren, auch symmetrische Verschlüsselungsalgorithmen und Hashverfahren. Er ist Koautor eines Lehrbuchs über Algorithmen. Seine home page ist `http://theory.lcs.mit.edu/~rivest/` .



ADI SHAMIR wurde 1952 in Tel Aviv geboren. Er studierte zunächst Mathematik an der dortigen Universität; nach seinem Bachelor wechselte er ans Weizmann Institut, wo er 1975 seinen Master und 1977 die Promotion in Informatik erhielt. Nach einem Jahr als Postdoc an der Universität Warwick und drei Jahren am MIT kehrte er ans Weizmann Institut zurück, wo er bis heute Professor ist. Außer für RSA ist er bekannt sowohl für die Entwicklung weiterer Kryptoverfahren als auch für erfolgreiche Angriffe gegen Kryptoverfahren. Er schlug auch einen optischen Spezialrechner zur Faktorisierung großer Zahlen vor. Seine home page ist erreichbar unter `http://www.wisdom.weizmann.ac.il/math/profile/scientists/shamir-profile.html`

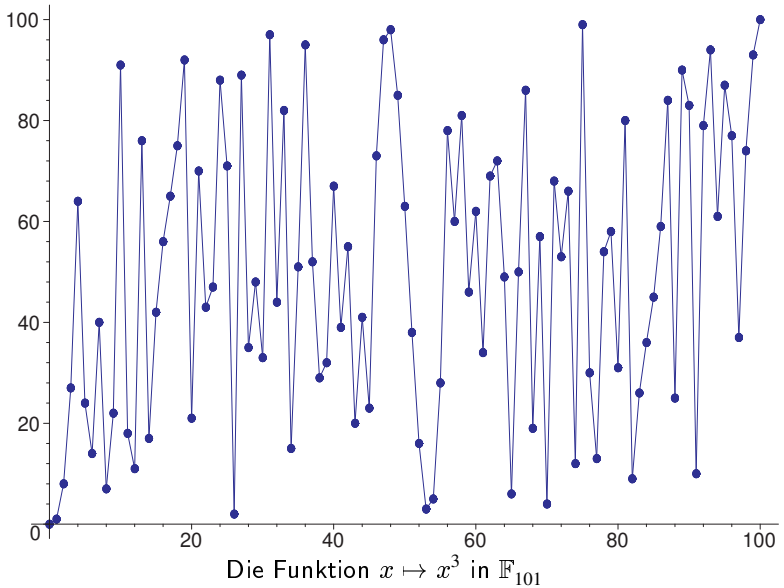


LEONARD ADLEMAN wurde 1945 in San Francisco geboren. Er studierte in Berkeley, wo er 1968 einen BS in Mathematik und 1976 einen PhD in Informatik erhielt. Thema seiner Dissertation waren zahlentheoretische Algorithmen und ihre Komplexität. Von 1976 bis 1980 war er an der mathematischen Fakultät des MIT; seit 1980 arbeitet er an der University of Southern California in Los Angeles. Seine Arbeiten beschäftigen sich mit Zahlentheorie, Kryptographie und Molekularbiologie. Er führte nicht nur 1994 die erste Berechnung mit einem „DNS-Computer“ durch, sondern arbeitete auch auf dem Gebiet der Aidsforschung. Heute hat er einen Lehrstuhl für Informatik und Molekularbiologie. `http://www.usc.edu/dept/molecular-science/fm-adleman.htm`

RSA ist übrigens identisch mit dem von COCKS vorgeschlagenen System, so daß einige Historiker auch Zweifel an den Behauptungen der GCHQ haben. Die Beschreibung durch RIVEST, SHAMIR und ADLEMAN erschien 1978 unter dem Titel *A method for obtaining digital signatures and public-key cryptosystems* in Comm. ACM **21**, 120–126.

§2: Das RSA-Verfahren

Für eine natürliche Zahl e ist die reelle Funktion $x \mapsto x^e$ für positive x monoton ansteigend und bijektiv; ihre Umkehrfunktion $x \mapsto \sqrt[e]{x}$ läßt sich mit etwa demselben Aufwand berechnen wie die Funktion selbst. Betrachten wir $x \mapsto x^e$ allerdings als Funktion von $\mathbb{Z}/N \to \mathbb{Z}/N$, so erhalten wir einen sehr chaotisch aussehenden Graphen und können uns daher Hoffnungen machen, daß diese Funktion vielleicht als Grundlage einer kryptographischen Verschlüsselung brauchbar sein könnte.



Dazu muß sie natürlich zunächst einmal injektiv sein. Da die Ordnung eines Elements von $(\mathbb{Z}/N\mathbb{Z})^\times$ Teiler von $\varphi(N)$ ist, muß insbesondere

e teilerfremd zu $\varphi(N)$ sein. Dann lassen sich mit dem erweiterten EUKLIDischen Algorithmus Zahlen $d, k \in \mathbb{N}$ finden, so daß $de - k\varphi(N) = 1$ ist, d.h. für jedes zu N teilerfremde x ist

$$(x^e)^d = x^{ed} = x^{1+k\varphi(N)} \equiv x \bmod N\,.$$

Somit sind die Funktionen

$$\begin{cases} (\mathbb{Z}/N\mathbb{Z})^\times \to (\mathbb{Z}/N\mathbb{Z})^\times \\ x \mapsto x^e \end{cases} \quad \text{und} \quad \begin{cases} (\mathbb{Z}/N\mathbb{Z})^\times \to (\mathbb{Z}/N\mathbb{Z})^\times \\ x \mapsto x^d \end{cases}$$

zueinander invers.

Die Beschränkung auf prime Restklassen ist für kryptographische Anwendungen ungünstig: Am einfachsten wäre es, wenn wir jede Bitfolge, deren Länge kleiner ist als die der Binärdarstellung von N, als Zahl zwischen 0 und $N-1$ auffassen, verschlüsseln und übertragen könnten. Der Empfänger könnte dann die Zahl entschlüsseln, als Bitfolge hinschreiben und daraus die Nachricht rekonstruieren. Zum Glück ist das zumindest für *quadratfreie* Zahlen N, d.h. Zahlen, die durch kein Primzahlquadrat teilbar sind, auch möglich:

Satz: Für eine quadratfreie natürliche Zahl N sind die beiden Funktionen

$$\begin{cases} \mathbb{Z}/N\mathbb{Z} \to \mathbb{Z}/N\mathbb{Z} \\ x \mapsto x^e \end{cases} \quad \text{und} \quad \begin{cases} \mathbb{Z}/N\mathbb{Z} \to \mathbb{Z}/N\mathbb{Z} \\ x \mapsto x^d \end{cases}$$

bijektiv und invers zueinander.

Beweis: Als quadratfreie Zahl ist N ein Produkt verschiedener Primzahlen p_i, und $\varphi(N)$ ist das Produkt der zugehörigen $\varphi(p_i) = p_i - 1$. Somit ist auch $ed \equiv 1 \bmod \varphi(p_i)$ für alle i, und nach dem chinesischen Restesatz genügt es, wenn wir den Satz für die einzelnen p_i beweisen. Ist $N = p$ prim, so ist die Null das einzige Element von $\mathbb{Z}/p$, das nicht in $(\mathbb{Z}/p)^*$ liegt, und sie wird von beiden Funktionen auf sich selbst abgebildet. ∎

Jeder, der e und N kennt, kann auch d mit berechnen, allerdings muß er dazu als erstes $\varphi(N)$ bestimmen. Nach der Formel aus Kapitel 1, §7

ist das möglich, wenn er die Primfaktorzerlegung von N kennt. Einfachere alternative Verfahren sind nicht bekannt, und wie wir im Kapitel über Faktorisierungsalgorithmen sehen werden, ist diese Zerlegung mit heutigen Mitteln für ein Produkt zweier gut gewählter Primzahlen nicht mehr mit realistischem Aufwand möglich, wenn dieses Produkt mehr als etwa 250 Dezimalstellen hat. Natürlich wird diese Schranke im Laufe der Jahrzehnte ansteigen, und wahrscheinlich können die Geheimdienste schon heute etwas mehr als der Rest der Welt. Es ist aber unwahrscheinlich, daß bei einem schon seit Jahrhunderten untersuchten Problem wie der Faktorisierung ganzer Zahlen ausgerechnet einem Geheimdienst ein Durchbruch gelingen sollte, von dem der Rest der Welt nichts bemerkt. Die Effekte einer leistungsfähigeren Hardware lassen sich durch großzügige Sicherheitszuschläge kompensieren.

Für eine Primzahl $N = p$ kann natürlich jeder ganz einfach $\varphi(p) = p - 1$ berechnen; ist $N = pq$ dagegen das Produkt zweier Primzahlen, so ist die Bestimmung von

$$\varphi(N) = (p-1)(q-1) = N - (p+q) + 1$$

äquivalent zur Kenntnis der Faktorisierung: Kennt man nämlich das Produkt $N = pq$ sowie die Summe $N + 1 - \varphi(N) = p + q$ der beiden Primzahlen, so kann man sie einfach berechnen als Lösungen der quadratischen Gleichung

$$(x-p)(x-q) = x^2 - \big(N + 1 - \varphi(N)\big)\,x + N = 0\,.$$

Auch bei Produkten von mehr als drei Primzahlen ist keine Methode zur Berechnung der EULERschen φ-Funktion bekannt, die effizienter wäre als der Umweg über die Faktorisierung, allerdings wird die Faktorisierung bei konstanter Größenordnung von N tendenziell einfacher, wenn die Anzahl der Faktoren steigt, da wir es dann zumindest teilweise mit kleineren Faktoren zu tun haben.

Zur praktischen Durchführung des RSA-Verfahrens wählt sich daher jeder Teilnehmer zwei verschiedene Primzahlen p, q, die unbedingt geheim gehalten werden müssen, und eine natürliche Zahl e, die keinen gemeinsamen Teiler mit $(p-1)(q-1)$ hat. Die Zahlen $N = pq$ und e sind sein öffentlicher Schlüssel, der beispielsweise in einem Verzeichnis publiziert werden kann.

Des weiteren berechnet er zu $\varphi(N) = (p-1)(q-1)$ nach dem EUKLIDischen Algorithmus eine natürliche Zahl d, so daß $de + k\varphi(N) = 1$ ist für ein gewisses $k \in \mathbb{Z}$. Diese Zahl d ist sein geheimer Schlüssel; da

$$\left(a^e\right)^d \equiv a \bmod N$$

für alle a, läßt sich damit die Entschlüsselung rückgängig machen.

Jeder, der den öffentlichen Schlüssel (N, e) kennt, kann damit Nachrichten verschlüsseln: Er bricht die Nachricht auf in Blöcke, die durch ganze Zahlen zwischen 0 und $N-1$ dargestellt werden können, berechnet für jeden so dargestellten Block a den Chiffretext $b = a^e \bmod N$ und schickt diesen an den Inhaber des geheimen Schlüssels. Dieser berechnet $b^d \bmod N = a^{ed} \bmod N = a$, und da er dazu seinen geheimen Schlüssel braucht, kann dies niemand außer ihm auf diese Weise berechnen.

Es wäre allerdings verfrüht, daraus schon auf die Sicherheit des RSA-Verfahrens zu schließen: Der Gegner, vor dem eine Nachricht geschützt werden soll, ist schließlich frei in der Wahl seiner Mittel und kann sich auch andere Formen der Attacke überlegen. Soweit entsprechende Strategien bekannt sind, muß man sich auch dagegen schützen. In der bislang dargestellten sogenannten „Lehrbuchversion“ bietet RSA, wie wir im Verlauf der Vorlesung noch mehrfach sehen werden, eine ganze Reihe von Angriffsmöglichkeiten, die gelegentlich erheblich schneller ans Ziel führen als die Faktorisierung des Moduls N.

Beispielsweise wird in der Praxis oft der öffentliche Exponent $e = 3$ verwendet (was natürlich voraussetzt, daß die verwendeten Primzahlen kongruent zwei modulo drei sind), so daß zumindest die Verschlüsselung recht schnell geht. Außerdem läßt sich dann der private Exponent d sehr einfach bestimmen: Nach der allgemeinen Theorie gibt es Zahlen d, k, so daß $de - k\varphi(N) = 1$ ist. Durch Addition eines Vielfachen der Gleichung $\varphi(N)e - e\varphi(N) = 0$ läßt sich dabei erreichen, daß $1 \le k \le e-1$ ist. Für $e = 3$ kommen also nur $k = 1$ und $k = 2$ in Frage. Man muß daher nur $d_k = \left(1 + k\varphi(N)\right)/3$ für diese beiden Werte berechnen, und erhält d als das ganzzahlige der beiden Ergebnisse.

Bei so vielen Vorteilen muß es auch Nachteile geben, darunter einen ganz offensichtlichen: Ist nämlich die Nachricht x kleiner als die dritte Wurzel aus N, so ist $x^3 < N$, d.h. der Chiffretext ist einfach x^3.

Die Kubikwurzel aus dieser ganzen Zahl kann natürlich leicht gezogen werden.

Auch für größere Exponenten e sind kurze Nachrichten problematisch. Ein Grund liegt darin, daß die Verschlüsselungsfunktion mit der Multiplikation verträglich ist: Auch im Ring $\mathbb{Z}/NZ$ ist $(yz)^e = y^e \cdot z^e$.

Nehmen wir an, unsere Nachricht x habe höchstens 2ℓ Bit, sei also kleiner als $2^{2\ell}$. Dann gibt es eine nicht vernachlässigbare Chance, daß sich $x = yz$ als Produkt zweier Zahlen darstellen läßt, die beide kleiner als 2^ℓ (oder als eine etwas größere Schranke M) sind. Wir können für alle Zahlen y von Null bis M deren Verschlüsselung $y^e \bmod N$ berechnen und die Ergebnisse in einer Tabelle notieren. Um nun vom Chiffretext $c = x^e \bmod N$ auf x zurückzuschließen, berechnen wir für jedes dieser Ergebnisse in $\mathbb{Z}/N$ den Quotienten c/y^e. (Falls sich dieser Quotient nicht bilden läßt, ist y^e nicht teilerfremd zu $N = pq$, und wir erhalten sogar eine Faktorisierung von N; das ist aber für gut gewählte, große N extrem unwahrscheinlich.) Falls einer dieser Quotienten als Eintrag $z^e \bmod N$ in unserer Tabelle auftaucht, haben wir eine Relation der Form $c \equiv y^e \cdot z^e = (yz)^e \bmod N$ gefunden, und damit kennen wir $x = yz$.

Um diese Attacke zu verhindern, muß bei 128-Bit-Sicherheit $\ell \ge 128$ sein, die übermittelten Nachrichten müssen also mindestens 256 Bit lang sein. Auch dann sind wir allerdings noch nicht unbedingt auf der sicheren Seite, denn wenn nur wenige Nachrichten in Frage kommen, kann ein Gegner die einfach alle mit dem öffentlichen Schlüssel chiffrieren und schauen, wann das Ergebnis mit dem Chiffretext übereinstimmt. Beim praktischen Einsatz von RSA werden daher nie einfach die zu übermittelnden Nachrichten übertragen, sondern Blöcke eines vorher festgelegten Formats. Diese Formate sehen vor, daß alle unbenutzten Positionen mit Zufallsbits gefüllt werden und daß jeder Block mindestens 128 Zufallsbits enthält. Auf dem Niveau der 128-Bit-Sicherheit ist dann jede Entschlüsselung durch systematisches Probieren ausgeschlossen, denn Nachrichtenlängen größer 256 Bit sind bei den heute üblichen Parameterwerten ohnehin selbstverständlich.

Falls eine Nachricht an mehrere Empfänger geschickt wird, müssen

– vor allem bei kleinen Verschlüsselungsexponenten wie $e = 3$ – die Zufallsbits für jeden Empfänger neu erzeugt werden, denn wenn jedes Mal derselbe Block x verschlüsselt wird und dabei – wie dies häufig in der Praxis der Fall ist – stets mit drei potenziert wird, kennt ein Gegner anschließend x^3 mod N_i für die Moduln N_i der sämtlichen Empfänger, kann also nach dem chinesischen Restesatz x^3 modulo dem Produkt der N_i berechnen, und da dieses Produkt bei mindestens drei Empfängern größer ist als x^3, kennt er x^3 und damit auch x.

Bei der Wahl der Schlüsseldaten geht man stets aus vom öffentlichen Exponenten e und berechnet dann dazu nach dem EUKLIDischen Algorithmus den privaten Exponenten d. Dadurch ist praktisch sichergestellt, daß dieser in der Größenordnung von N liegen wird, und das muß auch so sein: Im Kapitel über Kettenbrüche werden wir sehen, daß N leicht faktorisiert werden kann, wenn d zu klein ist.

Bei der Berechnung von x^d mod N und x^e mod N darf man natürlich nicht erst x^d *bzw.* x^e berechnen und dann modulo N reduzieren: Schon für rund dreißigstellige Exponenten würde das zu Zwischenergebnissen führen, mit denen selbst die leistungsfähigsten heutigen Supercomputer nicht mehr fertig würden. Um die Länge der Zwischenergebnisse in Grenzen zu halten, muß nach jeder Multiplikation sofort modulo N reduziert werden.

Für große Exponenten e ist es auch nicht mehr möglich, die Potenz durch sukzessive Multiplikation mit x zu berechnen, die benötigten $d - 1$ Multiplikationen lägen ebenfalls weit jenseits der Rechenleistung selbst von Supercomputern.

Zum Glück gibt es eine erheblich effizientere Alternative: Um beispielsweise x^{32} zu berechnen brauchen wir keine 31 Multiplikationen, sondern erhalten das Ergebnis über die Formel

$$x^{32} = \left(\left(\left(\left(x^2\right)^2\right)^2\right)^2\right)^2$$

mit nur fünf Multiplikationen (genauer: Quadrierungen).

Entsprechend können wir für jede gerade Zahl $n = 2m$ die Potenz x^n als Quadrat von x^m berechnen. Für einen ungeraden Exponenten e ist

$e - 1$ gerade, wenn wir also x^e als Produkt von x und x^{e-1} berechnen, können wir zumindest im nächsten Schritt wieder die Formel für gerade Exponenten verwenden. Somit reichen pro Binärziffer des Exponenten ein bis zwei Multiplikationen; der Aufwand wächst also nur proportional zur Stellenzahl von e. Für den ebenfalls recht populären Verschlüsselungsexponenten $e = 2^{16} + 1 = 65537$ beispielsweise braucht man nur 17 Multiplikationen, nicht 65536.

§3: Weitere Anwendungen des RSA-Verfahrens

Im Gegensatz zu symmetrischen Kryptoverfahren endet die Nützlichkeit des RSA-Verfahrens nicht mit der bloßen Möglichkeit einer Verschlüsselung; es erlaubt noch eine ganze Reihe weiterer Anwendungen:

a) Identitätsnachweis

Hier geht es darum, in Zugangskontrollsystemen, vor Geldautomaten oder bei einer Bestellung im Internet die Identität einer Person zu beweisen: Mit RSA ist das beispielsweise dadurch möglich, daß nur der Inhaber des geheimen Schlüssels d zu einer gegebenen Zahl a eine Zahl b berechnen kann, für die $b^e \equiv a \bmod N$ ist. Letzteres wiederum kann jeder überprüfen, der den öffentlichen Schlüssel (N, e) kennt.

Falls also der jeweilige Gegenüber eine Zufallszahl a erzeugt und als Antwort das zugehörige b verlangt, kann er anhand eines öffentlichen Schlüsselverzeichnisses die Richtigkeit von b überprüfen und sich so von der Identität seines Partners überzeugen. Im Gegensatz zu Kreditkarteninformation oder Paßwörtern ist dieses Verfahren auch immun gegen Abhören: Falls jedesmal ein neues zufälliges a erzeugt wird, nützt ein einmal abgehörtes b nichts.

Grundsätzlich bräuchte man hier kein Kryptosystem mit öffentlichen Schlüsseln; in der Tat funktionierten die ersten Freund-/Feinderkennungssysteme für Flugzeuge zur Zeit des zweiten Weltkriegs nach diesem Prinzip, aber damals natürlich mit einem klassischen symmetrischen Kryptosystem, wobei alle Teilnehmer mit demselben Schlüssel arbeiteten. Der Vorteil eines asymmetrischen Systems besteht darin, daß sich keiner der Teilnehmer für einen anderen ausgeben kann, was

beispielsweise wichtig ist, wenn man sich gegenüber weniger vertrauenswürdigen Personen identifizieren muß.

Trotzdem ist das Verfahren in dieser Form nicht als Ersatz zur Übertragung von rechtlich bindender Information geeignet, da der Gegenüber anhand des öffentlichen Schlüssels jederzeit zu einer willkürlich gewählten Zahl b die Zahl $a = b^e \bmod N$ erzeugen kann um dann zu behaupten, er habe b als Antwort darauf empfangen. Daher kann der Inhaber des geheimen Schlüssels zwar seine Identität beweisen, aber sein Gegenüber kann später nicht beispielsweise vor Gericht beweisen, daß er dies (zum Beispiel bei einer Geldabhebung oder Bestellung) getan hat. Falls dies eventuell nötig werden könnte, ist das hier vorgestellte Verfahren also ungeeignet; es funktioniert nur zwischen Personen, die einander vertrauen können.

Eine mögliche Modifikation bestünde darin, daß man beispielsweise noch zusätzlich verlangt, daß die Zahl a eine spezielle Form hat, etwa daß die vordere Hälfte der Ziffernfolge identisch mit der hinteren Hälfte ist. Ohne Kenntnis von d hat man praktisch keine Chancen eine Zahl b zu finden, für die $b^e \bmod N$ eine solche Gestalt hat: Bei Zahlen mit $2r$ Ziffern liegt die Wahrscheinlichkeit dafür bei 10^{-r}.

b) Elektronische Unterschriften

Praktische Bedeutung hat vor allem eine andere Variante: die elektronische Unterschrift. Hier geht es darum, daß der Empfänger erstens davon überzeugt wird, daß eine Nachricht tatsächlich vom behaupteten Absender stammt, und daß er dies zweitens auch einem Dritten gegenüber beweisen kann. (In Deutschland sind solche elektronischen Unterschriften, sofern gewisse formale Voraussetzungen erfüllt sind, rechtsverbindlich.)

Um einen Nachrichtenblock a mit $0 \le a < N$ zu unterschreiben, berechnet der Inhaber des öffentlichen Schlüssels (N, e) mit seinem geheimen Schlüssel d die Zahl

$$b = a^d \bmod N$$

und sendet das Paar (a, b) an den Empfänger. Dieser überprüft, ob

$$b^e \equiv a \bmod N\,;$$

falls ja, akzeptiert er dies als unterschriebene Nachricht a. Da er ohne Kenntnis des geheimen Schlüssels d nicht in der Lage ist, den Block (a, b) zu erzeugen, kann er auch gegenüber einem Dritten beweisen, daß der Absender selbst die Nachricht a unterschrieben hat.

Für kurze Nachrichten ist dieses Verfahren in der vorgestellten Form praktikabel; in vielen Fällen kann man sogar auf die Übermittlung von a verzichten, da $b^e \bmod N$ für ein falsch berechnetes b mit an Sicherheit grenzender Wahrscheinlichkeit keine sinnvolle Nachricht ergibt.

Falls die übermittelte Nachricht geheimgehalten werden soll, müssen a und b natürlich noch vor der Übertragung mit dem öffentlichen Schlüssel des Empfängers oder nach irgendeinem anderen Kryptoverfahren verschlüsselt werden.

Bei langen Nachrichten ist die Verdoppelung der Nachrichtenlänge nicht mehr akzeptabel, und selbst, wenn man auf die Übertragung von a verzichten kann, ist das Unterschreiben jedes einzelnen Blocks sehr aufwendig. Deshalb unterschreibt man meist nicht die Nachricht selbst, sondern einen daraus extrahierten Hashwert. Dieser Wert muß natürlich erstens von der gesamten Nachricht abhängen, und zweitens muß es für den Empfänger (praktisch) unmöglich sein, zwei Nachrichten zu erzeugen, die zum gleichen Hashwert führen. Letzteres bedeutet wegen des sogenannten *Geburtstagsparadoxons,* daß für n-Bit Sicherheit Hashwerte der Länge $2n$ erforderlich sind. Die immer noch recht verbreiteten Hashalgorithmen, die 160 Bit liefern, haben somit nur eine Sicherheit von etwa 80 Bit, was heute nicht mehr als wirklich sicher gelten kann. Neuere Hashalgorithmen liefern Werte mit 224 oder 256 Bit, was einer 112- oder 128-Bit Sicherheit entspricht. Die Algorithmen funktionieren ähnlich wie symmetrische Kryptoverfahren; sie versuchen durch Konfusion und Diffusion ein Ergebnis zu berechnen, dessen sämtliche Bits in einer nicht offensichtlichen Weise von jedem einzelnen Nachrichtenbit abhängen.

Eine wichtige Anwendung elektronischer Unterschriften ist übrigens auch die Veröffentlichung von RSA-Schlüsseln: Falls es einem Angreifer gelingt, einem Teilnehmer A einen falschen öffentlichen Schlüssel von Teilnehmer B unterzuschieben, kann (nur) der Angreifer die Nach-

richten von A an B lesen, und er kann sich gegenüber A mittels elektronischer Unterschrift als B ausgeben. Daher sind öffentliche Schlüssel meist unterschrieben von einer Zertifizierungsstelle. Auch deren Unterschrift muß natürlich gegen Manipulationen gesichert sein, beispielsweise indem sie von der nächsthöheren Zertifizierungsstelle unterschrieben ist. An der Spitze der Zertifizierungshierarchie stehen Stellen, deren elektronische Unterschrift jeder Teilnehmer kennen sollte, weil es sich entweder um staatliche Stellen handelt, deren elektronische Unterschriften auf leicht zugänglichen Webseiten verifiziert werden können, oder aber – in der Praxis häufiger – weil die Unterschriften dieser Stellen in Mail- und Browserprogramme eingebaut sind. Letzteres bietet selbstverständlich keine Sicherheit gegen manipulierte Browserprogramme aus dubiosen Quellen, die möglicherweise auch die Mafia als Zertifizierungsstelle anerkennen.

c) Blinde Unterschriften und elektronisches Bargeld

Einer der erfolgversprechendsten Ansätze zum Aushebeln eines Kryptosystems besteht darin, sich auf die Dummheit seiner Mitmenschen zu verlassen.

So sollte es durch gutes Zureden nicht schwer sein, jemanden zu Demonstrationszwecken zum Unterschreiben einer sinnlosen Nachricht zu bewegen: Eine Folge von Nullen und Einsen ohne sinnvolle Interpretation hat schließlich keine rechtliche Wirkung.

Nun muß eine sinnlose Nachricht aber nicht unbedingt eine Zufallszahl sein: Sie kann sorgfältig präpariert sein. Sei dazu etwa m eine Nachricht, die ein Zahlungsversprechen enthält, (N, e) der öffentliche Schlüssel des Opfers und r eine Zufallszahl zwischen 2 und $N - 2$. Dann wird

$$x = m \cdot r^e \bmod N$$

wie eine Zufallsfolge aussehen, für die man eine Unterschrift

$$u = x^d \bmod N = (mr^e)^d \bmod N = m^d r \bmod N$$

bekommt. Multiplikation mit r^{-1} macht daraus eine Unterschrift unter die Zahlungsverpflichtung m.

Das angegebene Verfahren kann nicht nur von Trickbetrügern benutzt werden; blinde Unterschriften sind auch die Grundlage von *digitalem Bargeld.*

Zahlungen im Internet erfolgen meist über Kreditkarten; die Kreditkartengesellschaften haben also einen recht guten Überblick über die Ausgaben ihrer Kunden und machen teilweise auch recht gute Geschäfte mit Kundenprofilen.

Digitales Bargeld will die Anonymität von Geldscheinen mit elektronischer Übertragbarkeit kombinieren und so ein anonymes Zahlungssystem z.B. für das Internet bieten.

Es wir ausgegeben von einer Bank, die für jede angebotene Stückelung einen öffentlichen Schlüssel (N, e) bekanntgibt. Eine Banknote ist eine mit dem zugehörigen geheimen Schlüssel unterschriebene Seriennummer.

Die Seriennummer kann natürlich nicht einfach *jede* Zahl sein; sonst wäre jede Zahl kleiner N eine Banknote. Andererseits dürfen die Seriennummern aber auch nicht von der Bank vergeben werden, denn sonst wüßte diese, welcher Kunde Scheine mit welchen Seriennummern hat. Als Ausweg wählt man Seriennummern einer sehr speziellen Form: Ist $N > 10^{150}$, kann man etwa als Seriennummer eine 150-stellige Zahl wählen, deren Ziffern spiegelsymmetrisch zur Mitte sind, d.h. ab der 76. Ziffer werden die vorherigen Ziffern rückwärts wiederholt. Die Wahrscheinlichkeit, daß eine zufällige Zahl x nach Anwendung des öffentlichen Exponenten auf so eine Zahl führt, ist 10^{-75} und damit vernachlässigbar.

Seriennummern werden von den Kunden zufällig erzeugt. Für jede solche Seriennummer m erzeugt der Kunde eine Zufallszahl r, schickt $mr^e \bmod N$ an die Bank und erhält (nach Belastung seines Kontos) eine Unterschrift u für diese Nachricht zurück. Wie oben berechnet er daraus durch Multiplikation mit r^{-1} die Unterschrift $v = m^d \bmod N$ für die Seriennummer N, und mit diesem Block kann er bezahlen.

Der Zahlungsempfänger berechnet $v^e \bmod N$; falls dies die Form einer gültigen Seriennummer hat, kann er sicher sein, einen von der Bank

unterschriebenen Geldschein vor sich zu haben. Er kann allerdings noch nicht sicher sein, daß dieser Geldschein nicht schon einmal ausgegeben wurde.

Deshalb muß er die Seriennummer an die Bank melden, die mit ihrer Datenbank bereits ausbezahlter Seriennummern vergleicht. Falls sie darin noch nicht vorkommt, wird sie eingetragen und der Händler bekommt sein Geld; andernfalls verweigert sie die Zahlung.

Bei 10^{75} möglichen Nummern liegt die Wahrscheinlichkeit dafür, daß zwei Kunden, die eine (wirklich) zufällige Zahl wählen, dieselbe Nummer erzeugen, bei etwa $10^{-37,5}$. Die Wahrscheinlichkeit, mit jeweils einem Spielschein fünf Wochen lang hintereinander sechs Richtige im Lotto zu haben, liegt dagegen bei $\binom{49}{6}^{-5} \approx 5 \cdot 10^{-35}$, also etwa um den Faktor sechzig höher. Zwei gleiche Seriennummern sind also praktisch auszuschließen, wenn auch theoretisch möglich.

Falls wirklich einmal zufälligerweise zwei gleiche Seriennummern erzeugt worden sein sollten, kann das System nur funktionieren, wenn der zweite Geldschein mit derselben Seriennummer nicht anerkannt wird, so daß der zweite Kunde sein Geld verliert. Dies muß als eine zusätzliche Gebühr gesehen werden, die mit an Sicherheit grenzender Wahrscheinlichkeit nie fällig wird, aber trotzdem nicht ausgeschlossen werden kann.

Da digitales Bargeld nur in kleinen Stückelungen sinnvoll ist (Geldscheinen im Millionenwert wären auf Grund ihrer Seltenheit nicht wirklich anonym und würden wegen der damit verbundenen Möglichkeiten zur Geldwäsche auch in keinem seriösen Wirtschaftssystem akzeptiert), wäre der theoretisch mögliche Verlust ohnehin nicht sehr groß.

Digitales Bargeld der gerade beschriebenen Form wurde 1982 von DAVID CHAUM vorgestellt; 1990 gründete er eine Firma namens DigiCash, die es kommerziell vermarkten sollte. Zu deren Kunden gehörte beispielsweise auch die Deutsche Bank, die allerdings nur 27 Kunden fand, die Zahlungen damit akzeptierten. DigiCash wurde 1998 zahlungsunfähig; derzeit gibt es meines Wissens keine ähnlichen Zahlungssysteme.

d) Bankkarten mit Chip

In Deutschland und den meisten anderen Ländern hat eine Bankkarte einen Magnetstreifen, auf dem die wichtigsten Informationen wie Kontenname und -nummer, Bankleitzahl, Gültigkeitsdauer *usw.* gespeichert sind; dazu kommt verschlüsselte Information, die unter anderem die Geheimzahl enthält, die aber auch von den obengenannten Daten abhängt. Zur Verschlüsselung verwendet man hier ein konventionelles, d.h. symmetrisches Kryptoverfahren; derzeit noch meist Triple-DES.

Der Schlüssel, mit dem dieses arbeitet, muß natürlich streng geheimgehalten werden: Wer ihn kennt, kann problemlos die Geheimzahlen fremder Karten ermitteln und eigene Karten zu beliebigen Konten erzeugen.

Um eine Karte zu überprüfen, muß daher eine Verbindung zu einem Zentralrechner aufgebaut werden, an den sowohl der Inhalt des Magnetstreifens als auch die vom Kunden eingetippte Zahl übertragen werden; dieser wendet Triple-DES mit dem Systemschlüssel an und meldet dann, wie die Prüfung ausgefallen ist.

In Frankreich haben die entsprechenden Karten zusätzlich zum Magnetstreifen noch einen Chip, in dem ebenfalls die Kontendaten gespeichert sind sowie, in einem auslesesicheren Register, Informationen über die Geheimzahl. Dort wird die ins Lesegerät eingetippte Geheimzahl nicht an den Zentralrechner übertragen, sondern an den Chip, der sie überprüft und akzeptiert oder auch nicht.

Da frei programmierbare Chipkarten relativ billig sind, muß dafür Sorge getragen werden, daß ein solches System nicht durch einen *Yes-Chip* unterlaufen werden kann, der ebenfalls die Konteninformationen enthält, ansonsten aber ein Programm, das ihn *jede* Geheimzahl akzeptieren läßt. Das Terminal muß also, bevor es überhaupt eine Geheimzahl anfordert, zunächst einmal den Chip authentisieren, d.h. sich davon überzeugen, daß es sich um einen vom Bankenkonsortium ausgegebenen Chip handelt.

Aus diesem Grund sind die Kontendaten auf dem Chip mit dem privaten RSA-Schlüssel des Konsortiums unterschrieben. Die Terminals

kennen den öffentlichen Schlüssel dazu und können so die Unterschrift überprüfen.

Diese Einzelheiten und speziell deren technische Implementierung wurden vom Bankenkonsortium zunächst streng geheimgehalten. Trotzdem machte sich 1997 ein elsässischer Ingenieur namens SERGE HUMPICH daran, den Chip genauer zu untersuchen. Er verschaffte sich dazu ein (im freien Verkauf erhältliches) Terminal und untersuchte sowohl die Kommunikation zwischen Chip und Terminal als auch die Vorgänge innerhalb des Terminals mit Hilfe eines Logikanalysators. Damit gelang es ihm nach und nach, die Funktionsweise des Terminals zu entschlüsseln und in ein äquivalentes PC-Programm zu übersetzen. Durch dessen Analyse konnte er die Authentisierungsprozedur und die Prüflogik entschlüsseln und insbesondere auch feststellen, daß hier mit RSA gearbeitet wurde.

Blieb noch das Problem, den Modul zu faktorisieren. Dazu besorgte er sich ein japanisches Programm aus dem Internet, das zwar eigentlich für kleinere Zahlen gedacht war, aber eine Anpassung der Wortlänge ist natürlich auch für jemanden, der den Algorithmus hinter dem Programm nicht versteht, kein Problem. Nach sechs Wochen Laufzeit hatte sein PC damit den Modul faktorisiert:

$$\begin{aligned} &213598703592091008239502270499962879705109534182\backslash \\ &641740644252416500858395774644508840500943086599 9 \\ &= 111395432514882798792549017547702484407092284484 3 \\ &\times 191748170252450443937578626823086218069693418929 3 \end{aligned}$$

Als er seine Ergebnisse über einen Anwalt dem Bankenkonsortium mitteilte, zeigte sich, was dieses sich unter Sicherheitsstandards vorstellt: Es erreichte, daß HUMPICH wegen des Eindringens in ein DV-System zu zehn Monaten Haft auf Bewährung sowie einem Franc Schadenersatz plus Zinsen verurteilt wurde; dazu kamen 12 000 F Geldstrafe. Einzelheiten findet man in seinem Buch

SERGE HUMPICH: Le cerveau bleu, Xo, 2001

Seit November 1999 haben neu ausgegebene Bankkarten noch ein zusätzliches Feld mit einer Unterschrift, die im Gegensatz zum obigen 320-Bit-Modul einen 768-Bit-Modul verwendet. Natürlich können damit erzeugte Unterschriften nur von neueren Terminals überprüft werden, so daß viele Transaktionen weiterhin nur über den 320-Bit-Modul mit inzwischen wohlbekannter Faktorisierung „geschützt“ sind.

§4: Wie groß sollten die Primzahlen sein?

Das Beispiel der französischen Bankkarten zeigt, daß RSA höchstens dann sicher ist, wenn die Primzahlen p und q hinreichend groß gewählt werden. Als erstes müssen wir uns daher die Frage stellen, wie groß eine „hinreichend große“ Zahl heute sein muß.

Ein treu sorgender Staat läßt seine Bürgern bei einer derart wichtigen Frage natürlich nicht allein: Zwar gibt es noch keine oberste Bundesbehörde für Primzahlen, aber das Bundesamt für Sicherheit in der Informationstechnik (BSI) und die Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen publizieren jedes Jahr ein Dokument mit dem Titel *Geeignete Algorithmen zur Erfüllung der Anforderungen nach §17 Abs. 1 bis 3 SigG vom 22. Mai 2001 in Verbindung mit Anlage 1 Abschnitt I Nr. 2 SigV vom 22. November 2001.*

SigV steht für die aufgrund des Signaturgesetzes SigG erlassene Signaturverordnung; beide gemeinsam legen fest, daß elektronische Unterschriften in Deutschland grundsätzlich zulässig und rechtsgültig sind, sofern gewisse Bedingungen erfüllt sind. Zu diesen Bedingungen gehört unter anderem, daß das Verfahren und die Schlüssellänge gemeinsam einen „geeigneten Kryptoalgorihmus“ im Sinne der jeweils gültigen Veröffentlichung der Bundesnetzagentur ist.

Da Rechner immer schneller und leistungsfähiger werden und auch auf der mathematisch-algorithmischen Seite fast jedes Jahr kleinere oder größere Fortschritte zu verzeichnen sind, gelten die jeweiligen Empfehlungen nur für etwa sechs Jahre. Für Dokumente, die länger gültig sein sollen, sind elektronische Unterschriften also nicht vorgesehen.

Die neuesten Richtlinien stammen vom 22. Dezember 2010 und wurden am 1. Februar 2011 im Bundesanzeiger Nr. 17, S. 383. veröffentlicht. Sie empfehlen 2048 Bit, aber wirklich verbindlich sind nur 1976.

(Der minimale Unterschied hängt mit Implementierungsproblemen beim Chipkarten-Betriebssystem SECCOS zusammen.)

Die beiden Primfaktoren p, q sollen zufällig und unabhängig voneinander erzeugt werden und aus einem Bereich stammen, in dem

$$\varepsilon_1 < |\log_2 p - \log_2 q| < \varepsilon_2$$

gilt. Als *Anhaltspunkte* werden dabei die Werte $\varepsilon_1 = 0{,}1$ und $\varepsilon_2 = 30$ vorgeschlagen; ist p die kleinere der beiden Primzahlen, soll also $2^{-10}p < q < 2^{30}p \approx 10^9 p$ gelten. Die beiden Primzahlen sollten somit zwar ungefähr dieselbe Größenordnung haben, aber nicht *zu nahe* beieinander liegen. Der Grund dafür ist ein von FERMAT entdecktes Faktorisierungsverfahren auf Grundlage der dritten binomischen Formel: Falls für eine Zahl N und eine natürliche Zahl y die Zahl $N + y^2$ eine Quadratzahl x^2 ist, ist $N = x^2 - y^2 = (x + y)(x - y)$, womit zwei Faktoren gefunden sind. Probiert man alle kleinen natürlichen Zahlen y systematisch durch, führt dieses Verfahren offensichtlich umso schneller zum Erfolg, je näher die beiden Faktoren von N beieinander liegen. Wir werden uns in Kapitel über Faktorisierung noch genauer damit befassen. Außerdem wird *empfohlen,* $e \geq 2^{16} + 1$ zu wählen.

§5: Verfahren mit diskreten Logarithmen

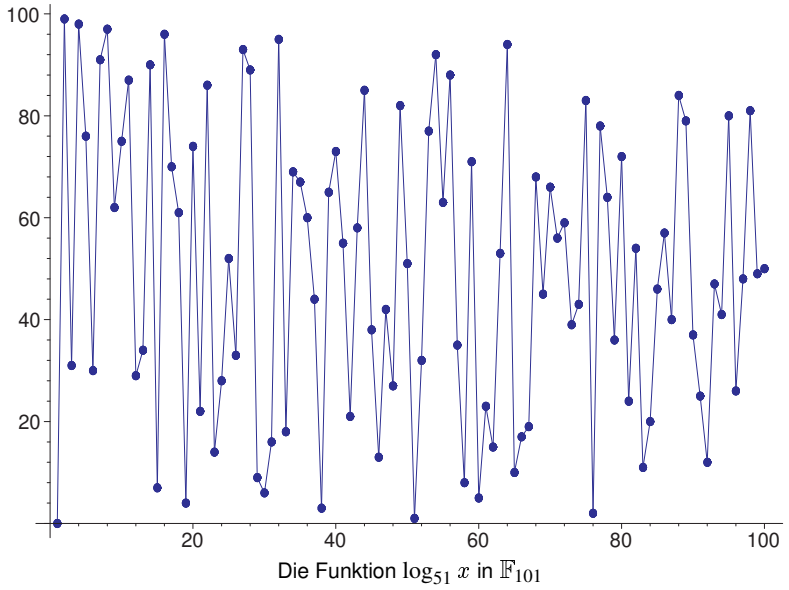
Kurz nach der Veröffentlichung des RSA-Algorithmus fanden auch DIFFIE und HELLMAN ein Verfahren, das im Gegensatz zu RSA sogar ganz ohne vorvereinbarte Schlüssel auskommt: Zwei Personen vereinbaren über eine unsichere Leitung einen Schlüssel, den anschließend nur sie kennen.

Ausgangspunkt ist wieder das Potenzieren im Körper $\mathbb{F}_p$; hier betrachten wir aber die Exponentialfunktion $x \mapsto a^x$ zu einer geeigneten Basis a. Ihre Umkehrfunktion bezeichnet man als *Index* oder *diskreten Logarithmus* zur Basis a:

$$y = a^x \Longrightarrow x = \log_a y\,.$$

Trotz dieser formalen Übereinstimmung gibt es es allerdings große Unterschiede zwischen reellen Logarithmen und ihren Analoga in endlichen Körpern: Während reelle Logarithmen sanft ansteigende stetige

Funktionen sind, die man leicht mit beliebig guter Genauigkeit annähern kann, sieht der diskrete Logarithmus typischerweise so aus, wie es in der Abbildung zu sehen ist. Auch ist im Reellen der Logarithmus zur Basis $a > 1$ für jede positive Zahl definiert; in endlichen Körpern ist es viel schwerer zu entscheiden, ob ein bestimmter Logarithmus existiert: Modulo sieben etwa sind 2, 4 und 1 die einzigen Zweierpotenzen, so daß 3, 5 und 6 keine Zweierlogarithmen haben. Ein Satz aus der Algebra besagt allerdings, daß es stets Elemente a gibt, für die a^x jeden Wert außer der Null annimmt, die sogenannten primitiven Wurzeln. In $\mathbb{F}_7$ wären dies etwa drei und fünf.



Die Funktion $\log_{51} x$ in $\mathbb{F}_{101}$

Die Berechnung der Potenzfunktion durch sukzessives Quadrieren und Multiplizieren ist auch in endlichen Körpern einfach, für ihre Umkehrfunktion, den diskreten Logarithmus gibt es aber derzeit nur deutlich schlechtere Verfahren. Die derzeit besten Verfahren zur Berechnung von diskreten Logarithmen in Körpern mit N Elementen erfordern etwa denselben Aufwand wie die Faktorisierung eines RSA-Moduls der

Größenordnung N. Diese Diskrepanz zwischen Potenzfunktion und Logarithmen kann kryptologisch ausgenutzt werden.

Als Körper verwendet man entweder Körper von Zweipotenzordnung, die wir weiter unten betrachten werden, oder Körper von Primzahlordnung. Da es für viele interessante Körper von Zweipotenzordnung bereits Chips gibt, die dort diskrete Logarithmen berechnen, dürften Körper von Primzahlordnung bei ungefähr gleicher Elementanzahl wohl etwas sicherer sein: Es gibt einfach viel mehr Primzahlen als Zweierpotenzen, und jeder Fall erfordert einen neuen Hardwareentwurf. Falls man die Primzahlen hinreichend häufig wechselt, dürfte sich dieser Aufwand für kaum einen Gegner lohnen.

Da Körper von Primzahlordnung auch einfacher sind als solche von Primzahlpotenzordnung, wollen wir uns zunächst auf diese beschränken; die spätere Übertragung des Algorithmus auf Körper von Zweipotenzordnung sollte dem Leser keine Schwierigkeiten machen.

Beim DIFFIE-HELLMAN-Verfahren, dem ältesten auf der Grundlage diskreter Logarithmen, geht es wie gesagt darum, daß zwei Teilnehmer, die weder über gemeinsame Schlüsselinformation noch über eine sichere Leitung verfügen, einen Schlüssel vereinbaren wollen.

Dazu einigen sie sich zunächst (über die unsichere Leitung) auf eine Primzahl p und eine natürliche Zahl a derart, daß die Potenzfunktion $x \mapsto a^x$ möglichst viele Werte annimmt.

Als nächstes wählt Teilnehmer A eine Zufallszahl $x < p$ und B entsprechend $y < p$; A schickt $u = a^x \bmod p$ an B und erhält dafür $v = a^y \bmod p$.

Sodann berechnet A die Zahl $v^x \bmod p = \left(a^y\right)^x \bmod p = a^{xy} \bmod p$ und B entsprechend $u^y \bmod p = \left(a^x\right)^y \bmod p = a^{xy} \bmod p$. Beide haben also auf verschiedene Weise dieselbe Zahl berechnet, die sie nun als Schlüssel in einem klassischen Kryptosystem verwenden können, wobei sie sich wohl meist auf einen Teil der Bits beschränken müssen, da solche Schlüssel typischerweise eine Länge von 128 bis 256 Bit haben, während die Primzahl p erheblich größer sein muß.

Ein Gegner, der den Datenaustausch abgehört hat, kennt die Zahlen p, a, u und v; um $a^{xy} \bmod p$ zu finden, muß er den diskreten Logarithmus von u oder v berechnen.

Mit den besten heute bekannten Algorithmen ist die möglich, wenn p eine Primzahl von bis zu etwa 200 Dezimalstellen ist; dies entspricht etwa 665 Bit. Auch in diesem Fall dauert die Berechnung allerdings selbst bei massiver Parallelisierung über das Internet mehrere Monate, gefolgt von einer Schlußrechnung auf einem Supercomputer.

Natürlich gibt es keine Garantie, daß kein Gegner mit einem besseren als den bislang bekannten Verfahren diskrete Logarithmen oder Faktorisierungen auch in weitaus größeren Körpern berechnen kann. Dazu bräuchte er allerdings einen Durchbruch entweder auf der mathematischen oder auf der technischen Seite, für den weit und breit keine Grundlage zu sehen ist.

Trotzdem gibt es einen verhältnismäßig einfachen Angriff auf den Schlüsselaustausch nach DIFFIE und HELLMAN, die sogenannte *man in the middle attack.* Dabei unterbricht der Angreifer die Verbindung zwischen A und B und gibt sich gegenüber A als B aus und umgekehrt. So kann er mit beiden Teilnehmern je einen Schlüssel vereinbaren, und die damit verschlüsselte Kommunikation kann (nur) von ihm gelesen und gegebenenfalls manipuliert werden. In der vorgestellten Form funktioniert das Verfahren also nur, wenn man sicher sein weiß, mit wem man kommuniziert.

§6: DSA

DSA steht für *Digital Signature Algorithm,* ein Algorithmus der im *Digital Signature Standard* DSS der USA festgelegt ist und neben RSA auch zu den von der Bundesnetzagentur empfohlenen „Geeigneten Algorithmen“ zählt. Seine Sicherheit beruht auf diskreten Logarithmen, allerdings wird das klassische Verfahren dadurch modifiziert, daß die Sicherheit zwar auf dem diskreten Logarithmenproblem in einem großen Körper beruht, die Rechenoperationen bei der Anwendung des Algorithmus aber nur eine deutlich kleinere Untergruppe verwenden.

Für diese kleine Untergruppe wählt man eine Primzahl q mit einer Länge von mindestens 224 Bit, ab 2016 mindestens 256. (Das entspricht der Länge der Hashwerte, die in der Praxis anstelle des Texts unterzeichnet werden.) Zu dieser Primzahl q sucht man eine Primzahl $p \equiv 1 \bmod q$, für deren Länge die Bundesnetzagentur mindestens 2048 Bit vorschreibt.

Daß die mit der empfohlenen RSA-Modullänge übereinstimmt, ist kein Zufall: Auch wenn kein direkter Zusammenhang zwischen Faktorisierung und der Berechnung diskreter Logarithmen bekannt ist, hat bislang doch jede neue Idee für einen Faktorisierungsalgorithmus auch zu einem Algorithmus zur Berechnung diskreter Logarithmen geführt, und auch die Laufzeiten dieser Algorithmen sind bei gleicher Zahlenlänge ungefähr gleich.

Als nächstes muß ein Element g gefunden werden, dessen Potenzen im Körper $\mathbb{F}_p$ eine Gruppe der Ordnung q bilden. Das ist einfach: Man starte mit irgendeinem Element $g_0 \in \mathbb{F}_p \setminus \{0\}$ und berechne seine $(p-1)/q$-te Potenz. Falls diese ungleich eins ist, muß sie wegen $g_0^{p-1} = 1$ die Ordnung q haben; andernfalls muß ein neues g_0 betrachtet werden.

Die so bestimmten Zahlen q, p und g werden veröffentlicht und können auch in einem ganzen Netzwerk global eingesetzt werden. Geheimer Schlüssel jedes Teilnehmers ist eine Zahl x zwischen eins und $q-1$; der zugehörige öffentliche Schlüssel ist $y = g^x \bmod p$.

Unterschreiben lassen sich mit diesem Verfahren Nachrichtenblöcke m mit $0 \le m < q$, insbesondere also 224 Bit lange Hashwerte. Dazu wählt man für jede Nachricht eine Zufallszahl k mit $0 < k < q$ und berechnet

$$r = (g^k \bmod p) \bmod q\,.$$

Da q eine Primzahl ist, hat k ein multiplikatives Inverses modulo q; man kann also durch k dividieren und erhält eine Zahl s, für die

$$sk \equiv m + xr \bmod q$$

ist; die Unterschrift unter die Nachricht m besteht dann aus den beiden Zahlen r und s modulo q. Sie kann nur berechnet werden von jemanden, der den geheimen Schlüssel x kennt.

Überprüfen kann die Unterschrift allerdings jeder: Ist t das multiplikative Inverse zu s modulo q, so ist $k \equiv tsk \equiv tm + xtr \bmod q$, also, da g die Ordnung q hat,

$$r \equiv g^k \equiv g^{tm} g^{xtr} \equiv g^{tm} y^{tr} \bmod p\,.$$

In dieser Gleichung sind die linke wie auch die rechte Seite *modulo* q öffentlich bekannt, die Gleichung kann also modulo q überprüft werden. Die Unterschrift wird anerkannt, wenn beide Seiten modulo q gleich sind. Ein Angreifer müßte sich x aus y verschaffen, müßte also ein diskretes Logarithmenproblem modulo der großen Primzahl p lösen.

§7: Anwendungen bei SSL/TLS

SSL steht für *secure socket layer,* TLS für *transport layer security;* Zweck ist jeweils der Aufbau einer sicheren Internetverbindung. Wie im Internet üblich, können dazu die verschiedensten Verfahren benutzt werden; die auf Grundlage von RSA zählen derzeit zu den populärsten.

Natürlich ist RSA zu aufwendig, um damit eine längere Kommunikation wie beispielsweise eine *secure shell* Sitzung zu verschlüsseln; tatsächlich dient RSA daher nur zur Übertragung eines Schlüssels für ein konventionelles Kryptoverfahren wie AES, IDEA oder Triple-DES, auf das sich die Beteiligten unter SSL/TLS ebenfalls einigen müssen.

Am einfachsten wäre es, wenn der Client einen Schlüssel für ein solches Verfahren wählt und dann diesen mit dem RSA-Schlüssel des Servers verschlüsselt an diesen schickt – vorausgesetzt, er kennt diesen RSA-Schlüssel. Letzteres ist im allgemeinen nicht der Fall; daher muß zunächst der Server dem Client seinen Schlüssel mitteilen.

Da der Client nicht sicher sein kann, mit dem richtigen Server verbunden zu sein, schickt er diesen Schlüssel meist zusammen mit einem Zertifikat, das sowohl seine Identität als auch seinen RSA-Schlüssel enthält und von einer Zertifizierungsstelle unterschrieben ist.

Die öffentlichen Schlüssel der gängigen Zertifizierungsstellen sind in die Browserprogramme eingebaut; bei weniger bekannten Zertifizierungsstellen wie etwa dem Rechenzentrum der Universität Mannheim

fragt der Browser den Benutzer, ob er das Zertifikat anerkennen will oder nicht. Bei *secure shell* schließlich, wo die Gegenseite typischerweise keinerlei Zertifikat vorweisen kann, fragt das Programm beim ersten Verbindungsaufbau zu einem Server, ob dessen Schlüssel anerkannt werden soll und speichert dann einen sogenannten *fingerprint* davon; dieser wird bei späteren Verbindungen zur Identitätsfeststellung benutzt.

§8: Ausblick

Dieses kurze Kapitel konnte selbstverständlich keine umfassende Übersicht über die Kryptographie oder auch nur die asymmetrische Kryptographie geben: Auch das RSA-Verfahren kann mit anderen Methoden angegriffen werden als der direkten Faktorisierung des Moduls. Eine dieser Methoden werden wir im Kapitel über Kettenbrüche kennenlernen, und auch sonst werden im weiteren Verlauf der Vorlesungen noch gelegentlich Themen aus der Kryptologie angeschnitten werden.

Mit Ausnahme von Verfahren wie dem sogenannten *one time pad* gibt es für keines der heute benutzten Kryptoverfahren einen Sicherheitsbeweis, nicht einmal in dem Sinn, daß man den Aufwand eines Gegners zum Knacken des Verfahrens in irgendeiner realistischen Weise nach unten abschätzen könnte. Seriöse Kryptographie außerhalb des Höchstsicherheitsbereichs muß sich daher damit begnügen, daß die Verantwortlichen für den Einsatz eines Verfahrens und der Wahl seiner Parameter (wie den Primzahlen bei RSA) darauf achten, auf dem neuesten Stand der Forschung zu bleiben und ihre Wahl so treffen, daß nicht nur die bekannten Angriffsmethoden versagen, sondern daß auch noch ein recht beträchtlicher Sicherheitszuschlag für künftige Entwicklungen und für nicht publizierte Entwicklungen bleibt.

Auf ewige Sicherheit kann man mit Verfahren wie RSA ohnehin nicht hoffen: Als RSA 1977 von MARTIN GARDNER im *Scientific American* vorgestellt wurde, bekam er von RIVEST, SHAMIR und ADLEMAN die 129-stellige Zahl

$$114381625757888867669235779976146612010218296721242362562561842935706935245733897830597123563958705058989075147599290026879543541$$

(seither bekannt als RSA-129) und eine damit verschlüsselte Nachricht, für deren Entschlüsselung die drei einen Preis von hundert Dollar ausgesetzt hatten. Sie schätzten, daß eine solche Entschlüsselung etwa vierzig Quadrillionen ($4 \cdot 10^{25}$) Jahre dauern würde. (Heute sagt RIVEST, daß dies auf einem Rechenfehler beruhte.) Tatsächlich wurde der Modul 1994 faktorisiert in einer gemeinsamen Anstrengung von 600 Freiwilligen, deren Computer immer dann, wenn sie nichts besseres zu tun hatten, daran arbeiteten. Nach acht Monaten war die Faktorisierung gefunden: Die obige Zahl ist gleich

$$\begin{aligned}&4905295108476509491478496199038981334177646384933878439908205773276913299326670954996198819083446141317764296799294253979828853\\ &\times 3\,.\end{aligned}$$

Mit dem Schema $A = 01$ bis $Z = 26$ und Zwischenraum gleich 00 ergab sich die Nachricht *The Magic Words are Squeamish Ossifrage.*

Auch bei den heute als sicher geltenden symmetrischen Kryptoverfahren rechnet niemand ernsthaft damit, daß sie noch in hundert Jahren sicher sind: Diese Verfahren werden üblicherweise so gewählt, daß man auf eine Sicherheit für etwa dreißig Jahren hoffen kann – garantieren kann aber auch das niemand.

Falls sich sogenannte *Quantencomputer* realisieren lassen, werden alle heute bekannten Verfahren der Kryptographie mit öffentlichen Schlüsseln, egal ob mit diskreten Logarithmen, RSA oder elliptischen Kurven, unsicher sein. Bislang können Quantencomputer kaum mit acht Bit rechnen, und nicht alle Experten sind davon überzeugt, daß es je welche geben wird, die mit mehreren Tausend Bit rechnen können.

Wer mehr über Kryptographie wissen will, findet einen ersten Überblick beispielsweise bei

BUCHMANN: Einführung in die Kryptographie, *Springer,* ^{5}2011

oder natürlich auch in der hier immer wieder angebotenen Kryptologie-Vorlesung.

Mehr über die Geschichte der Kryptographie mit öffentlichen Schlüsseln ist (mathematikfrei) zu finden in

STEVEN LEVY: **crypto:** how the rebels beat the government – saving privacy in the digital age, *Penguin Books,* 2002

Kapitel 3
Primzahlen

Wie wir aus dem ersten Kapitel wissen, sind Primzahlen die Grundbausteine für die multiplikative Struktur der ganzen Zahlen, und aus Kapitel zwei wissen wir, daß sie auch wichtige Anwendungen außerhalb der Zahlentheorie haben. Es lohnt sich also auf jeden Fall, sie etwas genauer zu untersuchen.

§1: Die Verteilung der Primzahlen

Als erstes stellt sich die Frage, wie viele Primzahlen es gibt. Die Antwort finden wir schon in EUKLIDs Elementen; der dort gegebene Beweis dürfte immer noch der einfachste sein: Es gibt unendlich viele Primzahlen, denn gäbe es nur endlich viele Primzahlen $p_1, \ldots, p_n$, so könnten wir deren Produkt P bilden und die Primzerlegung von $P+1$ betrachten. Da P durch alle p_i teilbar ist, ist $P+1$ durch kein p_i teilbar, im Widerspruch zur Existenz der Primfaktorzerlegung. Somit muß es noch weitere, also unendlich viele Primzahlen geben.

Um nicht ganz auf dem Stand von vor rund zweieinhalb Jahrtausenden stehen zu bleiben, wollen wir uns noch einen zweiten, auf EULER zurückgehenden Beweis ansehen.

Dazu betrachten wir für eine reelle Zahl $s > 1$ die unendliche Reihe

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} .$$

Als erstes müssen wir uns überlegen, daß diese Reihe konvergiert. Da alle Summanden positiv sind, müssen wir dafür nur zeigen, daß es eine

gemeinsame obere Schranke für alle Teilsummen gibt. Da die Funktion $x \mapsto 1/x^s$ für $x > 0$ monoton fallend ist, haben wir für $n-1 \le x \le n$ die Abschätzung $1/n^s \le 1/x^s$, d.h.

$$\sum_{n=1}^{N} \frac{1}{n^s} = 1 + \sum_{n=2}^{N} \frac{1}{n^s} \le 1 + \int_1^N \frac{dx}{x^s}$$

$$< 1 + \int_1^{\infty} \frac{dx}{x^s} = 1 + \frac{1}{s-1} = \frac{s}{s-1} .$$

Somit ist $\zeta(s)$ für alle $s > 1$ wohldefiniert.

Einen Zusammenhang mit Primzahlen liefert der folgende

Satz: *a)* Für $s > 1$ ist $\zeta(s) = \prod_{p \text{ prim}} \frac{1}{1 - \frac{1}{p^s}}$.

b) Für alle $N \in \mathbb{N}$ und alle reellen $s > 0$ ist $\sum_{n=1}^{N} \frac{1}{n^s} \le \prod_{\substack{p \le N \\ p \text{ prim}}} \frac{1}{1 - \frac{1}{p^s}}$.

Beweis: Wir beginnen mit *b)*. Für $N = 1$ steht hier die triviale Formel $1 \le 1$; sei also $N \ge 2$, und seien $p_1, \ldots, p_r$ die sämtlichen Primzahlen kleiner oder gleich N. Nach der Summenformel für die geometrische Reihe ist

$$\frac{1}{1 - \frac{1}{p_k^s}} = \sum_{\ell=0}^{\infty} \frac{1}{p_k^{\ell s}} ,$$

und das Produkt der rechtsstehenden Reihen über $k = 1$ bis r ist wegen der Eindeutigkeit der Primzerlegung die Summe über alle jene $1/n^s$, für die n keinen Primteiler größer N hat. Darunter sind insbesondere alle $n \le N$, womit *b)* bewiesen wäre.

Die Differenz zwischen $\zeta(s)$ und dem Produkt auf der rechten Seite von *b)* ist gleich der Summe über alle $1/n^s$, für die n mindestens einen Primteiler größer N haben. Diese Summe ist natürlich höchstens gleich der Summe aller $1/n^s$ mit $n > N$, und die geht wegen der Konvergenz von $\zeta(s)$ gegen null für $N \to \infty$. Damit ist auch *a)* bewiesen. ∎

Auch daraus folgt, daß es unendlich viele Primzahlen gibt: Gäbe es nämlich nur endlich viele, so stünde auf der rechten Seite von *b)* für jedes hinreichend große N das Produkt über die *sämtlichen* Primzahlen. Da es nur endlich viele Faktoren hat, wäre es auch für $s = 1$ endlich, und damit müßte

$$\sum_{n=1}^{\infty} \frac{1}{n}$$

kleiner oder gleich dieser Zahl sein, im Widerspruch zur Divergenz der harmonischen Reihe.

Verglichen mit dem Beweis aus EUKLIDs Elementen ist EULERs Methode erheblich komplizierter. Um trotzdem ihre Existenzberechtigung zu haben, sollte sie uns daher auch mehr Informationen liefern. In welchem Maße sie dies tatsächlich leistet, geht wahrscheinlich sogar noch deutlich über alles hinaus, was EULER seinerzeit träumen konnte.

Zunächst einmal können wir Teil *b)* für $s = 1$ zu einer quantitativen Abschätzung bezüglich der Anzahl $\pi(N)$ der Primzahlen kleiner oder gleich N umformulieren: Wie oben im Konvergenzbeweis für $\zeta(s)$ können wir aus der Monotonie der Funktion $x \mapsto 1/x$ folgern, daß für alle $N \in \mathbb{N}$ gilt

$$\log(N+1) = \int_1^{N+1} \frac{dx}{x} < \sum_{n=1}^{N} \frac{1}{n} < 1 + \int_1^{N} \frac{dx}{x} = 1 + \log N\,.$$

Zur Abschätzung der rechten Seite beachten wir einfach, daß der Faktoren $1/(1-1/p)$ für $p = 2$ gleich zwei ist, ansonsten aber kleiner. Somit ist

$$\log(N+1) < \sum_{n=1}^{N} \frac{1}{n} < \prod_{\substack{p \le N \\ p \text{ prim}}} \frac{1}{1-\frac{1}{p}} \le 2^{\pi(N)}$$

und damit

$$\pi(N) \ge \frac{\log\log(N+1)}{\log 2}\,.$$

Wie wir bald sehen werden, ist das allerdings eine sehr schwache Abschätzung.

EULERs Methode erlaubt uns auch, die Dichte der Primzahlen zu vergleichen mit der Dichte beispielsweise der Quadratzahlen: Wie wir oben gesehen haben, konvergiert

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

für alle $s > 1$, insbesondere also konvergiert die Summe der inversen Quadratzahlen. EULER konnte mit seiner Methode zeigen, daß die Summe der inversen Primzahlen *divergiert,* so daß die Primzahlen zumindest in diesem Sinne dichter liegen als die Quadratzahlen und alle anderen Potenzen mit (reellem) Exponenten $x > 1$.

Zum Beweis fehlt uns nur noch eine Analysis I Übungsaufgabe: Wir wollen uns überlegen, daß für alle $0 \le x \le \frac{1}{2}$ gilt $(1-x) \ge 1/4^x$. An den Intervallenden stimmen beide Funktionen überein und $1-x$ ist eine lineare Funktion. Es reicht daher, wenn wir zeigen, daß $1/4^x$ eine konvexe Funktion ist, daß also ihre zweite Ableitung überall im Intervall positiv ist. Das ist aber klar, denn die ist einfach $\log(1/4)^2/4^x$. Für jede Primzahl p ist daher

$$\left(1-\frac{1}{p}\right)^{-1} \le 4^{1/p}$$

und damit

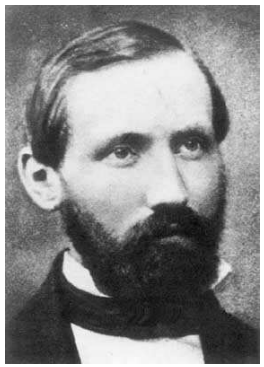
$$\log(N+1) < \prod_{\substack{p \le N \\ p \text{ prim}}} \left(1-\frac{1}{p}\right)^{-1} \le \prod_{\substack{p \le N \\ p \text{ prim}}} 4^{1/p} = 4^{\sum_{\substack{p \le N \\ p \text{ prim}}} \frac{1}{p}}\,.$$

Da $\log(N+1)$ für $N \to \infty$ gegen unendlich geht, muß somit auch die Summe der inversen Primzahlen divergieren.

Mit diesen Bemerkungen fängt allerdings die Nützlichkeit der Funktion $\zeta(s)$ für das Verständnis der Funktion $\pi(N)$ gerade erst an: Ein Jahrhundert nach EULER erkannte RIEMANN, daß die Funktion $\zeta(s)$ ihre wahre Nützlichkeit für das Studium von $\pi(N)$ erst zeigt, wenn man sie auch für komplexe Argumente s betrachtet. Jeder, der sich ein bißchen mit Funktionen einer komplexer Veränderlichen auskennt, kann leicht zeigen, daß $\zeta(s)$ auch für komplexe Zahlen mit Realteil größer ein konvergiert: Der

Imaginärteil des Exponenten führt schließlich nur zu einem Faktor vom Betrag eins.

RIEMANNs wesentliche Erkenntnis war, daß sich $\zeta(s)$ fortsetzen läßt zu einer analytischen Funktion auf der gesamten Menge der komplexen Zahlen mit Ausnahme der Eins (wo die ζ-Funktion wegen der Divergenz der harmonischen Reihe keinen endlichen Wert haben kann).



GEORG FRIEDRICH BERNHARD RIEMANN (1826-1866) war Sohn eines lutherischen Pastors und schrieb sich 1946 auf Anraten seines Vaters an der Universität Göttingen für das Studium der Theologie ein. Schon bald wechselte an die Philosophische Fakultät, um dort unter anderem bei GAUSS Mathematikvorlesungen zu hören. Nach Promotion 1851 und Habilitation 1854 erhielt er dort 1857 einen Lehrstuhl. Trotz seines frühen Todes initiierte er grundlegende auch noch heute fundamentale Entwicklungen in der Geometrie, der Zahlentheorie und über abelsche Funktionen. Wie sein Nachlaß zeigte, stützte er seine 1859 aufgestellte Vermutung über die Nullstellen der ζ-Funktion auf umfangreiche Rechnungen.

Für Leser, die nicht mit dem Konzept der analytischen Fortsetzung vertraut sind, möchte ich ausdrücklich darauf hinweisen, daß dies selbstverständlich nicht bedeutet, daß die definierende Summe der ζ-Funktion für reelle Zahlen kleiner eins oder komplexe Zahlen mit Realteil kleiner oder gleich eins konvergiert: Analytische Fortsetzung besteht darin, daß eine differenzierbare Funktion (die im Komplexen automatisch beliebig oft differenzierbar ist und um jeden Punkt in eine TAYLOR-Reihe entwickelt werden kann) via TAYLOR-Reihen über ihren eigentlichen Definitionsbereich hinweg ausgedehnt wird. Man kann beispielsweise zeigen, das $\zeta(-1) = -\frac{1}{12}$ ist. Setzt man $s = -1$ in die für $s > 1$ gültige Reihe ein, erhält man die Summe aller natürlicher Zahlen, die selbstverständlich nicht gleich $-\frac{1}{12}$ ist, sondern divergiert. Entsprechend hat $\zeta(s)$ Nullstellen bei allen geraden negativen Zahlen, obwohl auch hier die entsprechenden Reihen divergieren. Diese Nullstellen bezeichnet man als die sogenannten *trivialen* Nullstellen der ζ-Funktion, da sie sich sofort aus einer bei der Konstruktion der analytischen Fortsetzung zu beweisenden Funktionalgleichung ablesen lassen. Für die Primzahlverteilung spielen vor allem die übrigen, die sogenannten nichttrivialen Nullstellen, eine große Rolle.

Wie wir gerade gesehen haben, liegen die Primzahlen zumindest in einem gewissen Sinne dichter als die Quadratzahlen. Zur Einstimmung auf das Problem der Primzahlverteilung wollen wir uns kurz mit der (deutlich einfacheren) Verteilung der Quadratzahlen beschäftigen.

Die Folge der Abstände zwischen zwei aufeinanderfolgenden Quadratzahlen ist einfach die Folge der ungeraden Zahlen, denn $(n+1)^2 - n^2 = 2n + 1$. Zwei aufeinanderfolgende Quadratzahlen $Q < Q'$ haben daher die Differenz $Q' - Q = 2\sqrt{Q} - 1$.

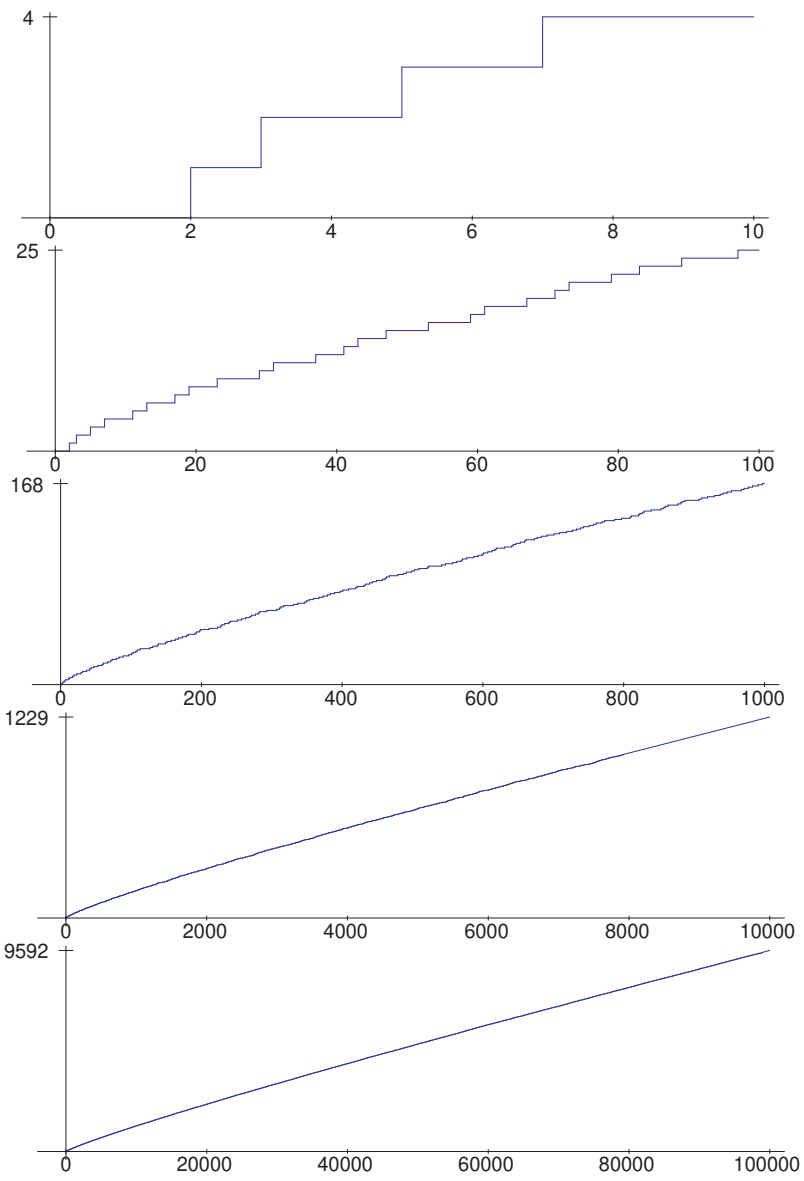
Bei den Primzahlen ist die Situation leider sehr viel unübersichtlicher: EULER meinte sogar, die Verteilung der Primzahlen sei ein Geheimnis, das der menschliche Verstand nie erfassen werde. Der kleinstmögliche Abstand zwischen zwei verschiedenen Primzahlen ist offensichtlich eins, der Abstand zwischen zwei und drei. Er kommt nur an dieser einen Stelle vor, denn außer der Zwei sind schließlich alle Primzahlen ungerade.

Der Abstand zwei ist schon deutlich häufiger: Zwei ist beispielsweise der Abstand zwischen drei und fünf, aber auch der zwischen den Primzahlen $10^{100} + 35737$ und $10^{100} + 35739$. Seit langer Zeit wird vermutet, daß es unendlich viele solcher *Primzahlzwillinge* gibt; experimentelle Untersuchungen deuten sogar darauf hin, daß ihre Dichte für Zahlen der Größenordnung n bei ungefähr $1 : (\log n)^2$ liegen sollte, aber bislang konnte noch niemand auch nur beweisen, daß es unendlich viele gibt.

Eine obere Grenze für den Abstand zwischen zwei aufeinanderfolgenden Primzahlen gibt es genauso wenig wie bei den Quadratzahlen: Ist $n \geq 2$ und $2 \leq i \leq n$, so ist die Zahl $n! + i$ durch i teilbar und somit keine Primzahl. Der Abstand zwischen der größten Primzahl kleiner oder gleich $n! + 1$ und ihrem Nachfolger ist somit mindestens n.

Um einen ersten Eindruck von der Verteilung der Primzahlen zu bekommen, betrachten wir den Graphen der Funktion

$$\pi\colon \begin{cases} \mathbb{R}_{>0} \to \mathbb{N}_0 \\ \quad x \mapsto \text{Anzahl der Primzahlen} \leq x \end{cases}.$$



Die Abbildungen auf der vorigen Seite zeigen ihn für die Intervalle von null bis 10^i für $i = 1, \ldots, 5$. Wie man sieht, werden die Graphen immer glatter, und bei den beiden letzten Bilder könnte man glauben, es handle sich um den Graphen einer differenzierbaren Funktion; daher auch die Schreibweise $\pi(x)$ statt – wie bisher – $\pi(N)$.

Auf den ersten Blick sieht diese Funktion fast linear aus.; sieht man sich allerdings die Zahlenwerte genauer an, so sieht man schnell, daß $\pi(x)$ etwas langsamer wächst als eine lineare Funktion; die Funktion $x/\log x$ ist eine deutlich bessere Approximation. In der Tat können wir auch mit unseren sehr elementaren Mitteln eine entsprechende Aussage beweisen:

Satz: Es gibt Konstanten $c_1, c_2 > 0$, so daß gilt:

$$c_1 \frac{x}{\log x} < \pi(x) < c_2 \frac{x}{\log x} .$$

Beweis: Wir betrachten die neue Funktion

$$\vartheta(x) = \sum_{p \le x} \log p ,$$

wobei ein Summationsindex p hier wie stets in diesem Beweis bedeuten soll, daß wir über alle *Primzahlen* mit der jeweils angegebenen Eigenschaft summieren.

Dann ist einerseits

$$\pi(x) = \sum_{p \le x} \frac{\log p}{\log p} \ge \sum_{p \le x} \frac{\log p}{\log x} = \frac{\vartheta(x)}{\log x} ,$$

andererseits ist

$$\begin{aligned} \vartheta(x) = \sum_{p \le x} \log p &\ge \sum_{\sqrt{x} < p \le x} \log p > \log\left(\sqrt{x}\right) \left(\pi(x) - \pi\left(\sqrt{x}\right)\right) \\ &= \frac{1}{2} \log(x) \left(\pi(x) - \pi\left(\sqrt{x}\right)\right) \end{aligned}$$

und damit auch $\pi(x) < \dfrac{2\vartheta(x)}{\log x} + \pi\left(\sqrt{x}\right) < \dfrac{2\vartheta(x)}{\log x} + \sqrt{x}$. Wenn wir also zeigen können

1. Es gibt Konstanten $c_1, c_3 > 0$, so daß $c_1 x < \vartheta(x) < c_3 x$
2. $\displaystyle\sum_{p\le x} \frac{\log p}{p} = \log x + O(1)$,

dann folgt die Behauptung des Satzes.

Zum Beweis der ersten Aussage betrachten wir die Primzerlegung

$$n! = \prod_{p\le n} p^{e_p}$$

von $n!$. Unter den natürlichen Zahlen bis n sind $\left[\frac{n}{p}\right]$ durch p teilbar, $\left[\frac{n}{p^2}\right]$ durch p^2, *usw.;* daher ist

$$e_p = \sum_{k\ge 1} \left[\frac{n}{p^k}\right] \quad \text{und} \quad \log n! = \sum_{p\le n} e_p \log p = \sum_{p\le n}\sum_{k\ge 1} \left[\frac{n}{p^k}\right] \log p\,.$$

Die Summanden mit $k > 1$ liefern dabei nur einen kleinen Beitrag:

$$\sum_{p\le n}\sum_{k\ge 2} \left[\frac{n}{p^k}\right] \log p \le \sum_{p\le n} \left(\log p \cdot \sum_{k\ge 2} \frac{n}{p^k} \right) = n \sum_{p\le n} \frac{\log p}{p(p-1)}$$

nach der Summenformel für die geometrische Reihe:

$$\sum_{k\ge 2} \frac{1}{p^k} = \frac{1}{p^2} \cdot \frac{1}{1-\frac{1}{p}} = \frac{1}{p^2-p} = \frac{1}{p(p-1)}\,.$$

Zur weiteren Abschätzung ersetzen wir die Summe über alle Primzahlen kleiner oder gleich n durch die Summe über alle Zahlen bis n und beachten, daß für reellen $x \ge 2$ gilt $\log x < \sqrt{x}$, also

$$\frac{\log x}{x(x-1)} < \frac{\sqrt{x}}{x^2} = \frac{1}{x^{3/2}}:$$

$$\sum_{p\le n} \frac{\log p}{p(p-1)} = \sum_{i=1}^{n} \frac{\log i}{i(i-1)} = \sum_{i=2}^{n} \frac{\log i}{i(i-1)} \le \sum_{i=2}^{n} \frac{1}{i^{3/2}}\,.$$

Da $\sum_{i=1}^{\infty} \frac{1}{i^s}$ für alle $s > 1$ konvergiert, konvergiert die rechts stehende Summe für $n \to \infty$ gegen einen endlichen Wert (ungefähr 1,612375), ist also $O(1)$, und damit ist $\sum_{k\ge 2} \frac{1}{p^k} = O(n)$. Setzen wir dies in die Formel

für $\log n!$ ein, erhalten wir nach allen bislang bewiesenen Abschätzungen, daß

$$\log n! = \sum_{p\le n} \left[\frac{n}{p}\right] \log p + O(n)\,.$$

Dies können wir vergleichen mit der STIRLINGschen Formel

$$\log n! = n \log n - n + O(\log n)\,,$$

deren Beweis für Leser, die ihn noch nicht kennen, im Anhang zu diesem Paragraphen skizziert ist. Kombinieren wir dies mit der gerade bewiesenen Formel, ist also

$$\sum_{p\le n} \left[\frac{n}{p}\right] \log p = n \log n + O(n)\,. \qquad (*)$$

Damit ist

$$\begin{aligned}\sum_{p\le 2n} \left(\left[\frac{2n}{p}\right] - 2\left[\frac{n}{p}\right] \right) \log p &= 2n \log 2n - 2n \log n + O(2n) \\ &= 2n \log 2 + O(n) = O(n)\,.\end{aligned}$$

Hier ist $\left[\frac{2n}{p}\right] - 2\left[\frac{n}{p}\right]$ stets entweder null oder eins; speziell für die Primzahlen p mit $n < p < 2n$ ist $\left[\frac{n}{p}\right] = 0$ und $\left[\frac{2n}{p}\right] = 1$. Somit ist

$$\vartheta(2n) - \vartheta(n) = \sum_{n<p<2n} \log p \le \sum_{p\le 2n} \left(\left[\frac{2n}{p}\right] - 2\left[\frac{n}{p}\right] \right) \log p = O(n)\,.$$

Die Formel $\vartheta(2n) - \vartheta(n) = O(n)$ bleibt gültig, wenn wir n durch eine reelle Zahl x ersetzen; somit ist

$$\vartheta(x) = \sum_{i=0}^{\infty} \left(\vartheta\left(\frac{x}{2^i}\right) - \vartheta\left(\frac{x}{2^{i+1}}\right) \right) = O\left(\sum_{i=0}^{\infty} \frac{x}{2^i} \right) = O(x)\,,$$

womit die obere Schranke für $\vartheta(x)$ bewiesen wäre.

Bevor wir uns der unteren Schranke zuwenden, beweisen wir zunächst die zweite Aussage. Natürlich ist $\frac{n}{p} = \left[\frac{n}{p}\right] + O(1)$, also ist nach $(*)$

$$\begin{aligned}\sum_{p\le n} \frac{n}{p} \log p &= \sum_{p\le n} \left[\frac{n}{p}\right] \log p + O\left(\sum_{p\le n} \log p \right) \\ &= n \log n + O(n) + O\left(\vartheta(n)\right) = n \log n + O(n)\,,\end{aligned}$$

denn wie wir gerade gesehen haben ist $\vartheta(n) = O(n)$. Kürzen wir die obige Formel durch n, erhalten wir die gewünschte Aussage

$$\sum_{p\le n} \frac{\log p}{p} = \log n + O(1)\,,$$

die natürlich auch dann gilt, wenn wir n durch eine reelle Zahl x ersetzen: Der Term $O(1)$ schluckt alle dabei auftretenden zusätzlichen Fehler.

Für $0 < \alpha < 1$ ist daher

$$\sum_{\alpha x<p\le x} \frac{\log p}{p} = \log x - \log \alpha x + O(1) = \log \frac{1}{\alpha} + O(1)\,,$$

wobei der Fehlerterm $O(1)$ nicht von α abhängt.

Da $\log \frac{1}{\alpha}$ für $\alpha \to 0$ gegen ∞ geht, ist für hinreichend kleine Werte von α und $x > c/\alpha$ für irgendein $c > 2$ beispielsweise

$$\sum_{\alpha x<p\le x} \frac{\log p}{p} > 10\,,$$

und für solche Werte von α und c ist dann

$$10 < \sum_{\alpha x<p\le x} \frac{\log p}{p} \le \frac{1}{\alpha x} \sum_{\alpha x<p\le x} \log p \le \frac{\vartheta(x)}{\alpha x}\,.$$

Somit ist $10\alpha x < \vartheta(x)$, womit auch die untere Schranke aus der ersten Behauptung bewiesen wäre und damit der gesamte Satz. ∎

Der bewiesene Satz ist nur ein schwacher Abglanz dessen, was über die Funktion $\pi(x)$ bekannt ist. Zum Abschluß des Kapitels seien kurz einige der wichtigsten bekannten und vermuteten Eigenschaften von $\pi(x)$ zusammengestellt. Diese knappe Übersicht folgt im wesentlichen dem Artikel *Primzahlsatz* aus

> DAVID WELLS: Prime Numbers – The Most Mysterious Figures in Math, *Wiley,* 2005,

einer Zusammenstellung im Lexikonformat von interessanten Tatsachen und auch bloßen Kuriosa aus dem Umkreis der Primzahlen.

GAUSS kam 1792, im Alter von 15 Jahren also, durch seine Experimente zur Vermutung, daß $\pi(x)$ ungefähr gleich dem sogenannten *Integrallogarithmus* von x sein sollte:

$$\pi(x) \approx \mathrm{Li}(x) \underset{\text{def}}{=} \int_2^x \frac{d\xi}{\log \xi}\,.$$

Auch LEGENDRE versuchte, $\pi(x)$ anhand experimenteller Daten anzunähern. Er stellte dazu eine Liste aller Primzahlen bis 400 000 zusammen, das sind immerhin 33 860 Stück, und suchte eine glatte Kurve, die den Graphen von π möglichst gut annähert. In seinem 1798 erschienenen Buch *Essai sur la théorie des nombres* gab er sein Ergebnis an als

$$\pi(x) \approx \frac{x}{\log x - 1{,}08366}\,.$$

Über ein halbes Jahrhundert später gab es den ersten Beweis einer Aussage: PAFNUTIJ L'VOVIČ ČEBYŠĒV (1821–1894), in der Numerik meist bekannt in der Schreibweise Tschebytscheff, zeigte 1851: *Falls*

$$\lim_{x\to\infty} \frac{\pi(x)}{x/\log x}$$

existiert, dann muß er den Wert eins haben.

1852 bewies er dann ein deutlich schärferes Resultat: Für *hinreichend große* Werte von x ist

$$c_1 \cdot \frac{x}{\log x} < \pi(x) < c_2 \cdot \frac{x}{\log x} \quad \text{mit} \quad c_1 \approx 0{,}92 \ \text{ und } \ c_2 \approx 1{,}105\,.$$

1896 schließlich zeigten der französische Mathematiker JACQUES SALOMON HADAMARD (1865–1963) und sein belgischer Kollege CHARLES JEAN GUSTAVE NICOLAS BARON DE LA VALLÉE POUSSIN (1866–1962) unabhängig voneinander die Aussage, die heute als **Primzahlsatz** bekannt ist:

$$\pi(x) \sim \frac{x}{\log x}\,.$$

Dies bedeutet nun freilich nicht, daß damit die Formeln von GAUSS und von LEGENDRE überflüssig wären: Die Tatsache, daß der Quotient zweier

Funktionen asymptotisch gleich eins ist, erlaubt schließlich immer noch beträchtliche Unterschiede zwischen den beiden Funktionen: Nur der *relative* Fehler muß gegen null gehen.

Offensichtlich ist für jedes $a \in \mathbb{R}$

$$\lim_{x\to\infty} \frac{x/\log x}{x/(\log x - a)} = \lim_{x\to\infty} \frac{\log x - a}{\log x} = 1 - \lim_{x\to\infty} \frac{a}{\log x} = 1\,,$$

und es ist auch nicht schwer zu zeigen, daß

$$\lim_{x\to\infty} \frac{x/\log x}{\mathrm{Li}(x)} = 1$$

ist. Nach dem Primzahlsatz ist daher auch für jedes $a \in \mathbb{R}$

$$\pi(x) \sim \frac{x}{\log x - a} \quad \text{und} \quad \pi(x) \sim \mathrm{Li}(x)\,.$$

Wie DE LA VALLÉE POUSSIN zeigte, liefert der Wert $a = 1$ unter allen reellen Zahlen a die beste Approximation an $\pi(x)$, aber $\mathrm{Li}(x)$ liefert eine noch bessere Approximation. Für kleine Werte von x sieht man das auch in der folgenden Tabelle, in der alle reellen Zahlen zur nächsten ganzen Zahl gerundet sind. Wie kaum anders zu erwarten, liefert LEGENDREs Formel für 10^4 und 10^5 die besten Werte:

n	$\pi(n)$	$\frac{n}{\log n}$	$\frac{n}{\log n - 1}$	$\frac{n}{\log n - 1{,}08366}$	$\mathrm{Li}(n)$
10^3	168	145	169	172	178
10^4	1 229	1 086	1 218	1 231	1 246
10^5	9 592	8 686	9 512	9 588	9 630
10^6	78 489	72 382	78 030	78 534	78 628
10^7	664 579	620 420	661 459	665 138	664 918
10^8	5 761 455	5 428 681	5 740 304	5 769 341	5 762 209
10^9	50 847 478	48 254 942	50 701 542	50 917 519	50 849 235

Wenn wir genaue Aussagen über $\pi(x)$ machen wollen, sollten wir also etwas über die Differenz $\mathrm{Li}(x) - \pi(x)$ wissen. Hier kommen wir in das Reich der offenen Fragen, und nach derzeitigem Verständnis hängt alles ab von der oben erwähnten RIEMANNschen Zetafunktion. Nach einer berühmten Vermutung von RIEMANN haben alle nichttrivialen Nullstellen von $\zeta(s)$ den Realteil ein halb. Falls dies stimmt, ist

$$\pi(x) = \mathrm{Li}(x) + O(\sqrt{x}\log x)\,.$$

Die RIEMANNsche Vermutung ist eines der wichtigsten ungelösten Probleme der heutigen Mathematik; sie war 1900 eines der HILBERTschen Probleme und ist auch eines der sieben *Millennium problems,* für deren Lösung das CLAY Mathematics Institute in Cambridge, Mass. 2000 einen Preis von jeweils einer Million Dollar ausgesetzt hat; für Einzelheiten siehe `http://www.claymath.org/millennium/` .

Anhang: Die Eulersche Summenformel und die Stirlingsche Formel

Die EULERsche Summenformel erlaubt es, eine endliche Summe auf ein Integral zurückzuführen und dadurch in vielen Fällen erst rechnerisch handhabbar zu machen. Wir betrachten eine reellwertige differenzierbare Funktion f, deren Definitionsbereich das Intervall $[1,\, n]$ enthält.

Für eine reelle Zahl x bezeichnen wir weiterhin mit $[x]$ die größte ganze Zahl kleiner oder gleich x; außerdem führen wir noch die Bezeichnung $\{x\} \underset{\text{def}}{=} x - [x]$ ein für den gebrochenen Anteil von x. Für eine ganze Zahl k ist somit $\{x\} = x - k$ für alle x aus dem Intervall $[k,\, k+1)$.

Partielle Integration führt auf die Gleichung

$$\int_k^{k+1} \big(\{x\} - \tfrac{1}{2}\big) f'(x)\,dx = \big(x - k - \tfrac{1}{2}\big) f(x) \Big|_k^{k+1} - \int_k^{k+1} f(x)\,dx$$

$$= \frac{f(k+1) + f(k)}{2} - \int_k^{k+1} f(x)\,dx\,.$$

Addition aller solcher Gleichungen von $k = 1$ bis $k = n - 1$ liefert

$$\int_1^n \big(\{x\} - \tfrac{1}{2}\big) f'(x)\,dx = \frac{f(1)}{2} + \sum_{k=2}^{n-1} f(k) + \frac{f(n)}{2} - \int_1^n f(x)\,dx\,,$$

womit man die Summe der $f(k)$ berechnen kann:

Satz (EULERsche Summenformel): Für eine differenzierbare Funktion $f\colon D \to \mathbb{R}$, deren Definitionsbereich das Intervall $[1,\, n]$ umfaßt,

ist

$$\sum_{k=1}^{n} f(k) = \int_1^n f(x)\,dx + \frac{f(1)+f(n)}{2} + \int_1^n \left(\{x\} - \tfrac{1}{2}\right) f'(x)\,dx\,.$$

∎

Für die Abschätzung von $n!$ interessiert uns speziell der Fall, daß $f(x) = \log x$ der natürliche Logarithmus ist; hier wird die EULERsche Summenformel zu

$$\begin{aligned}\log n! &= \int_1^n \log x\,dx + \frac{\log n}{2} + \int_1^n \frac{\{x\}-\frac{1}{2}}{x}\,dx \\ &= x(\log x - 1)\Big|_1^n + \frac{\log n}{2} + \int_1^n \frac{\{x\}-\frac{1}{2}}{x}\,dx \\ &= n(\log n - 1) + 1 + \frac{\log n}{2} + \int_1^n \frac{\{x\}-\frac{1}{2}}{x}\,dx\,.\end{aligned}$$

In dieser Formel stört noch das rechte Integral; dieses können wir wie folgt abschätzen: Für eine natürliche Zahl k ist

$$\begin{aligned}\int_k^{k+1} \frac{\{x\}-\frac{1}{2}}{x}\,dx &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{x}{k+\frac{1}{2}+x}\,dx \\ = \int_0^{\frac{1}{2}} \left(\frac{x}{k+\frac{1}{2}+x} - \frac{x}{k+\frac{1}{2}-x}\right) dx &= \int_0^{\frac{1}{2}} \frac{-2x^2}{\left(k+\frac{1}{2}\right)^2 - x^2}\,dx\,.\end{aligned}$$

Im Intervall von 0 bis $\frac{1}{2}$ ist der Integrand monoton fallend, d.h.

$$0 \geq \frac{-2x^2}{\left(k+\frac{1}{2}\right)^2 - x^2} \geq \frac{-\frac{1}{2}}{\left(k+\frac{1}{2}\right)^2 - \frac{1}{4}} = \frac{-2}{(2k+1)^2 - 1} \geq -\frac{1}{2k^2}\,,$$

und damit ist

$$0 \geq \int_k^{k+1} \frac{\{x\}-\frac{1}{2}}{x}\,dx = \int_0^{\frac{1}{2}} \frac{-2x^2}{\left(k+\frac{1}{2}\right)^2 - x^2}\,dx \geq -\frac{1}{4k^2}\,,$$

denn wir können das Integral abschätzen durch das Produkt aus der Länge des Integrationsintervalls und dem Minimum des Integranden. Summation von $k = 1$ bis $n-1$ schließlich gibt die Abschätzung

$$0 \geq \int_1^n \frac{\{x\}-\frac{1}{2}}{x}\,dx \geq -\sum_{k=1}^{n-1} \frac{1}{4k^2} > -\frac{1}{4}\sum_{k=1}^{\infty} \frac{1}{k^2}$$

für das störende Integral aus der obigen Formel. Da die Summe rechts konvergiert, konvergiert auch das Integral für $n \to \infty$ gegen einen Grenzwert I. Somit ist

$$\log n! = n(\log n - 1) + \frac{\log n}{2} + I + 1 + o(1)\,,$$

also folgt insbesondere die Abschätzung

$$\log n! = n \log n + O(n)\,,$$

die wir im Beweis des Satzes über $\pi(n)$ verwendet haben.

§2: Das Sieb des Eratosthenes

Das klassische Verfahren zur Bestimmung aller Primzahlen unterhalb einer bestimmten Schranke geht zurück auf ERATOSTHENES im dritten vorchristlichen Jahrhundert. Es funktioniert folgendermaßen:

Um alle Primzahlen kleiner oder gleich einer Zahl N zu finden, schreibe man zunächst die Zahlen von Eins bis N in eine Reihe.

Eins ist nach Definition keine Primzahl – für griechische Mathematiker wie EUKLID war die Eins nicht einmal eine Zahl. Also streichen wir die Eins durch. Die Zwei ist prim, aber ihre echten Vielfachen sind natürlich keine Primzahlen, werden also durchgestrichen. Dazu müssen wir nicht von jeder Zahl nachprüfen, ob sie durch zwei teilbar ist, sondern wir streichen einfach nach der Zwei jede zweite Zahl aus der Liste durch.

Die erste nichtdurchgestrichene Zahl der Liste ist dann die Drei. Sie muß eine Primzahl sein, denn hätte sie einen von eins verschiedenen kleineren Teiler, könnte das nur die Zwei sein, und alle Vielfachen von zwei (außer der Zwei selbst) sind bereits durchgestrichen.

Auch die echten Vielfachen der Drei sind keine Primzahlen, werden also durchgestrichen. Auch dazu streichen wir wieder einfach jede dritte Zahl aus der Liste durch, unabhängig davon, ob sie bereits durchgestrichen ist oder nicht. (Alle durch sechs teilbaren Zahlen sind offensichtlich schon durchgestrichen.)

Genauso geht es weiter mit der Fünf *usw.;* nach jedem Durchgang durch die Liste muß offenbar die erste noch nicht durchgestrichene Zahl eine Primzahl sein, denn alle Vielfache von kleineren Primzahlen sind bereits durchgestrichen, und wenn eine Zahl überhaupt einen echten Teiler hat, dann ist sie natürlich auch durch eine echt kleinere Primzahl teilbar.

Wie lange müssen wir dieses Verfahren durchführen? Wenn eine Zahl x Produkt zweier echt kleinerer Faktoren u, v ist, können u und v nicht beide größer sein als $\sqrt{x}$: Sonst wäre schließlich $x = uv$ größer als x. Also ist einer der beiden Teiler u, v kleiner oder gleich $\sqrt{x}$, so daß x mindestens einen Teiler hat, dessen Quadrat kleiner oder gleich x ist. Damit ist eine zusammengesetzte Zahl x durch mindestens eine Primzahl p teilbar mit $p^2 \leq x$.

ERATOSTHENES (Ερατοσθένες) wurde 276 v.Chr. in Cyrene im heutigen Libyen geboren, wo er zunächst von Schülern des Stoikers ZENO ausgebildet wurde. Danach studierte er noch einige Jahre in Athen, bis ihn 245 der Pharao PTOLEMAIOS III als Tutor seines Sohns nach Alexandrien holte. 240 wurde er dort Bibliothekar der berühmten Bibliothek im Museion.
Heute ist er außer durch sein Sieb vor allem durch seine Bestimmung des Erdumfangs bekannt. Er berechnete aber auch die Abstände der Erde von Sonne und Mond und entwickelte einen Kalender, der Schaltjahre enthielt. 194 starb er in Alexandrien, nach einigen Überlieferungen, indem er sich, nachdem er blind geworden war, zu Tode hungerte.

Für das Sieb des ERATOSTHENES, angewandt auf die Zahlen von eins

bis N heißt das, daß wir aufhören können, sobald die erste nichtdurchgestrichene Zahl p ein Quadrat $p^2 > N$ hat; dann können wir sicher sein, daß jede zusammengesetzte Zahl $x \leq N$ bereits einen kleineren Primteiler als p hat und somit bereits durchgestrichen ist. Die noch nicht durchgestrichenen Zahlen in der Liste sind also Primzahlen.

Damit lassen sich leicht von Hand alle Primzahlen bis hundert finden, mit etwas Fleiß auch die bis Tausend, aber sicher nicht die hundertstelligen.

Trotzdem kann uns ERATOSTHENES helfen, zumindest zu zeigen, daß gewissen Zahlen nicht prim sind: Wenn wir Primzahlen in einem Intervall $[a,\ b]$ suchen, d.h. also Primzahlen p mit

$$a \leq p \leq b\,,$$

so können wir ERATOSTHENES auf dieses Intervall fast genauso anwenden wie gerade eben auf das Intervall $[1,\ N]$:

Wir gehen aus von einer Liste $p_1, \ldots, p_r$ der ersten Primzahlen; dabei wählen wir r so, daß die Chancen auf nicht durch p_r teilbare Zahlen im Intervall $[a,\ b]$ noch einigermaßen realistisch sind, d.h. wir gehen bis zu einer Primzahl p_r, die ungefähr in der Größenordnung der Intervallänge $b - a$ liegt.

Nun können wir mit jeder der Primzahlen p_i sieben wie im klassischen Fall; wir müssen nur wissen, wo wir anfangen sollen.

Dazu berechnen wir für jedes p_i den Divisionsrest $r_i = a \bmod p_i$. Dann ist $a - r_i$ durch p_i teilbar, liegt allerdings nicht im Intervall $[a,\ b]$. Die erste Zahl, die wir streichen müssen, ist also $a - r_i + p_i$, und von da an streichen wir einfach, ohne noch einmal dividieren zu müssen, wie gehabt jede p_i-te Zahl durch.

Was nach r Durchgängen noch übrig bleibt, sind genau die Zahlen aus $[a,\ b]$, die durch keine der Primzahlen p_i teilbar sind. Sie können zwar noch größere Primteiler haben, aber wichtig ist, daß wir mit minimalem Aufwand für den Großteil aller Zahlen aus $[a,\ b]$ gesehen haben, daß sie keine Primzahlen sind. Für den Rest brauchen wir andere Verfahren, aber die sind allesamt erheblich aufwendiger als ERATOSTHENES, so daß sich diese erste Reduktion auf jeden Fall lohnt.

§3: Fermat-Test und Fermat-Zahlen

Nach dem kleinen Satz von FERMAT gilt für jede Primzahl p und jede nicht durch p teilbare Zahl a die Formel $a^{p-1} \equiv 1 \bmod p$. Im Umkehrschluß folgt sofort:

Falls für eine natürliche Zahl $1 \le a \le p-1$ *gilt* $a^{p-1} \not\equiv 1 \bmod p$, *kann* p *keine Primzahl sein.*

Beispiel: Ist $p = 129$ eine Primzahl? Falls ja, ist nach dem kleinen Satz von FERMAT $2^{128} \equiv 1 \bmod 129$. Tatsächlich ist aber

$$2^7 = 128 \equiv -1 \bmod 129\,,$$

also hat die Zwei in $(\mathbb{Z}/129)^\times$ die Ordnung 14. Da 14 kein Teiler von 128 ist, kann $2^{128} \bmod 129$ nicht eins sein. (Wegen $128 \equiv 2 \bmod 14$ ist $2^{128} \equiv 2^2 = 4 \bmod 129$.) Somit ist 129 keine Primzahl.

Dieses Ergebnis hätten wir natürlich auch durch Probedivisionen leicht gefunden: Da 129 die Quersumme 12 hat, ist die Zahl durch drei teilbar; ihre Primzerlegung ist $129 = 3 \cdot 43$.

Keine Kopfrechenaufgabe ist die Frage, ob $F_{20} = 2^{2^{20}} + 1$ eine Primzahl ist. Falls ja, wäre nach dem kleinen Satz von FERMAT insbesondere

$$3^{F_{20}-1} = 1 \bmod F_{20}, \quad \text{also} \quad 3^{(F_{20}-1)/2} \equiv \pm 1 \bmod F_{20}\,.$$

Nachrechnen zeigt, daß dies nicht der Fall ist, allerdings ist das „Nachrechnen“ bei dieser 315 653-stelligen Zahl natürlich keine Übungsaufgabe für Taschenrechner: 1988 brauchte eine Cray X-MP dazu 82 Stunden, der damals schnellste Supercomputer Cray-2 immerhin noch zehn; siehe

JEFF YOUNG, DUNCAN A. BUELL: The Twentieth Fermat Number is Composite, *Math. Comp.* **50** (1988), 261–263.

Damit war gezeigt, daß F_{20} keine Primzahl ist. (Die anscheinend etwas weltabgewandt lebenden Autoren meinten, dies sei die aufwendigste bis dahin produzierte 1-Bit-Information.)

Umgekehrt können wir leider nicht folgern, daß p eine Primzahl ist, wenn für ein $a \in \mathbb{N}$ mit $1 < a < p-1$ gilt $a^{p-1} \equiv 1 \bmod p$. So ist

beispielsweise $18^{322} \equiv 1 \bmod 323$, aber $323 = 17 \cdot 19$ ist zusammengesetzt. Immerhin gibt es nicht viele $a \le 323$ mit $a^{322} \equiv 1 \bmod 323$: Die einzigen Möglichkeiten sind $a = \pm 1$ und $a = \pm 18$.

Es kann nicht vorkommen, daß für eine zusammengesetzte Zahl n und alle $1 \le a \le n$ gilt $a^{n-1} \equiv 1 \bmod n$, denn ist p ein Primteiler von n, so ist für jedes Vielfache a von p natürlich auch a^{n-1} durch p teilbar, kann also nicht kongruent eins modulo des Vielfachen n von p sein. Zumindest für die a mit $\mathrm{ggT}(a, n) > 1$ kann die Gleichung also nicht erfüllt sein.

Bei großen Zahlen n mit nur wenigen Primfaktoren ist die Chance, ein solches a zu erwischen, recht klein; wenn dies die einzigen Gegenbeispiele sind, wird uns der FERMAT-Test also fast immer in die Irre führen.

Definition: Eine natürliche Zahl n heißt CARMICHAEL-Zahl, wenn sie keine Primzahl ist, aber trotzdem für jede natürliche Zahl a mit $\mathrm{ggT}(a, n) = 1$ gilt: $a^{n-1} \equiv 1 \bmod n$.

ROBERT DANIEL CARMICHAEL (1879–1967) war ein amerikanischer Mathematiker, der unter anderem Bücher über die Relativitätstheorie, über Zahlentheorie, über Analysis und über Gruppentheorie veröffentlichte. Ab 1915 lehrte er an der University of Illinois. Er zeigte 1910, daß 561 die gerade definierte Eigenschaft hat und publizierte auch später noch eine Reihe von Arbeiten über solche Zahlen.

Lemma: Eine CARMICHAEL-Zahl n ist ein Produkt von mindestens drei paarweise verschiedenen ungeraden Primzahlen p_i; für jede davon ist $p_i - 1$ ein Teiler von $n - 1$.

Beweis: Für gerade Zahlen n ist $(n-1)^{n-1} \equiv -1 \bmod n$; somit ist jede CARMICHAEL-Zahl ungerade.

Als nächstes wollen wir uns überlegen, daß n Produkt verschiedener Primzahlen sein muß: Angenommen, n wäre teilbar durch ein Primzahlquadrat p^2. Da $p+1$ teilerfremd zu n ist, muß $(p+1)^{n-1} \equiv 1 \bmod n$ sein, erst recht gilt also $(p+1)^{n-1} \equiv 1 \bmod p^2$. Die Ordnung von $p+1$ in $\left(\mathbb{Z}/p^2\right)^\times$ ist daher ein Teiler von $n-1$.

Nach dem Lemma von LAGRANGE *(Kap. 1, §7)* muß diese Ordnung aber auch die im gleichen Paragraphen berechnete Gruppenordnung

$\varphi(p^2) = p(p-1)$ teilen. Da p ein Teiler von n ist, kann es kein Teiler von $n-1$ sein; jeder echte Teiler von $n-1$ ist daher prim zu p. Somit muß die Ordnung von $p+1$ modulo p^2 ein Teiler von $p-1$ sein, d.h. $(p+1)^{p-1} \equiv 1 \bmod p^2$ und $(p+1)^p \equiv p+1 \bmod p^2$.

Berechnen wir aber

$$(p+1)^p = \sum_{k=0}^{p} \binom{p}{k} p^k$$

nach der binomischen Formel, so sind alle Terme dieser Summe außer dem für $k = 0$ durch p^2 teilbar, d.h. $(p+1)^p \equiv 1 \bmod p^2$. Dies widerspricht der oben hergeleiteten Kongruenz. Somit ist n Produkt verschiedener Primzahlen.

Für jeden Primteiler p von n muß $p-1$ ein Teiler von $n-1$ sein, denn nach dem chinesischen Restesatz ist $(\mathbb{Z}/n)^\times$ das Produkt der Gruppen $(\mathbb{Z}/p)^\times$, es gibt also in $(\mathbb{Z}/n)^\times$ eine primitive Wurzel a modulo p. Da $a^{n-1} \equiv 1 \bmod n$ und damit insbesondere modulo p ist, muß $n-1$ ein Vielfaches der Ordnung $p-1$ von a modulo p sein.

Schließlich müssen wir uns noch überlegen, daß n ein Produkt von mindestens drei Primzahlen ist: Da n nach Definition keine Primzahl ist, wäre $n = pq$ sonst das Produkt zweier Primzahlen. Wie wir gerade gesehen haben, müßte

$$n - 1 = pq - 1 = (p-1)q + (q-1)$$

sowohl durch $p-1$ als auch durch $q-1$ teilbar sein, also müßten $p-1$ und $q-1$ durcheinander teilbar sein, d.h. $p = q$, was wir bereits ausgeschlossen haben. ∎

Als Beispiel können wir ein Produkt $n = (6t+1)(12t+1)(18t+1)$ mit drei primen Faktoren betrachten, z.B.

$$1729 = 7 \times 13 \times 19 \quad \text{für } t = 1 \quad \text{oder} \quad 294409 = 37 \times 73 \times 109$$

für $t = 6$. Hier ist $n - 1 = 1296t^3 + 396t^2 + 36t = 36t \cdot (36t^2 + 11t + 1)$ offensichtlich durch $6t$, $12t$ und $18t$ teilbar, n ist also eine CARMICHAEL-Zahl.

Natürlich muß nicht jede CARMICHAEL-Zahl von dieser Form sein; die kleinste CARMICHAEL-Zahl beispielsweise ist $561 = 3 \cdot 11 \cdot 17$ und hat keinen einzigen Primfaktor kongruent eins modulo sechs.

Eine größte CARMICHAEL-Zahl gibt es nicht, denn nach

W.R. ALFORD, ANDREW GRANVILLE, CARL POMERANCE: There are infinitely many Carmichael numbers, *Ann. Math,* **140** (1994), 703–722

gibt es unendlich viele. Konkret zeigen sie, daß die Anzahl der CARMICHAEL-Zahlen kleiner oder gleich x für große Zahlen x mindestens gleich $x^{2/7}$ ist. Die tatsächliche Anzahl dürfte wohl deutlich größer sein, ist aber immer noch sehr viel kleiner als die der Primzahlen.

Für große Zahlen p wird es zunehmend unwahrscheinlich, daß sie auch nur für ein a den FERMAT-Test bestehen, ohne Primzahl zu sein. Rechnungen von

> SU HEE KIM, CARL POMERANCE: The probability that a Random Probable Prime is Composite, *Math. Comp.* **53** (1989), 721–741

geben folgende obere Schranke für die Wahrscheinlichkeit ε, daß eine zufällig gewählte Zahl p der angegebenen Größenordnung den FERMAT-Test mit einem vorgegebenen a besteht und trotzdem keine Primzahl ist:

$p \approx 10^{60}$	10^{70}	10^{80}	10^{90}	10^{100}
$\varepsilon \le 7{,}16 \cdot 10^{-2}$	$2{,}87 \cdot 10^{-3}$	$8{,}46 \cdot 10^{-5}$	$1{,}70 \cdot 10^{-6}$	$2{,}77 \cdot 10^{-8}$
$p \approx 10^{120}$	10^{140}	10^{160}	10^{180}	10^{200}
$\varepsilon \le 5{,}28 \cdot 10^{-12}$	$1{,}08 \cdot 10^{-15}$	$1{,}81 \cdot 10^{-19}$	$2{,}76 \cdot 10^{-23}$	$3{,}85 \cdot 10^{-27}$
$p \approx 10^{300}$	10^{400}	10^{500}	10^{600}	10^{700}
$\varepsilon \le 5{,}8 \cdot 10^{-29}$	$5{,}7 \cdot 10^{-42}$	$2{,}3 \cdot 10^{-55}$	$1{,}7 \cdot 10^{-68}$	$1{,}8 \cdot 10^{-82}$
$p \approx 10^{800}$	10^{900}	10^{1000}	10^{2000}	10^{3000}
$\varepsilon \le 5{,}4 \cdot 10^{-96}$	$1{,}0 \cdot 10^{-109}$	$1{,}2 \cdot 10^{-123}$	$8{,}6 \cdot 10^{-262}$	$3{,}8 \cdot 10^{-397}$

(Sie geben natürlich auch eine allgemeine Formel an, jedoch ist diese zu grausam zum Abtippen.)

Selbst wenn wir noch mit 1024-Bit-Moduln arbeiten und somit etwa 155-stellige Primzahlen brauchen, liegt also die Fehlerwahrscheinlichkeit bei nur etwa 10^{-15}; falls man sie erniedrigen möchte, testet man einfach mit mehreren zufällig gewählten Basen und hat dann etwa bei zwei verschiedenen Basen eine Wahrscheinlichkeit von höchsten etwa 10^{-30}, daß beide Tests das falsche Ergebnis liefern. Dies sollte für die meisten Anwendungen genügen: Die Bundesnetzagentur empfiehlt bei probabilistischen Primzahltests eine Irrtumswahrscheinlichkeit von höchsten $2^{-80} \approx 8{,}27 \cdot 10^{-25}$ zuzulassen, die hier deutlich unterschritten wäre. Ab etwa zweihundertstelligen Primzahlen reicht sogar bereits ein einziger FERMAT-Test.

Einige Leute reden bei Zahlen, die einen FERMAT-Test bestanden haben, von „wahrscheinlichen Primzahlen“. Das ist natürlich Unsinn: Eine Zahl ist entweder *sicher* prim oder *sicher* zusammengesetzt; für Wahrscheinlichkeiten gibt es hier keinen Spielraum. Besser ist der ebenfalls gelegentlich zu hörende Ausdruck „industrial grade primes“, also „Industrieprimzahlen“, der ausdrücken soll, daß wir zwar nicht *bewiesen* haben, daß die Zahl wirklich prim ist, daß sie aber für (manche) „industrielle Anwendungen“ gut genug ist.

Zumindest grundsätzlich läßt sich der FERMAT-Test allerdings auch ausbauen zu einem echten Primtest:

Satz: Ist für zwei natürliche Zahlen p, a zwar $a^{p-1} \equiv 1 \bmod p$, aber für jeden Primteiler q von $p-1$ $a^{(p-1)/q} \not\equiv 1 \bmod p$, so ist p eine Primzahl.

Beweis: Offensichtlich muß dann die Ordnung von a in $(\mathbb{Z}/p\mathbb{Z})^\times$ gleich $p-1$ sein. Wie wir aus Kapitel 1, §8 wissen, hat $(\mathbb{Z}/p\mathbb{Z})^\times$ die Ordnung $\varphi(p)$, und für jede zusammengesetzte Zahl folgt aus der dort angegebenen Formel leicht, daß $\varphi(p) < p-1$ ist. Also muß p prim sein. ∎

Der Nachteil dieses Satzes ist natürlich, daß wir alle Primteiler von $p-1$ kennen müssen; wenn wir Primzahlen mit mehreren hundert Dezimalstellen suchen, ist das meist keine sehr realistische Annahme. Für Zahlen spezieller Bauart kann dieser Test jedoch sehr nützlich sein: Wenn wir

von einer Zahl n mit wenigen, uns bekannten Primteilern ausgehen, läßt sich so testen, ob $n+1$ eine Primzahl ist.

Die einfachsten Kandidaten für n sind natürlich Primzahlpotenzen $n = p^r$; für eine ungerade Primzahl p ist allerdings $n+1$ gerade und somit keine Primzahl. Auf den Fall $p = 2$ werden wir gleich zurückkommen.

Bei kleinen Zahlen geraden Zahlen b mit wenigen Primfaktoren gibt es gelegentlich Chancen, daß $b^r + 1$ eine Primzahl ist, allerdings kommt auch das nur selten vor. Ein Beispiel wäre etwa

$$p = 24^4 + 1 = 331\,777\,.$$

Hier hat $p - 1 = 24^4$ nur 2 und 3 als Primteiler, wir müssen also eine Zahl a finden, so daß

$$a^{p-1} \equiv 1 \bmod p, \quad a^{(p-1)/2} \not\equiv 1 \bmod p \quad \text{und} \quad a^{(p-1)/3} \not\equiv 1 \bmod p$$

ist. Dazu berechnen wir am besten zunächst $x = a^{(p-1)/6} \bmod p$, sodann

$$y = a^{(p-1)/3} \bmod p = x^2 \bmod p \quad \text{und} \quad z = a^{(p-1)/3} \bmod p = x^3 \bmod p\,.$$

Wenn p eine Primzahl ist, muß offensichtlich $z \equiv -1 \bmod p$ sein, und wenn dies gilt, ist auch $a^{(p-1)} \equiv 1 \bmod p$.

Für $a = 2$ erhalten wir $x = 92553$, $y = 239223$ und $z = 1$; das beweist nichts. Für $a = 3$ ist sogar bereits $x = 1$, aber für $a = 5$ wird $x = 92554$, $y = 92553$ und $z = 331776 \equiv -1 \bmod p$. Dies *beweist,* daß p eine Primzahl ist.

Wie sieht es aus mit den Zahlen 2^r+1? Falls $r > 1$ eine ungerade Zahl ist, muß diese Zahl durch drei teilbar sein, denn $2^r \equiv 2 \equiv -1 \bmod 3$. Auch wenn r nur durch eine ungerade Zahl $u > 1$ teilbar ist, kann $2^r + 1$ nicht prim sein, denn ist $r = uv$, so ist $2^r = (2^v)^u \equiv (-1)^u \equiv -1 \bmod 2^v + 1$, so daß $2^v + 1$ ein nichttrivialer Teiler ist. Die einzigen Kandidaten für Primzahlen sind daher die Zahlen

$$F_n = 2^{2^n} + 1\,,$$

bei denen der Exponent r eine Zweierpotenz ist. Sie heißen FERMAT-Zahlen, weil FERMAT zwischen 1630 und 1640 in mehreren Briefen, unter anderem an PASCAL und an MERSENNE, die Vermutung äußerte, diese Zahlen seien allesamt prim.

Für $F_0 = 3, F_1 = 5, F_2 = 17$ sieht man das mit bloßem Auge und mit auch $F_3 = 257$ gibt es keine nennenswerten Schwierigkeiten. Für größere Werte von n können wir den obigen Test anwenden; da 2 der einzige Primteiler von $F_n - 1$ ist, wird er hier einfach einfach zur folgenden Aussage:

Lemma: F_n ist genau dann eine Primzahl, wenn es ein a gibt mit

$$a^{(F_n-1)/2} \equiv -1 \bmod F_n\,.$$

■

Für $F_4 = 2^{16} + 1 = 65\,537$ ist $(F_n - 1)/2 = 2^{15}$, wir müssen also ein a finden, so daß $a^{2^{15}} \equiv -1 \bmod F_4$ ist. Für $a = 2$ bekommen wir nach 15 Quadrierungen das nutzlose Ergebnis eins, für $a = 3$ aber die gewünschte -1. Damit ist F_4 als Primzahl erkannt – was auch FERMAT wußte.

Für $F_5 = 2^{32} + 1 = 429\,4967\,297$ müssen wir $a^{2^{31}} \bmod F_5$ berechnen, brauchen also schon 31 Quadrierungen. Für $a = 2$ erhalten wir wieder die nutzlose Eins, für $a = 3$ aber das Ergebnis 10 324 303, das sich offenbar auch modulo F_5 deutlich von ± 1 unterscheidet. Somit ist F_5 *keine* Primzahl und FERMATs obige Vermutung ist widerlegt. Sie ist übrigens die einzige seiner vielen Vermutungen, die sich als falsch erwies.

Der erste, der dies erkannte, war EULER, den GOLDBACH in Sankt Petersburg auf FERMATs Vermutung aufmerksam machte. Nachdem EULERs Beweisversuche erfolglos blieben, fand er 1732 schließlich die Faktorisierung $F_5 = 641 \cdot 6\,700\,417$. Obwohl EULER viel rechnete, fand er diese Faktorisierung natürlich nicht durch systematisches Ausprobieren aller Primzahlen bis zur Quadratwurzel von F_5: dafür hätte er im schlimmsten Fall immerhin durch 6542 Primzahlen dividieren müssen!

Stattdessen benutzte EULER den kleinen Satz von FERMAT, um zunächst Aussagen über mögliche Teiler von FERMAT-Zahlen zu bekommen. Er fand

Lemma: Jeder Primteiler p von F_n ist kongruent eins modulo 2^{n+1}.

Beweis: Aus $2^{2^n} \equiv -1 \bmod F_n$ folgt insbesondere, daß $2^{2^n} \equiv -1 \bmod p$ ist, also $2^{2^{n+1}} \equiv 1 \bmod p$. Die Ordnung der Zwei in $\mathbb{F}_p^\times$ ist somit ein Teiler von 2^{n+1}, also eine Zweierpotenz. Wäre diese kleiner als 2^{n+1}, wäre sie ein Teiler von 2^n, so daß $2^{2^n} \equiv +1 \bmod p$ sein müßte. Somit ist 2^{n+1} die genaue Ordnung. Diese muß aber nach dem Satz von LAGRANGE ein Teiler der Gruppenordnung sein, d.h. 2^{n+1} teilt $p - 1$, was die Behauptung beweist.

■

Somit wußte EULER, daß jeder Primteiler von F_5 die Form $p = 64k + 1$ haben muß; Primzahlen dieser Form gibt es nur 209 unterhalb von 2^{15}, was das Problem schon viel handhabbarer erscheinen läßt. Dazu kam, daß er Glück hatte: Schon für $k = 10$ bekam er einen Teiler. Nach 193, 257, 449 und 577 ist 641 bereits die fünfte Primzahl dieser Form!

Tatsächlich aber machte er sich trotzdem zuviel Arbeit, denn wie der französische Mathematiker EDOUARD LUCAS (1842–1891) fand, gilt sogar

Lemma: Für $n \geq 3$ ist jeder Primteiler p von F_n kongruent eins modulo 2^{n+2}.

Beweis: Wir gehen genauso vor wie EULER, suchen aber ein Element von $\mathbb{F}_p^\times$, dessen Ordnung 2^{n+2} ist. Ein solches Element haben wir gefunden, wenn wir in $\mathbb{F}_p^\times$ ein x finden mit $x^2 = 2$. Wegen der Beziehung

$$\left(2^{2^{n-1}} + 1\right)^2 = 2^{2^n} + 1 + 2^{2^{n-1}+1} \equiv 2^{2^{n-1}+1} \bmod F_n$$

haben wir zunächst eine Zahl y gefunden mit $y^2 \equiv 2^u \bmod F_n$, wobei $u = 2^{n-1} + 1$ eine ungerade Zahl ist. Mit dem erweiterten EUKLIDischen Algorithmus können wir daher ganze Zahlen v, w finden mit $uv + 2w = 1$. Damit ist auch

$$2 = 2^{uv+2w} = (2^u)^v \cdot (2^w)^2 = y^{2v} \cdot 2^{2w} = (y^v \cdot 2^w)^2$$

ein Quadrat in $\mathbb{F}_p^\times$.

■

Mit dieser Verschärfung seines Lemmas hätte sich EULER auf Primzahlen der Form $128k + 1$ beschränken können und hätte bereits im zweiten

Anlauf (nach einem vergeblichen Versuch mit 257) seinen Faktor gefunden.

Auch wenn er nie etwas darüber publiziert hat, hätte er auch beweisen können und bewies vielleicht auch, daß sein Kofaktor $q = 6\,700\,417$ ebenfalls eine Primzahl ist: Da jeder Primteiler p von q insbesondere auch F_5 teilt, muß $p \equiv 1 \bmod 128$ sein, und wenn es einen echten Teiler gibt, gibt es auch einen der kleiner ist als $\sqrt{q} \approx 2588{,}5$. Es gibt nur zwanzig Zahlen der Form $128k + 1$ unterhalb von $\sqrt{q}$, und nur fünf davon sind prim: Neben den bereits bekannten Kandidaten 257 und 641 sind das noch 769, 1153 und 1409. Fünf einfache Divisionen mit Rest zeigen, daß q durch keine dieser Zahlen teilbar ist; somit haben wir bewiesen, daß auch q prim sein muß.

Die Suche nach FERMAT-Primzahlen sowie nach Faktoren von FERMAT-Zahlen beschäftigt auch heute noch eine ganze Reihe von Mathematikern; die jeweils neuesten Ergebnisse sind zum Beispiel auf den Webseiten von `www.fermatsearch.org` zu finden. Unter anderem ist bewiesen, daß F_n für $5 \le n \le 32$ sowie viele andere Werte von n zusammengesetzt ist; oft sind auch zumindest einige Faktoren bekannt. FERMATsche Primzahlen mit $n > 4$ wurden bislang keine gefunden, allerdings ist auch noch nicht bewiesen, daß es unendlich viele FERMAT-Zahlen gibt, die *keine* Primzahlen sind.

Der oben angegebene Beweis von LUCAS zeigt übrigens auch, warum wir beim Primzahltest für F_4 mit der Basis zwei keinen Erfolg hatten: Da 2 modulo F_4 ein Quadrat ist, *muß* die Potenz mit Exponent $(F_4 - 1)/2$ gleich eins sein, denn sie ist schließlich die $(F_4 - 1)$-te Potenz der Wurzel aus 2. Aus diesem Grund auch wurde bei der Überprüfung von F_{20} nicht mit $a = 2$ gearbeitet, obwohl dies rechnerisch sehr viel effizienter gewesen wäre. Wie wir im Kapitel über quadratische Reste sehen werden, konnten sich die Autoren *vor* Beginn ihrer Rechnung leicht davon überzeugen, daß drei modulo F_{20} *keine* Quadratzahl ist; der Aufwand für die Entscheidung, ob eine Zahl a Quadrat modulo F_{20} ist, erfordert einen Aufwand, der ungefähr dem für die Anwendung des EUKLIDischen Algerithmus auf a und F_{20} entspricht, im Falle $a = 3$ also nicht viel mehr als die Berechnung von F_{20} mod 3, was man im Kopf als $(-1)^{2^{20}} + 1 = 2$ berechnen kann.

§4: Der Test von Miller und Rabin

Der Test von MILLER und RABIN ist eine etwas strengere Version des Tests von FERMAT: Um zu testen, ob p eine Primzahl sein kann, schreiben wir $p - 1$ zunächst als Produkt $2^n u$ einer Zweierpotenz und einer ungeraden Zahl; sodann berechnen wir $a^u \bmod p$. Falls wir das Ergebnis eins erhalten, ist erst recht $a^{p-1} \equiv 1 \bmod p$, und wir können nicht folgern, daß p zusammengesetzt ist.

Andernfalls quadrieren wir das Ergebnis bis zu n-mal modulo p. Falls dabei nie eine Eins erscheint, folgt nach FERMAT, daß p zusammengesetzt ist. Falls vor der ersten Eins eine von -1 (*bzw.* $p - 1$) verschiedene Zahl erscheint, folgt das auch, denn im Körper $\mathbb{F}_p$ hat die Eins nur die beiden Quadratwurzeln ± 1. In allen anderen Fällen erfahren wir nicht mehr als bei FERMAT.

Algorithmisch funktioniert der Test also folgendermaßen:

Schritt 0: Wähle ein zufälliges a, schreibe $p - 1 = 2^n u$ mit einer ungeraden Zahl u und berechne $b = a^u \bmod p$. Falls dies gleich Eins ist, endet der Algorithmus und wir konnten nicht zeigen, daß p eine zusammengesetzte Zahl ist; sie kann prim sein.

Schritt $i, 1 \le i \le n$**:** Falls $b \equiv -1 \bmod p$, endet der Algorithmus und wir können nicht ausschließen, daß p prim ist. Falls $b = 1$ ist (was frühestens im zweiten Schritt der Fall sein kann), ist p zusammengesetzt und der Algorithmus endet. Andernfalls wird b durch $b^2 \bmod p$ ersetzt und es geht weiter mit Schritt $i + 1$.

Schritt $n + 1$**:** Der Algorithmus endet mit dem Ergebnis, daß p zusammengesetzt ist.

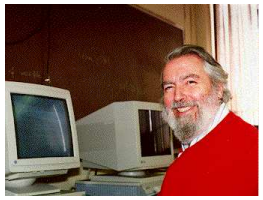
Beispiel: Ist 247 eine Primzahl? Wir wählen $a = 77$, und da $77^{246} \bmod 247 = 1$ ist, können wir mit FERMAT nicht ausschließen, daß 247 prim ist. Da aber $77^{123} \bmod 247 = 77$ ist, sagt uns der Algorithmus von MILLER und RABIN im zweiten Schritt, wenn wir $77^2 \equiv 1 \bmod 247$ betrachten, daß die Zahl zusammengesetzt sein muß.

Hätten wir allerdings mit $a = 87$ gearbeitet, hätten wir im nullten Schritt $87^{123} \equiv 1 \bmod 247$ berechnet und hätten $247 = 13 \cdot 19$ nicht als zusammengesetzt erkannt.

GARY L. MILLER entwickelte diesen Test 1975 im Rahmen seiner Dissertation (in Informatik) an der Universität von Berkeley. Dabei ging es ihm nicht um einen probabilistischen Test, sondern um einen Test, der immer die richtige Antwort liefert. Er konnte zeigen, daß dies hier beim Test von hinreichend vielen geeigneten Basen der Fall ist **vorausgesetzt** die bis heute immer noch offene verallgemeinerte RIEMANN-Vermutung ist richtig. Er lehrte später zunächst einige Jahre an der University of Waterloo, inzwischen an der Carnegie Mellon University. Seine späteren Arbeiten stammen hauptsächlich aus dem Gebiet der rechnerischen Geometrie. www.cs.icmu.edu/∼glmiller

MICHAEL O. RABIN wurde 1931 in Breslau geboren. Die Familie wanderte nach Israel aus, wo er an der hebräischen Universität von Jerusalem Mathematik studierte. Nach seinem Diplom 1953 ging er nach Princeton, wo er 1957 promovierte. Seit 1958 lehrt er an der hebräischen Universität, wo er unter anderem auch Dekan der mathematischen Fakultät und Rektor war. Seit 1983 ist er zusätzlich Inhaber des THOMAS J. WATSON-Lehrstuhls für Informatik an der Harvard University. Seine Forschungen, für die er u.a. 1976 den TURING-Preis erhielt, beschäftigen sich mit der Komplexität mathematischer Operationen und der Sicherheit von Informationssystemen. Seine home page in Harvard ist zu finden unter www.seas.harvard.edu/ourfaculty/profile/Michael_Rabin .

Anscheinend wurde der Test von MILLER und RABIN bereits 1974, also vor MILLERs Veröffentlichung, von SELFRIDGE verwendet; daher sieht man gelegentlich auch die korrektere Bezeichnung *Test von* MILLER, RABIN *und* SELFRIDGE.



Der amerikanische Mathematiker JOHN L. SELFRIDGE promovierte 1958 an der University of California in Los Angelos. Bis zu seiner Emeritierung lehrte er an der Northern Illinois University. Seine Arbeiten befassen sich vor allem mit der analytischen sowie der konstruktiven Zahlentheorie. Vierzehn davon schrieb er mit PAUL ERDŐS. math.niu.edu/faculty/index.php?cmd=detail&id=91

§5: Der Test von Agrawal, Kayal und Saxena

Im August 2002 stellten MANINDRA AGRAWAL, NEERAJ KAYAL und NITIN SAXENA, zwei Bachelor-Studenten am Indian Institute of Technology in Kanpur und ihr Professor, einen Primzahltest vor, der ebenfalls auf dem kleinen Satz von FERMAT beruht, aber (natürlich auf Kosten eines erheblich größeren Aufwands) immer die richtige Antwort liefert; er ist inzwischen erschienen in

> MANINDRA AGRAWAL, NEERAJ KAYAL, NITIN SAXENA: PRIMES is in P, *Annals of Mathematics* **160** (2004), 781-793.

Selbstverständlich war dies nicht der erste Primzahltest, der deutlich schneller als Probedivisionen zeigt, ob eine gegebene Zahl prim ist oder nicht; es ist auch bei weitem nicht der schnellste solche Test. Er hat aber gegenüber anderen solchen Tests zwei Besonderheiten:

1. Zu seinem Verständnis ist – nach einigen in der letzten Zeit gefundenen Vereinfachungen –nur elementare Zahlentheorie notwendig.
2. Es ist der bislang einzige Test, von dem man beweisen kann, daß seine Laufzeit für n-stellige Zahlen durch ein Polynom in n begrenzt werden kann.

Für uns ist vor allem der erste Punkt wichtig; der zweite ist zwar ein für Komplexitätstheoretiker sehr interessantes Ergebnis, hat aber keinerlei praktische Bedeutung: Im Buch

> VICTOR SHOUP: A computational Introduction to Number Theory and Algebra, *Cambridge University Press,* $^{2}2008$ (*Volltext* unter `http://shoup.net/ntb/`),

dem dieser Paragraph im wesentlichen folgt, argumentiert SHOUP, daß alternative Algorithmen, so man sich auf Zahlen von weniger als 2^{256} Bit beschränkt, durch eine vergleichbare Schranke abgeschätzt werden können, und natürlich sind die Zahlen, mit denen wir es üblicherweise zu tun haben, deutlich kleiner. In der Praxis sind die alternativen Algorithmen deutlich schneller.

(2^{256} liegt knapp über 10^{77}; derzeitige Schätzungen für die Anzahl der Nukleonen im Universum liegen bei etwa 10^{80}. Damit ist klar, daß

MANINDRA AGRAWAL erhielt 1986 seinen BTech und 1991 seinen PhD in Informatik am Indian Institute of Technology in Kanpur, wo er –abgesehen von Gastaufenthalten in Madras, Ulm, Princeton und Singapur – seither als Professor lehrt. Seine Arbeiten befassen sich hauptsächlich mit der Komplexität von Schaltungen und von Algorithmen. Für die Arbeit mit KAYAL und SAXENA erhielt er gemeinsam mit diesen unter anderem den GÖDEL-Preis 2006 für die besten Zeitschriftenveröffentlichung auf dem Gebiet der Theoretischen Informatik.

http://www.cse.iitk.ac.in/users/manindra/

NEERAJ KAYAL wurde 1979 geboren. Er erhielt 2002 seinen BTech und 2006 seinen PhD bei MANINDRA AGRAWAL am Indian Institute of Technology in Kanpur. Neuere Arbeiten beschäftigen unter anderem sich mit der Komplexität des Isomorphieproblems bei endlichen Ringen sowie der Lösbarkeit von bivariaten Polynomgleichungen über endlichen Körpern. Nach einem kurzen Aufenthalt an der Rutgers University arbeiter er inzwischen bei Microsoft Research.

http://research.microsoft.com/en-us/people/neeraka/

NITIN SAXENA wurde 1981 geboren. Er erhielt 2002 seinen Bachelor of Technology und 2006 seinen PhD bei MANINDRA AGRAWAL am Indian Institute of Technology in Kanpur. Während der Arbeit an seiner Dissertation über die Anwendung von Ringhomomorphismen auf Fragen der Komplexitätstheorie besuchte er jeweils ein Jahr lang die Universitäten Princeton und Singapur, danach arbeitete er als Postdoc in der Gruppe *Quantum Computing and Advanced Systems Research* am Centrum voor Wiskunde en Informatica in Amsterdam. Derzeit ist er als *Bonn Junior Fellow* am HAUSDORFF CENTER der Universität Bonn. Sein Interesse gilt algorithmischen Verfahren der Algebra und Zahlentheorie sowie Fragen der Komplexitätstheorie.

http://www.math.uni-bonn.de/sinsaxena/

kein Computer, der mit irgendeiner Art von heute üblicher Technologie

arbeitet, je eine solche Zahl speichern kann, geschweige denn damit rechnen.)

Im folgenden wird es daher nur um eine mathematische Betrachtung des Algorithmus von AGRAWAL, KAYAL und SAXENA gehen; für einen (kurzen und elementaren) Beweis der Komplexitätsaussage sei beispielsweise auf das zitierte Buch von SHOUP verwiesen.

Die Grundidee des Algorithmus steckt im folgenden

Satz: $n > 1$ sei eine natürliche Zahl und $a \in \mathbb{N}$ sei dazu teilerfremd. n ist genau dann prim, wenn im Polynomring über $\mathbb{Z}/N$ gilt:

$$(X + a)^n = X^n + a\,.$$

Beweis: Nach dem binomischen Lehrsatz ist

$$(X + a)^n = X^n + a^n + \sum_{i=1}^{n-1} \binom{n}{i} a^i X^{n-i}\,.$$

Für eine Primzahl n gilt nach dem kleinen Satz von FERMAT in $\mathbb{Z}/n$ die Gleichung $a^n = a$. Außerdem ist für $1 \leq i \leq n-1$ der Binomialkoeffizient

$$\binom{n}{i} = \frac{n(n-1)\cdots(n-i+1)}{i!}$$

durch n teilbar, da n Faktor des Zählers, nicht aber des Nenners ist. Somit verschwinden in $\mathbb{Z}/n$ alle diese Binomialkoeffizienten, und die Gleichung aus dem Satz ist bewiesen.

Umgekehrt sei n eine zusammengesetzte Zahl und p ein Primteiler von n. Genauer sei $n = p^k m$ mit einer zu p teilerfremden Zahl m. Dann ist der Zähler von $\binom{n}{p}$ genau durch p^k teilbar, denn die Faktoren $(n-1), \ldots, (n-p+1)$ sind allesamt teilerfremd zu p, und der Nenner ist genau durch p teilbar. Somit ist $\binom{n}{p}$ zwar durch p^{k-1} teilbar, nicht aber durch p^k und damit erst recht nicht durch n. Wenn wir $(X + a)^n$ über $\mathbb{Z}/n$ ausmultiplizieren, kann daher der Summand $\binom{n}{p} a^p X^{n-p}$ nicht verschwinden, und damit kann die Gleichung aus dem Satz nicht gelten. ∎

In dieser Form führt der Satz allerdings noch nicht zu einem praktikablen Primzahltest: Das Ausmultiplizieren von $(X+a)^n$ führt schließlich auf $n+1$ Summanden, der Aufwand ist also proportional zu n und damit vergleichbar damit, daß wir für jede natürliche Zahl $1 < m < n$ nachprüfen, ob n ohne Rest durch m teilbar ist. Die wesentliche neue Idee von AGRAWAL, KAYAL und SAXENA besteht darin zu zeigen, daß es bereits reicht, Gleichungen der im Satz genannten Art modulo einem geeigneten Polynom $X^r - 1$ mit einem relativ kleinen Grad r nachzuprüfen.

Konkret geht ihr Algorithmus folgendermaßen vor:

n sei die zu testende natürliche Zahl und $\ell(n) = [\log_2 n] + 1$ die Anzahl ihrer Binärziffern.

1. Schritt: Stelle sicher, daß n keine Potenz einer anderen natürlichen Zahl ist.

Das läßt sich beispielsweise dadurch bewerkstelligen, daß man die Quadratwurzel, Kubikwurzel *usw.* von n soweit ausrechnet bis man erkennt, daß es sich um keine natürliche Zahl handelt. Der ungünstigste Fall ist offenbar der, daß n eine Zweierpotenz sein könnte; man muß also bis zur $[\log_2 n]$-ten Wurzel gehen.

2. Schritt: Finde die kleinste natürliche Zahl $r > 1$ mit der Eigenschaft, daß entweder $\mathrm{ggT}(n, r) > 1$ ist oder aber $\mathrm{ggT}(n, r) = 1$ ist und $n \bmod r$ in $(\mathbb{Z}/r)^\times$ eine größere Ordnung als $4\ell(n)^2$ hat.

Dies geschieht einfach dadurch, daß man die Zahlen $r = 2, 3, \ldots$ allesamt durchprobiert, bis zum ersten mal eine der beiden Bedingungen erfüllt ist. Die Bedingung über die Ordnung der Restklasse von n in $(\mathbb{Z}/r)^\times$ prüft man nach, indem man nacheinander ihre Potenzen ausrechnet, bis man entweder eine Eins gefunden hat oder aber der Exponent größer als $4\ell(n)^2$ ist.

3. Schritt: Falls $r = n$, ist n prim und der Algorithmus endet.

In der Tat: Dann haben wir für alle $r < n$ überprüft, daß $\mathrm{ggT}(n, r) = 1$ ist. Wenn der Algorithmus etwas taugt, darf er natürlich höchstens für sehr kleine Werte von n mit diesem Schritt enden.

4. Schritt: Falls im zweiten Schritt ein r gefunden wurde, für das der ggT von n und r größer als eins ist, muß n zusammengesetzt sein und der Algorithmus endet.

Denn dann haben wir einen Teiler von n gefunden.

5. Schritt: Teste für $j = 1, \ldots, \ell \underset{\text{def}}{=} 2\ell(n)[\sqrt{r}] + 1$, ob über $\mathbb{Z}/n$

$$(X + j)^n \equiv X^n + j \bmod (X^r - 1)\,.$$

Sobald ein j gefunden wird, für das dies nicht erfüllt ist, endet der Algorithmus mit dem Ergebnis *n ist zusammengesetzt.*

Falls nämlich n eine Primzahl ist, stimmen $(X + j)^n$ und $X^n + j$ als Polynome mit Koeffizienten aus $\mathbb{Z}/n$ nach obigem Satz überein, sind also erst recht auch gleich modulo $(X^r - 1)$.

6. Schritt: Wenn alle Tests im fünften Schritt bestanden sind, ist n eine Primzahl.

Dies zu beweisen ist die Hauptarbeit dieses Paragraphen.

Nach den Kommentaren zu den einzelnen Schritten ist klar, daß der Algorithmus für eine Primzahl n stets das richtige Ergebnis liefert; wir müssen zeigen, daß er auch zusammengesetzte Zahlen stets erkennt.

Sei also n eine zusammengesetzte Zahl. Falls n Potenz einer anderen natürlichen Zahl ist, wird dies im ersten Schritt erkannt; wir können und werden im folgenden daher annehmen, daß dies nicht der Fall ist.

Das r aus dem zweiten Schritt ist auf jeden Fall echt kleiner als n, denn als zusammengesetzte Zahl hat n insbesondere einen Teiler $r < n$. Der Algorithmus kann daher nicht im dritten Schritt mit der Antwort „n ist prim“ enden. Falls er im vierten Schritt endet, lieferte der zweite Schritt einen Teiler von n, und wir erhalten die richtige Antwort „n ist zusammengesetzt“.

Für den Rest des Paragraphen können wir somit annehmen, daß der zweite Schritt auf ein r führte, für das $\mathrm{ggT}(n, r) = 1$ ist. Wir müssen zeigen, daß einer der Tests im fünften Schritt scheitert, daß es also eine natürliche Zahl j gibt mit

$$1 \le j \le \ell \quad \text{und} \quad (X + j)^n \not\equiv X^n + j \bmod (X^r - 1) \text{ in } \mathbb{Z}/n[X]\,.$$

Wir nehmen an, das sei nicht der Fall, und betrachten einen Primteiler p von n. Dieser muß größer als r sein, denn sonst hätte der Algorithmus bereits mit dem vierten Schritt spätestens bei $r = p$ geendet.

Jede Kongruenz modulo n ist erst recht eine Kongruenz modulo p; wir können daher davon ausgehen, daß für alle j mit $1 \le j \le \ell$ gilt

$$(X + j)^n \equiv X^n + j \bmod (X^r - 1) \text{ in } \mathbb{F}_p[X] .$$

Wenn wir zum Faktorring $R = \mathbb{F}_p[X]/(X^r - 1)$ übergehen, ist dort also

$$(X + j)^n = X^n + j \quad \text{falls} \quad 1 \le j \le \ell .$$

Um diese seltsame Relation genauer zu untersuchen, betrachten wir für jede zu r teilerfremde natürliche Zahl k die Abbildung

$$\widehat{\sigma}_k : \begin{cases} \mathbb{F}_p[X] \to R \\ \qquad g \mapsto g(X^k) \bmod (X^r - 1) \end{cases} ,$$

die in jedem Polynom g die Variable X überall durch X^k ersetzt.

Lemma: $\widehat{\sigma}_k$ ist surjektiv und sein Kern besteht genau aus den Vielfachen des Polynoms $X^r - 1$.

Beweis: Wir betrachten $\widehat{\sigma}_k$ nur für Indizes k, die zu r teilerfremd sind. Zu jedem solchen Index gibt es daher ein k', so daß $kk' \equiv 1 \bmod r$ ist, und modulo $X^r - 1$ ist damit $X^{kk'} \equiv X$. Für ein beliebiges Polynom $g \in \mathbb{F}_p[X]$ und $h(X) = g(X^{k'})$ ist daher in R

$$\widehat{\sigma}_k(h) = h(X^k) = g(X^{kk'}) = g(X) = g ,$$

die Abbildung ist also surjektiv.

Was ihren Kern betrifft, so enthält er auf jeden Fall $X^r - 1$ und alle seine Vielfachen, denn

$$\widehat{\sigma}_k(X^r - 1) = (X^{kr} - 1) \bmod (X^r - 1) = 1^k - 1 = 0 ,$$

da $X^r \equiv 1 \bmod (X^r - 1)$.

Umgekehrt sei g irgendein Polynom aus dem Kern von $\widehat{\sigma}_k$. Dann ist das Polynom $h(X) = g(X^k)$ modulo $X^r - 1$ gleich dem Nullpolynom,

ist also ein Vielfaches von $X^r - 1$. Konkret sei $h = (X^r - 1)f$. Im Faktorring R ist dann

$$g(X) = g(X^{kk'}) = h(X^{k'}) = (X^{k'r} - 1)f(X^{k'}) = 0 ,$$

denn wegen $X^r = 1$ in R ist dort $X^{k'r} - 1 = 0$.

In $\mathbb{F}_p[X]$ muß $g(X)$ daher ein Vielfaches von $X^r - 1$ sein, und genau das war die Behauptung über den Kern von $\widehat{\sigma}_k$. ∎

Da alle Vielfachen von $X^r - 1$ im Kern von $\widehat{\sigma}_k$ liegen, definiert $\widehat{\sigma}_k$ eine Abbildung σ_k von R nach R, die jedem Polynom $g \bmod (X^r - 1)$ aus R das Element $\widehat{\sigma}_k(g)$ zuordnet; nach dem gerade bewiesenen Lemma hängt dieses wirklich nur von der Restklasse $g \bmod (X^r - 1)$ ab. Außerdem zeigt das Lemma, daß σ_k sowohl surjektiv als auch injektiv ist, denn der Kern von $\widehat{\sigma}_k$ ist gleich dem Kern der Restklassenabbildung von $\mathbb{F}_p[X]$ nach R. Damit ist σ_k ein bijektiver Homomorphismus von R nach R, ein sogenannter *Automorphismus* von R. Wir haben damit für jede zu r teilerfremde natürliche Zahl k einen Automorphismus $\sigma_k : R \to R$, der jedem Polynom in X das entsprechende Polynom in X^k zuordnet. Da wir in R rechnen, werden natürlich alle Polynome modulo $X^r - 1$ betrachtet.

Unmittelbar aus der Definition folgt, daß die verschiedenen Automorphismen σ_k miteinander kommutieren; genauer ist

$$\sigma_k \circ \sigma_{k'} = \sigma_{k'} \circ \sigma_k = \sigma_{kk'} ,$$

denn in allen drei Fällen wird im Endeffekt X durch $X^{kk'}$ ersetzt.

Speziell für das Element $X + j$ aus R ist $\sigma_k(X + j) = X^k + j$. Für $k = n$ und $j = 1, \ldots, \ell$ ist andererseits auch

$$\sigma_n(X + j) = (X + j)^n ,$$

denn für diese j wurde ja nach unserer Annahme der Test im fünften Schritt bestanden.

Wir wollen genauer untersuchen, wann die Gleichung $\sigma_k(f) = f^k$ erfüllt ist. Dazu definieren zwei Arten von Mengen:

$$C(f) = \left\{ k \in (\mathbb{Z}/r)^\times \;\middle|\; \sigma_k(f) = f^k \right\} \quad \text{für alle } f \in R \qquad \text{und}$$

$$D(k) = \left\{ f \in R \;\middle|\; \sigma_k(f) = f^k \right\} \qquad \text{für alle } k \in (\mathbb{Z}/r)^\times$$

Beide Mengen enthalten mit zwei Elementen auch deren Produkt, denn für zwei Elemente $k, k' \in C(f)$ ist

$$\sigma_{kk'}(f) = \sigma_k(f)\sigma_{k'}(f) = \sigma_k\big(\sigma_{k'}(f)\big) = \sigma_k(f^{k'}) = \sigma_k(f)^{k'} = f^{kk'} ,$$

und für $f, g \in D(k)$ ist

$$\sigma_k(fg) = \sigma_k(f)\sigma_k(g) = f^k g^k = (fg)^k .$$

Der Rest des Beweises besteht darin, daß wir die „Größe“ der Menge $D(n)$ auf zwei verschiedene Weisen abschätzen und daraus einen Widerspruch herleiten zur Annahme, daß n zusammengesetzt ist, aber trotzdem vom Algorithmus als Primzahl klassifiziert wird. Wir definieren zunächst zwei neue Zahlen:

- s sei die Ordnung der Restklasse von p in $(\mathbb{Z}/r)^\times$. Dann ist r ein Teiler von $p^s - 1$, denn $p^s \equiv 1 \bmod r$.
- t sei die Ordnung der von den Restklassen von p und n erzeugten Untergruppe von $(\mathbb{Z}/r)^\times$, d.h also die Ordnung der kleinsten Untergruppe, die beide Restklassen enthält. Da diese Untergruppe insbesondere die Restklasse von p und deren Potenzen enthält, ist t ein Vielfaches von s.

Als nächstes betrachten wir einen Körper K mit p^s Elementen. Einen solchen Körper kann man konstruieren, indem man den Vektorraum $\mathbb{F}_p^s$ identifiziert mit dem Vektorraum aller Polynome vom Grad kleiner s mit Koeffizienten aus $\mathbb{F}_p$ und dort eine Multiplikation einführt, die zwei Polynomen deren Produkt modulo einem festen irreduziblen Polynom vom Grad s über $\mathbb{F}_p$ zuordnet. Man kann zeigen (siehe Algebra-Vorlesung oder entsprechendes Lehrbuch), daß es für jedes s ein solches Polynom gibt, und daß zwei verschiedene irreduzible Polynome vom Grad s zu isomorphen Körpern führen.

Aus Kapitel 1 wissen wir, daß die multiplikative Gruppe jedes endlichen Körpers zyklisch ist; $K^\times$ ist also eine zyklische Gruppe der Ordnung $p^s - 1$. Diese Zahl ist, wie wir gerade gesehen haben, ein Vielfaches von r; somit gibt es in $K^\times$ (mindestens) ein Element ζ der Ordnung r.

Für irgendein solches Element definieren wir einen Homomorphismus

$$\widehat{\tau}\colon \begin{cases} \mathbb{F}_p[X] \to K \\ \qquad g \mapsto g(\zeta) \end{cases} .$$

Da $\widehat{\tau}(X^r - 1) = \zeta^r - 1$ verschwindet, induziert $\widehat{\tau}$ einen Ringhomomorphismus $\tau\colon R \to K$. Die angekündigten Abschätzungen der „Größe“ von $D(n)$ beziehen sich auf die Mächtigkeit der Menge $S = \tau\big(D(n)\big)$:

Lemma: $S = \tau\big(D(n)\big)$ hat höchstens $n^{2[\sqrt{t}]}$ Elemente.

Beweis: Wir gehen davon aus, daß n weder eine Primzahl noch eine Primzahlpotenz ist; daher gibt es außer dem Primteiler p noch mindestens einen weiteren Primteiler q. Wenn wir (in $\mathbb{N}$) Potenzen der Form $n^u p^v$ und $n^{u'} p^{v'}$ mit $u, u', v, v' \in \mathbb{N}_0$ betrachten, sind diese daher genau dann gleich, wenn $(u, v) = (u', v')$ ist: Ist nämlich $u \neq u'$, so tritt g in der Primzerlegung der beiden Elemente mit verschiedenen Exponenten auf, und ist $u = u'$, aber $v \neq v'$, so gilt entsprechendes für p. Daher hat die Menge

$$I = \left\{ n^u p^v \;\middle|\; 0 \le u, v \le [\sqrt{t}] \right\}$$

mindestens $\big([\sqrt{t}]+1\big)^2$ Elemente, und diese Zahl ist offensichtlich größer als t.

Nun war aber t definiert als die Ordnung der Untergruppe von $(\mathbb{Z}/r)^\times$, die von den Restklassen von n und von p erzeugt wird; daher muß es mindestens zwei Elemente

$$k = n^u p^v \quad \text{und} \quad k' = n^{u'} p^{v'}$$

aus I geben, die dieselbe Restklasse in $(\mathbb{Z}/r)^\times$ definieren, für die also gilt: $k \equiv k' \bmod r$. Da die Exponenten u, u', v, v' höchstens gleich $[\sqrt{t}]$ sind und p ein Teiler von n ist, können wir $n^{2[\sqrt{t}]}$ als (sehr grobe) obere Schranke für k und k' nehmen.

Nun sei $f \in R$ ein Element von $D(n)$. Nach Definition der Mengen $C(f)$ und $D(n)$ ist dann auch n ein Element von $C(f)$. Außerdem enthält $C(f)$ stets die Eins und nach dem kleinen Satz von FERMAT auch die Primzahl p, denn Potenzieren mit p ist über $\mathbb{F}_p$ ein Homomorphismus. Da

mit zwei Elementen stets auch deren Produkt in $C(f)$ liegt, liegen daher die Restklassen modulo r aller Elemente von I in $C(f)$. Insbesondere sind daher k und k' Elemente von $C(f)$, d.h.

$$\sigma_k(f) = f^k \quad \text{und} \quad \sigma_{k'}(f) = f^{k'} .$$

Wegen $k \equiv k' \bmod r$ ist aber σ_k dieselbe Abbildung wie $\sigma_{k'}$; daher ist $f^k = f^{k'}$ für jedes $f \in D(n)$. Somit sind die Bilder $\tau(f)$ aller $f \in R$ Nullstellen des Polynoms $X^k - X^{k'}$. Dessen Grad ist das Maximum von k und k', und da $\tau(f)$ im Körper K liegt, gibt es höchstens so viele Nullstellen, wie der Grad angibt. Aufgrund der obigen Abschätzung für k und k' hat das Polynom daher höchstens $n^{2[\sqrt{t}]}$ Nullstellen, und damit kann auch S nicht mehr Elemente enthalten. ∎

Als untere Grenze für die Elementanzahl von S erhalten wir

Lemma: S enthält mindestens $2^{\min(t,\ell)} - 1$ Elemente.

Beweis: Wegen der bestandenen Tests in Schritt 5 liegt $\tau(X + j)$ in $D(n)$ für $j = 1, \ldots, \ell$. Da $p > r > t \geq m$ ist, sind die Zahlen von 1 bis m auch modulo p paarweise verschieden. Die Teilmenge

$$P = \left\{ \prod_{j=1}^{m} (X + j)^{e_j} \;\middle|\; e_j \in \{0, 1\} \text{ und } \sum_{j=1}^{m} e_j < m \right\}$$

von $\mathbb{F}_p[X]$ enthält daher $2^m - 1$ Polynome.

Aus diesen Polynomen können wir Elemente von R *bzw.* K machen, indem wir für die Variable X die Restklasse $\eta = X \bmod (X^r - 1)$ *bzw.* das oben gewählte Element ζ der Ordnung r einsetzen; wir erhalten Teilmengen

$$P(\eta) = \left\{ f(\eta) \;\middle|\; f \in P \right\} \subseteq R \quad \text{und} \quad P(\zeta) = \left\{ f(\zeta) \;\middle|\; f \in P \right\} \subseteq K .$$

Da sowohl n als auch p in $D(n)$ liegen und mit zwei Elementen auch deren Produkt, liegt $P(\eta)$ in $D(n)$ und damit $\tau\big(P(\eta)\big) = P(\zeta)$ in S. Das Lemma ist daher bewiesen, sobald wir gezeigt haben, daß $P(\zeta)$ mindestens $2^m - 1$ Elemente enthält.

Falls dies nicht der Fall wäre, müßte es in P zwei verschiedene Polynome g und h geben, für die $g(\zeta) = h(\zeta)$ wäre. Wir müssen also zeigen, daß $g(\zeta) = h(\zeta)$ nur dann gelten kann, wenn $g = h$ ist.

Wie im vorigen Lemma folgt, da $1, p$ und n alle drei sowohl in $C\big(g(\eta)\big)$ als auch in $C\big(h(\eta)\big)$ liegen, daß alle natürlichen Zahlen k der Form $k = n^u p^v$ in diesen beiden Mengen liegen.

Da $g(\zeta) = h(\zeta)$, gilt für jedes solche k

$$\begin{aligned} 0 &= g(\zeta)^k - h(\zeta)^k = \tau\big(g(\eta)\big)^k - \tau\big(h(\eta)\big)^k = \tau\big(g(\eta)^k\big) - \tau\big(h(\eta)^k\big) \\ &= \tau\big(g(\eta^k)\big) - \tau\big(h(\eta^k)\big) = g(\zeta^k) - h(\zeta^k) . \end{aligned}$$

Da ζ in K die Ordnung r hat, hängt ζ^k nur von $k \bmod r$ ab; die Anzahl verschiedener Restklassen der Form $n^u p^v$ modulo r hatten wir oben mit t bezeichnet. Somit hat die Differenz $g - h$ mindestens t Nullstellen. Andererseits sind aber g und h und damit auch ihre Differenz Polynome vom Grad höchstens $t - 1$, also muß $g - h$ das Nullpolynom sein, d.h. $g = h$. Somit enthält S mindestens $2^m - 1$ Elemente, wie behauptet. ∎

Zum Abschluß des Beweises, daß der Test von AGRAWAL, KAYAL und SAXENA stets die richtige Antwort liefert, müssen wir nun nur noch zeigen, daß die Schranken aus den beiden letzten Lemmata, die ja unter der Voraussetzung bewiesen wurde, daß eine zusammengesetzte Zahl als prim erkannt wird, einander widersprechen, daß also die untere Schranke größer ist als die obere:

Lemma: $2^{\min(t,\ell)} - 1 > n^{2[\sqrt{t}]}$.

Beweis: Da $\ell(n) > \log_2 n$, genügt es zu zeigen, daß

$$2^{\min(t,\ell)} - 1 > 2^{2\ell(n)[\sqrt{t}]} .$$

Da beide Exponenten natürliche Zahlen sind, genügt dazu wiederum, daß $\min(t, \ell) > 2\ell(n)[\sqrt{t}]$ ist, denn wenn sich die Exponenten um mindestens eins unterscheiden, ist die Differenz zwischen den Potenzen mindestens zwei. Wir müssen daher zeigen, daß sowohl t als auch ℓ größer sind als $2\ell(n)[\sqrt{t}]$.

Für $\ell = 2\ell(n)[\sqrt{r}] + 1$ ist das klar, da t die Ordnung einer Untergruppe von $(\mathbb{Z}/r)^\times$ bezeichnet und damit auf jeden Fall kleiner als r ist.

Die Ungleichung $t > 2\ell(n)[\sqrt{t}]$ ist sicherlich dann erfüllt, wenn sogar $t > 2\ell(n)\sqrt{t}$ ist, und dies wiederum ist äquivalent zur Ungleichung $t > 4\ell(n)^2$. Nun ist aber t die Ordnung jener Untergruppe von $(\mathbb{Z}/r)^\times$, die von den Restklassen von n und p erzeugt wird. Da wir im zweiten Schritt des Algorithmus sichergestellt haben, daß dort allein die Ordnung der Restklasse von n schon größer ist als $4\ell(n)^2$, ist auch die Ungleichung für t trivial. ∎

Damit ist die Korrektheit des Algorithmus vollständig bewiesen.

Kapitel 4
Faktorisierungsverfahren

Wie wir in §2 des letzten Kapitels gesehen haben, ist $M_{67} = 2^{67} - 1$ keine Primzahl, denn

$$13^{M_{67}-1} \equiv 81\,868\,480\,399\,682\,966\,751 \bmod M_{67} \not\equiv 1 \bmod M_{67}\,.$$

Somit ist M_{67} ein Produkt von mindestens zwei nichttrivialen Faktoren. Welche sind das?

FRANK NELSON COLE gab das Ergebnis am 31. Oktober 1903 auf einer Sitzung der American Mathematical Society bekannt: Er schrieb die Zahl

$$2^{67} - 1 = 147\,573\,952\,589\,676\,412\,927$$

auf eine der beiden Tafeln und

$$193\,707\,721 \times 761\,838\,257\,287$$

auf die andere. Dieses Produkt rechnete er wortlos aus nach der üblichen Schulmethode zur schriftlichen Multiplikation, und als er dieselbe Zahl erhielt, die auf der anderen Tafel stand, schrieb er ein Gleichheitszeichen zwischen die beiden Zahlen und setzte sich wieder. Das Ergebnis, d.h. die Faktorisierung von M_{67}, findet ein Computeralgebrasystem heute in weniger als einer Sekunde; für die damalige Zeit war sie eine Sensation! COLE gab später zu, daß er drei Jahre lang jeden Sonntag nachmittag daran gearbeitet hatte. Er versuchte M_{67} in der Form $x^2 - y^2$ darzustellen, wobei er mit Hilfe quadratischer Reste Kongruenzbedingungen für x modulo verschiedener relativ kleiner Primzahlen aufstellte und auch verwendete, daß jeder Teiler von M_{67} kongruent eins modulo 67 und kongruent ± 1 modulo acht sein muß. Dies führte zu einer ganzen Reihe

von Kongruenzen für x, die er in

$$x \equiv 1\,160\,932\,384 \bmod 1\,323\,536\,760$$

zusammenfassen konnte. Untersuchung quadratischer Reste zeigt, daß

$$x_k = 1\,323\,536\,760\ k + 1\,160\,932\,384$$

frühestens für $k = 287$ in Frage kommt, und mit $x = x_{287}$ ist tatsächlich

$$\begin{aligned} M_{67} &= 381\,015\,982\,504^2 - 380\,822\,274\,783^2 \\ &= 193\,707\,721 \times 761\,838\,257\,287\,. \end{aligned}$$

Für Einzelheiten siehe

F. N. COLE: On the factoring of large numbers, *Bull. Am. Math. Soc.* **10** (1903), 134–137 *oder* `http://www.ams.org/bull/1903-10-03/S0002-9904-1903-01079-9/home.html`

FRANK NELSON COLE (1861–1926) wurde in Massachusetts geboren. 1882 erhielt er seinen Bachelor in Mathematik von der Harvard University; danach konnte er dank eines Stipendiums drei Jahre lang bei FELIX KLEIN in Leipzig studieren. Mit einer von KLEIN betreuten Arbeit über Gleichungen sechsten Grades wurde er 1886 in Harvard promoviert. Nach verschiedenen Positionen in Harvard und Michigan bekam er 1895 eine Professor an der Columbia University in New York, wo er bis zu seinem Tod lehrte. Seine Arbeiten befassen sich hauptsächlich mit Primzahlen und mit der Gruppentheorie.

Der Auftritt von COLE schlug selbst außerhalb der Mathematik so große Wellen, daß seine Faktorisierung noch fast ein Jahrhundert später vorkommt in einer New Yorker (off-Broadway) Show von RINNE GROFF mit dem Titel *The five hysterical girls theorem.* Dort bringt sich ein junger Mathematiker um, weil er in einem Beweis von der *Primzahl* $2^{67} - 1$ ausgeht und die Tochter des Professors die obige Faktorisierung an die Tafel schreibt. Einzelheiten kann man, so man unbedingt möchte, unter `http://www.playscripts.com/play.php3?playid=551` nachlesen. (Die Show verschwand nach zwei Monaten Ende Mai 2000 in der Versenkung; sie wurde seither nur noch zweimal von Amateurgruppen aufgeführt.)

COLE konnte für seine Faktorisierung von M_{67} auf bekannte Tatsachen über die Struktur von Faktoren der MERSENNE-Zahlen zurückgreifen und auch bei den von ihm selbst gefundenen Eigenschaften potentieller Faktoren konnte er die spezielle Struktur von M_{67} ausnutzen. Ähnlich arbeiten auch heutige Mathematiker an der Faktorisierung spezieller Zahlen, beispielsweise im Rahmen des Cunningham-Projekts zur Faktorisierung von Zahlen der Form $b^n \pm 1$ für kleine Basen b. Für die Faktorisierung von RSA-Moduln kann man natürlich nicht mit solchen Techniken arbeiten. In diesem Kapitel soll es um Verfahren gehen, mit denen man eine zufällig gegebene Zahl ohne spezielle Struktur faktorisieren kann.

Es gibt kein „bestes“ Faktorisierungsverfahren; für Zahlen verschiedener Größenordnungen haben jeweils andere Verfahren ihre Stärken. Auch Vorwissen über die zu faktorisierende Zahl kann bei der Wahl eines geeigneten Verfahrens helfen: Bei einem RSA-Modul, der das Produkt zweier Primzahlen ähnlicher Größenordnung ist, wird man anders vorgehen als bei einer Zahl der Form $b^n \pm 1$. Mehr noch als bei Primzahltests gilt, daß asymptotische Komplexitätsaussagen als Auswahlkriterium nutzlos sind: Das für die Faktorsierung 150-stelliger RSA-Moduln heute optimale Verfahren, das Zahlkörpersieb, wird beim Versuch eine sechsstellige Zahl zu faktorisieren, oft nicht in der Lage sein die Faktoren zu trennen, und selbst in den Fällen, in denen es erfolgreich ist, braucht es erheblich länger als einfache Probedivisionen. Es gibt definitiv kein „bestes“ Faktorisierungsverfahren; je nach Größe der zu faktorisierenden Zahl und auch nach Größe der zu erwartenden Faktoren können ganz verschiedene Strategien angebracht sind, die oft nur in der Kombination zur vollständigen Primzerlegung führen.

Im folgenden sollen einige der einfachsten gebräuchlichen Verfahren vorgestellt werden.

§1: Die ersten Schritte

a) Test auf Primzahl

Der schlimmste Fall für praktisch jedes Faktorisierungsverfahren tritt dann ein, wenn die zu faktorisierende Zahl eine Primzahl ist: Gerade bei

den fortgeschrittenen Verfahren gibt es oft kein anderes Abbruchkriterium als das Auffinden eines Faktors. Daher sollte (außer eventuell bei ganz kleinen Zahlen) zu Beginn einer Faktorisierung immer ein Primzahltest stehen. Da auch das Testen auf Potenzen relativ einfach ist, läßt sich eventuell auch das noch durchführen – es sei denn, daß von der Situation her (beispielsweise bei RSA-Moduln) nicht mit einer Potenz zu rechnen ist.

b) Abdividieren kleiner Primteiler

Bei kleinen zusammengesetzten Zahlen n besteht die effizienteste Art der Faktorisierung im allgemeinen darin, einfach alle Primzahlen nacheinander durchzuprobieren, indem man sie der Reihe nach so lange abdividiert, wie es geht. Sobald der Quotient kleiner ist als das Quadrat der gerade betrachteten Primzahl, kann man sicher sein, daß auch er eine Primzahl ist und hat n vollständig faktorisiert.

Die genaue Vorgehensweise ist folgende: Wir nehmen an, daß eine Liste der Primzahlen bis zu einer gewissen Grenze zur Verfügung steht; gegebenenfalls muß diese zunächst nach ERATOSTHENES erzeugt werden.

1. Schritt: Setze M gleich der zu faktorisierende Zahl und $p = 2$.

2. Schritt: Solange M durch p teilbar ist, ersetze M durch M/p und notiere p als Faktor.

3. Schritt: Falls $M = 1$, sind alle Faktoren gefunden, und der Algorithmus endet. Ansonsten wird p auf die nächste Primzahl gesetzt. Ist dann $M < p^2$, muß M prim sein und wird als weiterer Faktor notiert; sodann endet auch in diesem Fall der Algorithmus. Andernfalls geht es zurück zum zweiten Schritt.

Als Beispiel wollen wir die Zahl 1 234 567 890 faktorisieren:

Im ersten Schritt werden $M = 1\,234\,567\,890$ und $p = 2$ initialisiert.

Im zweiten Schritt ist nun $p = 2$. Da M eine gerade Zahl ist, können wir durch p dividieren; wir notieren also die Zwei als Faktor und ersetzen M durch $M/2 = 617\,283\,945$. Diese Zahl ist ungerade; also geht es weiter zum dritten Schritt, wo offensichtlich keines der Abbruchkriterien erfüllt ist. Somit wird $p = 3$ und es geht zurück zum zweiten Schritt.

Das neue M ist durch drei teilbar, genauso der Quotient $M/3 = 205\,761\,315$. Also wird nochmals durch drei dividiert, und wir erhalten den nicht mehr durch drei teilbaren Quotienten 68 587 105. Somit werden zwei Faktoren drei notiert und es geht weiter zum dritten Schritt. Dort wird $p = 5$ gesetzt, und es geht wieder zurück zu Schritt 2.

Das aktuelle $M = 68\,587\,105$ ist durch fünf teilbar mit Quotient $M/5 = 13\,717\,421$. Dieser wird das neue M; und da er nicht durch fünf teilbar ist, notieren wir nur einen Faktor fünf.

Im dritten Schritt wird wieder p erhöht und es geht zurück zum zweiten Schritt. Dort passiert nun allerdings lange Zeit nichts, denn keine der Primzahlen zwischen sieben und 3 593 teilt das aktuelle M. Erst wenn p im dritten Schritt auf 3 607 gesetzt wird, finden wir wieder einen Faktor. Wir notieren ihn, ersetzen M durch $M/3\,607 = 3803$, was offensichtlich nicht durch 3 607 teilbar ist, und gehen weiter zum dritten Schritt.

Dort ist nun offensichtlich $M < p^2$, also ist M eine Primzahl, und

$$1\,234\,567\,890 = 2 \cdot 3^2 \cdot 5 \cdot 3\,607 \cdot 3\,803$$

ist vollständig faktorisiert.

Es ist klar, daß wir schon eine zwanzigstellige Zahl nur mit viel Glück auf diese Weise mit vertretbarem Aufwand vollständig faktorisieren können. Trotzdem ist Abdividieren selbst für noch viel größere Zahlen ein sinnvoller erster Schritt, denn die auf größere Faktoren spezialisierten Verfahren schaffen es im allgemeinen nicht, auch kleine Primfaktoren voneinander zu trennen.

Um Abdividieren statt zur vollständigen Faktorisierung nur zur Identifikation „kleiner“ Primfaktoren zu verwenden, ist lediglich eine kleine Modifikation des dritten Schritts notwendig: Wir legen eine Suchgrenze S fest und brechen im dritten Schritt auch dann ab, wenn $p > S$ ist. Im letzteren Fall können wir selbstverständlich nicht behaupten, daß das verbleibende M eine Primzahl ist; M muß dann mit anderen Verfahren weiter bearbeitet werden. Die Größe von S hängt natürlich etwas ab vom Arbeitsspeicher und der Geschwindigkeit des verwendeten Computers; ein minimaler Wert wäre etwa $2^{16} = 65\,536$; bei etwas besseren Computern läßt sich auch das Abdividieren bis zu einer Million und bei

derzeit aktuellen schnellen Computern auch einer Milliarde oder etwa 2^{30} in weniger als einer Sekunde durchführen.

§2: Die Verfahren von Pollard und ihre Varianten

In den Jahren um 1975 entwickelte der britische Mathematiker JOHN M. POLLARD mehrere recht einfache Algorithmen zur Faktorisierung ganzer Zahlen sowie zur Berechnung diskreter Logarithmen, die auch heute noch (teils in verbesserter Form) zu den Standardwerkzeugen der algorithmischen Zahlentheorie gehören. In diesem Paragraphen sollen die beiden bekanntesten vorgestellt werden; außerdem möchte ich zumindest kurz auf mathematisch anspruchsvollere Verallgemeinerungen eingehen. Die hier behandelten Verfahren haben im Gegensatz zu denen des nächsten Paragraphen die Eigenschaft, daß sie umso schneller zum Erfolg führen, je kleiner die gesuchten Primfaktoren sind, daß sie allerdings sehr kleine Primfaktoren oft nicht finden. Sie sind also die Verfahren der Wahl für die Weiterverarbeitung eines durch Abdividieren erhaltenen „Rests“, von dem man weiß, daß er keine allzu kleinen Primfaktoren mehr hat.

JOHN M. POLLARD ist ein britischer Mathematiker, der hauptsächlich bei British Telecom arbeitete. Er veröffentlichte zwischen 1971 und 2000 rund zwanzig mathematische Arbeiten, größtenteils auf dem Gebiet der algorithmischen Zahlentheorie. Bekannt ist er auch für seine Beiträge zur Kryptographie, für die er 1999 den RSA Award erhielt. Außer den hier vorgestellten Faktorisierungsalgorithmen entwickelte er unter anderem auch das Zahlkörpersieb, eine Variante des weiter hinten vorgestellten quadratischen Siebs, dessen Weiterentwicklungen derzeit die schnellsten Faktorisierungsalgorithmen für große Zahlen sind. Seine home page, um die er sich auch jetzt im Ruhestand noch kümmert, ist `jmptidcott.googlepages.com/index.html`.

Bei den in diesem und dem nächsten Paragraphen vorgestellten Verfahren besteht das Ziel immer darin, *irgendeinen* Faktor zu finden; sobald dies erreicht ist, bricht das Verfahren ab und der gefundene Faktor sowie sein Kofaktor werden für sich weiter untersucht – wobei natürlich immer an erster Stelle ein Primzahltest stehen sollte.

a) Die Monte-Carlo-Methode

Monte Carlo ist ein Stadtteil von Monaco, der vor allem für seine Spielbank bekannt ist. An deren Spieltischen sollen Roulette-Schüsseln idealerweise rein zufällig für jedes Spiel von neuem eine Zahl zwischen 0 und 36 bestimmen.

Eine ähnliche Idee läßt sich auch für die Faktorisierung einer ganzen Zahl N verwenden: Ausgehend von einer Folge $(x_i)_{i\in\mathbb{N}}$ zufällig gewählter Zahlen zwischen 1 und N (oder 0 und $N-1$) bildet man jeweils den ggT von x_i mit N in der Hoffnung, einen nichttrivialen Teiler zu finden.

Für einen Primteiler p von N können wir erwarten, daß im Mittel jedes p-te x_i durch p teilbar ist und wir somit einen durch p teilbaren ggT mit N erhalten, der allerdings möglicherweise auch noch andere Primteiler enthält.

Beim einfachen Abdividieren finden wir p, nachdem wir alle Primzahlen bis einschließlich p durchprobiert haben; wie wir im letzten Kapitel gesehen haben, sind dies etwa $p/\log p$ Stück. Für jede davon brauchen wir eine Division, verglichen mit den durchschnittlich $p/2$ EUKLIDischen Algorithmen bei der obigen Methode, die nicht einmal eine Garantie dafür bieten, den Faktor zu finden. Von daher hat die neue Methode zumindest in der bislang betrachteten Form ausschließlich Nachteile und ist keine sinnvolle Alternative zum Abdividieren.

POLLARDs Idee zur Beschleunigung beruht auf dem im Anhang genauer erklärten Geburtstagsparadoxon: Die Wahrscheinlichkeit dafür, daß eine gegebene Zufallszahl durch p teilbar ist, liegt zwar nur bei $1:p$, aber die Wahrscheinlichkeit, daß zwei der x_i modulo p gleich sind, steigt in der Nähe von etwa $\sqrt{p}$ Folgengliedern ziemlich steil von nahe null zu nahe eins. Wenn wir also anstelle der größten gemeinsamen Teiler von N mit den x_i die mit den Differenzen $x_i - x_j$ berechnen, haben wir bereits bei einer Folge der Länge um $\sqrt{p}$ gute Chancen, einen nichttrivialen ggT zu finden.

Auch in dieser Form ist das Verfahren noch nicht praktikabel: Wenn wir ein neues x_i mit $i \approx \sqrt{p}$ erzeugt haben, müssen wir für alle $j < i$ den ggT von $x_i - x_j$ berechnen, was noch einmal rund $\sqrt{p}$ Schritte sind, so

daß der Gesamtaufwand nicht proportional zu $\sqrt{p}$ ist, sondern eher zu

$$\int_0^{\sqrt{p}} x\,dx = \frac{p}{2}\,,$$

was keine große Ersparnis ist. Dazu kommt, daß alle bereits berechneten Folgenglieder gespeichert werden müssen, der Algorithmus hat also auch einen Platzbedarf in der Größenordnung $\sqrt{p}$.

Dieses Problem können wir umgehen, indem wir keine echten Zufallszahlen verwenden, sondern algorithmisch eine Folge sogenannter Pseudozufallszahlen erzeugen. Typischerweise verwendet man dazu eine Rekursionsvorschrift der Form $x_{i+1} = Q(x_i) \bmod N$ mit einem quadratischen Polynom Q. (Die bei Simulationen sehr beliebten Pseudozufallsgeneratoren nach der linearen Kongruenzmethode sind für die Monte-Carlo-Methode der Faktorisierung nicht geeignet.) Meist nimmt man einfach Polynome der Form $Q(x) = x^2 + c$, wobei allerdings $c \neq 0$ und $c \neq -2$ sein sollte, denn eine genauere Untersuchung zeigt, daß diese Wahlen keine guten Pseudozufallszahlen liefern. Daß die anderen Wahlen von c stets gute Generatoren liefern ist zwar nicht bewiesen, aber die praktischen Erfahrungen sind positiv.

Wegen der speziellen Form der Rekursion hängt die Restklasse von x_{i+1} modulo p nur ab von $x_i \bmod p$; insbesondere ist also $x_{i+1} \equiv x_{j+1} \bmod p$, falls $x_i \equiv x_j \bmod p$, und entsprechend stimmen auch für jedes $r \geq 0$ die Zahlen x_{i+r} und x_{j+r} modulo p überein, d.h. die Folge wird modulo p periodisch mit einer Periode π, die $|i-j|$ teilt.

Das Problem, Periodizität in einer Folge zu entdecken, tritt nicht nur in der Zahlentheorie auf, sondern beispielsweise auch in der Zeitreihenanalyse und anderen Anwendungen. Ein möglicher Algorithmus zu seiner Lösung, auch als Hase und Schildkröte Algorithmus bekannt, stammt von FLOYD (1967) und beruht auf folgender Beobachtung:

Wird eine Folge (y_i) irgendwann periodisch, so gibt es Indizes k derart, daß $y_k = y_{2k}$ ist.

In der Tat, ist $y_{i+\pi} = y_i$ für alle $i \geq r$, so können wir für k jedes Vielfache $\ell\pi$ der Periode nehmen, das mindestens gleich r ist.

ROBERT W. FLOYD (1936–2001) beendete seine Schulausbildung bereits im Alter von 14 Jahren, um dann mit einem Stipendium an der Universität von Chicago zu studieren, wo er mit 17 einen Bachelor in *liberal arts* bekam. Danach finanzierte er sich durch Arbeit ein zweites Bachelorstudium in Physik, das er 1958 abschloß. Damit war seine akademische Ausbildung beendet; er arbeitete als Operator in einem Rechenzentrum, brachte sich selbst Programmieren bei und begann einige Jahre später mit der Publikation wissenschaftlicher Arbeiten auf dem Gebiet der Informatik. Mit 27 wurde er Assistenzprofessor in Carnegie Mellon, fünf Jahre später erhielt er einen Lehrstuhl in Stanford. Zu den vielen Entwicklungen, die er initiierte, gehört die semantische Verifikation von Programmen, Design und Analyse von Algorithmen, Refactoring, dazu kommen Arbeiten über Graphentheorie und das FLOYD-STEINBERG dithering in der Computergraphik. 1978 erhielt er den TURING-Preis, die höchste Auszeichnung der Informatik. Stanfords Nachruf auf FLOYD ist zu finden unter `news-service.stanford.edu/news/2001/november7/floydobit-117.html`.

Damit sieht der Grobablauf der Monte-Carlo-Faktorisierung einer natürlichen Zahl N folgendermaßen aus:

Schritt 0: Man wähle ein quadratisches Polynom Q und einen Startwert x_0. Setze $x = y = x_0$.
Schritt $i, i > 0$**:** Ersetze x durch $Q(x)$ und y durch $Q\big(Q(y)\big)$; berechne dann ggT$(x - y, N)$. Falls dieser weder eins noch N ist, wurde ein Faktor gefunden.

Man beachte, daß hier im i-ten Schritt $x = x_i$ und $y = x_{2i}$ ist; wir erzeugen also die Folge der x_i (Schildkröte) und die der x_{2i} (Hase) simultan, ohne Zwischenergebnisse zu speichern.

Das Teuerste an diesem Algorithmus sind die EUKLIDischen Algorithmen zur ggT-Berechnung; da wir (sofern wir kleine Primfaktoren zuvor ausgeschlossen haben) nicht wirklich erwarten, daß hier häufig ein nichttriviales Ergebnis herauskommt, liegt es nahe, deren Anzahl möglichst zu reduzieren.

Eine Strategie dazu besteht darin, jeweils mehrere Differenzen $x_{2i} - x_i$ aufzumultiplizieren und dann erst für das Produkt den ggT des mit N zu

berechnen. Die „gewisse Anzahl“ darf nicht zu groß sein, denn sonst besteht die Gefahr, daß das Produkt nicht nur durch einen, sondern gleich durch mehrere Primteiler von N teilbar ist, es sollte aber aus Effizienzgründen auch nicht zu klein sein. Wenn alle kleinen Primteiler bereits ausgeschlossen sind, zeigt die Erfahrung, daß die Zusammenfassung von etwa hundert Differenzen ein guter Kompromiß ist; wenn bereits bei den „kleinen“ Faktoren mit einer hohen Suchgrenze gearbeitet wurde, bieten sich auch höhere Werte an.

Praktisch bedeutet das, daß wir eine neue Variable P einführen mit Anfangswert eins und dann im i-ten Schritt P durch $P \cdot (x - y) \bmod N$ ersetzen. Nur falls i durch die „gewisse Anzahl“ m teilbar ist, wird anschließend der ggT von N und P berechnet; andernfalls geht es gleich weiter mit dem $(i + 1)$-ten Schritt.

Die Monte-Carlo-Methode wird auch als ρ-Methode bezeichnet, da die Folge der X_i nicht von Anfang an periodisch sein muß. Sie muß aber, da es nur p Restklassen modulo p gibt, schließlich periodisch werden, d.h. sie beginnt auf dem unteren Ast des ρ und mündet irgendwann in den Kreis. Erfahrungsgemäs ist diese Methode sehr erfolgreich im Auffinden sechs- bis achtstelliger Faktoren; danach wird sie recht langsam, und kleine Faktoren kann sie oft nicht trennen.

Als Beispiel wollen wir die sechste FERMAT-Zahl $F_6 = 2^{64} + 1 = 18\,446\,744\,073\,709\,551\,617$ betrachten. Mit dem quadratischen Polynom $Q(x) = x^2 + 1$, dem Startwert $x_0 = 2$ und einem EUKLIDischen Algorithmus nach jeweils hundert Folgegliedern findet ein handelsüblicher PC nach 900 Iterationen in Sekundenbruchteilen den Faktor 274 177, der übrigens genau wie sein Kofaktor 67 280 421 310 721 prim ist. Damit ist F_6 vollständig faktorisiert.

Anhang: Das Geburtstagsparadoxon

Angenommen, in einem Raum befinden sich n Personen. Wie groß ist die Wahrscheinlichkeit dafür, daß zwei davon am gleichen Tag Geburtstag haben?

Um diese Frage wirklich beantworten zu können, müßte man die (recht inhomogene) Verteilung der Geburtstage über das Jahr kennen; wir

beschränken uns stattdessen auf ein grob vereinfachtes Modell ohne Schaltjahre mit 365 gleich wahrscheinlichen Geburtstagen. Dann ist die Wahrscheinlichkeit dafür, daß von n Personen keine zwei am gleichen Tag Geburtstage haben,

$$\prod_{k=0}^{n-1} \left(1 - \frac{k}{365}\right) ,$$

denn für eine Person ist das überhaupt keine Bedingung, und jede weitere Person muß die Geburtstage der schon betrachteten Personen vermeiden. (Da der Faktor mit $k = 365$ verschwindet, wird die Wahrscheinlichkeit für $n > 365$ zu null, wie es nach dem DIRICHLETschen Schubfachprinzip auch sein muß.)

Nachrechnen ergibt für $n = 23$ ungefähr den Wert 0,4927; bei 23 Personen liegt also die Wahrscheinlichkeit für zwei gleiche Geburtstage bei 50,7%. Tatsächlich dürfte sie noch deutlich höher liegen, denn bei Geburtstagen ist die Annahme einer Gleichverteilung sicherlich falsch.

Bei einer guten Folge von Zufallszahlen sollten die Restklassen modulo p in sehr guter Näherung gleichverteilt sein; die Wahrscheinlichkeit dafür, daß unter n Zufallszahlen keine zwei in der gleichen Restklasse liegen, ist somit

$$P_n = \prod_{k=0}^{n-1} \left(1 - \frac{k}{p}\right) .$$

Da wir uns für einigermaßen große Werte von p interessieren (die kleinen haben wir schon abdividiert), können wir davon ausgehen, daß

$$\left(1 - \frac{1}{p}\right)^p \approx e \quad \text{und} \quad \left(1 - \frac{1}{p}\right) \approx e^{-1/p}$$

ist; für nicht zu große Werte von k ist dann auch

$$\left(1 - \frac{k}{p}\right) \approx e^{-k/p} ,$$

und für nicht zu große Werte von n gilt

$$P_n = \prod_{k=0}^{n-1} \left(1 - \frac{k}{p}\right) \approx \prod_{k=0}^{n-1} e^{-k/p} = e^{-\frac{1}{p}\sum_{k=0}^{n-1} k} = e^{-\frac{n(n-1)}{2p}} .$$

Für $p = 365$ etwa ergibt dies den Näherungswert $p_{23} \approx 0{,}499998$ für den korrekten Wert 0,4927.

Wenn wir im Exponenten noch den Term $n(n-1)$ durch n^2 approximieren, können wir abschätzen, für welches n die Wahrscheinlichkeit P_n einen vorgegebenen Wert erreicht:

$$e^{-\frac{n^2}{2p}} = P \Longleftrightarrow \frac{n^2}{2p} = -\ln P \Longleftrightarrow n = \sqrt{-2p \ln P}\,.$$

Damit liegt P_n bei etwa 50%, falls $n \approx \sqrt{2p \ln 2} \approx 1{,}177\sqrt{p}$ ist; für $p = 365$ ergibt dies die immer noch recht gute Näherung 22,494.

Für $P = 1/1000$ ergibt sich $n \approx 3{,}717\sqrt{p}$, für $P = 999/1000$ entsprechend $n \approx 0{,}0447\sqrt{p}$. Die Wahrscheinlichkeit dafür, daß es unter n Zufallszahlen zwei mit derselben Restklasse modulo p gibt, wechselt also bei der Größenordnung $n \approx \sqrt{p}$ von sehr unwahrscheinlich zu sehr wahrscheinlich.

b) Die ($p-1$)-Methode

POLLARDs zweite Methode beruht auf dem kleinen Satz von FERMAT: Für einen Primteiler p von N, ein Vielfaches r von $p-1$ und eine zu p teilerfremde natürliche Zahl a ist $a^r \equiv 1 \bmod p$; der ggT von $(a^r - 1) \bmod N$ und N ist also durch p teilbar.

Natürlich ist $p-1$ nicht bekannt, wir können aber hoffen, daß $p-1$ nur durch vergleichsweise kleine Primzahlen teilbar ist. Sei etwa B eine Schranke mit der Eigenschaft, daß $p-1$ durch keine Primzahlpotenz größer B teilbar ist. Dann ist das Produkt r *aller* Primzahlpotenzen q^e, die höchstens gleich B sind, sicherlich ein Vielfaches von $p-1$, wenn auch ein extrem großes, das sich kaum mit realistischem Aufwand berechnen läßt. Für jedes konkrete a kann $a^r \bmod N$ jedoch verhältnismäßig einfach berechnet werden: Man potenziert einfach nacheinander für jede Primzahl $q \le B$ modulo N mit deren größter Potenz, die immer noch kleiner oder gleich B ist; mit dem Algorithmus zur modularen Exponentiation aus Kapitel 1 geht das auch für sechs- bis siebenstellige Werte von B noch recht flott.

Insgesamt funktioniert POLLARDs ($p-1$)-Methode zur Faktorisierung einer natürlichen Zahl N also folgendermaßen:

Schritt 0: Wähle eine Schranke B und eine Basis a zwischen 1 und N.

Schritt 1: Erstelle (z.B. nach ERATOSTHENES) eine Liste aller Primzahlen $q \le B$.

Schritt 2: Berechne für jede dieser Primzahlen q den größten Exponenten e derart, daß auch noch $q^e \le B$ ist, d.h. $e = [\log B / \log q]$. Ersetze dann den aktuellen Wert von a durch $a^{q^e} \bmod N$.

Schritt 3: Berechne ggT$(a-1, N)$. Falls ein Wert ungleich eins oder N gefunden wird, war das Verfahren erfolgreich, ansonsten nicht.

Es ist klar, daß der Erfolg dieses Verfahrens wesentlich davon abhängt, daß N einen Primteiler p hat mit der Eigenschaft, daß alle Primfaktoren von $p-1$ relativ klein sind. Ob dies der Fall ist, läßt sich im Voraus nicht sagen; die ($p-1$)-Methode liefert daher gelegentlich ziemlich schnell sogar 20- oder 30-stellige Faktoren, während sie andererseits deutlich kleinere Faktoren oft nicht findet.

Als Beispiel betrachten wir noch einmal $M_{67} = 2^{67} - 1$. Wenn wir mit der Basis $a = 17$ und der Schranke $B = 3\,000$ arbeiten, wird a modulo M_{67} potenziert zum neuen

$$a = 111\,153\,665\,932\,902\,146\,348 \text{ mit } \text{ggT}(a-1, M_{67}) = 193\,707\,721\,.$$

Damit ist (in Sekundenbruchteilen auf einem Standard-PC) eine nichttriviale Faktorisierung gefunden, und ein Primzahltest zeigt, daß sowohl der gefundene Faktor als auch sein Komplement prim sind.

Warum die Methode Erfolg hatte, sehen wir an der Faktorisierung der um eins verminderten Faktoren:

$$193\,707\,720 = 2^3 \cdot 3^3 \cdot 5 \cdot 67 \cdot 2\,677 \quad \text{und}$$

$$761\,838\,257\,286 = 2 \cdot 3^2 \cdot 29 \cdot 67 \cdot 2\,551 \cdot 8\,539\,.$$

Für jede Schranke $B \ge 2\,677$ ist also der erste Faktor ein Teiler des endgültigen $a-1$, aber für $B < 8\,539$ ist der zweite Faktor keiner.

c) Varianten

Falls $p-1$ nicht nur relativ kleine Primfaktoren hat, führt die ($p-1$)-Methode nicht zum Erfolg. In solchen Fällen hat dann aber vielleicht $p+1$

oder irgendeine andere Zahl in der Nähe von p nur kleine Primfaktoren. In solchen Fällen können Varianten der $(p-1)$-Methode zum Erfolg führen.

Um diese Varianten zu definieren, empfiehlt es sich, zunächst die $(p-1)$-Methode etwas abstrakter unter gruppentheoretischen Gesichtspunkten zu betrachten.

Wir rechnen in der primen Restklassengruppe $(\mathbb{Z}/N)^\times$ und damit implizit auch in $(\mathbb{Z}/p)^\times$ für jeden Primteiler p von N – egal ob wir ihn kennen, oder nicht. In $(\mathbb{Z}/p)^\times$ ist für jedes Element a die $(p-1)$-te Potenz gleich dem Einselement; genau dasselbe gibt für jede r-te Potenz, für die der Exponent r ein Vielfaches von $(p-1)$ ist. Bei der $(p-1)$-Methode wird ein r berechnet, das durch alle Primzahlpotenzen bis zu einer gewissen Schranke teilbar ist; falls in der Primzerlegung von $p-1$ keine Primzahlpotenz oberhalb der Schranke liegt, ist r ein Vielfaches von $p-1$.

Allgemeiner können wir statt in $(\mathbb{Z}/N)^\times$ und $(\mathbb{Z}/p)^\times$ auch in einem anderen Paar von Gruppen rechnen: Wir gehen aus von einer endlichen Gruppe G_n, deren Elemente sich in irgendeiner Weise als r-tupel über $(\mathbb{Z}/N)$ auffassen lassen; außerdem nehmen wir an, daß sich die Gruppenmultiplikation für zwei so dargestellte Elemente auf Grundrechenarten über $\mathbb{Z}/N$ zurückführen läßt. Dann können wir die Elemente von G_n zu Tupeln über $\mathbb{Z}/p$ reduzieren und die Menge aller so erhaltenen Tupel bildet eine Gruppe G_p. Wieder ist jede Rechnung in G_n implizit auch eine Rechnung in G_p.

Die Elementanzahl von G_p sei $N(p)$.

Wir wählen irgendein Element von G_n und potenzieren es mit demselben Exponenten r, mit dem wir bei der $p-1$-Methode die Zahl a modulo N potenziert haben. Falls r ein Vielfaches von $N(p)$ ist, erhalten wir ein Element $b \in G_n$, dessen Reduktion modulo p das Einselement von G_p ist. Ist daher b_i die i-te Koordinate von b und e_i die von e, so muß die Differenz $b_i - e_i$ durch p teilbar sein, und mit etwas Glück können wir p als ggT von n und $b_i - e_i$ bestimmen.

Bleibt nur noch das Problem, geeignete Gruppen zu finden. Bei der $(p-1)$-Methode ist $G_n = (\mathbb{Z}/n)^\times$ und $N(p) = p-1$.

Für die $(p+1)$-Methode benutzt POLLARD die Tatsache, daß es nicht nur zu jeder Primzahl p, sondern auch zu jeder Primzahlpotenz p^r einen Körper mit entsprechender Elementanzahl. Dieser Körper $\mathbb{F}_{p^r}$ ist natürlich verschieden vom Ring $\mathbb{Z}/p^r$; er ist ein r-dimensionaler Vektorraum über $\mathbb{F}_p$ mit geeignet definierter Multiplikation.

Speziell für $r = 2$ hat der Körper $\mathbb{F}_{p^2}$ eine multiplikative Gruppe $\mathbb{F}_{p^2}^\times$ der Ordnung $p^2 - 1 = (p+1)(p-1)$. Sie hat $\mathbb{F}_p^\times$ als Untergruppe und die Faktorgruppe $G_p = \mathbb{F}_{p^2}^\times / \mathbb{F}_p^\times$ hat die Ordnung $N(p) = p+1$. Das Rechnen in dieser Gruppe mit Repräsentanten modulo N ist etwas trickreich und benutzt die hier nicht behandelten LUCAS-Sequenzen.

Derzeit am populärsten ist eine andere Wahl von G_n und G_p: Wir nehmen für G_n eine elliptische Kurve über $\mathbb{Z}/n$. Dabei handelt es sich um die Menge aller Punkte $(x, y) \in (\mathbb{Z}/n)^2$, die einer vorgegebenen Gleichung $y^2 = x^3 - ax - b$ genügen, wobei a, b Elemente von $\mathbb{Z}_n$ sind, für die $\Delta = 4a^3 - 27b^2$ teilerfremd zu n ist; dazu kommt ein weiterer Punkt O, den wir formal als $(0, \infty)$ schreiben. G_p ist dann die entsprechende Punktmenge in $\mathbb{F}_p^2$ zusammen mit O. Nach einem Satz von HELMUT HASSE (1898–1979) ist

$$p + 1 - 2\sqrt{p} < N(p) < p + 1 + 2\sqrt{p}\,,$$

und wie man inzwischen weiß, kann man auch für jeden Wert, der diese Ungleichung erfüllt, Parameterwerte a und b finden, so daß $N(p)$ gleich diesem Wert ist. Wenn man mit hinreichend vielen verschiedenen Kurven arbeitet, ist daher die Chance recht groß, daß der Exponent r wenigstens für eine davon ein Vielfaches von $N(p)$ ist.

Die Multiplikation ist folgendermaßen definiert: Durch zwei Punkte (x_1, y_1) und (x_2, y_2) auf der Kurve geht genau eine Gerade; setzt man deren Gleichung $y = mx + c$ in die Kurvengleichung ein, erhält man ein Polynom dritten Grades in x. Dieses hat natürlich die beiden Nullstellen x_1, x_2, und daneben noch eine dritte Nullstelle x_3. Der dritte Schnittpunkt der Geraden mit der Kurve ist somit $(x_3, mx_3 + c)$; als Summe der beiden Punkte definiert man aber

$$(x_1, y_1) \oplus (x_2, y_2) = \big(x_3, -(mx_3 + c)\big)\,.$$

Man kann zeigen, daß dies die Menge der Kurvenpunkte zu einer Gruppe mit Neutralelement O macht, in der man genauso vorgehen kann wie bei der klassischen $(p-1)$-Methode.

Unter den Faktorisierungsmethoden, deren Rechenzeit von der Größe des zu findenden Faktors abhängt, ist die Faktorisierung mit elliptischen Kurven die für große Zahlen derzeit beste bekannte Methode; sie fand schon Faktoren mit bis zu 67 Stellen. Produkte zweier ungefähr gleich großer Primzahlen wie beispielsweise RSA-Moduln sind für solche Methoden allerdings der schlechteste Fall; hierfür sind andere Methoden, deren Aufwand nur von der Größe der zu faktorisierenden Zahl abhängt, meist besser geeignet.

§3: Das Verfahren von Fermat und seine Varianten

Die bisher betrachteten Verfahren funktionieren vor allem dann gut, wenn die zu faktorisierende Zahl mindestens einen relativ kleinen Primteiler hat. Das hier beschriebene Verfahren von FERMAT führt genau dann schnell ans Ziel, wenn sie sich als Produkt zweier fast gleich großer Faktoren schreiben läßt. In seiner einfachsten Form beruht es auf der dritten binomischen Formel $x^2-y^2=(x+y)(x-y)$: Ist $N=pq$ Produkt zweier ungerader Primzahlen, so ist

$$N=(x+y)(x-y) \quad \text{mit} \quad x=\frac{p+q}{2} \quad \text{und} \quad y=\frac{p-q}{2}\,;$$

Ausmultiplizieren führt auf die Beziehung $N+y^2=x^2$.

FERMAT berechnet für $y=0,1,2,\ldots$ die Zahlen $N+y^2$; falls er auf ein Quadrat x^2 stößt, hat er zwei Faktoren $x\pm y$ gefunden. Da y gleich der halben Differenz der beiden Faktoren ist, kommt er umso schneller ans Ziel, je näher die beiden Faktoren beieinander liegen; dies erklärt die Vorschrift der Bundesnetzagentur, daß die beiden Faktoren eines RSA-Moduls zwar ungefähr gleich groß sein sollten, daß sie aber doch einen gewissen Mindestabstand einhalten müssen.

Anstelle der Zahlen $N+y^2$ kann man auch für ein festes k die Zahlen $kN+y^2$ betrachten. Falls dies eine Quadratzahl x^2 ist, gilt entsprechend

$$kN=x^2-y^2=(x+y)(x-y)\,,$$

und wenn man Glück hat, sind $\mathrm{ggT}(x\pm y,N)$ echte Faktoren von N. Wenn man Pech hat, sind es freilich einfach die beiden Zahlen eins und N. Trotzdem lassen sich darauf sehr effiziente Faktorisierungsverfahren aufbauen, denn die obige Gleichung besagt ja auch, daß wir nur *irgendwie* zwei Zahlen x,y finden müssen mit $x^2\equiv y^2 \bmod N$ und dann eine Chance haben, daß $\mathrm{ggT}(x\pm y,N)$ uns zwei Faktoren von N liefert. Je mehr solche Paare (x,y) wir finden, desto größer sind die Erfolgschancen.

Der Grundalgorithmus zum Finden solcher Paare ist das sogenannte quadratische Sieb, mit dem wir uns als nächstes beschäftigen wollen.

In seiner einfachsten Variante wählen wir uns ein quadratisches Polynom, z.B. das Polynom

$$f(x)=\left(x+\left\lceil\sqrt{N}\right\rceil\right)^2-N\,.$$

Für jedes x ist dann $f(x)\equiv\left(x+\left\lceil\sqrt{N}\right\rceil\right)^2 \bmod N$, wobei links und rechts verschiedene Zahlen stehen. Insbesondere steht links im allgemeinen keine Quadratzahl.

Falls wir allerdings Werte $x_1,x_2,\ldots,x_r$ finden können, für die das Produkt der $f(x_i)$ eine Quadratzahl ist, dann ist

$$\prod_{i=1}^{r} f(x_i)\equiv\prod_{i=1}^{r}\left(x+\left\lceil\sqrt{N}\right\rceil\right)^2 \bmod N$$

eine Relation der gesuchten Art.

Diese Strategie erklärt die Wahl des Polynoms f: Wir betrachten sowohl die Zahlen $x+\left\lceil\sqrt{N}\right\rceil$ als auch

$$f(x)=\left(x+\left\lceil\sqrt{N}\right\rceil\right)^2-N=x^2+2x\left\lceil\sqrt{N}\right\rceil+\left\lceil\sqrt{N}\right\rceil^2-N\,.$$

Für x-Werte, die deutlich kleiner als $\sqrt{N}$ sind (und nur mit solchen werden wir es im folgenden zu tun haben) liegen beide Zahlen in der Größenordnung $\sqrt{N}$; bei den meisten anderen Polynomen, die ein Quadrat minus N produzieren, wäre entweder die zu quadrierende Zahl

oder deren Funktionswert deutlich größer. Natürlich kann anstelle von $\left\lceil\sqrt{N}\right\rceil$ genauso gut eine andere Zahl ähnlicher Größenordnung verwendet werden; es kommt nur darauf an, daß ihr Quadrat in der Nähe von N liegt.

Um geeignete x_i zu finden, betrachten wir eine Menge $\mathcal{B}$ von Primzahlen, die sogenannte Faktorbasis. Typischerweise enthält $\mathcal{B}$ für die Faktorisierung einer etwa hundertstelligen Zahl zwischen 100 und 120 Tausend Primzahlen, deren größte somit, wie die folgende Tabelle zeigt, im einstelligen Millionenbereich liegt.

n	n-te Primzahl	n	n-te Primzahl
100 000	1 299 709	600 000	8 960 453
200 000	2 750 159	700 000	10 570 841
300 000	4 256 233	800 000	12 195 257
400 000	5 800 079	900 000	13 834 103
500 000	7 368 787	1 000 000	15 485 863

Beim quadratischen Sieb interessieren nur x-Werte, für die $f(x)$ als Produkt von Primzahlen aus $\mathcal{B}$ (und eventuell auch Potenzen davon) darstellbar ist. Ist $f(x_i) = \prod_{p\in\mathcal{B}} p^{e_{ip}}$, so ist

$$\prod_{i=1}^{r} f(x_i)^{\varepsilon_i} = \prod_{p\in\mathcal{B}} p^{\sum_{i=1}^{r} e_{ip}\varepsilon_i}$$

genau dann ein Quadrat, wenn $\sum_{i=1}^{r} e_{ip}\varepsilon_i$ für alle $p \in \mathcal{B}$ gerade ist. Dies hängt natürlich nur ab von den ε_i mod 2 und den e_{ip} mod 2; wir können ε_i und e_{ip} daher als Elemente des Körpers mit zwei Elementen auffassen und bekommen dann über $\mathbb{F}_2$ die Bedingungen

$$\sum_{i=1}^{r} e_{ip}\varepsilon_i = 0 \quad \text{für alle } p \in \mathcal{B}\,.$$

Betrachten wir die ε_i als Variablen, ist dies ein homogenes lineares Gleichungssystem in r Variablen mit soviel Gleichungen, wie es Primzahlen in der Faktorbasis gibt. Dieses Gleichungssystem hat nichttriviale Lösungen, falls die Anzahl der Variablen die der Gleichungen übersteigt, falls es also mehr Zahlen x_i gibt, für die $f(x_i)$ über der Faktorbasis faktorisiert werden kann, als Primzahlen in der Faktorbasis.

Für jede nichttriviale Lösung ist

$$\prod_{i=1}^{r} f(x_i)^{\varepsilon_i} = \prod_{i=1}^{r} \left(x + \left\lceil\sqrt{N}\right\rceil\right)^{2\varepsilon_i} \bmod N$$

eine Relation der Form $x^2 \equiv y^2 \bmod N$, die mit einer Wahrscheinlichkeit von etwa ein halb zu einer Faktorisierung von N führt. Falls wir zehn linear unabhängige Lösungen des Gleichungssystems betrachten, führt also mit einer Wahrscheinlichkeit von etwa 99,9% mindestens eine davon zu einer Faktorisierung.

Da ε_i nur die Werte 0 und 1 annimmt, stehen in obigem Produkt natürlich keine echten Potenzen: Man multipliziert einfach nur die Faktoren miteinander, für die $\varepsilon_i = 1$ ist. Außerdem interessieren nicht die links- und rechtsstehenden Quadrate, sondern deren Quadratwurzeln; tatsächlich also berechnet man (hier natürlich in $\mathbb{N}_0$)

$$x = \prod_{p\in\mathcal{B}} p^{\frac{1}{2}\sum_{i=1}^{r} \varepsilon_i e_{ip}} \bmod N \quad \text{und} \quad y = \prod_{i=1}^{r} \left(x + \left\lceil\sqrt{N}\right\rceil\right)^{\varepsilon_i} \bmod N\,.$$

Zum besseren Verständnis des Grundprinzips wollen wir versuchen, damit die Zahl 15 zu faktorisieren. Dies ist zwar eine sehr untypische Anwendung, da das quadratische Sieb üblicherweise erst für mindestens etwa vierzigstellige Zahlen angewandt wird, aber zumindestens das Prinzip sollte auch damit klarwerden.

Als Faktorbasis verwenden wir die Menge

$$\mathcal{B} = \{2, 3, 7, 11\}\,;$$

die Primzahl fünf fehlt, da $3 \cdot 5 = 15$ ist und daher bei einer Faktorbasis, die sowohl drei als auch fünf enthält, die Gefahr zu groß ist, daß die linke wie auch die rechte Seite der Kongruenz durch fünfzehn teilbar ist. Bei realistischen Anwendungen muß man auf solche Überlegungen keine Rücksicht nehmen, denn dann sind die Elemente der Faktorbasis höchstens siebenstellig und somit erheblich kleiner als die gesuchten Faktoren.

Wir berechnen $f(x)$ für $x = 1, 2, \ldots$, bis wir einige Funktionswerte haben, die über der Faktorbasis faktorisiert werden können. Die faktorisierbaren Werte sind in folgender Tabelle zusammengestellt:

x	$x + \left\lceil\sqrt{N}\right\rceil$	$f(x)$	*Faktorisierung*
1	4	1	
3	6	21	$3 \cdot 7$
5	8	49	7^2
6	9	66	$2 \cdot 3 \cdot 11$
10	13	154	$2 \cdot 7 \cdot 11$
54	57	3234	$2 \cdot 3 \cdot 7^2 \cdot 11$

Die erste und die dritte Zeile sind selbst schon Relationen der gesuchten Art, nämlich

$$4^2 \equiv 1 \bmod 15 \quad \text{und} \quad 8^2 \equiv 7^2 \bmod 15\,.$$

Die zweite Relation ist nutzlos, denn $8 - 7 = 1$ und $8 + 7 = 15$. Die erste dagegen führt zur Faktorisierung, denn

$$\mathrm{ggT}(4 + 1, 15) = 5 \quad \text{und} \quad \mathrm{ggT}(4 - 1, 15) = 3\,.$$

Da dies aber ein Zufall ist, der bei großen Werten von N so gut wie nie vorkommt, wollen wir das ignorieren und mit den Relationen zu $x = 3, 6, 10$ und 51 arbeiten:

$$\begin{aligned} 6^2 &\equiv 3 \cdot 7 && \bmod 15 \\ 9^2 &\equiv 2 \cdot 3 \cdot 11 && \bmod 15 \\ 13^2 &\equiv 2 \cdot 7 \cdot 11 && \bmod 15 \\ 57^2 &\equiv 2 \cdot 3 \cdot 7^2 \cdot 11 && \bmod 15 \end{aligned}$$

Multipliziert man die ersten drei dieser Relationen miteinander, folgt

$$(6 \cdot 9 \cdot 13)^2 \equiv (2 \cdot 3 \cdot 7 \cdot 11)^2 \bmod 15$$

oder $702^2 \equiv 462^2 \bmod 15$. Da

$$\mathrm{ggT}(702 - 462, 15) = \mathrm{ggT}(240, 15) = 15$$

ist, bringt das leider nichts.

Wir erhalten auch dann rechts ein Quadrat, wenn wir das Produkt der ersten, dritten und vierten Relation bilden; dies führt auf

$$(6 \cdot 13 \cdot 57)^2 \equiv (2 \cdot 3 \cdot 7^2 \cdot 11)^2 \bmod 15$$

oder $4446^2 \equiv 3234^2 \bmod 15$. Hier ist

$$\mathrm{ggT}(4446 - 3234, 15) = \mathrm{ggT}(1212, 15) = 3\,,$$

womit wir die Zahl 15 faktorisiert haben – wenn auch nicht unbedingt auf die einfachstmögliche Weise.

Bei realistischen Beispielen sind die Funktionswerte $f(x)$ deutlich größer als die Primzahlen aus der Faktorbasis; außerdem liegen die vollständig faktorisierbaren Zahlen viel dünner als hier: Bei der Faktorisierung einer hundertstelligen Zahl etwa muß man davon ausgehen, daß nur etwa jeder 10^9-te Funktionswert über der Faktorbasis zerfällt.

Daher ist es wichtig, ein Verfahren zu finden, mit dem diese wenigen Funktionswerte schnell und einfach bestimmt werden können. Das ist zum Glück möglich:

Der Funktionswert $f(x)$ ist genau dann durch p teilbar, wenn

$$f(x) \equiv 0 \bmod p$$

ist. Für ein Polynom f mit ganzzahligen Koeffizienten ist offensichtlich $f(x) \equiv f(y) \bmod p$, falls $x \equiv y \bmod p$ ist. Daher ist für ein x mit $f(x) \equiv 0 \bmod p$ auch

$$f(x + kp) \equiv 0 \bmod p \quad \text{für alle } k \in \mathbb{Z}\,.$$

Es genügt daher, im Bereich $0 \le x < p - 1$ nach Werten zu suchen, für die $f(x)$ durch p teilbar ist.

Dazu kann man f auch als Polynom über dem Körper mit p Elementen betrachten und nach Nullstellen in diesem Körper suchen. Für Polynome großen Grades und große Werte von p kann dies recht aufwendig sein; hier, bei einem quadratischen Polynom, müssen wir natürlich einfach eine quadratische Gleichung lösen: In $\mathbb{F}_p$ wie in jedem anderen Körper auch gilt

$$f(x) = \left(x + \left\lceil\sqrt{N}\right\rceil\right)^2 - N = 0 \Longleftrightarrow \left(x + \left\lceil\sqrt{N}\right\rceil\right)^2 = N\,,$$

und diese Gleichung ist genau dann lösbar, wenn es ein Element $w \in \mathbb{F}_p$ gibt mit Quadrat N, wenn also in $\mathbb{Z}$

$$w^2 \equiv N \bmod p$$

ist. Für $p > 2$ hat $f(x) = 0$ in $\mathbb{F}_p$ dann die beiden Nullstellen

$$x = -\left\lfloor\sqrt{N}\right\rfloor \pm w\,;$$

andernfalls gibt es keine Lösung.

Insbesondere kann also $f(x)$ nur dann durch p teilbar sein, wenn N modulo p ein Quadrat ist; dies ist für etwa die Hälfte aller Primzahlen der Fall. Offensichtlich sind alle anderen Primzahlen nutzlos, und sollten daher gar nicht erst in die Faktorbasis aufgenommen werden.

Im Kapitel über quadratische Reste werden wir sehen, daß sich auch für große N und p leicht und schnell entscheiden läßt, ob N modulo p ein Quadrat ist; der Aufwand entspricht ungefähr dem eines auf N und p angewandten EUKLIDischen Algorithmus. Wir werden dort auch sehen, daß sind für solche Quadrate relativ schnell die beiden Quadratwurzeln modulo p berechnen lassen.

Das eigentliche Sieben zum Auffinden der komplett über der Faktorbasis zerlegbaren Funktionswerte $f(x)$ geht dann folgendermaßen vor sich: Man legt ein Siebintervall $x = 0, 1, \ldots, M$ fest und speichert in einem Feld der Länge $M + 1$ für jedes x eine Approximation von $\log_2 |f(x)|$.

Für jede Primzahl p aus der Faktorbasis berechnet man dann die beiden Nullstellen $x_{1/2}$ von f modulo p im Intervall von 0 bis $p - 1$ und subtrahiert von jedem Feldelement mit Index der Form $x_1 + kp$ oder $x_2 + kp$ eine Approximation von $\log_2 p$.

Falls $f(x)$ über der Faktorbasis komplett faktorisierbar ist, sollte dann am Ende der entsprechende Feldeintrag bis auf Rundungsfehler gleich null sein; um keine Fehler zu machen, untersucht man daher für alle Feldelemente, die betragsmäßig unterhalb einer gewissen Grenze liegen, durch Abdividieren, ob sie wirklich komplett faktorisieren, und man bestimmt auf diese Weise auch *wie* sie faktorisieren. Damit läßt sich dann das oben erwähnte Gleichungssystem über $\mathbb{F}_2$ aufstellen und, falls genügend viele Relationen gefunden sind, nichttrivial so lösen, daß eine der daraus resultierenden Gleichungen $x^2 \equiv y^2 \bmod p$ zu einer nichttrivialen Faktorisierung von N führt.

Als zwar immer noch untypisch kleines Beispiel, das besser und schneller durch Abdividieren faktorisiert werden könnte, betrachten wir die Zahl $N = 5\,352\,499$. Wir nehmen als Faktorbasis alle Primzahlen kleiner hundert modulo derer N ein Quadrat ist; Nachrechnen zeigt, daß

$$\mathcal{B} = \{3, 5, 11, 13, 17, 19, 23, 31, 41, 43, 53, 59, 83, 89\}$$

dann 14 Elemente enthält. Für jedes davon müssen wir die quadratische Gleichung $f(x) \equiv 0 \bmod p$ lösen, was hier natürlich selbst durch Ausprobieren recht schnell möglich wäre. Die Lösungsmengen sind

$p =$	3	5	11	13	17	19	23
Lösungen	$\{1, 2\}$	$\{0, 4\}$	$\{0, 5\}$	$\{4, 11\}$	$\{6, 9\}$	$\{2, 8\}$	$\{0, 20\}$
$p =$	31	41	43	53	59	83	89
Lösungen	$\{27, 28\}$	$\{3, 4\}$	$\{26, 35\}$	$\{39, 52\}$	$\{2, 33\}$	$\{35, 70\}$	$\{23, 68\}$

Wenn wir damit das Intervall der natürlichen Zahlen von 1 bis 20 000 sieben, erhalten wir 18 Zahlen, die über der Faktorbasis komplett zerfallen:

i	x_i	$f(x_i)$	*Faktorisierung*
1	23	104397	$3 \cdot 17 \cdot 23 \cdot 89$
2	121	571857	$3 \cdot 11 \cdot 13 \cdot 31 \cdot 43$
3	533	2747217	$3 \cdot 11 \cdot 17 \cdot 59 \cdot 83$
4	635	3338205	$3 \cdot 5 \cdot 13 \cdot 17 \cdot 19 \cdot 53$
5	741	3974417	$31 \cdot 41 \cdot 53 \cdot 59$
6	895	4938765	$3 \cdot 5 \cdot 13 \cdot 19 \cdot 31 \cdot 43$
7	2013	13361777	$11 \cdot 13 \cdot 41 \cdot 43 \cdot 53$
8	2185	14879505	$3 \cdot 5 \cdot 17 \cdot 23 \cdot 43 \cdot 59$
9	2477	17591601	$3 \cdot 31 \cdot 43 \cdot 53 \cdot 83$
10	2649	19268945	$5 \cdot 19 \cdot 43 \cdot 53 \cdot 89$
11	4163	36586077	$3 \cdot 11 \cdot 19 \cdot 23 \cdot 43 \cdot 59$
12	4801	45256497	$3 \cdot 11 \cdot 13 \cdot 31 \cdot 41 \cdot 83$
13	5497	55643601	$3 \cdot 13 \cdot 17 \cdot 23 \cdot 41 \cdot 89$
14	6253	68023857	$3 \cdot 11 \cdot 19 \cdot 23 \cdot 53 \cdot 89$
15	10991	171643917	$3 \cdot 17 \cdot 23 \cdot 41 \cdot 43 \cdot 83$
16	11275	179281245	$3 \cdot 5 \cdot 11 \cdot 13 \cdot 19 \cdot 53 \cdot 83$
17	14575	279852045	$3 \cdot 5 \cdot 11 \cdot 17 \cdot 19 \cdot 59 \cdot 89$
18	18535	429286605	$3 \cdot 5 \cdot 11 \cdot 23 \cdot 31 \cdot 41 \cdot 89$

Ein Vektor $\vec{\varepsilon} \in \mathbb{F}_2^{18}$, für den $\prod f(x_i)^{\varepsilon_i}$ ein Quadrat ist, löst daher über $\mathbb{F}_2$ das lineare Gleichungssystem mit Matrix

$$\begin{pmatrix}
1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1
\end{pmatrix}.$$

Der GAUSS-Algorithmus führt auf die Lösungen

$$(\sigma, \mu+\rho, \mu+\rho, \mu+\rho, \sigma+\rho+\tau, \nu+\sigma, \lambda+\rho, \sigma,$$
$$\nu+\mu+\lambda, \nu+\sigma+\mu+\tau, \mu+\sigma+\tau, \lambda, \nu+\sigma, \mu, \nu, \rho, \sigma, \tau)$$

mit sechs freien Parametern $\lambda, \mu, \nu, \rho, \sigma, \tau \in \mathbb{F}_2$. Setzen wir zunächst $\lambda = \mu = \sigma = 1$, $\nu = \rho = \tau = 0$, so erhalten wir die Lösung

$$\vec{\varepsilon} = (1,1,1,1,1,1,1,1,0,0,0,1,1,1,0,0,1,0)\,.$$

Sie führt auf die beiden Zahlen

$$x = \prod_{p \in \mathcal{B}} p^{\frac{1}{2}\sum_{i=1}^{r} \varepsilon_i e_{ip}} \bmod N = 854\,237 \quad \text{und}$$

$$y = \prod_{i=1}^{r} \left(x + \left\lceil \sqrt{N} \right\rceil\right)^{\varepsilon_i} \bmod N = 3\,827\,016\,.$$

Leider ist die Differenz dieser beiden Zahlen teilerfremd zu N.

Setzen wir in einem zweiten Versuch $\nu = 1$ statt $\nu = 0$, so erhalten wir die weitere Lösung

$$\vec{\varepsilon} = (1,1,1,1,1,0,1,1,1,1,0,1,0,1,1,0,1,0)\,,$$

die uns die Zahlen

$$x = \prod_{p \in \mathcal{B}} p^{\frac{1}{2}\sum_{i=1}^{r} \varepsilon_i e_{ip}} \bmod N = 1\,020\,903 \quad \text{und}$$

$$y = \prod_{i=1}^{r} \left(x + \left\lceil \sqrt{N} \right\rceil\right)^{\varepsilon_i} \bmod N = 4\,093\,611$$

liefert. Nun ist der ggT der Differenz mit N gleich 1 237, womit wir die Faktorisierung

$$5\,352\,499 = 1237 \times 4327$$

gefunden haben. In diesem Fall sind die beiden Faktoren sogar bereits Primzahlen; das wird natürlich im allgemeinen nicht der Fall sein – es sei denn, man startet wie hier mit dem Produkt zweier Primzahlen.

Natürlich hätte uns jede der bisher behandelten Methoden dieses Ergebnis mit erheblich geringerem Aufwand und auch erheblich schneller geliefert; das quadratische Sieb entwickelt seine Stärken erst bei erheblich größeren Zahlen, für die es dann oft tagelang rechnet.

Dabei verwendet man das quadratische Sieb meist nicht in der hier vorgestellten Einfachstsversion, sondern mit verschiedenen Optimierungen.

Bei realistischen Anwendungen wird der überwiegende Teil der Rechenzeit für das Sieben gebraucht. Dies läßt sich relativ einfach parallelisieren, indem man das Sieben für verschiedene Teilintervalle auf verschiedene Computer verteilt. Auf diese Weise können mehrere Tausend Computer jeweils ein Teilintervall sieben und anschließend die gefundenen Faktorisierungen an eine Zentrale melden. Sobald genügend viele eingegangen sind, kann diese ein lineares Gleichungssystem aufstellen und dieses lösen.

Eine weitere Verbesserung, die zu kürzeren Suchintervallen und damit auch kleineren Zahlen führt, besteht darin, anstelle des einen Polynoms f mehrere Polynome zu verwenden. Auch diese können wieder auf verschiedene Computer verteilt werden. Falls einige der Polynome auch negative Werte annehmen können, muß auch das berücksichtigt werden, indem man bei der Faktorisierung der nach dem Sieben übrig

gebliebenen Zahlen auch noch die -1 als zusätzliche „Primzahl“ in die Faktorbasis aufnimmt.

Ab etwa 120 bis 130 Stellen wird eine Variante schneller, bei der auch mit komplizierteren als nur quadratischen Polynomen gearbeitet wird, das sogenannte Zahlkörpersieb. Es hat seinen Namen daher, daß die dahinterstehende Theorie mit algebraischen Zahlkörpern arbeitet; konkret gerechnet wird allerdings weiterhin mit ganzen Zahlen. Dieses Zahlkörpersieb ist die derzeit beste bekannte Methode zur Faktorisierung von Zahlen, die Produkte zweier Primzahlen ähnlicher Größenordnung sind; von diesem Verfahren geht also die größte Gefahr für RSA aus. Der derzeitige Rekord für dieses Verfahren ist die Faktorisierung einer zweihundertstelligen *challenge number* der Firma RSA.

Kapitel 5
Kettenbrüche

§1: Der Kettenbruchalgorithmus

Der EUKLIDische Algorithmus läßt sich auch verwenden, um eine Zahl durch Brüche zu approximieren. Beginnen wir der Einfachheit halber mit einer rationalen Zahl $\alpha = \frac{n}{m}$ mit $n, m \in \mathbb{N}$. Der erste Schritt des EUKLIDischen Algorithmus dividiert n durch m:

$$n : m = q_0 \text{ Rest } r_1 \Longrightarrow \alpha = \frac{n}{m} = q_0 + \frac{r_1}{m} .$$

Falls $r_1 \neq 0$ ist, wird im zweiten Schritt m durch r_1 dividiert:

$$m : r_1 = q_1 \text{ Rest } r_2 \Longrightarrow \frac{m}{r_1} = q_1 + \frac{r_2}{r_1} \Longrightarrow \alpha = q_0 + \frac{1}{q_1 + \frac{r_2}{r_1}} .$$

Ist auch noch r_2 von Null verschieden, wird sodann r_1 durch r_2 dividiert:

$$r_1 : r_2 = q_2 \text{ Rest } r_3 \Longrightarrow \frac{r_1}{r_2} = q_2 + \frac{r_3}{r_2} \Longrightarrow \alpha = q_0 + \frac{1}{q_1 + \frac{1}{q_2 + \frac{r_3}{r_2}}} ,$$

und so weiter. Die Konstruktion muß nach endlich vielen Schritten abbrechen, denn die Folge der Reste r_i beim EUKLIDischen Algorithmus ist monoton fallend und muß daher schließlich Null erreichen. Damit ist α dargestellt als ein sogenannter **Kettenbruch**.

Wir können die Konstruktion auch so formulieren, daß sie nur von der Zahl $\alpha = \frac{n}{m}$ abhängt: Der Quotient bei der Division mit Rest von n

durch m ist $q_0 = [\alpha]$, und der durch m dividierte Rest ist $\alpha - q_0$. Dies führt zu folgender Formulierung des Algorithmus:

Setze zur Initialisierung $c_0 = [\alpha]$ und schreibe

$$\alpha = c_0 + \alpha_1 \quad \text{mit} \quad 0 \le \alpha_1 < 1\,.$$

Im i-ten Schritt, $i \ge 1$, bricht der Algorithmus ab, falls α_i verschwindet; andernfalls wird c_i definiert als größte ganze Zahl kleiner oder gleich $1/\alpha_i$ und α_{i+1} so, daß gilt

$$\frac{1}{\alpha_i} = c_i + \alpha_{i+1}\,.$$

Offensichtlich ist dann

$$\alpha = c_0 + \alpha_0 = c_0 + \frac{1}{c_1 + \alpha_1} = c_0 + \frac{1}{c_1 + \dfrac{1}{c_2 + \alpha_2}}$$

$$= \cdots = c_0 + \cfrac{1}{c_1 + \cfrac{1}{c_2 + \cfrac{1}{\ddots\, + \cfrac{1}{c_{r-1} + \cfrac{1}{c_r + \alpha_r}}}}}\,.$$

Da diese Darstellung sehr viel Platz verbraucht, verwendet man dafür oft auch die kompaktere Schreibweise

$$\alpha = [c_0, c_1, \ldots, c_r; \alpha_r]\,.$$

Falls der Algorithmus mit $\alpha_r = 0$ abbricht, steht im untersten Bruch natürlich nur c_r im Nenner, und wir schreiben den Kettenbruch kurz als $[c_0, c_1, \ldots, c_r]$.

So, wie der Algorithmus formuliert ist, können wir ihn aber auch auf irrationale Zahlen α anwenden. Dann kann kein α_r verschwinden, denn sonst hätten wir ja eine Darstellung von α als rationale Zahl. Wir können aber nach dem r-ten Schritt abbrechen und den Bruch betrachten, der entsteht, wenn wir $\alpha_r = 0$ setzen. Diesen Bruch bezeichnen wir als die r-te **Konvergente** der Kettenbruchentwicklung von α.

Als Beispiel betrachten wir $\alpha = \sqrt{2}$. Hier ist $c_0 = \lfloor\sqrt{2}\rfloor = 1$ und $\alpha_1 = \sqrt{2} - 1$. Also ist

$$\frac{1}{\alpha_1} = \frac{1}{\sqrt{2}-1} = \frac{\sqrt{2}+1}{(\sqrt{2}-1)(\sqrt{2}+1)} = \sqrt{2}+1\,,$$

d.h. $c_1 = \lfloor 1 + \sqrt{2}\rfloor = 2$ und $\alpha_2 = 1 + \sqrt{2} - 2 = \sqrt{2} - 1 = \alpha_1$. Damit wiederholt sich ab jetzt alles, d.h.

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2+\cdots}}}}}\,.$$

In Analogie zu periodischen Dezimalbrüchen schreibt man dies auch kurz in der Form

$$\sqrt{2} = [1, 2, 2, 2, \ldots] = [1, \overline{2}]\,.$$

Die ersten Partialbrüche sind

$$P_0 = 1, \quad P_1 = 1 + \frac{1}{2} = 1{,}5, \quad P_2 = 1 + \frac{1}{2 + \frac{1}{2}} = \frac{7}{5} = 1{,}4\,,$$

$$P_3 = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \frac{1}{2}}} = \frac{17}{12} = 1{,}41\overline{6} \quad \text{und} \quad P_4 = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \frac{1}{2}}}} = \frac{41}{29}\,,$$

was ungefähr gleich 1,4137931 ist. Die Fehler $\sqrt{2} - P_n$ sind, gerundet auf sechs Nachkommastellen, die Zahlen

$$0{,}414214,\ -0{,}085786,\ 0{,}014214,\ -0{,}002453 \quad \text{und} \quad 0{,}000420\,;$$

verglichen mit den kleinen Nenner $1, 2, 5, 12$ und 29 haben wir also erstaunlich gute Übereinstimmungen, und im übrigen ist auch die Kettenbruchentwicklung erheblich regelmäßiger als die Dezimalbruchdarstellung von $\sqrt{2}$.

Als zweites Beispiel betrachten wir $\alpha = \pi$; hier erhalten wir zunächst $c_0 = 3$ und $\alpha_1 = \pi - 3 \approx 0{,}14159$, sodann

$$c_1 = \left[\frac{1}{\pi - 3}\right] = 7 \quad \text{und} \quad \alpha_2 \approx 0{,}062513285\,.$$

Im nächsten Schritt ist $c_2 = \left[\frac{1}{\alpha_2}\right] = 15$ und $\alpha_3 \approx 0{,}99659976$. Weiter geht es mit $c_3 = 1$, $c_4 = 292$, $c_5 = c_6 = c_7 = 1$, $c_8 = 2$ und $c_9 = 1$. Ein Muster ist weder erkennbar, noch ist eines bekannt.

Die Kettenbruchentwicklung von π ist somit

$$\pi = [3, 7, 15, 1, 292, 1, 1, 1, 2, 1, \dots]\,.$$

Die ersten Partialbrüche und ihre Differenzen von π sind

3	$3\frac{1}{7}$	$3\frac{15}{106}$	$3\frac{16}{113}$	$3\frac{4687}{33102}$
$0{,}14$	$-0{,}0013$	$8{,}3 \cdot 10^{-5}$	$-2{,}7 \cdot 10^{-7}$	$5{,}8 \cdot 10^{-10}$

Auch hier haben wir wieder, verglichen mit der Größe des Nenners, exzellente Approximationseigenschaften.

§2: Geometrische Formulierung

Wir wollen uns zunächst überlegen, daß die Konvergenten der Kettenbruchentwicklung einer irrationalen Zahl stets die bei vorgegebener Größenordnung des Nenners bestmögliche rationale Approximation dieser Zahl liefern.

Dazu betrachten wir (im wesentlichen nach dem Ansatz von HAROLD STARK in seinem Buch *An Introduction to Number Theory,* MIT Press, 1978) das Problem der rationalen Approximation von der geometrischen Seite: Zur reellen Zahl $\alpha > 0$ haben wir die Gerade $y = \alpha x$ durch den Nullpunkt, und offensichtlich ist α genau dann rational, wenn auf dieser Geraden außer dem Nullpunkt noch ein weiterer Punkt (q, p) mit ganzzahligen Koordinaten liegt. Rationale Approximationen erhalten wir durch Punkte $(q, p) \in \mathbb{Z} \times \mathbb{Z}$, die in der Nähe der Geraden liegen.

(Die Reihenfolge der Koordinaten mag auf den ersten Blick verwundern; wir interessieren uns jedoch in erster Linie für die Steigung des Ortsvektors zum Punkt (q, p), und die ist der Bruch p/q.)

Die folgende Konstruktion liefert solche Punkte P_n. Sie liegen für gerade n stets unterhalb der Geraden $y = \alpha x$ und für ungerades n darüber.

Wir starten mit $P_{-2} = (1, 0)$ und $P_{-1} = (0, 1)$.

Zu zwei Punkten $P = (q, p)$ und $P' = (q', p')$, die auf verschiedenen Seiten der Geraden liegen, gibt es stets eine nichtnegative ganze Zahl $c \in \mathbb{N}_0$, so daß $P + cP'$ entweder auf der Geraden liegt oder aber auf derselben Seite wie P, während $P + (c+1)P'$ auf der anderen Seite liegt.

Liegt nämlich beispielsweise P unterhalb der Geraden, so ist $p/q < \alpha$, also $p - \alpha q < 0$. Für den oberhalb der Geraden liegenden Punkt P' ist entsprechend $p' - \alpha q' > 0$. Damit ist klar, daß

$$c = \left[\left|\frac{p - \alpha q}{p' - \alpha q'}\right|\right]$$

das Verlangte leistet. Man überlegt sich leicht, daß diese Formel auch gilt, wenn P oberhalb und P' unterhalb der Geraden liegt.

Zähler und Nenner des obigen Bruchs lassen sich einfach geometrisch interpretieren: $(q, \alpha q)$ hat dieselbe x-Koordinate wie $P = (q, p)$ und liegt auf der Geraden $y = \alpha x$; daher ist $p - \alpha q$ der (gerichtete) vertikale Abstand von P zur Geraden und $p' - \alpha q'$ entsprechend der von P'.

Ausgehend von $P = P_{-2} = (1, 0)$ und $P' = P_{-1} = (0, 1)$ definieren wir nun die Punkte P_n für $n \geq 0$ mit dem gerade konstruierten $c = c_n$ aus ihren beiden Vorgängern rekursiv als

$$P_n = P_{n-2} + c_n P_{n-1}\,.$$

Dann liegt P_n auf derselben Seite der Geraden wie P_{n-2}, für gerades n also unterhalb und für ungerades oberhalb – es sei denn, irgendwann einmal liegt ein P_n *auf* der Geraden. In diesem Fall ist α rational und wir brechen die Konstruktion ab. Für irrationales α erhalten wir eine unendliche Folge von Punkten P_n.

Bezeichnen wir mit $d_n = p_n - \alpha q_n$ den gerichteten vertikalen Abstand des Punktes $P_n = (q_n, p_n)$ von der Geraden $y = \alpha x$, so ist nach obiger Formel

$$c_n = \left[\left|\frac{d_{n-2}}{d_{n-1}}\right|\right] .$$

Daher verschwindet c_n genau dann, wenn $|d_{n-2}| < |d_{n-1}|$ ist.

Ist dagegen $|d_{n-1}| < |d_{n-2}|$, so ist $c_n \geq 1$, und da $P_n = P_{n-2} + c_n P_{n-1}$ auf derselben Seite der Geraden liegt wie P_{n-2}, ist auch

$$\begin{aligned} d_n = d_{n-2} + c_n d_{n-1} &= d_{n-2} + \left[\left|\frac{d_{n-2}}{d_{n-1}}\right|\right] d_{n-1} \\ &= d_{n-1}\left(\left[\left|\frac{d_{n-2}}{d_{n-1}}\right|\right] + \frac{d_{n-2}}{d_{n-1}}\right) \end{aligned}$$

betragsmäßig kleiner als d_{n-1}. (Man beachte, daß d_{n-1} und d_{n-2} verschiedene Vorzeichen haben!) Falls daher für einen Index n der Abstand von P_{n-1} zur Geraden $y = \alpha x$ kleiner ist als der von P_{n-2}, gilt dasselbe auch für alle folgenden Indizes, und ab dem Index n sind alle $c_i \geq 1$.

Die ersten beiden Abstände sind $d_{-2} = -\alpha$ und $d_{-1} = 1$; es hängt von α ab, welche der beiden Zahlen den größeren Betrag hat.

Der nächste Punkt ist $P_0 = (1, c_0)$ mit $c_0 = [\alpha]$, also ist $d_0 = [\alpha] - \alpha$, und der Betrag davon ist kleiner als $d_{-1} = 1$. Somit ist für alle $n \geq 1$ der Koeffizient c_n von Null verschieden und $|d_n| < |d_{n-1}|$.

Aus den Beziehungen $p_n = p_{n-2} + c_n p_{n-1}$ und $q_n = q_{n-2} + c_n q_{n-1}$ sehen wir daher, daß die Folge der q_n wie auch der p_n für $n \geq 1$ strikt monoton ansteigt, während die Folge der Differenzen

$$\left|\alpha - \frac{p_n}{q_n}\right| = \frac{|d_n|}{q_n}$$

strikt monoton fällt. Die Brüche p_n/q_n geben also immer bessere Annäherungen an α.

Wir können die obigen Rekursionsformeln zusammenfassen zur Matrixgleichung

$$\begin{pmatrix} p_n & q_n \\ p_{n-1} & q_{n-1} \end{pmatrix} = \begin{pmatrix} c_n & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} p_{n-1} & q_{n-1} \\ p_{n-2} & q_{n-2} \end{pmatrix} ;$$

wenden wir darauf den Multiplikationssatz für Determinanten an, erhalten wir die Formel

$$p_n q_{n-1} - q_n p_{n-1} = -(p_{n-1} q_{n-2} - q_{n-1} p_{n-2}) .$$

Für $n = 0$ ist $p_{-1}q_{-2} - q_{-1}p_{-2} = 0 \cdot 0 - 1 \cdot 1 = -1$; daraus folgt induktiv die Formel $p_n q_{n-1} - q_n p_{n-1} = (-1)^{n-1}$. Insbesondere sind die Zahlen p_n und q_n stets teilerfremd, p_n/q_n ist also ein gekürzter Bruch.

Als nächstes wollen wir uns überlegen, daß die Folge dieser Brüche gegen α konvergiert. Da P_n und P_{n+1} auf verschiedenen Seiten der Geraden $y = \alpha x$ liegen, ist für $n \geq 0$

$$\begin{aligned} \left|\alpha - \frac{p_n}{q_n}\right| \leq \left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| &= \left|\frac{p_{n+1}q_n - q_{n+1}p_n}{q_n q_{n+1}}\right| \\ = \frac{1}{q_n q_{n+1}} &= \frac{1}{q_n(q_{n-1} + c_{n+1} q_n)} \leq \frac{1}{q_n^2} . \end{aligned}$$

Da die Folge der q_n strikt monoton ansteigt, konvergiert die Folge der p_n/q_n somit gegen α, und dies sogar extrem gut: Ist p/q eine rationale Approximation einer irrationalen Zahl α, so kann der Fehler im allgemeinen bis zu $1/2q$ betragen; hier ist er höchstens $1/q^2$ und tatsächlich wohl, da wir recht grob abgeschätzt haben, meist deutlich kleiner. Wie wir gleich sehen werden, muß umgekehrt p/q eine Konvergente der Kettenbruchentwicklung von α sein, wenn $|\alpha - p/q| < 1/2q^2$ ist.

Zuvor sollten wir uns aber noch überlegen, daß die hier betrachteten Brüche p_n/q_n tatsächlich die Konvergenten der in §1 definierten Kettenbruchentwicklung sind und daß die hier betrachteten Zahlen c_i mit denen übereinstimmen, die der Kettenbruchalgorithmus liefert.

Dazu setzen wir

$$\alpha_n = \left|\frac{d_{n-1}}{d_{n-2}}\right| = -\frac{d_{n-1}}{d_{n-2}} ;$$

zumindest für $n \geq 1$ ist dann $\alpha_n < 1$. Wegen $c_n = \left[\left|d_{n-2}/d_{n-1}\right|\right]$ ist dann $c_n = [1/\alpha_n]$. Division der Beziehung $d_n = d_{n-2} + c_n d_{d-1}$ durch d_{n-1} führt auf

$$\frac{d_n}{d_{n-1}} = \frac{d_{n-2}}{d_{n-1}} + c_n \quad \text{oder} \quad -\alpha_{n+1} = -\frac{1}{\alpha_n} + c_n \quad \text{oder} \quad \frac{1}{\alpha_n} = c_n + \alpha_{n+1} ,$$

was zusammen mit $c_n = [1/\alpha_n]$ und dem Induktionsanfang $\alpha = c_0 + \alpha_1$ genau auf die zu Beginn des Abschnitts konstruierten Folgen der c_n und α_n führt.

Für spätere Anwendungen wollen wir noch eine Formel herleiten, wie sich α aus α_n sowie den Konvergenten p_{n-1}/q_{n-1} und p_{n-2}/q_{n-2} berechnet läßt: Nach Definition ist

$$\alpha_n = -\frac{d_{n-1}}{d_{n-2}} = -\frac{p_{n-1} - \alpha q_{n-1}}{p_{n-2} - \alpha q_{n-2}} .$$

Damit ist $\alpha_n(\alpha q_{n-2} - p_{n-2}) = p_{n-1} - \alpha q_{n-1}$, was durch Umordnung der Terme auf $\alpha(\alpha_n q_{n-2} + \alpha q_{n-1}) = \alpha_n p_{n-2} + p_{n-1}$ führt. Also ist

$$\alpha = \frac{\alpha_n p_{n-2} + p_{n-1}}{\alpha_n q_{n-2} + q_{n-1}} .$$

§3: Optimale Approximation

Nach den Vorbereitungen im letzten Paragraphen können wir nun beweisen, daß Kettenbrüche in der Tat bestmögliche Approximationen geben: Ist r/s irgendein Bruch, dessen Nenner s zwischen den Nennern q_{n-1} und q_n zweier Konvergenten der Kettenbruchentwicklung liegt, so ist p_{n-1}/q_{n-1} eine bessere Approximation als r/s:

Lemma: p_n/q_n seien die Konvergenten der Kettenbruchentwicklung einer reellen Zahl α. Falls α irrational ist oder rational mit einem Nenner echt größer q_n, $n \geq 2$, so ist für jede rationale Zahl r/s mit $s \leq q_n$ und $r/s \notin \{p_{n-1}/q_{n-1}, p_n/q_n\}$

$$\left|\alpha - \frac{r}{s}\right| > \left|\alpha - \frac{p_{n-1}}{q_{n-1}}\right| .$$

Beweis: Wir betrachten die Punkte $P_{n-1} = (q_{n-1}, p_{n-1})$, $P_n = (q_n, p_n)$ und $R = (s, r)$. Es genügt zu zeigen, daß der vertikale Abstand von P_{n-1} zur Geraden $y = \alpha x$ einen kleineren Betrag hat als der von R.

Wir schreiben R als ganzzahlige Linearkombination $R = kP_{n-1} + \ell P_n$ der Punkte P_{n-1} und P_n. Das ist möglich, denn die Determinante des

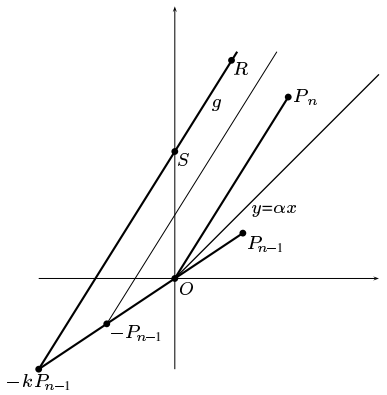
linearen Gleichungssystems

$$\begin{pmatrix} p_{n-1} & p_n \\ q_{n-1} & q_n \end{pmatrix} \begin{pmatrix} k \\ \ell \end{pmatrix} = \begin{pmatrix} r \\ s \end{pmatrix}$$

ist, wie wir oben gesehen haben, gleich $(-1)^{n-1}$; wenn wir es nach der CRAMERschen Regel lösen, erhalten wir also eine ganzzahlige Lösung (k, l).

Für das Folgende wollen wir uns auf den Fall $p_{n-1}/q_{n-1} < \alpha < p_n/q_n$ beschränken; der Fall $p_{n-1}/q_{n-1} > \alpha > p_n/q_n$ geht völlig analog.

Wir betrachten die Geraden g durch kP_{n-1} mit Steigungsvektor $\overrightarrow{OP_n}$; nach unserer Annahme ist ihre Steigung somit größer als α.



Im Fall $k < 0$ liegt der Punkt kP_{n-1} und damit die ganze Gerade g zumindest ab dem Punkt kP_{n-1} oberhalb der Geraden $y = \alpha x$, und wegen der größeren Steigung von g steigt der Abstand zwischen den beiden Geraden mit wachsendem x. Wir können den Abstand von R zur Geraden $y = \alpha x$ daher nach unten abschätzen durch den Abstand des Schnittpunkts S von g mit der y-Achse. Dessen Abstand wiederum können wir nach unten abschätzen, indem wir $k = -1$ setzen, denn in diesem Fall ist der Abstand von g zur Geraden $y = \alpha x$ am kleinsten. Der Punkt $-P_{n-1}$ hat (betragsmäßig) denselben Abstand von $y = \alpha x$ wie P_{n-1}, und da die Abszisse $x = 0$ von S größer ist als die von $-P_n$, hat somit S einen größeren Abstand von $y = \alpha x$ als P_{n-1}. Im Fall $k < 0$ ist damit die Behauptung bewiesen.

Als nächstes betrachten wir den Fall $k > 0$. Dann muß $\ell \leq 0$ sein, denn sonst wäre die x-Koordinate $s = kq_{n-1} + \ell q_n$ von R größer als q_n. Der Punkt kP_{n-1} liegt unterhalb der Geraden $y = \alpha x$ und die Gerade g nähert sich dieser mit steigender Abszisse immer mehr an. Da der Punkt R entweder dieselbe Abszisse wie kP_{n-1} hat oder eine kleinere, ist sein

Abstand somit höchstens gleich dem von kP_{n-1}, der wiederum das k-fache des Abstands von P_{n-1} ist. Für $k \geq 2$ erhalten wir damit die gewünschte strikte Ungleichung. Für $k = 1$ erhalten wir auch eine, denn wegen der Voraussetzung $R \neq P_{n-1}$ muß dann $\ell \geq 1$ sein.

Bleibt noch der Fall $k = 0$. Dann ist $R = \ell P_n$, wobei $\ell \neq 1$, da $R \neq P_n$. Anderseits kann ℓ auch nicht größer als eins sein, denn $s \leq q_n$. Somit kommt dieser Fall gar nicht vor.

∎

Als nächstes wollen wir uns überlegen, wann gute Approximationen Konvergenten der Kettenbruchentwicklung sein müssen. Wir wissen bereits, daß für die Konvergenten gilt

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2}\,.$$

Dies charakterisiert die Konvergenten allerdings noch nicht: Betrachten wir etwa die Kettenbruchentwicklung von $\alpha = \sqrt{3}$. Der Algorithmus liefert zunächst $c_0 = \lfloor\sqrt{3}\rfloor = 1$ und $\alpha_1 = \sqrt{3} - 1$. Der Kehrwert davon ist

$$\frac{1}{\sqrt{3}-1} = \frac{\sqrt{3}+1}{2} \Longrightarrow c_1 = 1 \quad \text{und} \quad \alpha_2 = \frac{\sqrt{3}-1}{2}\,.$$

Der Kehrwert davon ist

$$\frac{2}{\sqrt{3}-1} = \sqrt{3}+1 \Longrightarrow c_2 = 2 \quad \text{und} \quad \alpha_3 = \sqrt{3}-1 = \alpha_1\,.$$

Ab hier wiederholt sich also alles periodisch, d.h.

$$\sqrt{3} = 1 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{2 + \cdots}}}} = [1, \overline{1,2}]\,.$$

Die ersten Konvergenten der Kettenbruchentwicklung sind

$$1, \quad 2, \quad 1\frac{2}{3}, \quad 1\frac{3}{4}, \quad 1\frac{8}{11} \quad \text{und} \quad 1\frac{11}{15}\,;$$

da die Folge der Nenner monoton steigt, gibt es also keine Konvergente mit Nenner sieben. Trotzdem ist

$$\left|\sqrt{3} - 1\frac{5}{7}\right| \approx 0{,}017765 < 0{,}2 = \frac{1}{50} < \frac{1}{49} = \frac{1}{7^2}\,.$$

Dafür gilt aber

Satz: *a)* Eine irrationale Zahl α erfüllt für jedes $n \geq 2$ mindestens eine der beiden Ungleichungen

$$\left|\alpha - \frac{p_{n-1}}{q_{n-1}}\right| < \frac{1}{2q_{n-1}^2} \quad \text{und} \quad \left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{2q_n^2}\,.$$

b) Erfüllen zwei ganze Zahlen p, q die Ungleichung $\left|\alpha - \frac{p}{q}\right| < \frac{1}{2q^2}$, so ist $\frac{p}{q}$ eine Konvergente der Kettenbruchentwicklung von α.

Beweis: a) Angenommen, beide Ungleichungen sind falsch. Nach Multiplikation mit q_{n-1} *bzw.* q_n haben wir dann die beiden Relationen

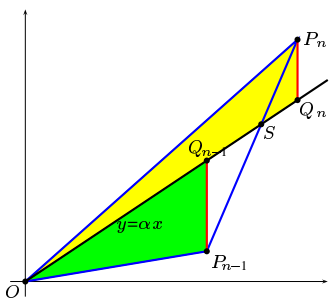
$$\left|q_{n-1}\alpha - p_{n-1}\right| \geq \frac{1}{2q_{n-1}} \quad \text{und} \quad \left|q_n\alpha - p_n\right| \geq \frac{1}{2q_n}\,.$$

Wir nehmen für den Beweis wieder an, daß $p_{n-1}/q_{n-1} < \alpha < p_n/q_n$ ist; der umgekehrte Fall geht völlig analog.

Nach unserer Annahme liegt der Punkt $P_{n-1} = (q_{n-1}, p_{n-1})$ unterhalb der Geraden $y = \alpha x$, und $P_n = (q_n, p_n)$ liegt darüber.

Das Kreuzprodukt (siehe Anhang) der Vektoren $\overrightarrow{OP_{n-1}}$ und $\overrightarrow{OP_n}$ hat als Betrag die Fläche des davon aufgespannten Parallelogramms; das Dreieck mit Ecken O, P_{n-1} und P_n ist halb so groß. Wegen der Beziehung $p_n q_{n-1} - q_n p_{n-1} = (-1)^{n-1}$ ist die Fläche dieses Dreiecks daher gleich $1/2$.

Als nächstes betrachten wir zu den Punkten $P_i = (q_i, p_i)$ ihre Projektionen $Q_i = (q_i, \alpha q_i)$ in y-Richtung auf die Gerade $y = \alpha x$ und die Dreiecke $\triangle OP_iQ_i$. Nach Voraussetzung ist die Länge der Seite P_iQ_i für $i = n-1$ und $i = n$ mindestens $1/2q_i$. Die darauf senkrecht stehende Höhe ist q_i, also ist die Fläche jedes der beiden Dreiecks mindestens $1/4$.



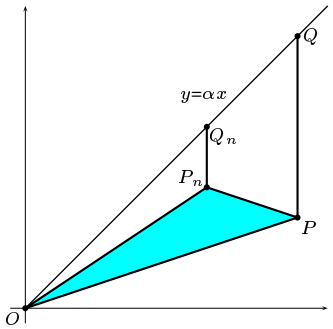
Ist S der Schnittpunkt der Geraden $y = \alpha x$ mit der Verbindungsstrecke von P_{n-1} und P_n, so ist das Dreieck $\triangle OP_{n-1}P_n$ die Vereinigung der Dreiecke $\triangle OP_{n-1}Q_{n-1}$, $\triangle OP_nQ_n$ und $\triangle P_{n-1}Q_{n-1}S$, minus dem Dreieck $\triangle SP_nQ_n$. Die Dreiecke beiden $\triangle P_{n-1}Q_{n-1}S$ und $\triangle SP_nQ_n$ sind ähnlich, und da jede Konvergente eine bessere Approximation liefert als ihre Vorgänger, ist das zweite dieser Dreiecke das kleinere. Daher ist die Fläche des Dreiecks $\triangle OP_{n-1}P_n$ größer als die Summe der Flächen der Dreiecke $\triangle OP_{n-1}Q_{n-1}$ und $\triangle OP_nQ_n$, also größer als $1/4 + 1/4 = 1/2$. Dies ist ein Widerspruch zur obigen direkten Berechnung dieser Fläche.

b) Wir können natürlich voraussetzen, daß der Bruch p/q gekürzt ist, denn für jede nichtgekürzte Darstellung ist die Bedingung echt schärfer.

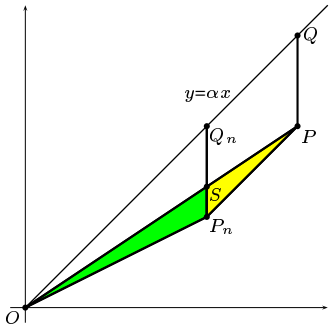
Da die Folge der Nenner q_n strikt monoton ansteigt, gibt es genau ein n, so daß $q_n \le q < q_{n+1}$ ist; wir müssen zeigen, daß $p/q = p_n/q_n$ ist. Andernfalls ist $pq_n - qp_n \neq 0$, also – da dies eine ganze Zahl ist – $|pq_n - qp_n| \ge 1$. Setzen wir $P = (q, p)$, so ist also die Fläche des Dreiecks $\triangle OPP_n$ mindestens gleich $1/2$.

Seien wieder $Q = (q, \alpha q)$ und $Q_n = (q_n, \alpha q_n)$ die Projektionen der betrachteten Punkte auf die Gerade $y = \alpha x$. Die Länge der Strecke $\overline{PQ}$ ist $|\alpha q - p|$, was nach Voraussetzung kleiner als $1/2q$ ist. Nach dem Lemma zu Beginn dieses Paragraphen ist die Strecke $\overline{P_nQ_n}$ kürzer als $\overline{PQ}$, also ebenfalls kleiner als $1/2q$ und damit erst recht kleiner als $1/2q_n$. Somit haben beide Dreiecke $\triangle OPQ$ und $\triangle OP_nQ_n$ Flächen, die kleiner sind als $1/4$.

Wir wollen uns überlegen, daß dann auch die Fläche des Dreiecks $\triangle OPP_n$ kleiner als $1/2$ sein muß, im Widerspruch zur obigen Rechnung. Die Geometrie hängt dabei stark davon ab, wie die Punkte P und P_n sowohl zueinander wie auch in Bezug auf die Gerade $y = \alpha x$ liegen.



Betrachten wir als erstes den Fall, daß p_n/q_n zwischen α und p/q liegt. Dann liegt der Punkt P_n im Innern des Dreiecks $\triangle OPQ$, also ist das gesamte Dreieck $\triangle OPP_n$ im Dreieck $\triangle OPQ$ enthalten. Da ersteres mindestens die Fläche $1/2$ hat, letzteres aber weniger als $1/4$, kann dieser Fall offensichtlich nicht vorkommen.

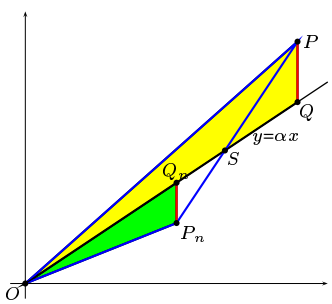


Als nächstes nehmen wir an, p/q liege zwischen α und p_n/q_n. Dann schneiden sich die Strecken $\overline{P_nQ_n}$ und $\overline{OP}$ in einem Punkt S, und das Dreieck $\triangle OPP_n$ ist die Vereinigung der beiden Dreiecke $\triangle OSP_n$ und $\triangle SPP_n$. Zur Flächenberechnung gehen wir aus von der gemeinsamen Kante $\overline{SP_n}$; die darauf senkrecht stehenden Höhen haben die Längen q_n und $q - q_n$. Somit ist die verdoppelte Fläche des gesamten Dreiecks $\triangle OPP_n$ gleich

$$\left|\overline{SP_n}\right| \cdot q_n + \left|\overline{SP_n}\right| \cdot (q - q_n) = \left|\overline{SP_n}\right| \cdot q \le \left|\overline{P_nQ_n}\right| \cdot q \le \left|\overline{PQ}\right| \cdot q\,,$$

denn da q zwischen q_n und q_{n+1} liegt, kann P_n nach obigem Lemma keinen größeren Abstand von der Geraden $y = \alpha x$ haben als P. Rechts steht aber die verdoppelte Fläche des Dreiecks $\triangle OPQ$, von der wir wissen, daß sie höchstens gleich $1/2$ ist, so daß auch dieser Fall nicht auftreten kann.

Bleibt noch der Fall, daß α zwischen p/q und p_n/q_n liegt, P und P_n also auf verschiedenen Seiten der Geraden $y = \alpha x$ liegen. Dann schneidet ihre Verbinungsstrecke $\overline{PP_n}$ diese Gerade in einem Punkt S. Damit sind wir in einer ähnlichen Situation wie beim Beweis von *a):* Das Dreieck $\triangle OPP_n$ ist gleich dem Dreieck $\triangle OP_nQ_n$ plus dem Dreieck $\triangle OPQ$ plus $\triangle SP_nQ_n$ minus $\triangle SPQ$. Die beiden letzteren Drei-



ecke sind ähnlich, und da $\overline{PQ}$ nicht kürzer sein kann als $\overline{P_nQ_n}$ ist das subtrahierte Dreieck mindestens genauso groß wie $\triangle SP_nQ_n$. Somit ist die Fläche von $\triangle OPP_n$ höchstens gleich der Summe der Flächen von $\triangle OPQ$ und $\triangle OP_nQ_n$, also kleiner als $1/4 + 1/4 = 1/2$. Damit haben wir auch hier einen Widerspruch, d.h. p/q muß gleich p_n/q_n sein. ∎

Anhang: Das Kreuzprodukt zweier Vektoren

Im $\mathbb{R}^3$ (und nur dort) gibt es eine bilineare Verknüpfung, die zwei Vektoren einen dritten zuordnet, das (vielleicht aus der Schule bekannte) Vektorprodukt oder Kreuzprodukt. Wie schon der Name sagt, ordnet es je zwei Vektoren v und w aus $\mathbb{R}^3$ einen *Vektor* zu, und dieser wird mit $v \times w \in \mathbb{R}^3$ bezeichnet. Er ist festgelegt durch folgende Eigenschaften:

- $v \times w$ hat die Länge $|v \times w| = |v|\,|w|\,|\sin \measuredangle(v,w)|$.
  Insbesondere ist also $v \times w = \vec{0}$, wenn v und w auf einer Geraden liegen, denn dann bilden sie einen Winkel von null oder 180 Grad, so daß der Sinus verschwindet.
- $v \times w$ steht senkrecht sowohl auf v als auch auf w.
  Falls $v \times w \neq \vec{0}$ ist, spannen v und w eine Ebene auf, auf der (da wir im $\mathbb{R}^3$ sind) genau ein eindimensionaler Unterraum senkrecht steht. Darin gibt es allerdings für jede vorgegebene positive Länge zwei Vektoren, die sich durch ihr Vorzeichen unterscheiden. Um $v \times w$ eindeutig festzulegen, brauchen wir daher noch eine weitere Bedingung:
- Die drei Vektoren v, w und $v \times w$ bilden ein Rechtssystem, d.h. wenn sich die Finger der *rechten* Hand so ausrichten lassen, daß der Daumen in Richtung von v zeigt, der Zeigefinger in Richtung von w und der Mittelfinger in Richtung von $v \times w$.

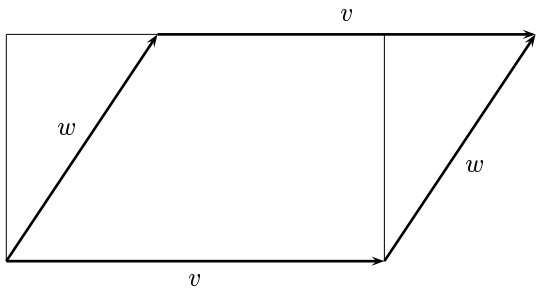
Alternativ kann man ein Rechtssystem auch so definieren, daß sich, ein von v nach w gedrehter Korkenzieher in Richtung $v \times w$ in den Kork

bohrt. Ähnlich geht es auch mit Schrauben; da es allerdings neben den (üblichen) Rechtsschrauben auch die (seltenen) Linksschrauben gibt, ist diese Definition eventuell zirkulär: Alles hängt davon ab, wie man Rechtsschrauben definiert.

Aus jeder dieser Regeln folgt sofort die *Antikommutativität* des Vektorprodukts:

$$v \times w = -w \times v\,.$$

Weitere Rechenregeln lassen sich leicht geometrisch ableiten: Da der Sinus eines Winkels gleich Gegenkathete durch Hypothenuse ist, ist in der von v und w aufgespannten Ebenen $|w|\,|\sin \measuredangle(v,w)|$ gleich der Länge des auf die senkrecht auf v stehenden Geraden projizierten Vektors w, das heißt also gleich der Höhe des in der Abbildung eingezeichneten Rechtecks. Die Länge des Vektors $v \times w$ ist daher gleich dem Flächeninhalt dieses Rechtecks und damit – wie eine Scherung zeigt – gleich der Fläche des von v und w aufgespannten Parallelogramms.



Daraus folgt nun sofort das Distributivgesetz

$$v \times (w+u) = v \times w + v \times u$$

für den zweiten Faktor, und wegen der Antikommutativität folgt daraus wiederum das für den ersten:

$$(u+v) \times w = u \times w + v \times w\,.$$

Um das Vektorprodukt in Koordinaten ausrechnen zu können, müssen wir zunächst die Produkte der Koordinateneinheitsvektoren e_i kennen.

Da sie allesamt die Länge eins haben und paarweise aufeinander senkrecht stehen, ist klar, daß das Produkt zweier verschiedener dieser Vektoren bis aufs Vorzeichen gleich dem dritten ist; das Vorzeichen hängt ab von der Orientierung des Koordinatensystems. Das Produkt eines Vektors e_i mit sich selbst ist natürlich, wie jedes Produkt eines Vektors mit sich selbst, gleich dem Nullvektor, denn der eingeschlossene Winkel ist null Grad.

Für die folgende Rechnung wollen wir annehmen, daß e_1, e_2 und e_3 in dieser Reihenfolge ein Rechtssystem bilden; das ist beispielsweise dann der Fall, wenn e_1 nach rechts, e_2 nach vorne und e_3 nach oben zeigt. Dann folgt sofort, daß

$$e_1 \times e_2 = e_3$$

ist, und nach einigen Fingerübungen auch findet man auch die Formeln

$$e_2 \times e_3 = e_1 \quad \text{und} \quad e_1 \times e_3 = -e_2\,.$$

Die Produkte mit vertauschten Faktoren sind natürlich gerade das negative davon, und $e_i \times e_i = 0$. Für

$$v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \quad \text{und} \quad w = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}$$

ist also

$$v \times w = (v_1e_1 + v_2e_2 + v_3e_3) \times (w_1e_1 + w_2e_2 + w_3e_3)\,,$$

nach den obigen Rechenregeln gleich

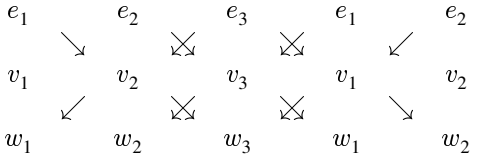
$$\sum_{i=1}^{3}\sum_{j=1}^{3} v_iw_j\, e_i \times e_j$$

$$= (v_2w_3 - v_3w_2)e_1 + (v_3w_1 - v_1w_3)e_2 + (v_1w_2 - v_2w_1)e_3\,,$$

d.h.

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \times \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} v_2w_3 - v_3w_2 \\ v_3w_1 - v_1w_3 \\ v_1w_2 - v_2w_1 \end{pmatrix}\,.$$

Dies läßt sich dadurch merken, daß man im Schema



von e_i ausgeht und als dessen Koeffizient das Zweierprodukt entlang der schrägen Linie nach rechts unten *positiv* und das entlang der schrägen Linie nach links unten *negativ* nimmt; man wendet also die SARRUSsche Regel an auf die „Determinante“

$$\begin{vmatrix} e_1 & e_2 & e_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}\,.$$

§4: Kettenbrüche und Kalender

Schon in den ältesten bekannten Kulturen richtete sich die Zeitrechnung nach astronomischen Gesetzmäßigkeiten: dem Umlauf der Erde um die Sonne, dem Umlauf des Mondes um die Erde sowie der Drehung der Erde um sich selbst.

Der Tag als Zeiteinheit ist ein so selbstverständlicher Teil unseres Lebensrhythmus, daß er als Zeiteinheit nie zur Debatte stand. Ebenso selbstverständlich war, daß als Tag nicht die Dauer einer vollen Drehung der Erde um ihre Achse genommen wurde, sondern der etwa vier Minuten längere Zeitraum, bis sie der Sonne wieder dieselbe Stelle zuwendet.

Als nächstgrößere Einheit führten wahrscheinlich die Babylonier die Woche ein; ob sie sich dabei vom ungefähren Abstand zwischen zwei Mondphasen leiten ließen, ist unbekannt; vielleicht wurde die Dauer von sieben Tagen auch einfach deshalb gewählt, weil die Sieben als heilige Zahl galt.

Definitiv vom Mond abgeleitet ist der Monat. Dieser dreht sich bekanntlich um die Erde; die Zeit für einen vollständigen Umlauf beträgt ungefähr 27,3 Tage. Dieser sogenannte *siderische* Monat spielt allerdings

für die Kalenderrechnung keine Rolle; für die Zeitbestimmung wurden seit Alters her die gut und einfach zu beobachtenden Mondphasen verwendet. Da der Mond nicht selbst leuchtet, sondern das Sonnenlicht reflektiert, hängen diese ab vom Winkelabstand zwischen Sonne und Mond; für den Kalender relevant ist daher der sogenannte *synodische* Monat von 29,53 Tagen, nach dem sich dieser Winkelabstand wiederholt. (Der tatsächliche Abstand zwischen zwei Neumonden ist wegen der komplizierten Mondbewegung keine Konstante; erst im Mittel kommt man auf den synodischen Monat.)

Da 29,53 keine ganze Zahl ist, lassen sich Monate nicht einfach als eine feste Anzahl von Tagen definieren sondern müssen in einem lunaren Kalender mal 29, mal 30 Tage haben.

Einer der einfachsten und zugleich einer der jüngsten dieser Kalender ist der islamische: Sobald mindestens zwei vertrauenswürdige Männer den neuen Mond gesehen haben, beginnt ein neuer Monat, bei bewölktem Himmel unabhängig davon dreißig Tage nach dem letzten Monatsanfang. Alle zwölf Monate beginnt ein neues Jahr.

Es ist klar, daß bei einer solchen Festlegung die Länge der Monate sowohl innerhalb eines Jahres als auch von Jahr zu Jahr schwankt, außerdem sind die Jahre nicht synchron zum Umlauf der Erde um die Sonne. Da der Kalender auf der arabischen Halbinsel entstand, wo Jahreszeiten keine Rolle spielen und auch die Landwirtschaft das ganze Jahr über konstante Bedingungen vorfindet, ist letzteres dort kein Nachteil.

Für Regionen mit ausgeprägten Jahreszeiten oder jährlich wiederkehrenden Ereignissen ist die Synchronisation des Kalenders mit der Sonne wichtiger als die mit dem Mond. Das wohl älteste Beispiel eines reinen Sonnenjahrs bietet der ägyptische Kalender. Da die für die Landwirtschaft fundamentalen jährlichen Nilüberschwemmungen ungefähr mit der ersten Sichtung des Sterns Sirius übereinstimmten, wurde dieses Ereignis als Beginn des neuen Jahrs genommen. Dies ist das sogenannte *siderische* Jahr mit einer Länge von 365,256 Tagen. Obwohl die Ägypter wußten, daß diese Länge ungefähr $365\frac{1}{4}$ Tage beträgt, legten Sie doch fest, daß jedes Jahr genau 365 Tage haben sollte, verteilt auf zwölf Monate zu jeweils dreißig Tagen sowie fünf Zusatztagen.

Ein Jahr, das wirklich synchron zu den Jahreszeiten ist, sollte allerdings nicht anhand des Fixsternhimmels definiert werden, sondern anhand jahreszeitlicher Phänomene wie beispielsweise der Tag- und Nachtgleiche oder dem Durchgang der Sonne durch den Frühlingspunkt. Das ist (im Mittel) das sogenannte *tropische* Jahr mit einer Länge von 356,2422 Tagen. Der Unterschied zum siderischen Jahr ist zwar gering, aber – wie wir gleich sehen werden – trotzdem relevant.

Viel bedeutender als dieser Unterschied war aber zunächst einmal die Tatsache, daß ein Jahr mit exakt 365 Tagen natürlich im Laufe der Jahrhunderte zu einem Verlust der Synchronisation des Kalenders mit den Jahreszeiten führt. Aus diesem Grund beauftragte GAIUS JULIUS CAESAR (100–44) den alexandrinischen Astronomen SOSIGENES mit einer Kalenderreform, die die damit verbundene Verschiebung des Jahresanfang (die sich in nur 120 Jahren auf einen Monat summiert) kompensieren sollte. Das Ergebnis, der *Julianischer Kalender* wurde offiziell eingeführt zum 1. Januar des Jahres 709 *ab urbe condita,* d.h. nach Gründung der Stadt Rom. In unserer heutigen Zeitrechnung handelt es sich dabei um das Jahr 45 v.Chr. Die Monatsnamen und -längen des Julianischen Kalenders sind die heute noch gebräuchlichen. Er unterscheidet sich vom klassischen ägyptischen Kalender durch die zusätzliche Regel, daß in jedem vierten Jahr ein Schalttag eingeführt wird, der 29. Februar.

Dabei blieb es bis ins sechzehnte Jahrhundert. Bis dahin hatte das astronomisch etwas zu lange Julianische Jahr zu einer Verschiebung beispielsweise der Frühjahrssonnenwende um rund elf Tage geführt. Um dies zu korrigieren, setzte Papst GREGOR XIII (UGO BONCOMPAGNI, 1502–1585, Papst ab 1572) eine Kommission ein, auf Grund von deren Empfehlungen in den Jahren, die durch hundert aber nicht durch vierhundert teilbar sind, auf den Schalttag verzichtet wird. Dieser *Gregorianische Kalender* trat in den katholischen Ländern am Freitag, dem 15. Oktober 1582 in Kraft; um die bis dahin akkumulierten Fehler des Julianischen Kalenders zu kompensieren, folgte dieser Tag auf den noch Julianischen Donnerstag, den 4. Oktober 1582. In nichtkatholischen Ländern galt der Julianische Kalender noch länger, wurde aber schließlich (mit den verschiedensten Übergangsregeln) überall durch den Gregorianischen ersetzt – zuletzt 1927 in der Türkei.

Um die Neuerung des Gregorianischen Kalenders zu verstehen, betrachten wir die Kettenbruchentwicklung von 365,2422; sie ist

$$[365, 4, 7, 1, 3, 4, 1, 1, 1, 2]$$

und hat die Konvergenten

$$365, \quad 365\frac{1}{4}, \quad 365\frac{7}{29}, \quad 365\frac{8}{33}, \quad 365\frac{31}{128}, \quad 365\frac{132}{545}, \quad \dots .$$

Der Julianische Kalender verwendet die einfach zu realisierende Konvergente $365\frac{1}{4}$. Unter den folgenden Konvergenten finden wir keine mit einem Nenner, der sich gut für eine einfache Kalenderregel eignen würde. Am ehesten kommt vielleicht noch der Nenner 33 in Frage, da er ungefähr ein Drittel von hundert ist. Arbeitet man mit der Approximation $365\frac{8}{33}$, so sollten unter 33 Jahren acht Schaltjahre sein, also 24 pro 99 Jahre. Nimmt man stattdessen 24 Schaltjahre pro Jahrhundert, was der Regel entspricht, daß durch hundert teilbaren Jahre *keine* Schaltjahre sind, so sorgt der Unterschied zwischen 99 und 100 dafür, daß nach jeweils 400 Jahren eine Vierjahresperiode fehlt. Auch diese hat Anspruch auf ein Schaltjahr, daher die Gregorianische Regel, daß durch hundert teilbare Jahre *keine* Schaltjahre sind, es sei denn, die Jahreszahl sei sogar durch vierhundert teilbar. Innerhalb einer jeden Periode von 400 Jahren gibt es also $100 - 3 = 97$ Schaltjahre; das Gregorianische Jahr hat somit eine Länge von $365\frac{97}{400}$ Tagen, eine praktikable Zahl in der Nähe der Konvergente $365\frac{8}{33}$.

Beginnen wir mit dem einfachsten Problem, den Wochentagen, und überlegen wir uns, auf welchen Wochentag der T-te Tag des M-ten Monats im Jahr J fällt.

Alle Wochen haben exakt sieben Tage, die Jahre haben nach einer relativ klaren Regel 365 oder 366 Tage; es ist daher relativ einfach, den Wochentag für den i-ten Tag des Jahres zu berechnen, sofern man ihn für irgendeinen anderen Tag dieses Jahres kennt: er hängt innerhalb eines Jahres schließlich nur ab von $i \bmod 7$.

Um auch die Abhängigkeit vom Jahr noch zu berücksichtigen, ist es am einfachsten, die Tage nicht vom 1. Januar des jeweils betrachteten Jahres aus zu zählen, sondern ab irgendeinem festen Datum. Historisch sinnvoll

wäre hier beispielsweise das Datum der Einführung des Gregorianischen Kalenders; da hier aber Jahr, Monat und Tag „krumme“ Zahlen sind, würde dies zu unnötig komplizierten mathematischen Formeln mit zu vielen willkürlich erscheinenden Konstanten führen.

Für die Mathematik ist es unerheblich, ob zum fiktiven Anfangspunkt bereits der Gregorianische Kalender in Gebrauch war oder nicht; wir können daher beispielsweise ausgehen von einem 1. Januar eines fiktiven Jahres Null. (Die Zählung der Jahre ab Christi Geburt wurde im sechsten Jahrhundert initiiert von DIONYSIUS EXIGUUS, der allerdings ein falsches Geburtsjahr 1 berechnete, auf das wir uns heute noch beziehen. Jahre davor interessierten ihn nicht; der erste der auch Jahre zuvor in Bezug auf Christi Geburt datierte, war wohl der angelsächsische Theologe und Historiker BEDA VENERABILIS (673–735), der – da er keine Null kannte – das Jahr vor dem Jahr eins *nach* Christus als eins *vor* Christus bezeichnete.)

Wenn wir dem fiktiven 1. Januar 0 die Nummer eins geben, können wir die Nummer des 31. Dezembers des Jahres $J - 1$ folgendermaßen berechnen: Bis dahin sind $J - 1$ Jahre verflossen, von denen jedes mindestens 365 Tage hatte; damit kommen schon einmal $365(J-1)$ Tage zusammen. Was noch fehlt sind die Schalttage: Im Julianischen Kalender hätte es davon $[(J-1)/4]$ gegeben, im Gregorianischen sind aber die durch 100, nicht aber durch 400 teilbaren Jahre keine Schaltjahre, also müssen wir $[(J-1)/100] - [(J-1)/400]$ subtrahieren. Der i-te Tag des Jahres J hat somit die Nummer

$$365(J-1) + [(J-1)/4] - [(J-1)/100] + [(J-1)/400] + i\,.$$

Um den zugehörigen Wochentag zu finden, müssen wir nun nur noch den Wochentag des Tags Nummer eins bestimmen. Dazu können wir von irgendeinem bekannten Datum ausgehen: Donnerstag, der 14. April 2011 ist der $(31 + 28 + 31 + 14) = 104$-te Tag des Jahres 2011, hat also die Nummer

$$365 \cdot 2010 + 502 - 20 + 5 + 104 = 734\,241\,.$$

Diese Zahl ist kongruent vier modulo sieben; somit war Tag eins ein Montag. Geben wir den Wochentagen, wie es die DIN- und ISO-Normen

vorsehen, von Montag ausgehend die Nummern eins bis sieben, so fällt der Tag mit Nummer i also auf den Wochentag mit Nummer i mod 7, wobei die Null dem normgemäß mit 7 bezeichneten Sonntag entspricht.

Um den Wochentag zu einem vorgegebene Datum bestimmen zu können, müssen wir nun nur noch berechnen, der wievielte Tag des Jahres der T-te Tag des M-ten Monats ist. Eine sehr einfache Methode besteht darin, daß wir zählen, wie viele Tage vor dem Ersten des jeweiligen Monats bereits vergangen sind. Dabei müssen wir natürlich zwischen Schaltjahren und gewöhnlichen Jahren unterscheiden: Für letztere seien dies t_M Tage, für erste s_M. Dann haben wir folgende Tabelle:

M =	1	2	3	4	5	6	7	8	9	10	11	12
t_M =	0	31	59	90	120	151	181	212	243	273	304	334
s_M =	0	31	60	91	121	152	182	213	244	274	305	335

Dann fällt der T-te Tag des M-ten Monats des Jahrs J auf den Wochentag mit der Nummer

$$365(J-1) + [(J-1)/4] - [(J-1)/100] + [(J-1)/400] + t_M + T$$

modulo sieben falls J kein Schaltjahr ist; andernfalls muß t_M durch s_M ersetzt werden. Da es uns nur auf Restklassen modulo sieben ankommt, kann der Faktor 365 auch weggelassen werden: $365 = 52 \cdot 7 + 1$. Auch genügt es natürlich, die Zahlen t_M oder s_M modulo 7 einzusetzen. Die entsprechenden Zahlen sind

M =	1	2	3	4	5	6	7	8	9	10	11	12
t_M mod 7 =	0	3	3	6	1	4	6	2	5	0	3	5
s_M mod 7 =	1	4	4	0	2	5	0	3	6	1	4	6

Schöner wäre es freilich, wenn wir eine Formel hätten, die ohne Wertetabelle auskommt. Um eine solche zu finden, ignorieren wir zunächst einmal die historisch überkommenen Monatslängen und tun so, als könnte ein Mathematiker am grünen Tisch festlegen, wie er 365 Tage auf zwölf Monate verteilt.

Für ihn wäre die am wenigsten irreguläre Verteilung der Monatslängen wohl die, bei der Tag i eines Jahres mit N Tagen genau dann im k-ten

Monat liegt, wenn gilt

$$\frac{(k-1)N}{12} < i \le \frac{kN}{12} \quad \text{oder} \quad k - 1 < \frac{12i}{N} \le k\,,$$

d.h. k ist die kleinste ganze Zahl größer oder gleich $12i/N$. Für ein Jahr mit N = 365 Tagen würde dies auf die Monatslängen

$$30, 30, 31, 30, 31, 30, 30, 31, 30, 31, 30, 31$$

führen, in einem Schaltjahr mit N = 366 hätten alle ungeraden Monate dreißig und alle geraden Monate 31 Tage. Ein Februar mit 28 oder 29 Tagen kann bei einer derartigen Strategie natürlich nie vorkommen.

Trotzdem läßt sich auch unser chaotisches Monatssystem fast auf eine solche Formel bringen: Nehmen wir an, wir hätten ein Jahr mit 367 Tagen und zwölf Monaten. Dann liefert uns die obige Vorgehensweise Monate der Längen

$$30, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31, 31\,,$$

wir haben also wie im wirklichen Kalender an zwei Stellen aufeinanderfolgende Monate mit 31 Tagen. Im Kalender sind dies Juli/August und Dezember/Januar, hier sind es die Monate 6 und 7 sowie 11 und 12. Wenn wir zyklisch um eine Position verschieben, so daß die hintere 31 an der ersten Stelle steht, stimmen die beiden Positionen überein, und abgesehen vom Februar, der hier dreißig Tage hat, haben wir genau die Folge der Monatslängen.

(Kurioserweise gab es 1712 in Schweden sogar ein Jahr mit 367 Tagen und einem 30. Februar: 1699 wurde beschlossen, langsam zum Gregorianischen Kalender überzugehen und dazu als erstes den Schalttag 1700 zu streichen. Danach wurde der Beschluß aufgegeben, und um wieder synchron zum Julianischen Kalender zu werden, gab es 1712 einen 30. Februar als zweiten Schalttag. 1753 wurde der Gregorianische Kalender dann endgültig und abrupt eingeführt.)

Die Anzahl der Tage vor dem Ersten des M-ten Monats wäre in unserem hypothetischen Kalender einfach gleich $[367M/12]$; durch die zyklische Verschiebung wird diese Formel freilich zerstört. Sie kann aber gerettet werden durch eine Verschiebung im Zähler: Wie explizites Nachrechnen

zeigt, sind in einem Jahr mit 367 Tagen, in dem der Februar dreißig Tage hat, vor dem Ersten des M-ten Monats gleich $[(367M-362)/12]$ Tage vergangen. Somit ist für unseren realen Kalender

$$t_M = \begin{cases} \left[\frac{367M-362}{12}\right] & \text{für } M \le 2 \\ \left[\frac{367M-362}{12}\right] - 2 & \text{für } M \ge 3 \end{cases} \quad \text{und}$$

$$s_M = \begin{cases} \left[\frac{367M-362}{12}\right] & \text{für } M \le 2 \\ \left[\frac{367M-362}{12}\right] - 1 & \text{für } M \ge 3 \end{cases} .$$

Der Wochentag des T-ten Tags im M-ten Monat des Jahrs J ist somit

$$(J-1) + \left[\frac{J-1}{4}\right] - \left[\frac{J-1}{100}\right] + \left[\frac{J-1}{400}\right] + \left[\frac{367M-362}{12}\right] + \delta_M + T$$

modulo sieben mit

$$\delta_M = \begin{cases} 0 & \text{falls } M \le 2 \\ -1 & \text{falls } M \ge 3 \text{ und } J \text{ Schaltjahr} \\ -2 & \text{falls } M \ge 3 \text{ und } J \text{ kein Schaltjahr} \end{cases} .$$

Diese Formel gilt selbstverständlich nur für Daten nach dem Gregorianischen Kalender; bei älteren Daten muß man zunächst wissen, auf welchem Kalender und welchem Jahresanfang sie beruhen.

Als beispielsweise der amerikanische Naturwissenschaftler, Philosoph und Politiker BENJAMIN FRANKLIN geboren wurde, zeigten die Kalender in seiner Heimatstadt Boston den 6. Januar 1705. Massachusetts war zu der Zeit noch britische Kolonie, und da Großbritannien den Gregorianischen Kalender erst 1752 einführte, ist das ein Julianisches Datum. Gregorianisch ist sein Geburtstag elf Tage später, d.h. am 17. Januar. Allerdings handelt es sich dabei nicht um den 17. Januar 1705, sondern um den des Jahres 1706: In Großbritannien begann das neue Jahr damals nämlich nicht am ersten Januar, sondern am 25. März. Auf den 31. Dezember 1705 folgte also der 1. Januar 1705 und auf den 24. März 1705 der 25. März 1706. Solche Besonderheiten bei der Interpretation alter Datumsangaben gibt es viele; hier ist also Vorsicht geboten.

Mindestens genauso wichtig wie eine verbesserte Schaltjahrregel war für Papst Gregor das Datum des Osterfests; auch darum sollte sich seine Kommission kümmern. Damit kam nun plötzlich auch der Mond in den Kalender, denn 325 beschloß das Konzil von Nicäa (bei Konstantinopel),

daß Ostern stets am ersten Sonntag nach dem ersten Vollmond am oder nach der Frühlings-Tag-und-Nacht-Gleiche zu feiern sei. (Man beachte, daß das erste *nach* im Sinne eines $>$, das zweite im Sinne eines $\ge$ definiert ist. Der Grund für das $>$ lag darin, daß so Ostern nur sehr selten gleichzeitig mit dem jüdischen Pascha-Fest begangen wird.)

Der erste Vollmond am oder nach der Frühlings-Tag-und-Nacht-Gleiche kann nicht einfach nach einer ähnlichen Regel wie der Monatsanfang im islamischen Kalender bestimmt werden, also etwa dann, wenn ihn mindestens zwei vertrauenswürdige Kardinäle gesehen haben, denn der österliche Festkreis beginnt bereits siebzig Tage vor Ostern. Das Datum mußte daher im Voraus berechnet werden. Der Mathematiker und Informatiker DONALD E. KNUTH sagt in Abschnitt 1.3.2, Aufgabe 14 seiner *Art of Computer Programming: There are many indications that the sole important application of arithmetic in Europe during the Middle Ages was the calculation of the Easter date, and so such algorithms are historically significant.* So ordnete etwa KARL DER GROSSE (747/8–814) bei seiner Neuordnung des Bildungssystems mehrfach an, daß in jeder Diözese mindestens ein Geistlicher in der Lage sein müsse, das Osterdatum zuverlässig zu berechnen. Schauen wir uns also an, wie Papst Gregor das Osterdatum berechnen ließ.

Wir brauchen Informationen über die Wochentage, über die Mondphasen und über die Tag-Nacht-Gleiche im Frühling. Letztere ist, da der Gregorianische Kalender das tropische Jahr recht genau approximiert, noch recht lange konstant am 21. März jedes Jahres; bei der bei der Berechnung des Osterdatums geht man daher stets von diesem Tag aus. Wie man Wochentage bestimmt, haben wir uns gerade überlegt; bleibt also noch das Problem mit den Mondphasen.

Die mittlere Zeitspanne zwischen zwei Neumonden, die synodische Umlaufzeit des Mondes, beträgt etwa 29,5306 Tage; ein tropisches Jahr mit seinen 365,2422 Tagen besteht also aus

$$365{,}2422 : 29{,}5306 \approx 12{,}3679$$

solchen Zykeln. Die Kettenbruchentwicklung dieses Quotienten ist

$$[12, 2, 1, 2, 1, 1, 4, 1, 81]$$

mit Konvergenten

$$12, \quad 12\frac{1}{3}, \quad 12\frac{3}{8}, \quad 12\frac{4}{11}, \quad 12\frac{7}{19}, \quad 12\frac{32}{87}, \quad 12\frac{39}{106}, \quad \dots .$$

Für einen Kalender, der sowohl mit der Sonne als auch mit dem Mond synchronisiert ist, könnte man also Jahre mit 12 und 13 Monaten kombinieren, wobei in erster Näherung jedes dritte Jahr 13 Monate hätte. Tatsächlich war man im fünften vorchristlichen Jahrhundert bereits erheblich weiter: Der um 440 v.Chr. lebende Athener Mathematiker und Astronomen METON verwendete die Konvergente mit Nenner 19. Ein Metonischer Zyklus besteht besteht demnach aus 19 Jahren, darunter zwölf *Gemeinjahren* aus zwölf Monaten und sieben *Schaltjahren* aus 13 Monaten. Die Monate hatten teils 29, teils 30 Tage. Der darauf basierende Kalender wurde in Griechenland bis 46 v.Chr. verwendet. Die Synchronisation zwischen Sonnen- und Mondzykeln ist fast perfekt:

$$19 \cdot 365{,}2422 = 6939{,}6018 \quad \text{und} \quad 235 \cdot 29{,}5306 = 6939{,}691\,,$$

der Fehler pro Zyklus liegt also bei nur etwa zwei Stunden. Seit Einführung des Gregorianischen Kalenders sind etwas über 22 Metonische Zykeln vergangen; der akkumulierte Fehler liegt also noch unter zwei Tagen.

Der Gregorianische Kalender geht deshalb bei der Bestimmung des Osterdatums nicht von astronomischen Beobachtungen aus, sondern von Metonischen Zykeln, allerdings mit einer Korrektur für den akkumulierten Fehler. Ebenfalls unberücksichtigt bleiben die Irregularitäten der realen Mondbewegung; gerechnet wird mit einer Approximation der *mittleren* Mondbewegung. Auf den ersten Blick seltsam erscheinen mag auch die Tatsache, daß bei der Fehlerkorrektur mit der *Julianischen* Jahreslänge von $365\frac{1}{4}$ Tagen gerechnet wird; der Grund lag wohl vor allem darin, daß Papst Gregor bisherige Praktiken nicht mehr als unbedingt notwendig ändern wollte.

Die wesentliche Größe, mit der die Mondphasen in unseren an der Sonne orientierten Kalender gebracht werden, ist der sogenannte *Epakt*. Mit diesem Wort bezeichneten die Griechen die Anzahl der Tage, die an Neujahr seit dem letzten Neumond des alten Jahres vergangen waren.

Gemäß dem Metonischen Zyklus sollte diese Zahl sich alle 19 Jahre wiederholen; in der Kalenderrechnung wird daher die um eins vermehrte Restklasse modulo 19 der Jahreszahl als die „Goldene Zahl“ bezeichnet. (Die Addition der Eins kommt natürlich daher, daß zur Zeit ihrer Einführung die Null in der europäischen Mathematik noch nicht vorkam.)

Wenn jedes Jahr genau 365 Tage hätte, könnten wir einfach mit den $12 \times 29{,}5 = 354$ Tagen eines Mondjahrs vergleichen und wüßten dann, daß sich die Mondphase an einem festen Datum jedes Jahr um elf Tage verschiebt. Als *Mondphase* bezeichnen wir dabei die Anzahl von Tagen, die seit dem letzten Neumond vergangen sind.

Eine der vielen Vereinfachungen in der Berechnung des Osterdatums liegt darin, daß man innerhalb des aktuellen Metonischen Zyklus im wesentlichen von dieser Formel ausgeht, die Schalttage also ignoriert.

Da die Schalttage Ende Februar eingeführt werden, wir uns aber für den Vollmond am oder nach dem 21. März interessieren, sollten wir nicht mit dem klassischen Epakt, der Mondphase des 1. Januar, rechnen: Sonst gäbe es schließlich algorithmisch unangenehme Fallunterscheidungen für die Schaltjahre. Aus Effizienzgründen bietet sich an, stattdessen mit einem *verschobenen* Epakt zu rechnen, d.h. mit der Mondphase eines geeigneten Datums, das näher bei Ostern liegt.

Die Länge eines lunaren Zyklus liegt bei ungefähr 29,5 Tagen; zum einfacheren Rechnen sollten wir das zumindest für den einen Zyklus, in den Ostern fällt, auf den ganzzahligen Wert 30 runden. Der erste Vollmond nach dem 21. März ist dann der letzte Vollmond vor dem 19. April, und sein Abstand zum 19. April ist, wenn wir den Vollmond als Tag mit Mondphase 14 betrachten, gleich dem Abstand des letzten Neumonds vor dem 5. April zum 5. April, also die Mondphase des 5. Aprils. Somit bietet sich an, als verschobenen Epakt die Mondphase des 5. Aprils zu nehmen. Gemäß unserer vereinfachenden Annahmen sollte auch sie sich alle 19 Jahre wiederholen und sollte sich zumindest innerhalb eines Metonischen Zyklus von Jahr zu Jahr modulo 30 um elf verschieben.

Damit brauchen wir nur noch für ein Jahr des Metonischen Zyklus

den tatsächlichen Wert der Mondphase des fünften Aprils kennen, um den verschobenen Epakt allgemein berechnen zu können. Die vor der Gregorianischen Reform gebräuchliche Formel berechnet ihn für das Jahr J als

$$E = \big(14 + 11 \cdot (J \bmod 19)\big) \bmod 30\,.$$

Der erste Vollmond am oder nach dem 21. März lag somit E Tage vor dem 19. April, und Ostern war der (echt) darauf folgende Sonntag. So wird Ostern noch heute in fast allen orthodoxen Kirchen berechnet; die einzige Ausnahme ist die finnische.

Der Gregorianische Kalender modifiziert diese Formel durch drei zusätzliche Terme: Zunächst berücksichtigt er, daß der Metonische Zyklus nicht wirklich exakt ist, insbesondere dann nicht, wenn man mit dem Julianischen Jahr arbeitet:

$$19 \cdot 365{,}25 = 6939{,}750 \quad \text{und} \quad 235 \cdot 29{,}5306 = 6939{,}691\,;$$

hier beträgt die Differenz also 0,059 Tage pro Zyklus und

$$0{,}059 \times \frac{100}{19} \approx 0{,}31$$

Tage pro Jahrhundert. Die Gregorianische Osterformel approximiert dies durch $8/25 = 0{,}32$, addiert allerdings im h-ten Jahrhundert nicht $[8h/25]$, sondern $\big[(5+8h)/25\big]$. Diese Modifikation soll in erster Linie dafür sorgen, daß Ostern möglichst selten mit dem jüdischen Paschafest zusammenfällt. Für das Jahrhundert wird dabei die gleiche Konvention benutzt wie für die Feier des Jahrtausendanfangs am 1. Januar 2000: Das h-te Jahrhundert beginnt mit dem Jahr $100(h-1)$, d.h. $h = [J/100]+1$.

Da der Gregorianische Kalender bei der Korrektur der Metonischen Zyklen mit Julianischen Jahren arbeitet, am Ende aber ein Gregorianisches Datum braucht, müssen als nächstes die unterschiedlichen Anzahlen von Schalttagen berücksichtigt werden, d.h. die „ausfallenden“ Schalttage des Gregorianischen Kalenders müssen subtrahiert werden. Das sind drei Stück pro 400 Jahre, also wird $[3h/4]$ subtrahiert. Dies ergäbe die neue Formel

$$E = \left(14 + 11 \cdot (J \bmod 19)\right) + \left[\frac{5+8h}{25}\right] - \left[\frac{3h}{4}\right) \bmod 30\,.$$

Tatsächlich gibt es noch eine weitere Modifikation, die dafür sorgen soll, daß die 19 Epakte eines Metonischen Zyklus alle verschieden sind und $E = 0$ nicht auftritt: Falls $E = 0$ ist oder falls $E = 1$ ist und $J \bmod 19 > 10$, wird E um eins erhöht. Der (berechnete) Vollmond ist dann E Tage vor dem 19. April, und Ostern wird weiterhin am darauf folgenden Sonntag gefeiert.

Das Jahr $J = 2011$ liegt im $h = 21$. Jahrhundert und $2011 \equiv 16 \bmod 19$. Somit ist

$$\begin{aligned} E &= \left(14 + 11 \cdot 16 + \left[\frac{173}{25}\right] - \left[\frac{63}{4}\right]\right) \bmod 30 \\ &= 181 \bmod 30 = 1\,, \end{aligned}$$

Da $E = 1$ ist und $J \bmod 19 = 16 > 10$, wird E nach der letzten Regel auf *zwei* erhöht; Der rechnerische Vollmond ist daher am 17. April. (Der tatsächliche ist erst am 18. April um 4.44 h.) Der 17. April 2011 ist ein Sonntag; also ist Ostern 2011 eine Woche später am 24. April.

§5: Eine kryptographische Anwendung

Beim RSA-Verfahren wählt man den öffentlichen Exponenten e oft ziemlich klein, z.B. $e = 3$ oder $e = 2^{16} + 1$. Dies hat den Vorteil, daß zumindest die Verschlüsselung ziemlich schnell geht und man nur zur Entschlüsselung mit einem Exponenten in der Größenordnung des Moduls arbeiten muß.

Für jemanden, der RSA hauptsächlich für elektronische Unterschriften verwendet, würde sich anbieten, stattdessen den privaten Exponenten d relativ klein zu wählen. Dann könnte er schnell viele Dokumente unterschreiben, und falls jeder Empfänger nur eines davon bekommt, fällt dessen höherer Aufwand bei der Überprüfung nicht so sehr ins Gewicht.

Natürlich kann man nicht $d = 3$ oder $d = 2^{16} + 1$ wählen: Der private Exponent muß schließlich geheim sein und es darf nicht möglich sein, ihn durch Probieren zu erraten.

Andererseits geht man heute bei symmetrischen Kryptoverfahren davon aus, daß ein Verfahren sicher ist, falls ein Gegner mindestens

2^{128} Möglichkeiten durchprobieren muß, so daß gängige Verfahren wie AES mit einer Schlüssellänge von 128 Bit auskommen. Verglichen damit erscheinen 2048 Bit für einen privaten Entschlüsselungsexponenten recht hoch.

Trotzdem läßt sich hier nicht wesentlich sparen, denn ein Gegner kann kurze private Exponenten nicht nur durch Ausprobieren bestimmen, sondern auch wesentlich schneller nach dem Kettenbruchalgorithmus.

Wir gehen aus von einem öffentlichen RSA-Schlüssel (N, e) sowie dem zugehörigen privaten Exponenten d. Dann gibt es bekanntlich eine natürliche Zahl k, so daß $ed - k\varphi(N) = 1$ ist. Dies können wir umschreiben als

$$\frac{e}{\varphi(N)} - \frac{k}{d} = \frac{1}{d\varphi(N)}\,.$$

Falls d sehr viel kleiner ist als $\varphi(N)$ haben wir hier einen Bruch mit dem großen Nenner $\varphi(N)$ sehr gut angenähert durch einen Bruch mit dem sehr viel kleineren Nenner d. Für hinreichend kleines d ist das nur möglich, wenn k/d eine Konvergente der Kettenbruchentwicklung von $e/\varphi(N)$ ist.

Das mag zunächst harmlos erscheinen, denn die Sicherheit von RSA beruht ja gerade darauf, daß niemand außer dem Inhaber des privaten Schlüssels d die Faktorisierung $N = pq$ und damit den Wert von

$$\varphi(N) = (p-1)(q-1) = N - (p+q) + 1$$

kennt. Dafür kennt aber jeder den Wert von N, und wie die obige Gleichung zeigt, liegt der recht nahe bei $\varphi(N)$: Die Primzahlen p und q sind schließlich nur von der Größenordnung $\sqrt{N}$. Damit sollte k/d auch eine gute Approximation für e/N liefern, und in der Tat ist

$$\begin{aligned}\left|\frac{e}{N} - \frac{k}{d}\right| &= \left|\frac{e}{N} - \frac{e}{(p-1)(q-1)} + \frac{e}{(p-1)(q-1)} - \frac{k}{d}\right| \\ &\le \left|\frac{e(p-1)(q-1) - epq}{N(p-1)(q-1)}\right| + \frac{1}{d(p-1)(q-1)} \\ &= \frac{e(p+q-1)}{N(p-1)(q-1)} + \frac{1}{d(p-1)(q-1)}\,.\end{aligned}$$

Falls dies kleiner ist als $1/2d^2$, muß k/d eine Konvergente der Kettenbruchentwicklung von e/N sein; um d zu berechnen, müssen wir also nur so lange Konvergenten p_n/q_n bestimmen, bis für einen der Nenner q_n die Exponentiation mit q_n modulo N invers ist zu der mit e. Falsche Kandidaten sollten dabei praktisch immer bereits beim ersten Versuch erkannt werden.

Eine einfache Abschätzung zeigt, daß dies für p und q von etwa gleicher Größe funktioniert, sofern d höchstens die Größenordnung von etwa $\sqrt[4]{N}$ hat; neuere, etwas aufwendigere Untersuchungen zeigen, daß man d auch noch für $d < N^{0{,}289}$ rekonstruieren kann. Fachleute erwarten, daß möglicherweise sogar alle $d < \sqrt{N}$ unsicher sind.

Private Exponenten müssen also immer groß sein. Falls man von einem vorgegebenen öffentlichen Exponenten ausgeht, ist das für realistische N mit an Sicherheit grenzender Wahrscheinlichkeit erfüllt; Vorsicht ist nur geboten, wenn man mit dem privaten Exponenten startet. Daher verlangen auch die Vorschriften der Bundesnetzagentur, daß man immer vom öffentlichen Exponenten e ausgehen muß, und erst daraus einen privaten Exponenten berechnet.

Kapitel 6
Quadratische Zahlkörper

Ein Zahlkörper ist ein Körper K, der den Körper $\mathbb{Q}$ der rationalen Zahlen enthält und als $\mathbb{Q}$-Vektorraum endlichdimensional ist. Im zweidimensionalen Fall reden wir von quadratischen Zahlkörpern. Die algebraische Zahlentheorie untersucht die (noch zu definierenden) ganzen Zahlen eines solchen Zahlkörpers. In dieser Vorlesung geht es zwar eher um elementare als um algebraische Zahlentheorie, jedoch werden wir im nächsten Kapitel sehen, daß ein Umweg über quadratische Zahlen auch bei rein ganzzahligen Problemen gelegentlich hilfreich sein kann.

§1: Grundbegriffe der Ringtheorie

Als erstes wollen wir uns überlegen, in welchen Zahlbereichen außer $\mathbb{Z}$ wir noch sinnvoll von Teilbarkeit und eventuell auch Division mit Rest reden können. Wir brauchen dazu selbstverständlich zumindest eine Addition und eine Multiplikation, d.h. einen der bereits im ersten Kapitel definierten *Ringe.* Wenn wir eindeutige Quotienten wollen, müssen wir aber noch zusätzlich voraussetzen, daß es keine sogenannte *Nullteiler* gibt, d.h. von null verschiedene Elemente r, s, deren Produkt gleich null ist. Ist nämlich $y = qs$, so ist dann auch $y = (q + r)s$, was unserer Vorstellung von Teilbarkeit mit eindeutig bestimmtem Quotienten widerspricht.

Definition: *a)* Ein Ring heißt *nullteilerfrei* wenn gilt: Ist $x \cdot y = 0$, so muß mindestens einer der beiden Faktoren x, y verschwinden. Ein nullteilerfreier kommutativer Ring heißt *Integritätsbereich* (englisch *domain*).

b) Wir sagen, ein Element u eines Integritätsbereichs R sei *Teiler* von $x \in R$, in Zeichen $u|x$, wenn es ein $q \in R$ gibt, so daß $x = q \cdot u$.
c) $u \in R$ heißt *größter gemeinsamer Teiler* von x und y, wenn u Teiler von x und von y ist und wenn für jeden anderen gemeinsamen Teiler v von x und y gilt: $v|u$.
d) Ein Element $e \in R$ heißt *Einheit,* falls es ein $e' \in R$ gibt mit $e \cdot e' = 1$. Die Menge aller Einheiten von R bezeichnen wir mit $R^\times$.
e) Zwei Elemente $x, y \in R$ heißen assoziiert, wenn es eine Einheit $e \in R$ gibt, so daß $y = e \cdot x$.

Der Prototyp eines kommutativen Rings ist der Ring $\mathbb{Z}$ der ganzen Zahlen; er ist ein Integritätsbereich mit ± 1 als einzigen Einheiten. Zwei ganze Zahlen sind somit genau dann assoziiert, wenn sie denselben Betrag haben.

Der Ring $\mathbb{Z}/m$ ist genau dann nullteilerfrei, wenn m prim ist; in diesem Fall ist er sogar ein Körper. Ist aber $m = ab$ eine Zerlegung (in $\mathbb{N}$) von m in ein Produkt mit $a, b > 1$, so ist in $\mathbb{Z}/m$ zwar $ab = 0$, aber $a, b \neq 0$.

Die Menge aller $n \times n$-Matrizen über einem Körper ist ein Beispiel eines nichtkommutativen Rings. Er ist nicht nullteilerfrei, enthält aber viele invertierbare Elemente.

Auch die Polynome über einem Körper k bilden einen Ring, den Polynomring $k[X]$. Allgemeiner gilt sogar:

Lemma: Ist R ein Integritätsbereich, so auch der Polynomring

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

Seine Einheiten sind genau die Einheiten von R.

Beweis: Wenn wir Addition und Multiplikation nach den üblichen Regeln definieren, ist klar, daß $R[X]$ alle Ringaxiome erfüllt. Um zu zeigen, daß $R[X]$ nullteilerfrei ist, betrachten wir zwei Polynome

$$f = \sum_{i=0}^{n} a_i X^i \quad \text{und} \quad g = \sum_{j=0}^{m} b_j X^j ,$$

die beide von Null verschieden sind. Wir können etwa annehmen, daß n und m so gewählt sind, daß a_n und b_m beide nicht verschwinden. Da R Integritätsbereich ist, kann dann auch das Produkt $a_n b_m$ nicht verschwinden, also ist der führende Term $a_n b_m X^{n+m}$ von fg von Null verschieden und damit auch fg selbst. Tatsächlich beweist dies sogar etwas mehr als die Nullteilerfreiheit, denn wir wissen nun, daß sich bei der Multiplikation zweier Polynome die Grade addieren.

Ist $f \in R[X]$ eine Einheit, so gibt es ein $g \in R[X]$ mit $fg = 1$; da das konstante Polynom 1 den Grad null hat, muß dasselbe auch für f und g gelten, d.h. $f, g \in R$ und damit in $R^\times$. ∎

Allgemein gilt:

Lemma: *a)* Die Menge $R^\times$ aller Einheiten von R ist eine abelsche Gruppe bezüglich der Multiplikation.
b) Ein kommutativer Ring R ist genau dann ein Integritätsbereich, wenn die folgende *Kürzungsregel* erfüllt ist: Gilt für $x, y, z \in R$ und $z \neq 0$ die Gleichung $xz = yz$, so ist $x = y$.
c) Zwei Elemente x, y eines Integritätsbereich R sind genau dann assoziiert, wenn $x|y$ und $y|x$.
d) Ein größter gemeinsamer Teiler, so er existiert, ist bis auf Assoziiertheit eindeutig bestimmt.

Beweis: a) Sind $e, f \in R$ Einheiten, so gibt es Elemente e', f' mit $ee' = ff' = 1$. Damit ist $(ef)(f'e') = e(ff')e' = ee' = 1$, d.h. auch ef ist eine Einheit. Außerdem ist jede Einheit invertierbar, denn offensichtlich ist e' ein multiplikatives Inverses zu e.
b) Ist R ein Integritätsbereich und $xz = yz$, so ist $(x - y)z = 0$; da $z \neq 0$ vorausgesetzt war, folgt $x - y = 0$, also $x = y$. Folgt umgekehrt aus $xz = yz$ und $z \neq 0$ stets $x = y$, so ist R nullteilerfrei, denn ist $xy = 0$ und $y \neq 0$, so ist $xy = 0y$, also $x = 0$.
c) Ist $y = ex$, so ist x ein Teiler von y. Da Einheiten invertierbar sind, ist auch $x = e^{-1}y$, d.h. $y|x$.
Gilt umgekehrt $x|y$ und $y|x$, so gibt es Elemente q, r mit $x = qy$ und $y = rx$. Damit ist $1x = x = (qr)x$, also $qr = 1$. Somit ist q eine Einheit.
d) Sind u, v zwei größte gemeinsame Teiler von x, y, so ist nach Definition u Teiler von v und v Teiler von u, also sind u und v assoziiert. ∎

In Integritätsbereichen können wir somit einen Teilbarkeitsbegriff einführen, der den üblichen, von $\mathbb{Z}$ her gewohnten Regeln genügt. Manchmal können wir auch, wie in $\mathbb{Z}$, von einer eindeutigen Primzerlegung reden:

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

Lemma: In einem faktoriellen Ring gibt es zu je zwei Elementen x, y einen größten gemeinsamen Teiler.

Beweis: Wir wählen zunächst aus jeder Klasse assoziierter irreduzibler Elemente einen Vertreter; für die Zerlegung eines Elements in ein Produkt irreduzibler Elemente reicht es dann, wenn wir nur irreduzible Elemente betrachten, die Vertreter ihrer Klasse sind.

Sind $x = u \prod_{i=1}^{r} p_i^{e_i}$ und $y = v \prod_{j=1}^{s} q_j^{f_j}$ mit $u, v \in R^\times$ und p_i, q_j irreduzibel die entsprechenden Zerlegungen von x und y in Primfaktoren, so können wir, indem wir nötigenfalls Exponenten null einführen, o.B.d.A. annehmen, daß $r = s$ ist und $p_i = q_i$ für alle i. Dann ist offenbar $\prod_{i=1}^{r} p_i^{\min(e_i, f_i)}$ ein ggT von x und y, denn $z = \prod_{i=1}^{r} p_i^{g_i}$ ist genau dann Teiler von x, wenn $g_i \leq e_i$ für alle i, und Teiler von y, wenn $g_i \leq f_i$. ∎

§2: Die Elemente quadratischer Zahlkörper

Ein quadratischer Zahlkörper ist ein Zahlkörper, der als $\mathbb{Q}$-Vektorraum betrachtet die Dimension zwei hat. Es gibt daher ein von der Eins linear unabhängiges Element α. Die drei Elemente $1, \alpha, \alpha^2$ müssen aber linear

abhängig sein; es gibt also rationale p, q, r, so daß $p\alpha^2 + q\alpha + r$ verschwindet. Indem wir mit dem Hauptnenner von p, q, r multiplizieren, erhalten wir eine entsprechende Gleichung mit ganzzahligen Koeffizienten, und wenn wir dann noch durch deren ggT dividieren, erhalten wir teilerfremde ganze Zahlen A, B, C, so daß $A\alpha^2 + B\alpha + C = 0$ ist.

Nach der Lösungsformel für quadratische Gleichungen folgt

$$\alpha = -\frac{B}{2A} \pm \frac{\sqrt{B^2 - 4AC}}{2A}\,.$$

Den Ausdruck $\Delta = B^2 - 4AC$ unter der Wurzel bezeichnen wir als die *Diskriminante* von α. Für $\alpha = \sqrt{W}$ mit $W \in \mathbb{Z}$ beispielsweise ist $A = 1$, $B = 0$ und $C = -W$, also $\Delta = 4W$. Für $\alpha = \frac{1}{3} + \frac{1}{5}\sqrt{2}$ haben wir die Gleichung

$$\alpha^2 - \frac{2}{3}\alpha + \frac{1}{9} - \frac{2}{25} = \alpha^2 - \frac{2}{3}\alpha + \frac{7}{225} = 0 \Longrightarrow 225\alpha^2 - 150\alpha + 7 = 0\,;$$

hier ist die Diskriminante somit $\Delta = 150^2 - 4 \cdot 225 \cdot 7 = 16200$.

Wegen der Irrationalität von α muß auch $\sqrt{\Delta}$ irrational sein, d.h. Δ ist kein Quadrat. Wegen der Eindeutigkeit der Primzerlegung in $\mathbb{Z}$ können wir ganze Zahlen $Q, D \in \mathbb{Z}$ finden, so daß $\Delta = Q^2 D$ und $\sqrt{\Delta} = Q\sqrt{D}$ ist mit einer quadratfreien Zahl D, d.h. einer Zahl D, die durch keine Quadratzahl ungleich eins teilbar ist. Somit läßt sich α in der Form $r + s\sqrt{D}$ schreiben mit $r, s \in \mathbb{Q}$. Da K als $\mathbb{Q}$-Vektorraum zweidimensional ist, läßt sich jedes Element von K so schreiben, als Vektorraum ist also $K = \mathbb{Q} \oplus \mathbb{Q}\sqrt{D}$.

Umgekehrt ist $\mathbb{Q} \oplus \mathbb{Q}\sqrt{D}$ für jedes Nichtquadrat D ein Körper, denn natürlich liegen Summe und Differenz zweier Elemente wieder in diesem Vektorraum und wegen

$$(r + s\sqrt{D})(u + v\sqrt{D}) = (ru + svD) + (rv + su)\sqrt{D}$$

auch das Produkt. Für den Quotienten können wir wie bei den komplexen Zahlen über die dritte binomische Formel argumentieren:

$$\frac{r + s\sqrt{D}}{u + v\sqrt{D}} = \frac{(r + s\sqrt{D})(u - v\sqrt{D})}{(u + v\sqrt{D})(u - v\sqrt{D})} = \frac{ru - svD}{u^2 - v^2 D} + \frac{su - rv}{u^2 - v^2 D}\,.$$

Wir bezeichnen diesen Körper kurz mit $K = \mathbb{Q}[\sqrt{D}]$.

Für $D > 0$ ist $\mathbb{Q}[\sqrt{D}]$ ein Teilkörper von $\mathbb{R}$; wir reden in diesem Fall von einem *reellquadratischen* Zahlkörper. Falls $D < 0$, gibt es in $\mathbb{Q}[\sqrt{D}]$ auch imaginäre Elemente; hier reden wir von einem *imaginärquadratischen* Zahlkörper.

§3: Die Hauptordnung eines Zahlkörpers

Jede rationale Zahl ist Lösung einer linearen Gleichung $aX + b = 0$ mit ganzzahligen Koeffizienten a, b, von denen der erste nicht verschwinden darf; sie ist genau dann eine ganze Zahl, wenn man $a = 1$ wählen kann.

Entsprechend ist jedes Element x eines Zahlkörpers K Lösung einer Polynomgleichung

$$a_n X^n + a_{n-1} X^{n-1} + \cdots + a_1 X + a_0 = 0 \quad \text{mit} \quad a_i \in \mathbb{Z}\,,$$

denn da K nach Definition ein endlichdimensionaler $\mathbb{Q}$-Vektorraum ist, können die Potenzen von x nicht allesamt linear unabhängig sein. Es gibt also für irgendein n eine lineare Abhängigkeit

$$\lambda_n x^n + \lambda_{n-1} x^{n-1} + \cdots + \lambda_1 x + \lambda_0 = 0 \quad \text{mit} \quad \lambda_i \in \mathbb{Q}\,.$$

Multiplikation mit dem Hauptnenner der Koeffizienten λ_i macht daraus eine Polynomgleichung mit ganzzahligen Koeffizienten.

Definition: Ein Element x eines Zahlkörpers K heißt *ganz,* wenn es einer Polynomgleichung

$$X^n + a_{n-1} X^{n-1} + \cdots + a_1 X + a_0 = 0$$

mit ganzzahligen Koeffizienten $a_i \in \mathbb{Z}$ und höchstem Koeffizienten eins genügt.

Man kann relativ einfach zeigen, daß die ganzen Zahlen in einem Zahlkörper K einen Ring bilden; da wir uns hier aber auf quadratische Zahlkörper beschränken, bei denen wir dies ganz explizit sehen können, sei hier auf einen solche Beweis verzichtet.

Wir betrachten also einen quadratischen Zahlkörper $K = \mathbb{Q}[\sqrt{D}]$. Ein Element $\alpha = r + s\sqrt{D}$ mit $r, s \in \mathbb{Z}$ ist genau dann ganz, wenn es einer Gleichung der Form $x^2 + ax + b$ mit $a, b \in \mathbb{Z}$ genügt. Da

$$x^2 = (r + s\sqrt{D})^2 = (r^2 + s^2 D) + 2rs\sqrt{D}$$

ist, genügt x der Gleichung

$$x^2 - 2rx + (r^2 - s^2 D) = 0\,.$$

Somit müssen $c = 2r$ und $d = r^2 - s^2 D$ ganze Zahlen sein.

Für $r \in \mathbb{Z}$ ist die erste Bedingung trivialerweise erfüllt und die zweite genau dann, wenn auch s eine ganze Zahl ist: Da D keinen Nenner hat, ist der Nenner von $r^2 - s^2 D$ in diesem Fall das Quadrat des Nenners von s.

Falls r keine ganze Zahl ist, muß es wegen der ersten Bedingung von der Form $r = c/2$ sein mit einer ungeraden Zahl c. Notwendige Bedingung für die Ganzheit von $r^2 - s^2 D$ ist dann, daß auch $s = e/2$ von dieser Form ist. Dann ist

$$r^2 - s^2 D = \frac{c^2 - e^2 D}{4} \in \mathbb{Z} \Longrightarrow c^2 - e^2 D \equiv 0 \bmod 4\,.$$

c und e sind ungerade Zahlen; ihre Quadrate sind also kongruent eins modulo vier. Somit ist $r^2 - s^2 D$ genau dann ganz, wenn $D \equiv 1 \bmod 4$ ist.

In $\mathbb{Q}[\sqrt{D}]$ ist ein Element $r + s\sqrt{D}$ daher für $D \not\equiv 1 \bmod 4$ genau dann ganz, wenn r und s beide ganz sind; die Menge der ganzen Zahlen ist also $\mathbb{Z} \oplus \mathbb{Z}\sqrt{D}$. Diese Menge ist offensichtlich eine abelsche Gruppe bezüglich der Addition, und da das Quadrat von $\sqrt{D}$ die ganze Zahl D ist, ist sie auch abgeschlossen bezüglich der Multiplikation; die ganzen Zahlen bilden also einen Ring.

Im Fall $D \equiv 1 \bmod 4$ ist $r + s\sqrt{D}$ auch noch dann ganz, wenn r und s beide die Hälfte einer ungeraden Zahl sind. Insbesondere ist also auch

$$\beta_D = \frac{1 + \sqrt{D}}{2}$$

eine ganze Zahl, und offensichtlich sind die ganzen Zahlen genau die Zahlen, die sich als $u + v\beta_D$ mit $u, v \in \mathbb{Z}$ schreiben lassen. Die Menge der ganzen Zahlen ist also $\mathbb{Z} \oplus \mathbb{Z}\beta_D$. Auch dies ist ein Ring, denn

$$\beta_D^2 = \frac{1 + 2\sqrt{D} + D}{4} = \frac{D-1}{4} + \frac{1 + \sqrt{D}}{2} = \frac{D-1}{4} + \beta_D$$

liegt wieder in dieser Menge, da $(D-1)/4$ im Fall $D \equiv 1 \bmod 4$ eine ganze Zahl ist.

Die ganzen Zahlen in $\mathbb{Q}[\sqrt{D}]$ bilden also in jedem Fall einen Ring; diesen Ring bezeichnen wir als die *Hauptordnung* $\mathcal{O} = \mathcal{O}_D$ von $\mathbb{Q}[\sqrt{D}]$. Wie wir gerade gesehen haben, ist also

$$\mathcal{O}_D = \begin{cases} \mathbb{Z} \oplus \mathbb{Z}\sqrt{D} & \text{falls } D \not\equiv 1 \bmod 4 \\ \mathbb{Z} \oplus \mathbb{Z}\beta_D \;\text{ mit }\; \beta_D = \frac{1}{2}(1 + \sqrt{D}) & \text{falls } D \equiv 1 \bmod 4 \end{cases}.$$

Beim Körper $K = \mathbb{Q}[i]$ der komplexen Zahlen mit rationalem Real- und Imaginärteil ist $D = -1 \equiv 3 \bmod 4$, also ist die Hauptordnung hier einfach $\mathcal{O}_{-1} = \mathbb{Z} \oplus \mathbb{Z}i$, die sogenannten ganzen GAUSSschen Zahlen. Für $D = -3 \equiv 1 \bmod 4$ dagegen ist auch $\beta_{-3} = \frac{1}{2}(1 + \sqrt{-3})$ eine ganze Zahl und $\mathcal{O}_{-3} = \mathbb{Z} \oplus \mathbb{Z}\beta_{-3}$.

Dieses Beispiel wirft die Frage auf, ob unsere Definition ganzer Zahlen wirklich so geschickt war: Wir hätten schließlich auch einfach definieren können, daß $r + s\sqrt{D}$ genau dann ganz heißen soll, wenn r und s ganze Zahlen sind.

Einer der Gründe ist sicherlich, daß wir in nichtquadratischen Zahlkörpern keine ausgezeichneten Elemente wie $\sqrt{D}$ haben, und selbst im quadratischen Fall ist $\sqrt{D}$ nicht immer das einzige ausgezeichnete Element. Im Falle $D = -3$ beispielsweise ist $\beta_{-3} = \frac{1}{2}(1 + \sqrt{-3})$ eine primitive sechste Einheitswurzel, und es gibt keinen Grund, diese als „weniger ganz“ oder „weniger ausgezeichnet“ zu betrachten als $\sqrt{-3}$.

Viel wichtiger ist aber, daß wir nur bei dieser Definition der Ganzheit eine Chance auf eindeutige Primzerlegung in der Hauptordnung haben:

Definition: *a)* Sind $R \leq S$ Integritätsbereiche, so heißt ein Element $x \in S$ ganz über R, wenn es einer Gleichung

$$x^n + r_{n-1}x^{n-1} + \cdots + r_1 x + r_0 = 0 \quad \text{mit} \quad r_i \in R$$

genügt.
b) R heißt ganzabgeschlossen oder normal, wenn jedes über R ganze Element des Quotientenkörpers K von R in R liegt.

Satz: Ein faktorieller Ring ist ganzabgeschlossen.

Beweis: Jedes Element x des Quotientenkörpers eines Rings R kann als Quotient $x = p/q$ mit $p, q \in R$ dargestellt werden. Falls R faktoriell ist, können wir dabei annehmen, daß p und q teilerfremd sind. x ist genau dann ganz über R, wenn es ein $n \in \mathbb{N}$ und Elemente $r_0, \ldots, r_{n-1} \in R$ gibt derart, daß

$$x^n = -r_{n-1}x^{n-1} - \cdots - r_1 x - r_0$$

ist. Multiplikation mit q^n macht daraus die Gleichung

$$p^n = -r_{n-1}p^{n-1}q - \cdots - r_1 pq^{n-1} - r_o q^n \,.$$

Hier ist die rechte Seite durch q teilbar, also auch die linke. Da p und q als teilerfremd vorausgesetzt war, ist das nur möglich, wenn q eine Einheit ist, d.h. $x = p/q$ liegt in R. ∎

§4: Normen und Spuren in quadratischen Zahlkörpern

Beginnen wir mit einem Beispiel: Die Hauptordnung von $K = \mathbb{Q}[\sqrt{-5}]$ ist $\mathcal{O}_{-5} = \mathbb{Z} \oplus \mathbb{Z}[\sqrt{-5}]$, und dort haben wir die beiden Produktzerlegungen

$$6 = 2 \cdot 3 = (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5}) \,.$$

Folgt daraus, daß $\mathcal{O}_{-5}$ nicht faktoriell ist?

Bevor wir diese Frage beantworten können, müssen wir zunächst wissen, ob möglicherweise die Faktoren auf der rechten Seite noch weiter zerlegt werden können. Solche Fragen lassen sich oft entscheiden, indem man die *Normen* der beteiligten Elemente betrachtet.

Definition: *a)* Für ein Elements $\alpha = r + s\sqrt{D}$ von $K = \mathbb{Q} \oplus \mathbb{Q}\sqrt{D}$ heißt $\overline{\alpha} = r - s\sqrt{D}$ das zu α konjugierte Element.
b) Die Norm von α ist

$$\mathrm{N}(\alpha) = \alpha\overline{\alpha} = (r + s\sqrt{D})(r - s\sqrt{D}) = r^2 - s^2 D \in \mathbb{Q} \,.$$

c) Die Spur von α ist $\quad \mathrm{Sp}(\alpha) = \alpha + \overline{\alpha} = 2r \,.$

Lemma: *a)* Für $\alpha, \beta \in \mathbb{Q}[\sqrt{D}]$ ist $\overline{\alpha\beta} = \overline{\alpha} \cdot \overline{\beta}$.
b) Für $\alpha, \beta \in \mathbb{Q}[\sqrt{D}]$ ist $\mathrm{N}(\alpha\beta) = \mathrm{N}(\alpha) \cdot \mathrm{N}(\beta)$.
c) $\alpha \in \mathbb{Q}[\sqrt{D}]$ ist Wurzel der quadratischen Gleichung

$$X^2 - \mathrm{Sp}(\alpha)X + \mathrm{N}(\alpha) = 0 \,.$$

d) $\alpha \in \mathbb{Q}[\sqrt{D}]$ ist genau dann ganz, wenn $\mathrm{N}(\alpha)$ und $\mathrm{Sp}(\alpha)$ in $\mathbb{Z}$ liegen.
e) $\alpha \in \mathcal{O}_D$ ist genau dann eine Einheit, wenn $\mathrm{N}(\alpha) = \pm 1$ ist.

Beweis: a) Folgt sofort durch direktes Nachrechnen: Für $\alpha = r + s\sqrt{D}$ und $\beta = u + v\sqrt{D}$ ist

$$\begin{aligned}\overline{\alpha\beta} &= \overline{(ru + svD) + (rv + su)\sqrt{D}} = (ru + svD) - (rv + su)\sqrt{D} \\ &= (r - s\sqrt{D})(u - v\sqrt{D}) = \overline{\alpha}\overline{\beta} \,.\end{aligned}$$

b) Nach Definition ist

$$\mathrm{N}(\alpha\beta) = \alpha\beta \cdot \overline{\alpha\beta} = \alpha\beta\overline{\alpha}\overline{\beta} = \alpha\overline{\alpha} \cdot \beta\overline{\beta} = \mathrm{N}(\alpha) \cdot \mathrm{N}(\beta) \,.$$

c) Ist offensichtlich, denn nach dem Satz von VIÈTE sind α und $\overline{\alpha}$ Nullstellen der Gleichung

$$(X - \alpha)(X - \overline{\alpha}) = X^2 - \mathrm{Sp}(\alpha)X + \mathrm{N}(\alpha) = 0 \,.$$

d) folgt sofort aus *c)* und der Definition der Ganzheit.

e) Ist $\alpha \in \mathcal{O}_D^\times$ eine Einheit, so gibt es ein dazu inverses ganzes Element $\beta \in \mathcal{O}_D$, und wegen $\alpha\beta = 1$ ist auch $\mathrm{N}(\alpha) \cdot \mathrm{N}(\beta) = \mathrm{N}(\alpha\beta) = 1$. Die Norm ist also eine Einheit von $\mathbb{Z}$, d.h. $\mathrm{N}(\alpha) = \pm 1$.

Ist umgekehrt $\mathrm{N}(\alpha) = \alpha\overline{\alpha} = \pm 1$, so ist $\alpha \cdot (\pm\overline{\alpha}) = 1$, wir haben also ein ganzes Inverses. ∎

Das können wir beispielsweise anwenden auf die eingangs betrachteten Zerlegungen $6 = 2 \cdot 3 = (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5})$. In $\mathbb{Q}[\sqrt{-5}]$ ist

$$\mathrm{N}(2) = 2 \cdot 2 = 4, \quad \mathrm{N}(3) = 3 \cdot 3 = 9, \quad \mathrm{N}(1 \pm \sqrt{-5}) = 1 + 5 = 6 \,.$$

Echte Primteiler einer dieser Zahlen müßten also Norm ± 2 oder ± 3 haben. Wegen

$$\mathrm{N}(a + b\sqrt{-5}) = a^2 + 5b^2$$

müßte für solche Elemente $b = 0$ und $a^2 = 2$ oder 3 sein, was für ein $a \in \mathbb{Q}$ offensichtlich nicht möglich ist. Somit sind die Elemente $2, 3$ und $1 \pm \sqrt{-5}$ allesamt irreduzibel, und die Zahl sechs läßt sich auf zwei verschiedene Weisen als Produkt irreduzibler Elemente schreiben. (Es ist klar, daß 2 und 3 nicht zu $1 \pm \sqrt{-5}$ assoziiert sein können, denn die Normen assoziierter Elemente unterscheiden sich höchstens im Vorzeichen.)

Damit haben wir gezeigt, daß die Hauptordnung von $\mathbb{Q}[\sqrt{-5}]$ nicht faktoriell ist.

§5: Euklidische Ringe

In Kapitel I bewiesen wir die eindeutige Primzerlegung in $\mathbb{Z}$ mit Hilfe des EUKLIDischen Algorithmus. Wenn wir Beispiele für faktorielle Ringe $\mathcal{O}_D$ suchen, liegt es daher nahe, nach Ringen zu suchen, in denen es einen EUKLIDischen Algorithmus gibt. Solche Ringe heißen EUKLIDische Ringe.

Wie wir gesehen haben, ist die Division mit Rest das wichtigste Werkzeug beim EUKLIDischen Algorithmus, und wie sich in diesem Abschnitt herausstellen wird, brauchen wir kein weiteres. Wir definieren daher

Definition: Ein EUKLIDischer Ring ist ein Integritätsbereich R zusammen mit einer Abbildung $\nu\colon R \setminus \{0\} \to \mathbb{N}_0$, so daß gilt: Ist $x|y$, so ist $\nu(x) \leq \nu(y)$, und zu je zwei Elementen $x, y \in R$ gibt es Elemente $q, r \in R$ mit

$$x = qy + r \quad \text{und} \quad r = 0 \quad \text{oder} \quad \nu(r) < \nu(y)\,.$$

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

Das Standardbeispiel ist natürlich der Ring $\mathbb{Z}$ der ganzen Zahlen mit $\nu(x) = |x|$. Ein anderes Beispiel ist der Polynomring $k[X]$ über einem Körper k: Hier können wir $\nu(f)$ für ein Polynom $f \neq 0$ als den Grad von f definieren; dann erfüllt auch die Polynomdivision mit Rest die Forderung an einen EUKLIDischen Ring.

Wie angekündigt, gilt

Lemma: In einem EUKLIDischen Ring R gibt es zu je zwei Elementen $x, y \in R$ einen ggT. Dieser kann nach dem EUKLIDischen Algorithmus berechnet werden und läßt sich als Linearkombination mit Koeffizienten aus R von x und y darstellen

Beweis: In jedem Integritätsbereich folgt aus der Gleichung $x = qy + r$ mit $x, y, q, r \in R$, daß die gemeinsamen Teiler von x und y gleich denen von y und r sind. Speziell in einem EUKLIDischen Ring können wir dabei r als Divisionsrest wählen und, wie beim klassischen EUKLIDischen Algorithmus, danach y durch r dividieren *usw.,* wobei wir eine Folge (r_i) von Divisionsresten erhalten mit der Eigenschaft, daß in jedem Schritt die gemeinsamen Teiler von x und y gleich denen von r_{i-1} und r_i sind. Außerdem ist stets entweder $r_i = 0$ oder $\nu(r_i) < \nu(r_{i-1})$, so daß die Folge nach endlich vielen Schritten mit einem $r_n = 0$ abbrechen muß. Auch hier sind die gemeinsamen Teiler von r_{n-1} und $r_n = 0$ genau die gemeinsamen Teiler von x und y. Da jede Zahl Teiler der Null ist, sind die gemeinsamen Teiler von r_{n-1} und Null aber genau die Teiler von r_{n-1}, und unter diesen gibt es natürlich einen größten, nämlich r_{n-1} selbst. Somit haben auch x und y einen größten gemeinsamen Teiler, nämlich den nach dem EUKLIDischen Algorithmus berechneten letzten von Null verschiedenen Divisionsrest r_{n-1}.

Auch die lineare Kombinierbarkeit folgt wie im klassischen Fall: Bei jeder Division mit Rest ist der Divisionsrest als Linearkombination von Dividend und Divisor darstellbar; beim EUKLIDischen Algorithmus beginnen wir mit Dividend x und Divisor y, die natürlich beide als Linearkombinationen von x und y darstellbar sind, und induktiv folgt, daß auch alle folgenden Dividenden und Divisoren sind als Reste einer vorangegangenen Division Linearkombinationen von x und y sind, also ist es auch ihr Divisionsrest. Insbesondere ist auch der ggT als letzter nichtverschwindender Divisionsrest Linearkombination von x und y, und die Koeffizienten können wie in Kapitel I mit dem erweiterten EUKLIDischen Algorithmus berechnet werden. ∎

Satz: Jeder EUKLIDische Ring ist faktoriell.

Beweis: Wir müssen zeigen, daß jedes Element $x \neq 0$ aus R bis auf Reihenfolge und Assoziiertheit eindeutig als Produkt aus einer Einheit und

geeigneten Potenzen irreduzibler Elemente geschrieben werden kann. Wir beginnen damit, daß sich x überhaupt in dieser Weise darstellen läßt.

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

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

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

Andernfalls läßt sich $x = yz$ als Produkt zweier Elemente schreiben, die beide keine Einheiten sind. Da y und z Teiler von x sind, sind $\nu(y), \nu(z) \leq \nu(x)$. Wir wollen uns überlegen, daß sie tatsächlich sogar echt kleiner sind.

Dazu dividieren wir y mit Rest durch x; das Ergebnis sei q Rest r, d.h. $y = qx+r$ mit $r = 0$ oder $\nu(r) < \nu(x)$. Wäre $r = 0$, wäre y ein Vielfaches von x, es gäbe also ein $u \in R$ mit $y = ux = u(yz) = (uz)y$. Damit wäre $uz = 1$, also z eine Einheit, im Widerspruch zur Annahme. Somit ist $\nu(r) < \nu(x)$.

Als Teiler von x ist y auch Teiler von $r = y - qx = y(1 - qz)$, also muß $\nu(y) \leq \nu(r) < \nu(x)$ sein. Genauso folgt, daß auch $\nu(z) < \nu(x)$ ist.

Nach Induktionsvoraussetzung lassen sich daher y und z als Produkte von Einheiten und Potenzen irreduzibler Elemente schreiben, und damit läßt sich auch $x = yz$ so darstellen.

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

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

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

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

als Linearkombination von p und x schreiben. Multiplikation mit y macht daraus $y = \alpha px + \beta xy$, und hier sind beide Summanden auf der rechten Seite durch p teilbar: Bei αpx ist das klar, und bei βxy folgt es daraus, daß nach Voraussetzung p ein Teiler von xy ist. Also ist p Teiler von y, und die Zwischenbehauptung ist bewiesen.

Induktiv folgt sofort:

Falls ein irreduzibles Element p ein Produkt $\prod_{i=1}^{r} x_i$ teilt, teilt es mindestens einen der Faktoren x_i.

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

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

Seien nun

$$x = u \prod_{i=1}^{r} p_i^{e_i} = v \prod_{j=1}^{s} q_j^{f_j}$$

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

Bemerkung: Die Umkehrung dieses Satzes gilt nicht: Beispielsweise sind nach einem Satz von GAUSS auch $\mathbb{Z}[X]$ sowie Polynomringe in mehr als einer Veränderlichen über $\mathbb{Z}$ oder einem Körper faktoriell, aber

keiner dieser Ringe ist EUKLIDisch, da sich weder der ggT eins von 2 und X in $\mathbb{Z}[X]$ noch der ggT eins von X und Y in $k[X,Y]$ als Linearkombination der Ausgangselemente schreiben läßt.

Wir interessieren uns in diesem Kapitel vor allem für quadratische Zahlkörper; daher wollen wir uns fragen, wann die Hauptordnung eines solchen Körpers EUKLIDisch ist.

Für einen EUKLIDischen Ring brauchen wir zunächst eine Abbildung ν nach $\mathbb{N}_0$. Für $\mathbb{Z}$ konnten wir einfach den Betrag nehmen; für die Hauptordnung eines quadratischen Zahlkörpers können wir unser Glück versuchen mit dem Betrag der Norm.

Falls die Hauptordnung $\mathcal{O}_D$ von $\mathbb{Q}[\sqrt{D}]$ zusammen mit dieser Abbildung ein EUKLIDischer Ring ist, muß es zu je zwei Elementen $r, s \in \mathcal{O}_D$ mit $s \neq 0$ ein Element $q \in \mathcal{O}_D$ geben, so daß $|\mathrm{N}(r - sq)| < |\mathrm{N}(s)|$ ist. Division durch s macht daraus die Ungleichung

$$\left|\mathrm{N}\left(\frac{r}{s} - q\right)\right| < |\mathrm{N}(1)| = 1 \,.$$

Da sich jedes Element von $\mathbb{Q}[\sqrt{D}]$ als so ein Quotient r/s darstellen läßt, muß es also zu jedem $x \in \mathbb{Q}[\sqrt{D}]$ ein $q \in \mathcal{O}_D$ geben, so daß $|\mathrm{N}(x - q)| < 1$ ist. Dies zeigt auch, wie man im EUKLIDischen Fall die Division mit Rest durchführt: Man berechnet den Quotienten x/y zunächst im Körper $\mathbb{Q}[\sqrt{D}]$ und nimmt dann das bezüglich der Norm nächstgelegene Element von $\mathcal{O}_D$.

Betrachten wir als Beispiel die Division von $23 + 9i$ durch $2 - 3i$ im Ring $\mathbb{Z}[i]$ der GAUSSschen Zahlen. In $\mathbb{Q}[i]$ ist

$$\frac{23+9i}{2-3i} = \frac{(23+9i)(2+3i)}{13} = \frac{19}{13} + \frac{87}{13}i \,.$$

Da $19 : 13 = 1$ Rest 6 und $87 : 13 = 6$ Rest 9 ist, liegt das Element $1 + 7i$ aus $\mathbb{Z}[i]$ am nächsten bei dieser Zahl. Die Norm von

$$\frac{19}{13} + \frac{87}{13}i - (1+7i) = \frac{6}{13} - \frac{4}{13}i$$

ist $(6^2 + 4^2)/13^2 = 52/169$ und damit deutlich kleiner als eins. Somit ist

$$(23+9i) : (2-3i) = (1+7i) \text{ Rest } -2i$$

ein mögliches Ergebnis der Division mit Rest. Ein anderes wäre

$$(23+9i) : (2-3i) = (1+6i) \text{ Rest } 3 \,,$$

denn auch die Norm von 3 ist kleiner als die von $2+3i$. (Der Rest wurde jeweils als Dividend minus Divisor mal Quotient berechnet.) Da in der Definition eines EUKLIDischen Rings von Eindeutigkeit keine Rede war, ist dies kein Problem. (Auch beim EUKLIDischen Algorithmus wird nie gebraucht, daß das Ergebnis der Division mit Rest eindeutig ist; in der Tat läßt sich der sogar für $\mathbb{Z}$ gelegentlich dadurch etwas beschleunigen, daß man bei der Division mit Rest auch negative Reste zuläßt und stets das Ergebnis nimmt, bei dem der Betrag des Rests minimal ist.)

Um zu sehen, in welchen der Ringe $\mathcal{O}_D$ eine solche Division mit Rest stets möglich ist, betrachten wir die Situation geometrisch. Wir beschränken uns dabei zunächst auf den imaginärquadratischen Fall.

Um besser zu sehen, welche Terme in den folgenden Rechnungen positiv und welche negativ sind, schreiben wir den Körper als $\mathbb{Q}[\sqrt{-D}]$ mit $D > 0$; seine Elemente lassen sich dann in der Form $x+iy\sqrt{D}$ darstellen, wobei $i = \sqrt{-1}$ die imaginäre Einheit bezeichnet.

Wir betrachten $\mathbb{Q}[\sqrt{-D}]$ als Teilmenge der komplexen Zahlenebene $\mathbb{C}$; dann ist

$$\mathrm{N}(r + is\sqrt{D}) = (r + is\sqrt{D})(r - is\sqrt{D}) = r^2 + s^2 D = \left|r + is\sqrt{D}\right|^2$$

einfach das Quadrat des üblichen komplexen Betrags. $\mathcal{O}_{-D}$ ist also genau dann ein EUKLIDischer Ring mit der Norm als Betragsfunktion, wenn es zu jedem Element $x \in \mathbb{Q}[\sqrt{-D}]$ ein $q \in \mathcal{O}_{-D}$ gibt, so daß $|x - q| < 1$ ist. Da $\mathbb{Q}[\sqrt{-D}]$ dicht in $\mathbb{C}$ liegt, müssen dazu die Kreisscheiben mit Radius eins um die Punkte aus $\mathcal{O}_{-D}$ die ganze komplexe Zahlenebene überdecken. Bei den Punkten, die nur auf Rändern solcher Kreisscheiben liegen, muß zudem überprüft werden, daß sie nicht in $\mathbb{Q}[\sqrt{-D}]$ liegen: Andernfalls sind das Körperelemente, für die obige Ungleichung nicht erfüllt ist.

Die Punkte aus $\mathcal{O}_D$ bilden ein Gitter in $\mathbb{C}$; für jeden der Gitterpunkte $q \in \mathcal{O}_D$ definieren wir dessen *Wirkungsbereich* oder VORONOI-Bereich

als den Abschluß der Menge aller $z \in \mathbb{C}$, die näher bei q liegen als bei jedem der anderen Gitterpunkte:

$$W(q) = \left\{ z \in \mathbb{C} \;\middle|\; \forall q' \in \mathcal{O}_{-D} : |z-q| \le |z-q'| \right\}$$

Offensichtlich liegt jedes $z \in \mathbb{C}$ in mindestens einem dieser Wirkungsbereiche, und falls

$$W(q) \subseteq \left\{ z \in \mathbb{C} \;\middle|\; |z-q| < 1 \right\},$$

folgt insbesondere, daß jedes Element von $\mathbb{Q}[\sqrt{-D}]$ im Innern einer Kreisscheibe mit Radius eins um einen Gitterpunkt liegt: Dann ist der Ring $\mathcal{O}_{-D}$ EUKLIDisch.

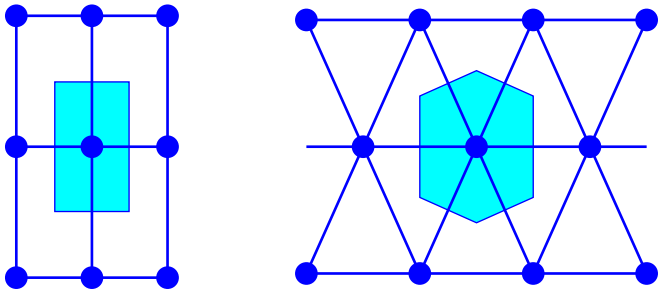
Der Wirkungsbereich eines Gitterpunkts z unterscheidet sich von dem des Nullpunkts nur durch eine Verschiebung um z; entsprechendes gilt auch für die Kreise mit Radius eins um die beiden Punkte. Daher reicht es, zu untersuchen, wann der Wirkungsbereich des Nullpunkts ganz im Innern des Einheitskreises liegt.

Die Struktur des Wirkungsbereichs hängt ab von $D \bmod 4$: Falls $D \not\equiv -1 \bmod 4$, d.h. $D \not\equiv 3 \bmod 4$, ist $\mathcal{O}_{-D} = \mathbb{Z} \oplus \mathbb{Z}[\sqrt{-D}]$. In der komplexen Zahlenebene bilden diese Punkte ein Rechteckgitter mit den Gitterpunkten $q = r + is\sqrt{D}$ zu $r, s \in \mathbb{Z}$. Der Wirkungsbereich des Nullpunkts ist daher das Rechteck mit Ecken $\pm\frac{1}{2} \pm \frac{i}{2}\sqrt{D}$, und die am weitesten von der Null entfernte Punkte sind die Ecken mit Abstand

$$\sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{\sqrt{D}}{2}\right)^2} = \frac{\sqrt{1+D}}{2}.$$

Dies ist genau dann echt kleiner als eins, wenn $D \le 2$ ist, d.h. $D = 1$ oder $D = 2$.

Für $D = 3$ überdecken zwar die abgeschlossenen Kreisscheiben mit Radius eins um die Gitterpunkte ganz $\mathbb{C}$, aber die gerade betrachteten Eckpunkte sind Elemente des Körpers $\mathbb{Q}[\sqrt{-3}]$, die in keiner offenen Kreisscheibe um einen Gitterpunkt liegen. Das ist allerdings hier kein Problem, denn in $\mathbb{Q}[\sqrt{-3}]$ sind diese Eckpunkte ja selbst Gitterpunkte: Für $D \equiv 3 \bmod 4$ gibt es schließlich mehr ganze Zahlen in $\mathbb{Q}[\sqrt{-D}]$.



Hier wird das Gitter $\mathcal{O}_{-D}$ erzeugt von der Eins und von $\frac{1}{2}(1 + i\sqrt{D})$. Der Nullpunkt hat somit sechs nächste Nachbarn, nämlich ± 1 und $\pm\frac{1}{2} \pm \frac{i}{2}\sqrt{D}$. Die Wirkungsbereiche der Null und von ± 1 werden getrennt durch die Geraden $x = \pm\frac{1}{2}$, und auch für die vier anderen Punkte müssen wir die Mittelsenkrechte zur Verbindungsstrecke betrachten. Diese geht durch den Streckenmittelpunkt, also durch $\pm\frac{1}{4} \pm \frac{i}{4}\sqrt{D}$, und sie steht senkrecht auf dieser Strecke.

Eine Drehung um 90° kann in der komplexen Zahlenebene realisiert werden durch Multiplikation mit i; wir haben also die vier Geraden

$$\left\{ \left(\pm\frac{1}{2} \pm \frac{i}{2}\sqrt{D}\right) + \left(\mp\frac{\sqrt{D}}{2} \pm \frac{i}{2}\right) t \;\middle|\; t \in \mathbb{R} \right\}.$$

Zwei der Ecken des Wirkungsbereichs liegen (aus Symmetriegründen) auf der imaginären Achse; Einsetzen in die Geradengleichungen ergibt, daß deren Imaginärteile gleich $\pm\frac{1}{4}(\sqrt{D} + 1/\sqrt{D})$ sind. Die restlichen vier Ecken liegen auf den Geraden $x = \pm\frac{1}{2}$, haben also Realteil $\pm\frac{1}{2}$; hier führt die Rechnung auf die Imaginärteile $\pm\frac{1}{4}(\sqrt{D} - 1/\sqrt{D})$.

Der Abstand dieser Punkte vom Nullpunkt ist

$$\sqrt{\left(\frac{1}{2}\right)^2 + \frac{(\sqrt{D} - 1/\sqrt{D})^2}{16}} = \frac{1}{4}\sqrt{4 + D - 2 + \frac{1}{D}} = \frac{1}{4}\sqrt{2 + D + \frac{1}{D}};$$

dies ist genau dann kleiner als eins, wenn gilt

$$2 + D + \frac{1}{D} < 4^2 = 16 \quad \text{oder} \quad D + \frac{1}{D} < 14.$$

Die einzigen $D \equiv 3 \bmod 4$, die dies erfüllen, sind $D = 3$, $D = 7$ und $D = 11$. Für diese ist auch $\frac{1}{4}(\sqrt{D} + 1/\sqrt{D}) < 1$, so daß dann und nur dann der gesamte Wirkungsbereich der Null im Einheitskreis liegt.

Die einzigen imaginärquadratischen Zahlkörper $\mathbb{Q}[\sqrt{D}]$, deren Hauptordnung bezüglich der Norm EUKLIDsch ist, sind somit die mit

$$D \in \{-1, -2, -3, -7, -11\} ;$$

von diesen wissen wir damit auch, daß ihre Hauptordnung faktoriell ist.

Es ist nicht bekannt, ob es andere $D < 0$ gibt, für die die Hauptordnung bezüglich einer anderen Funktion $\nu: \mathcal{O}_D \setminus \{0\} \to \mathbb{N}_0$ EUKLIDisch ist. Bekannt ist aber, daß die einzigen weiteren faktoriellen Hauptordnungen $\mathcal{O}_D$ die sind mit $D \in \{-19, -43, -67, -163\}$: *siehe* H. STARK: A complete determination of the complex fields of class numbers one, *Michigan J. of Math.* **14** (1967), 1–27. Die Methoden seines Beweises liegen deutlich über dem Niveau dieser Vorlesung.

Im reellquadratischen Fall wird die Ungleichung $|\mathrm{N}(z - q)| - 1$ für $z = x + y\sqrt{D}$ und $q = r + s\sqrt{D}$ zu

$$\left|(x - r)^2 - (y - v)^2 D\right| < 1 .$$

Betrachten wir für festes $q = r + s\sqrt{D} \in \mathcal{O}_D$ die Menge Z_q aller $(x, y) \in \mathbb{R}^2$, für die $z = x + y\sqrt{D}$ diese Ungleichung erfüllt, erhalten wir also einen Bereich, der durch Hyperbeln begrenzt wird, und wir müssen zeigen, daß die Vereinigung aller Z_q für $q \in \mathcal{O}_D$ ganz $\mathbb{R}^2$ ist. Durch mühsames Abhaken vieler Einzelfälle folgt aus einer ganzen Reihe von Arbeiten, daß dies genau dann der Fall ist, wenn

$$D \in \{2, 3, 5, 6, 7, 11, 13, 17, 19, 21, 29, 33, 37, 41, 57, 73\} .$$

Die letzten offenen Fälle wurden 1950 untersucht in H. CHATLAND, H. DAVENPORT: Euclid's algorithm in real quadratic fields, *Canadian J. Math.* **2** (1950), 289–296; dort sind auch die weiteren Arbeiten zitiert, aus denen zusammen schließlich das obige Ergebnis folgt.

Genau für diese D ist also $\mathcal{O}_D$ EUKLIDisch bezüglich der Norm. Es gibt zahlreiche weitere positive D, für die $\mathcal{O}_D$ faktoriell ist; vermutungsweise sind es sogar unendlich viele. Ob einige dieser Ringe möglicherweise

bezüglich einer anderen Abbildung $\nu: R \setminus \{0\} \to \mathbb{N}_0$ EUKLIDisch sind, ist nicht bekannt, und die Nichtexistenz einer solchen Abbildung ist natürlich nur schwer zu beweisen.

§6: Einheiten in quadratischen Zahlkörpern

Ist $x + y\sqrt{D}$ eine Einheit in $\mathcal{O}_D$ (man spricht auch kurz, aber schlampig, von einer Einheit des Zahlkörpers $\mathbb{Q}[\sqrt{D}]$), so muß die Norm $x^2 - Dy^2$ eine Einheit in $\mathbb{Z}$ sein, also gleich ± 1.

Im imaginärquadratischen Fall ist $x^2 - Dy^2$ die Summe zweier positiver Terme; hier kommt also nur der Wert +1 in Frage. Die einzigen ganzzahligen Lösungen sind offensichtlich $(x, y) = (\pm 1, 0)$, sowie im Fall $D = -1$ der GAUSSschen Zahlen $(x, y) = (0, \pm 1)$. Für $D \equiv 1 \bmod 4$ sind auch echt halbzahlige Werte für sowohl x als auch y zugelassen; dies führt offensichtlich nur für $D = -3$ zu weiteren Lösungen, nämlich $x = \pm\frac{1}{2}$ und $y = \pm\frac{1}{2}$. Damit haben wir gezeigt:

Lemma: In einem imaginärquadratischen Zahlkörper $\mathbb{Q}[\sqrt{D}]$ gibt es für $D \neq -1$ und $D \neq -3$ nur die Einheiten ± 1. In $\mathbb{Q}[i]$ gibt es zusätzlich noch die Einheiten $\pm i$, und in $\mathbb{Q}[\sqrt{-3}]$ sind die Einheiten genau die sechsten Einheitswurzeln ± 1 und $\pm\frac{1}{2} \pm \frac{1}{2}\sqrt{-3}$. ∎

In reellquadratischen Körpern führt die Bedingung $\mathrm{N}(x) = \pm 1$ auf die Gleichung $x^2 - Dy^2 = \pm 1$ mit einem positiven D; hier können wir nicht ausschließen, daß es unendlich viele Lösungen gibt.

Betrachten wir zunächst den Fall, daß $x^2 - Dy^2 = 1$ ist. Diese Gleichung bezeichnet man als die PELLsche Gleichung.

JOHN PELL (1611–1685) wurde im englischen Sussex geboren und ging auch dort zur Schule. Bereits 1624 begann er sein Studium an der Universität Cambridge; 1628 erhielt er seinen Bachelor und 1630 seinen Master. Danach arbeitete er meist als Lehrer. Von 1654–1658 war er als Diplomat im Auftrag CROMWELLs in Zürich. In einem dort von JOHANN HEINRICH RAHN (1622–1676) verfaßten Buch, an dem PELL wesentlich mitwirkte, ist ein Beispiel der obigen Gleichung zu finden, weshalb sie EULER (1707–1783) nach PELL benannte. Tatsächlich wurde sie wohl erstmalig von dem indischen Mathematiker und Astronom BRAHMAGUPTA (598–670) untersucht; die vollständige Theorie dazu geht

zurück auf LAGRANGE (1736–1813), der die Gleichung als ein Problem bezeichnet, das FERMAT den englischen Mathematikern stellte. Nach seiner Rückkehr aus Zürich wurde PELL Priester. 1663 wählte ihn die Royal Society zum Mitglied, 1675 wurde er deren Vizepräsident.

Mit der PELLschen Gleichung werden wir uns im nächsten Kapitel genauer beschäftigen, und wir werden sehen, daß sie stets unendlich viele Lösungen hat. Als Vorbereitung dazu wollen wir uns hier etwas genauer mit der Struktur der Einheitengruppe beschäftigen. Dazu betrachten wir die Abbildung

$$\lambda\colon \begin{cases} \mathcal{O}_D^\times \to \mathbb{R}^2 \\ \quad \alpha \mapsto (\log|\alpha|, \log|\overline{\alpha}|) \end{cases}$$

Da eine Einheit Norm ± 1 hat, ist $|\alpha| \cdot |\overline{\alpha}| = 1$, das Bild von λ liegt also auf der zweiten Winkelhalbierenden $y = -x$ von $\mathbb{R}^2$. Außerdem sind α und $\overline{\alpha}$ reell, so daß α genau dann im Kern von λ liegt, wenn $\alpha = \pm 1$ ist.

Das Bild von λ ist diskret, denn hat $\lambda(\alpha)$ höchstens den Abstand M vom Nullpunkt, so ist $\log|\alpha| \le M$ und $\log|\overline{\alpha}| \le M$. Ist $\log R = M$, so ist also $|\alpha| \le R$ und $|\overline{\alpha}| \le R$. Damit ist $|\mathrm{Sp}(\alpha)| \le 2R$ und $|\mathrm{N}(\alpha)| \le R^2$. Da Norm und Spur ganzzahlig sind, gibt es also für beide nur endlich viele Möglichkeiten, und da für ein ganzes Element Norm und Spur zusammen mit dem führenden Koeffizienten eins die Koeffizienten der quadratischen Gleichung sind, gibt es auch nur endlich viele quadratische Gleichungen und damit nur endlich viele Möglichkeiten für α.

Somit gibt es im Bild von λ ein Element $\lambda(\alpha) = (r, -r)$ mit *minimalem* $r > 0$. Wir wollen uns überlegen, daß das jeder andere Punkt im Bild ein ganzzahliges Vielfaches davon ist. Da mit $(s, -s)$ auch $(-s, s)$ im Bild liegt, können wir uns dabei auf Punkte $(s, -s)$ mit $s \ge 0$ beschränken.

Für einen solchen Punkt $\lambda(\beta) = (s, -s)$ gibt es jedenfalls ein größtes $n \in \mathbb{N}_0$, so daß $nr \le s$ ist. Dann ist

$$\lambda(\beta\alpha^{-n}) = \lambda(\beta) - n\lambda(\alpha) = (s, -s) - n(r, -r) = (s - nr, nr - s)\,,$$

so daß auch dieser Punkt im Bild liegt. Nach Wahl von n ist aber $0 \le s - nr < r$; wegen der Minimalität von r ist also $s - nr = 0$, d.h. $s = nr$ und $\beta = \alpha^n$.

Damit haben wir bewiesen

Satz: Falls es im reellquadratischen Zahlkörper $K = \mathbb{Q}[\sqrt{D}]$ ein Element aus $\mathcal{O}_D^\times$ gibt, dessen Norm größer als eins ist, gibt es auch ein entsprechendes Element α mit kleinster Norm, und die Einheiten von $\mathcal{O}_D$ sind genau die Elemente $\pm\alpha^n$ mit $n \in \mathbb{Z}$. Insbesondere ist dann die Einheitengruppe unendlich.

∎

Im nächsten Kapitel werden wir sehen, daß jeder reellquadratische Zahlkörper eine solche „Grundeinheit“ α hat; die Einheitengruppe eines reellquadratischen Zahlkörpers besteht also stets genau aus den Elemente der Form $\pm\alpha^n$ mit $n \in \mathbb{Z}$ und $\alpha \in \mathcal{O}_D^\times$.

Bevor wir das im einzelnen untersuchen, wollen wir zum Abschluß dieses Kapitels und zur Vorbereitung auf das nächste noch ein Beispiel einer nichtkommutativen Variante eines Zahlkörpers betrachten.

§7: Quaternionen

Nachdem durch die komplexen Zahlen $\mathbb{R}^2$ mit der Struktur eines Körpers versehen war, versuchten viele Mathematiker ähnliches auch für $\mathbb{R}^3$ zu erreichen. Natürlich kann weder $\mathbb{R}^3$ noch sonst ein $\mathbb{R}^n$ mit $n > 2$ zu einem Körper gemacht werden, denn ein solcher Körper wäre eine algebraische Erweiterung von $\mathbb{R}$; da aber der algebraische Abschluß von $\mathbb{R}$ gleich $\mathbb{C}$ ist, muß dann $n = 1$ oder $n = 2$ sein.

Die damaligen Mathematiker waren jedoch bescheidener: Ihnen genügte es, einfach irgendeine Art von Multiplikation zu finden, die nicht unbedingt allen Körperaxiomen genügte – von Körpern sprach damals ohnehin noch niemand.

Erst 1940 konnte HEINZ HOPF (1894–1971) (auf dem Umweg über Vektorfelder auf Sphären) zeigen, daß das nicht möglich ist: Selbst eine bilineare Abbildung $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ kann nur dann existieren, wenn n eine Zweierpotenz ist, und 1958 zeigten dann unabhängig voneinander und mit verschiedenen Methoden JOHN MILNOR und MICHEL KERVAIRE, daß auch noch $n \le 8$ sein muß, so daß nur die vier Möglichkeiten $n = 1, 2, 4$ und 8 in Frage kommen. Genau in diesen Fällen waren auch bereits entsprechende Produkte bekannt:

Für $n = 1$ und 2 haben wir natürlich die reelle *bzw.* komplexe Multiplikation. Den Fall $n = 4$ löste HAMILTON 1843: Er fand eine Multiplikation auf $\mathbb{R}^4$, die zwar nicht kommutativ ist, ansonsten aber alle Körperaxiome erfüllt. Man spricht in so einem Fall von einem *Schiefkörper* oder, in der neueren Literatur, einer *Divisionsalgebra.* HAMILTON bezeichnete seine vierdimensionalen Zahlen als *Quaternionen.* Kurz danach konstruierte ARTHUR CAYLEY (1821–1895) ein nicht-assoziatives Produkt auf $\mathbb{R}^8$; die so erhaltenen „Zahlen" nannte er *Oktaven.*

HAMILTON wählte eine Basis von $\mathbb{H} = \mathbb{R}^4$, die aus der Eins sowie drei „imaginären Einheiten" $\mathbf{i}, \mathbf{j}, \mathbf{k}$ besteht, d.h. $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$. Außerdem postulierte er, daß $\mathbf{ij} = -\mathbf{ji} = \mathbf{k}$ sein sollte; daraus lassen sich dann über das Assoziativgesetz auch die anderen Produkte imaginärer Einheiten berechnen.

Damit ist, wenn man die Gültigkeit des Distributivgesetzes postuliert, eine Multiplikation auf $\mathbb{R}^4$ definiert; der Beweis, daß hierbei alle Körperaxiome außer der Kommutativität der Multiplikation erfüllt sind, enthält wie üblich nur einen etwas schwierigeren Punkt, die Existenz von Inversen; der Rest ist mühsame Abhakerei.

WILLIAM ROWEN HAMILTON (1805–1865) wurde in Dublin geboren; bereits mit fünf Jahren sprach er Latein, Griechisch und Hebräisch. Mit dreizehn begann er, mathematische Literatur zu lesen, mit 21 wurde er, noch als Student, Professor der Astronomie am Trinity College in Dublin. Er verlor allerdings schon bald sein Interesse an der Astronomie und beschäftigte sich stattdessen mit mathematischen und physikalischen Problemen. Am bekanntesten ist er für seine Entdeckung der Quaternionen, vorher publizierte er aber auch bedeutende Arbeiten über Optik, Dynamik und Algebra.

Zum Glück fand CAYLEY 1858 einen einfacheren Weg: Die vier komplexen 2×2-Matrizen

$$E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ I = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},\ J = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \text{ und } K = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

erfüllen dieselben Relationen

$$I^2 = J^2 = K^2 = -E \quad \text{und} \quad IJ = -JI = K\,;$$

wir können also die Quaternion $a + bi + cj + dk$ identifizieren mit der Matrix

$$aE + bI + cJ + dK = \begin{pmatrix} a + di & b + ci \\ -b + ci & a - di \end{pmatrix} \in \mathbb{C}^{2\times 2}\,.$$

Da für Matrizen das Assoziativgesetz wie auch das Distributivgesetz gelten, ist klar, daß das Produkt zweier solcher Matrizen wieder von derselben Form ist und daß auch die Quaternionenmultiplikation Assoziativ- und Distributivgesetz erfüllt.

Die Quaternionen entsprechen somit genau den komplexen 2×2-Matrizen der Form

$$\begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \quad \text{mit} \quad \alpha = a + di, \beta = b + ci\,.$$

Die Determinante dieser Matrix ist

$$\alpha\overline{\alpha} + \beta\overline{\beta} = a^2 + b^2 + c^2 + d^2\,.$$

Definieren wir in Analogie zum Fall der quadratischen Zahlkörper wieder das konjugierte Element zu $\gamma = a + bi + cj + dk$ als die Quaternion $\overline{\gamma} = a - bi - cj - dk$, so entspricht $\overline{\gamma}$ der Matrix

$$\begin{pmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \begin{pmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{pmatrix} = (\alpha\overline{\alpha} + \beta\overline{\beta})E\,.$$

Damit folgt insbesondere, daß $\gamma\overline{\gamma}$ eine reelle Zahl ist, die genau dann verschwindet, wenn $\gamma = 0$ ist. Wir bezeichnen diese Zahl wieder als die *Norm* $\mathrm{N}(\gamma)$ der Quaternion γ, und wieder ist $\overline{\gamma}/\mathrm{N}(\gamma)$ das multiplikative Inverse zu γ – sowohl für die Links- wie auch die Rechtsmultiplikation.

$\mathrm{N}(\gamma)$ ist gleichzeitig die Determinante der γ zugeordneten Matrix; aus dem Multiplikationssatz für Determinanten folgt daher sofort die Formel

$$\mathrm{N}(\gamma\delta) = \mathrm{N}(\gamma)\mathrm{N}(\delta)\,.$$

Kapitel 7
Quadratische Formen

Eine quadratische Form ist ein Ausdruck der Form

$$F(x,y) = Ax^2 + Bxy + Cy^2 \quad \text{mit} \quad A,B,C \in \mathbb{Z}\,;$$

die Zahlentheorie interessiert sich vor allem dafür, welche Werte $F(x,y)$ für $x,y \in \mathbb{Z}$ annehmen kann.

§1: Summen zweier Quadrate

Der einfachste Fall ist die Form $F(x,y) = x^2 + y^2$. Er hängt eng zusammen mit der Hauptordnung $\mathbb{Z}[i]$ von $\mathbb{Q}[i]$, denn

$$x^2 + y^2 = (x+iy)(x-iy)$$

ist die Norm von $x+iy$. Eine ganze Zahl n ist also genau dann als Summe zweier Quadrate darstellbar, wenn sie die Norm einer GAUSSschen ganzen Zahl ist.

Modulo vier ist $0^2 \equiv 2^2 \equiv 0$ und $1^2 \equiv 3^2 \equiv 1$; somit ist jede Summe zweier Quadrate kongruent null, eins oder zwei modulo vier. Eine Zahl kongruent drei modulo vier kann somit nicht als Summe zweier Quadratzahlen auftreten.

Auf der Suche nach positiven Ergebnissen können wir uns auf Primzahlen beschränken, denn wie FIBONACCI bereits im dreizehnten Jahrhundert zeigte, gilt:

Lemma: Sind zwei Zahlen $n,m \in \mathbb{N}$ darstellbar als Summen zweier Quadrate, so gilt dasselbe für ihr Produkt nm.

Beweis: Wenn n und m als Summen zweier Quadrate darstellbar sind, gibt es $\alpha,\beta \in \mathbb{Z}[i]$, so daß $n = \mathrm{N}(\alpha)$ und $m = \mathrm{N}(\beta)$ ist. Wegen der Multiplikativität der Norm ist dann $nm = \mathrm{N}(\alpha\beta)$ ebenfalls eine Norm und damit als Summe zweier Quadrate darstellbar. ∎

(FIBONACCI bewies dieses Lemma natürlich nicht über den Umweg mit Normen GAUSSscher Zahlen; er fand eine explizite Formel für die Darstellung des Produkts als Summe zweier Quadrate. Es handelt sich dabei um dieselbe Formel, zu der wir auch durch Ausmultiplizieren der Gleichung $\mathrm{N}(\alpha)\cdot\mathrm{N}(\beta) = \mathrm{N}(\alpha\beta)$ für $\alpha = a+ib$ und $\beta = c+id$ kämen.)

Da $2 = 1^2+1^2$ als Summe zweier Quadrate darstellbar ist, müssen wir nur die ungeraden Primzahlen untersuchen, und hier wissen wir bereits, daß Zahlen kongruent drei modulo vier nicht als solche Summen auftreten.

Satz: Eine ungerade Primzahl p ist genau dann darstellbar als Summe zweier Quadrate, wenn $p \equiv 1 \bmod 4$. Diese Darstellung ist eindeutig bis auf die Reihenfolge der Summanden.

Beweis: Aus Kapitel I, §8 wissen wir, daß die multiplikative Gruppe des Körper $\mathbb{F}_p$ von einem einzigen Element g erzeugt wird. Für $p = 4k+1$ ist dann $g^{4k} = 1$, also $g^{2k} = -1$. Somit ist $-1 = p-1$ in $\mathbb{F}_p$ ein Quadrat.

In $\mathbb{Z}$ gibt es daher Zahlen x, für die $x^2 \equiv -1 \bmod p$ ist oder, anders ausgedrückt, $x^2+1 = \ell p$ für ein $\ell \in \mathbb{N}$. Da jede Restklasse modulo p einen Vertreter mit Betrag kleiner $p/2$ enthält, können wir dabei annehmen, daß $|x| < p/2$ ist; dann ist mit einer geeigneten natürlichen Zahl ℓ

$$x^2 + 1 = \ell p < \frac{p^2}{4} + 1 < \frac{p^2}{2} \Longrightarrow \ell < p\,.$$

Es gibt also ein $\ell < p$, so daß ℓp Summe zweier Quadrate ist. Das kleinste solche ℓ sei m; wir müssen zeigen, daß $m = 1$ ist.

Zunächst ist klar, daß m eine ungerade Zahl sein muß, denn aus der Formel $x^2+y^2 = mp$ mit geradem m folgt, daß x und y entweder beide gerade oder beide ungerade sind; $x \pm y$ sind also gerade und

$$\left(\frac{x+y}{2}\right)^2 + \left(\frac{x-y}{2}\right)^2 = \frac{x^2+y^2}{2} = \frac{m}{2}p\,,$$

im Widerspruch zur Minimalität von m.

Falls die Behauptung falsch wäre, müßte somit $m \geq 3$ sein. Wir definieren dann zwei neue Zahlen $u, v \in \mathbb{Z}$ durch die Bedingungen

$$|u| < \frac{m}{2}, \quad |v| < \frac{m}{2}, \quad u \equiv x \bmod m \quad \text{und} \quad v \equiv y \bmod m\,.$$

Offensichtlich können nicht beide dieser Zahlen verschwinden, denn sonst wären x und y beide durch m teilbar, also wäre $x^2 + y^2 = mp$ durch m^2 teilbar. Das kann aber nicht sein, denn p ist prim und $m < p$. Weiter ist

$$u^2 + v^2 \equiv x^2 + y^2 = mp \equiv 0 \bmod m\,,$$

also gibt es eine natürliche Zahl r, so daß $u^2 + v^2 = rm$ ist. Da $u^2 + v^2$ kleiner ist als $\frac{1}{2}m^2$, ist $r < \frac{m}{2}$.

Nach der zu Beginn des Paragraphen zitierten Formel von FIBONACCI, d.h. also durch explizite Berechnung von $(u+iv)(x+iy)$ und Berechnung der Norm davon, erhalten wir die Formel.

$$(rm)(mp) = (u^2 + v^2)(x^2 + y^2) = (xu + yv)^2 + (xv - yu)^2\,.$$

Dabei ist nach Definition von u und v

$$xu + yv \equiv x^2 + y^2 \equiv 0 \bmod m \text{ und } xv - yu \equiv xy - yx \equiv 0 \bmod m\,,$$

beide Zahlen sind also durch m teilbar. Somit gibt es natürliche Zahlen X, Y mit

$$(rm)(mp) = m^2rp = (mX)^2 + (mY)^2 \quad \text{oder} \quad rp = X^2 + Y^2\,.$$

Da $r < \frac{m}{2}$, widerspricht dies der Minimalität von m.

Damit haben wir gezeigt, daß $m = 1$ sein muß, d.h. p läßt sich als Summe zweier Quadrate darstellen. Wir müssen uns noch überlegen, daß diese Darstellung bis auf die Reihenfolge der Faktoren eindeutig ist.

Angenommen, es gibt zwei Darstellungen $p = x^2 + y^2 = u^2 + v^2$. In $\mathbb{Z}[i]$ ist dann

$$p = (x + iy)(x - iy) = (u + iv)(u - iv)\,.$$

Alle Faktoren haben Norm p und sind somit irreduzibel, und aus dem vorigen Kapitel §5 wissen wir, daß $\mathbb{Z}[i]$ ein EUKLIDischer, insbesondere also faktorieller Ring ist. Daher unterscheiden sich die beiden Zerlegungen nur durch Einheiten von $\mathbb{Z}[i]$. Auch diese kennen wir aus Kapitel 6:

Nach dem Lemma aus §6 sind es genau die Elemente ± 1 und $\pm i$. Somit ist entweder $x^2 = u^2$ und $y^2 = v^2$ oder umgekehrt, womit die Eindeutigkeit bewiesen wäre. ∎

Als erste Anwendung davon können wir die Primzahlen im Ring $\mathbb{Z}[i]$ der GAUSSschen Zahlen bestimmen:

Korollar: Eine Primzahl $p \in \mathbb{N}$ ist genau dann irreduzibel in $\mathbb{Z}[i]$, wenn $p \equiv 3 \bmod 4$. Andernfalls zerfällt sie in das Produkt zweier konjugiert komplexer irreduzibler Elemente $r \pm is$ mit $r^2 + s^2 = p$.

Beweis: $p = 2 = (1 + i)(1 - i)$ zerfällt offensichtlich, und dies ist bereits die Primzerlegung, denn $N(1 \pm i) = 2$ hat keine echten Teiler.

Falls eine ungerade Primzahl p einen echten Teiler $r + is$ hat, ist sie auch durch $r - is$ teilbar. Da die Norm von p gleich p^2 ist und $r \pm is$ keine Einheiten, muß $N(r \pm is) = p$ sein. Damit folgt zunächst, daß $r \pm is$ prim sind, denn ein echter Teiler müßte als Norm einen echten Teiler von p haben. Außerdem folgt, daß sich $(r + is)(r - is) = r^2 + s^2$ höchstens durch eine Einheit von p unterscheidet. Da beides positive Zahlen sind, muß diese gleich eins sein, d.h. die Primzerlegung von p in $\mathbb{Z}[i]$ ist

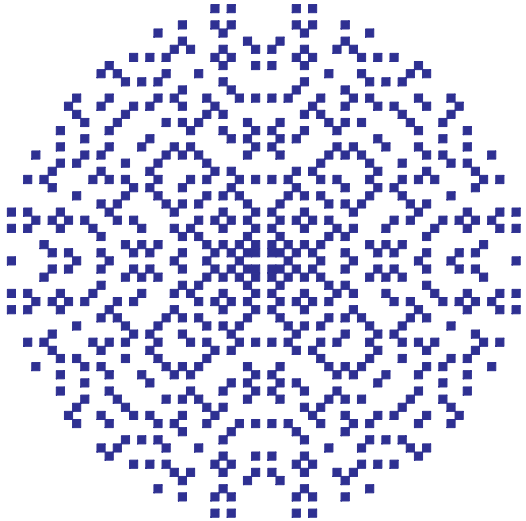
$$p = (r + is)(r - is) = r^2 + s^2\,.$$

Nach dem Satz ist daher $p \equiv 1 \bmod 4$.

Ist umgekehrt $p \equiv 1 \bmod 4$, so gibt es nach dem Satz zwei ganze Zahlen r, s, so daß $p = r^2 + s^2$ ist, d.h. $p = (r + is)(r - is)$ zerfällt in $\mathbb{Z}[i]$, und das Argument aus dem vorigen Abschnitt zeigt, daß dies die Primzerlegung ist.

Somit zerfallen genau die Primzahlen $p \equiv 1 \bmod 4$ und die Zwei, d.h. genau die $p \equiv 3 \bmod 4$ bleiben prim. ∎

In der Abbildung sind die GAUSSschen Primzahlen $a + ib$ der Norm höchstens 1000 durch Quadrate um den Punkt $(a, b) \in \mathbb{R}^2$ dargestellt. Mancher Leser wird hier ein gelegentlich von Designern verwendetes Muster erkennen.



Kehren wir zurück zur Ausgangsfrage: Wann kann eine vorgegebene natürliche Zahl als Summe zweier Quadrate dargestellt werden?

Satz: Eine natürliche Zahl n läßt sich genau dann als Summe zweier Quadrate schreiben, wenn jeder Primteiler $p \equiv 3 \bmod 4$ mit einer geraden Potenz in der Primzerlegung von n auftritt.

Beweis: Zunächst ist die Bedingung hinreichend, denn da mit n auch jedes Produkt $c^2 n$ als Summe zweier Quadrate darstellbar ist, können wir die Primteiler $p \equiv 3 \bmod 4$ ignorieren. Nach dem gerade bewiesenen Satz wissen wir, daß jede Primzahl $p \equiv 1 \bmod 4$ Summe zweier Quadrate ist, und natürlich gilt dies auch für $2 = 1^2 + 1^2$. Damit ist nach dem obigen Lemma auch jedes Produkt solcher Primzahlen als Summe zweier Quadrate darstellbar.

Umgekehrt sei $n = x^2 + y^2$ und $d = \mathrm{ggT}(x, y)$. Mit $x = du$, $y = dv$ und $n = d^2 m$ ist dann $m = u^2 + v^2$, und m enthält genau dann einen Primteiler $p \equiv 3 \bmod 4$ in ungerader Potenz, wenn dies für n der Fall ist. Ein solcher Primteiler p teilt auch $u^2 + v^2 = (u + iv)(u - iv)$ im Ring $\mathbb{Z}[i]$ der GAUSSschen Zahlen. Falls p auch dort eine Primzahl ist, muß es

mindestens einen der beiden Faktoren teilen; komplexe Konjugation zeigt, daß es dann auch den anderen teilt. Damit teilt es auch deren Summe $2u$ und Differenz $2iv$; da p ungerade ist und i eine Einheit, teilt p also die zueinander teilerfremden Zahlen u und v, ein Widerspruch.

Somit ist p in $\mathbb{Z}[i]$ keine Primzahl; nach obigem Korollar muß daher $p = 2$ oder $p \equiv 1 \bmod 4$ sein. Damit ist jeder Primteiler $p \equiv 3 \bmod 4$ von n zugleich ein Teiler von d und tritt in n daher mit einer geraden Potenz auf. ∎

Für zusammengesetzte Zahlen ist die Darstellung als Summe zweier Quadrate im allgemeinen nicht mehr eindeutig. Über die Primzerlegung in $\mathbb{Z}[i]$ läßt sich die Anzahl verschiedener Darstellungen leicht erkennen: Natürlich entsprechen auch für eine beliebige natürliche Zahl n die Darstellungen als Summe zweier Quadrate den Darstellungen von n als Norm eines Elements von $\mathbb{Z}[i]$, wobei assoziierte Elemente auf dieselbe Zerlegung führen.

Aus der Primzerlegung von n in $\mathbb{Z}$ können wir leicht auf die Primzerlegung in $\mathbb{Z}[i]$ schließen: Primzahlen kongruent drei modulo vier bleiben nach obigem Korollar auch in $\mathbb{Z}[i]$ irreduzibel, die kongruent eins modulo vier sind Produkte zweier konjugierter Elemente $x \pm iy$. Die beiden Faktoren sind nicht assoziiert, denn sonst wäre $|x| = |y|$ und $p = x^2 + y^2$ wäre gerade. Die Zwei schließlich ist Produkt der beiden irreduziblen Elemente $1 \pm i$, und die sind assoziiert zueinander, denn $(1 - i) \cdot i = 1 + i$.

Wir sortieren daher in der Primzerlegung von n nach den Kongruenzklassen modulo vier der Primfaktoren:

$$n = 2^e \prod_{j=1}^{r} p_j^{f_j} \prod_{k=1}^{s} q_k^{2g_k} \quad \text{mit} \quad p_j \equiv 1 \bmod 4, \quad q_k \equiv 3 \bmod 4 \,.$$

Für jedes p_j wählen wir ein $\pi_j \in \mathbb{Z}[i]$ derart, daß $\pi_j \cdot \overline{\pi}_j = p_j$ ist; dann ist n in $\mathbb{Z}[i]$ assoziiert zu

$$(1 + i)^{2e} \prod_{j=1}^{r} \pi_j^{f_j} \prod_{j=1}^{r} \overline{\pi}_j^{f_j} \prod_{k=1}^{s} q_k^{2g_k} \,.$$

Ein Element $\alpha \in \mathbb{Z}[i]$, für das $\mathrm{N}(\alpha) = n$ sein soll, hat daher bis auf eine Einheit die Form

$$\alpha = (1+i)^e \prod_{j=1}^{r} \pi_j^{h_j} \prod_{j=1}^{r} \overline{\pi}_j^{f_j - h_j} \prod_{k=1}^{s} q_k^{g_k} ,$$

mit $0 \le h_j \le f_j$. Die Anzahl verschiedener Möglichkeiten ist somit gleich dem Produkt der $(f_j + 1)$, wobei hier allerdings die Darstellungen $n = x^2 + y^2$ und $n = y^2 + x^2$ für $x \neq y$ als verschieden gezählt werden.

Die im Vergleich zur Größe von n meisten verschiedenen Darstellungen gibt es offenbar dann, wenn n ein Produkt verschiedener Primzahlen ist, die allesamt kongruent eins modulo vier sind. In diesem Fall ist die Anzahl der Darstellungen gleich zwei hoch Anzahl der Faktoren.

§2: Anwendung auf die Berechnung von π

Aus der Analysis I ist bekannt, daß gilt

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \frac{x^{11}}{11} + \frac{x^{13}}{13} - \frac{x^{15}}{15} + \cdots \ ;$$

falls es jemand nicht mehr weiß: Die Ableitung des Arkustangens ist $1/(1+x^2)$, und nach der Summenformel für die geometrische Reihe ist

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + x^8 - x^{10} + x^{12} - x^{14} + \cdots \ .$$

Durch gliedweise Integration folgt wegen $\arctan 0 = 0$ die obige Formel. Eine bekannte Anwendung davon ist der Spezialfall $x = 1$:

$$\frac{\pi}{4} = \arctan 1 = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \frac{1}{13} - \frac{1}{15} + \cdots \ .$$

Zur praktischen Berechnung von π ist diese Formel allerdings völlig unbrauchbar und der Alptraum eines jeden Numerikers: Zunächst einmal sind alternierende Summen grundsätzlich problematisch, allerdings ist das hier vergleichsweise harmlos: Wenn wir jeden negativen Summanden von seinem Vorgänger subtrahieren, bekommen wir eine Reihe

$$\frac{\pi}{4} = \frac{2}{1\cdot 3} + \frac{2}{5\cdot 7} + \frac{2}{9\cdot 11} + \frac{2}{13\cdot 15} + \cdots$$

mit lauter positiven Gliedern. Die Summanden sind jedoch immer noch monoton fallend, so daß die Rundungsfehler der ersten Additionen bei hinreichend langer Summation größer sind als die hinteren Summanden. Man muß also, wenn man eine endliche Teilsumme berechnen will, von hinten nach vorne summieren und damit bereits vor Beginn der Rechnung die Anzahl der Terme festlegen. Bei jeder Erhöhung der Anzahl der Summanden muß die gesamte Rechnung von vorne beginnen.

Dazu kommt, daß die Reihe extrem langsam konvergiert: Berechnet man für

$$\frac{\pi}{8} = \sum_{n=0}^{\infty} \frac{1}{(4n+1)(4n+3)}$$

die Teilsummen

$$S_N = \sum_{n=0}^{N} \frac{1}{(4n+1)(4n+3)} ,$$

so erhält man für die ersten Zehnerpotenzen N die folgenden Fehler:

N	10	100	1 000	10 000
$\pi - 8S_N$	$4{,}5 \cdot 10^{-2}$	$5{,}0 \cdot 10^{-3}$	$5{,}0 \cdot 10^{-4}$	$5{,}0 \cdot 10^{-5}$
N	100 000	1 000 000	10 000 000	100 000 000
$\pi - 8S_N$	$5{,}0 \cdot 10^{-6}$	$5,0 \cdot 10^{-7}$	$5{,}0 \cdot 10^{-8}$	$5{,}0 \cdot 10^{-9}$

Für eine zusätzliche Dezimalstelle muß also der Rechenaufwand ziemlich genau verzehnfacht werden. Angesichts der Tatsache, daß heute mehrere Billionen Ziffern von π bekannt sind ist klar, daß es bessere Wege zur Berechnung von π geben muß.

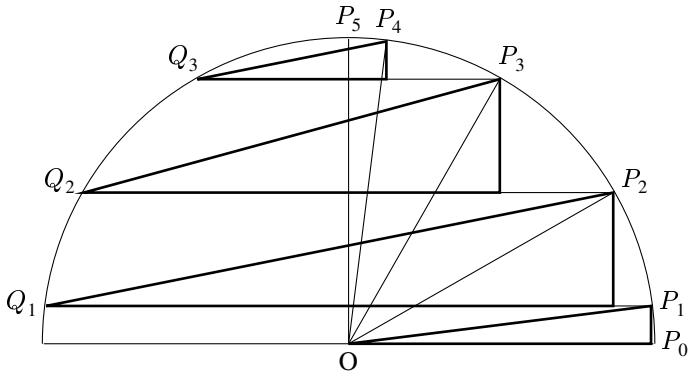
Einer davon benutzt Zahlen mit einer großen Anzahl verschiedener Darstellungen als Summen zweier Quadrate. Die Reihe für den Arkustangens konvergiert sicherlich umso besser, je kleiner der Wert von x ist. Wenn wir also den Winkel $\frac{\pi}{4}$ aufteilen können in mehrere kleine Winkel, deren Tangens wir kennen, sollten bessere Ergebnisse zu erwarten sein. Genau das können wir mit solchen Zahlen erreichen.

Angenommen, wir haben für eine Zahl n die r verschiedenen Darstellungen

$$n = x_1^2 + y_1^2 = \cdots = x_r^2 + y_r^2$$

als Summen von Quadraten, wobei $y_1 < \cdots < y_r$ sei. Dann ist $x_i = y_{r-i}$, denn wir können ja in jeder Darstellung die Reihenfolge der Faktoren vertauschen. Wir wollen außerdem voraussetzen, daß n nicht das Doppelte eines Quadrats ist, so daß stets $x_i \neq y_i$ und somit r eine gerade Zahl ist.

Die Punkte $P_i = (x_i, y_i)$ und $Q_i = (-x_i, y_i)$ für $i = 1, \ldots, r$ liegen auf der Kreislinie $x^2 + y^2 = n$ um den Nullpunkt O, genauso die drei Punkte $P_0 = \left(\sqrt{n}, 0\right)$, $Q_0 = \left(-\sqrt{n}, 0\right)$ und $P_{r+1} = \left(0, \sqrt{n}\right)$.



Da die y-Koordinaten y_i der P_i der Größe nach geordnet sind, ist

$$\frac{\pi}{2} = \sum_{i=0}^{r} \measuredangle OP_iP_{i+1} = 2 \sum_{i=0}^{r/2-1} \measuredangle OP_iP_{i+1} + \measuredangle OP_{r/2}P_{r/2+1} \,.$$

Leider ist keines der Dreiecke $\triangle OP_iP_{i+1}$ rechtwinklig, so daß uns die ganzzahligen Koordinaten der (meisten) P_i bei der Berechnung der Winkel $\measuredangle OP_iP_{i+1}$ nichts nützen.

Nun lehrt uns aber ein Satz der Elementargeometrie, der (im Anhang zu diesem Paragraphen bewiesene) Satz vom Innenwinkel, daß der Winkel $\measuredangle OP_iP_{i+1}$ doppelt so groß ist wie der Winkels $\measuredangle Q_iP_iP_{i+1}$. Letzterer gehört zu einem rechtwinkligen Dreieck, denn natürlich ändert sich nichts am Winkel, wenn wir den Punkt P_i ersetzen durch die senkrechte

Projektion $P_i' = (x_{i+1}, y_i)$ von P_{i+1} auf die Gerade Q_iP_i. Somit ist

$$\frac{\pi}{2} = 2\measuredangle OP_0'P_1 + 4 \sum_{i=1}^{r/2-1} \measuredangle Q_iP_i'P_{i+1} + 2\measuredangle Q_{r/2}P_{r/2}'P_{r/2+1} \,.$$

Division durch zwei macht daraus

$$\frac{\pi}{4} = \measuredangle OP_0'P_1 + 2 \sum_{i=1}^{r/2-1} \measuredangle Q_iP_i'P_{i+1} + \measuredangle Q_{r/2}P_{r/2}'P_{r/2+1} \,.$$

In dieser Darstellung sind die drei Punkte, die den Winkel bestimmen, in allen Fällen die Eckpunkte eines rechtwinkligen Dreiecks, sie haben allesamt ganzzahlige Koordinaten, und zumindest die Katheten der Dreiecke haben ganzzahlige Längen. Somit können wir alle auftretenden Winkel ausdrücken durch Arkustangenswerte rationaler Zahlen.

Als Beispiel betrachten wir das kleinste Produkt dreier verschiedener Primzahlen kongruent eins modulo vier, also $n = 5 \cdot 13 \cdot 17 = 1105$. Aus den Darstellungen

$$5 = 1^2 + 2^2, \quad 13 = 2^2 + 3^2 \quad \text{und} \quad 17 = 1^2 + 4^2$$

verschafft man sich leicht die vier Darstellungen

$$1105 = 4^2 + 33^2 = 9^2 + 32^2 = 12^2 + 31^2 = 23^2 + 24^2 \,,$$

zu denen natürlich auch noch vier mit vertauschten Faktoren kommen. Wir haben also

$$P_1 = (33, 4)\,, \;\; P_2 = (32, 9)\,, \;\; P_3 = (31, 12)\,, \;\; P_4 = (24, 23)\,,$$
$$P_8 = (4, 33)\,, \;\; P_7 = (9, 32)\,, \;\; P_6 = (12, 31)\,, \;\; P_5 = (23, 24)\,;$$

dazu kommen noch die beiden Randpunkte $P_0 = (\sqrt{1105}, 0)$ sowie $P_9 = (0, \sqrt{1105})$.

Die Q_i für $1 \leq i \leq 8$ unterscheiden sich von den P_i nur durch das Vorzeichen der Abszisse. Damit können wir die Tangenten aller Winkel bei O berechnen:

$$\tan \measuredangle OP_0P_1 = \tan \measuredangle OP_8P_9 = \frac{y_1}{x_1} = \frac{4}{33}$$

$$\tan\measuredangle OP_1P_2 = \tan\measuredangle OP_7P_8 = \tan 2\measuredangle Q_1P_1P_2 = \frac{y_2-y_1}{x_1+x_2} = \frac{5}{65} = \frac{1}{13}$$

$$\tan\measuredangle OP_2P_3 = \tan\measuredangle OP_6P_7 = \tan 2\measuredangle Q_2P_2P_3 = \frac{y_3-y_2}{x_2+x_3} = \frac{3}{63} = \frac{1}{21}$$

$$\tan\measuredangle OP_3P_4 = \tan\measuredangle OP_5P_6 = \tan 2\measuredangle Q_3P_3P_4 = \frac{y_4-y_3}{x_3+x_4} = \frac{11}{55} = \frac{1}{5}$$

$$\tan\measuredangle OP_4P_5 = \tan 2\measuredangle Q_4P_4P_5 = \frac{y_5-y_4}{x_4+x_5} = \frac{1}{47}$$

Die Summe aller dieser Winkel ist

$$\frac{\pi}{4} = \arctan\frac{4}{33} + 2\arctan\frac{1}{13} + 2\arctan\frac{1}{21} + 2\arctan\frac{1}{5} + \arctan\frac{1}{47}\,.$$

Approximieren wir dies, indem wir jede der TAYLOR-Reihen durch das TAYLOR-Polynom vom Grad n ersetzen, erhalten wir die folgenden betragsmäßigen Abweichungen Δ_n zwischen π und dem Vierfachen dieser Summe:

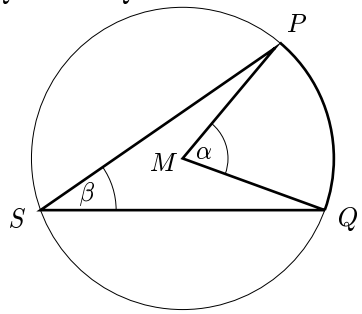
n	1	3	5	7	9
Δ_n	$2{,}5\cdot10^{-2}$	$5{,}2\cdot10^{-4}$	$1{,}4\cdot10^{-5}$	$4{,}4\cdot10^{-7}$	$1{,}4\cdot10^{-8}$
n	11	13	15	17	19
Δ_n	$5\cdot10^{-10}$	$4{,}9\cdot10^{-10}$	$6\cdot10^{-13}$	$2{,}1\cdot10^{-14}$	$7{,}7\cdot10^{-16}$

Die Verbesserung gegenüber der Berechnung via $\frac{\pi}{4} = \arctan 1$ ist in der Tat dramatisch: Die oben betrachtete Teilsumme S_N entspricht der Auswertung des TAYLOR-Polynoms vom Grad $n = 4N+3$, und selbst wenn wir N auf hundert Millionen setzen, haben wir noch einen Fehler von $5\cdot10^{-7}$. Mit dem neuen Ansatz kommen wir bereits mit TAYLOR-Polynome vom Grad neun auf einen Fehler, der gerade mal ein Zehntel davon beträgt. An Stelle von hundert Millionen Summanden mußten wir dazu nur fünf TAYLOR-Polynome mit jeweils fünf Summanden auswerten.

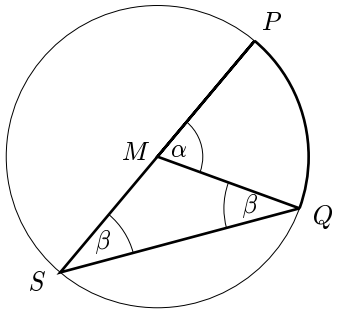
Anhang: Der Satz vom Innenwinkel

Da der Satz vom Innenwinkel in Deutschland anscheinend nicht zum Standardstoff im Geometrieunterricht der Schulen zählt, sei er hier noch einmal genauer formuliert und bewiesen:

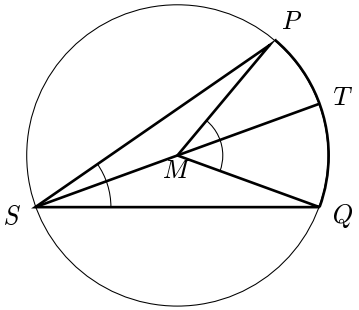
Satz: P, Q, S seien Punkte auf einer Kreislinie mit Mittelpunkt M. Dann ist $\angle MPQ = 2\angle SPQ$.



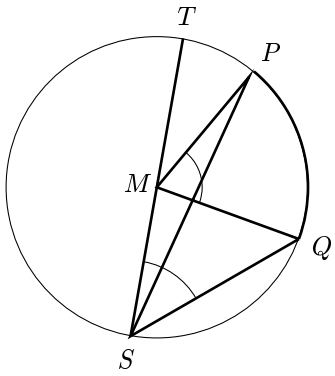
Beweis: Am einfachsten ist der Fall, daß M auf der Verbindungsstrecke von S mit einem der beiden Punkte P und Q liegt; wir nehmen an, er liege auf $\overline{SP}$. (Der andere Fall ist spiegelsymmetrisch dazu und geht genauso.) Dann ist das Dreieck $\triangle MSQ$ gleichschenklig, d.h. wir haben bei S und bei Q denselben Winkel β. Der verbleibende Dreieckswinkel bei M ist somit $\pi - 2\beta$. Andererseits ist dies aber der Komplementärwinkel zu $\alpha = \angle MPQ$, also ist $\alpha = 2\beta$, wie behauptet.



Der allgemeine Fall kann auf diesen Spezialfall zurückgeführt werden: Liegen P und Q auf verschiedenen Seiten des Durchmessers durch S, dessen anderer Endpunkt T sei, so erfüllen auch die Punkte S, P, T, M sowie die Punkte S, Q, T, M die Voraussetzung des Satzes, und in beiden Fällen sind wir in der Situation des bereits bewiesenen Spezialfalls. Addition der Ergebnisse für diese beiden Fälle liefert das Ergebnis für die Punkte S, P, Q, M.



Bleibt noch der Fall, daß P und A auf derselben Seite des Durchmessers $\overline{ST}$ liegen. Auch in diesem Fall erfüllen wieder sowohl die Punkte S, P, T, M als auch die Punkte S, Q, T, M die Voraussetzungen des Satzes, und beides Mal sind wir in der Situation des eingangs bewiesenen Spezialfalls. Dieses Mal führt die Subtraktion dieser beiden Ergebnisse zum gewünschten Resultat für die Ausgangssituation mit den Punkten S, P, Q, M.



Damit ist der Satz vollständig bewiesen. ∎

§3: Der Satz von Lagrange

Es ist nicht möglich, eine beliebige natürliche Zahl als Summe von höchstens drei Quadratzahlen zu schreiben; das kleinste Gegenbeispiel ist die Sieben. Wie EULER vermutete und LAGRANGE bewies, kommt man aber immer mit höchstens vier Quadratzahlen aus.

Einer der vielen Beweise dieses Satzes ist recht ähnlich zu dem des Zweiquadratesatzes aus §1; statt mit dem Ring $\mathbb{Z}[i]$ der GAUSSschen Zahlen arbeiten wir aber mit dem Ring

$$\mathbb{Z} \oplus \mathbb{Z}\mathbf{i} \oplus \mathbb{Z}\mathbf{j} \oplus \mathbb{Z}\mathbf{k}$$

der ganzen Quaternionen. Auch hier haben wir eine Normabbildung, und eine ganze Zahl n ist offensichtlich genau dann als Summe von vier Quadraten darstellbar, wenn sie Norm einer ganzen Quaternion ist. Wegen der Multiplikativität der Norm reicht es also wieder, wenn wir Primzahlen p betrachten.

Zur Vorbereitung zeigen wir zunächst

Lemma: Zu jeder Primzahl p gibt es ganze Zahlen $x, y, z \in \mathbb{Z}$ und eine natürliche Zahl $m < p$, so daß gilt: $mp = x^2 + y^2 + z^2$

Beweis: Für $p = 2$ ist $2 = 1^2 + 1^2 + 0^2$; sei also p ungerade.

Von den Zahlen a^2 mit $0 \le a \le \frac{1}{2}(p-1)$ sind keine zwei kongruent modulo p, denn $a^2 - b^2 = (a+b)(a-b)$, und falls $0 \le a, b < \frac{1}{2}(p-1)$ sind beide Faktoren kleiner als p. Damit gibt es auch in den Mengen

$$\mathcal{M}_1 = \left\{ -a^2 \;\middle|\; 0 \le a \le \tfrac{1}{2}(p-1) \right\}$$

und

$$\mathcal{M}_2 = \left\{ 1 + a^2 \;\middle|\; 0 \le a \le \tfrac{1}{2}(p-1) \right\}$$

keine zwei Elemente, die modulo p kongruent sind. Da die beiden Mengen disjunkt sind und jede davon $\frac{1}{2}(p+1)$ Elemente hat, enthält ihre Vereinigung $p+1$ Elemente; hier muß es also mindestens zwei Elemente geben, die modulo p kongruent sind. Es gibt also Zahlen $x, y \in \mathbb{Z}$ mit $-x^2 \equiv 1 + y^2 \bmod p$, d.h. $x^2 + y^2 + 1 == x^2 + y^2 + 1^2 = mp$ ist durch p teilbar. Da $x, y \le \frac{1}{2}(p-1)$, ist dabei $m < p$ und das Lemma ist bewiesen. ∎

Lemma: Jede Primzahl p läßt sich als Summe von höchstens vier Quadraten schreiben.

Beweis: Für $p = 2$ wissen wir das; sei also p wieder ungerade. Nach dem vorigen Lemma gibt es eine natürliche Zahl $m < p$ derart, daß mp als Summe von sogar höchstens drei Quadraten darstellbar ist; k sei die kleinste natürliche Zahl, für die kp als Summe von höchstens vier Quadraten darstellbar ist. Natürlich ist dann auch $k < p$.

Wäre k eine gerade Zahl, so wäre auch die Summe der vier Quadrate gerade, und dazu gibt es drei Möglichkeiten: Entweder alle Summanden sind gerade oder alle sind ungerade oder zwei davon sind gerade, der Rest ungerade. Im letzteren Fall wollen wir die vier Zahlen w, x, y, z so anordnen, daß w und x gerade sind, y und z dagegen ungerade. Dann sind in allen drei Fällen $w \pm x$ und $y \pm z$ gerade, und

$$\left(\frac{w+x}{2}\right)^2 + \left(\frac{w-x}{2}\right)^2 + \left(\frac{y+z}{2}\right)^2 + \left(\frac{y-z}{2}\right)^2 = \frac{w^2+x^2+y^2+z^2}{2} = \frac{k}{2}p\,,$$

im Widerspruch zur Minimalität von k. Also ist k ungerade, und falls das Lemma falsch wäre, müßte $k \ge 3$ sein.

Wir betrachten die modulo k zu w, x, y, z kongruenten ganzen Zahlen W, X, Y, Z vom Betrag kleiner $k/2$. Wie schon beim Zwei-Quadrate-Satz können diese nicht allesamt verschwinden, denn sonst wären w, x, y, z durch k teilbar, also ihre Quadratsumme kp durch k^2, was wegen $k < p$ für eine Primzahl p nicht möglich ist.

Somit ist $0 < W^2 + X^2 + Y^2 + Z^2 < 4 \cdot \left(\frac{k}{2}\right)^2 = k^2$. Andererseits ist aber

$$W^2 + X^2 + Y^2 + Z^2 \equiv w^2 + x^2 + y^2 + z^2 \equiv 0 \bmod k\,;$$

also ist

$$W^2 + X^2 + Y^2 + Z^2 = k\ell \quad \text{mit} \quad 1 \le \ell < k\,.$$

Damit haben die Quaternionen

$$q = w + ix + jy + kz \quad \text{und} \quad Q = W + iX + jY + kZ$$

die Normen $\mathrm{N}(q) = kp$ und $\mathrm{N}(Q) = k\ell$, ihre Produkt hat also die Norm $k^2\ell p$. Zumindest von der Norm her spricht also nichts dagegen, daß dieses Produkt durch k teilbar sein könnte.

Tatsächlich ist $q\overline{Q}$ durch k teilbar, und das sieht man am schnellsten durch brutales Nachrechnen: In

$$\begin{aligned} q\overline{Q} = {} & (wW + xX + yY + zZ) \\ & + (-wX + xW - yZ + zY)\mathbf{i} \\ & + (-wY + yW - zX + xZ)\mathbf{j} + (-wZ + zW - xY + yX)\mathbf{k} \end{aligned}$$

sind alle vier Klammern durch k teilbar, denn modulo k sind alle Großbuchstaben gleich den entsprechenden Kleinbuchstaben, so daß die Koeffizienten von $\mathbf{i}, \mathbf{j}, \mathbf{k}$ trivialerweise modulo k verschwinden, und für den Realteil haben wir

$$wW + xX + yY + zZ \equiv w^2 + x^2 + y^2 + z^2 = kp \equiv 0 \bmod k\,.$$

Somit ist

$$\frac{q\overline{Q}}{k} = A + B\mathbf{i} + C\mathbf{j} + D\mathbf{k}$$

eine Quaternion mit ganzzahligen Koeffizienten, und

$$A^2 + B^2 + C^2 + D^2 = \mathrm{N}\left(\frac{q\overline{Q}}{k}\right) = \frac{\mathrm{N}(q\overline{Q})}{k^2} = \frac{\mathrm{N}(q)\mathrm{N}(Q)}{k^2} = \ell p\,.$$

Dies widerspricht aber der Minimalität von k.

Somit muß $k = 1$ sein, und der Satz ist bewiesen. ∎

Satz (LAGRANGE)**:** Jede natürliche Zahl läßt sich als Summe von höchstens vier Quadraten schreiben.

Beweis: Wie wir in Kapitel 6, §7 gesehen haben, läßt sich eine Zahl n genau dann als Summe von höchstens vier Quadraten schreiben, wenn sie Norm einer ganzen Quaternion ist. Da wir gerade gesehen haben, daß sich jede Primzahl als Summe von höchstens vier Quadraten schreiben läßt (und die Eins natürlich auch), folgt die Behauptung aus der Multiplikativität der Norm. ∎

§4: Quadratische Formen und Matrizen

Nachdem wir in den vorigen Paragraphen gesehen haben, daß die spezielle quadratische Form $x^2 + y^2$ vielfältige Beziehungen sowohl zum Zahlkörper $\mathbb{Q}[i]$ als auch zu Anwendungen außerhalb der Zahlentheorie haben, wollen wir uns nun etwas mit der allgemeinen Theorie solcher Formen beschäftigen. In den nächsten Paragraphen werden wir sie dann auf quadratische Zahlkörper und die PELLsche Gleichung anwenden.

Viele abstrakte Aussagen über quadratische Formen werden einfacher, wenn wir sie in lineare Algebra übersetzen. In Matrixschreibweise ist

$$Ax^2 + Bxy + Cy^2 = \begin{pmatrix} x & y \end{pmatrix} Q \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{mit} \quad Q = \begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix}\,,$$

die quadratische Form kann also auch durch die symmetrische Matrix Q beschrieben werden.

Die Determinante von Q ist $AC - \frac{1}{4}B^2$; bis auf einen Faktor -4 ist das die Zahl $B^2 - 4AC$, die wir in Kapitel 6, §2 als Diskriminante eines Elements eines quadratisches Zahlkörpers definiert haben. Wir können also hoffen, daß uns die lineare Algebra via Determinantentheorie Aussagen über die Werte einer quadratischen Form sowie über Zusammenhänge zwischen den Diskriminanten verschiedener Elemente eines quadratischen Zahlkörpers gibt.

Die reellen Werte, die eine quadratische Form annehmen kann, hängen nicht davon ab, in welcher Basis des $\mathbb{R}^2$ wir das Argument $\binom{x}{y}$ darstellen; wir können die Basis daher bei Bedarf beliebig ändern.

Das ist zum Beispiel nützlich bei der Frage, wann eine quadratische Form nur positive oder nur negative Werte annimmt:

Definition: Eine symmetrische Matrix $Q \in \mathbb{R}^{2\times 2}$ und die dadurch definierte quadratische Form $f_Q(x,y) = (x\ y)Q\binom{x}{y}$ heißen $\left\{\begin{matrix}\text{positiv}\\ \text{negativ}\end{matrix}\right\}$ semidefinit, wenn $f_Q(x,y)\ \left\{\begin{matrix}\geq\\ \leq\end{matrix}\right\}\ 0$ für alle $x, y \in \mathbb{R}$. Sie heißen $\left\{\begin{matrix}\text{positiv}\\ \text{negativ}\end{matrix}\right\}$ definit, wenn zusätzlich $f(x,y)$ nur für $x = y = 0$ verschwindet.

Ein typisches Beispiel einer positiv definiten quadratischen Form ist x^2+y^2; hier ist Q einfach die Einheitsmatrix. Die negative Einheitsmatrix führt auf die negativ definite Form $-x^2 - y^2$, die Diagonalmatrix mit Einträgen +1 und -1 auf $x^2 - y^2$, was sowohl positive als auch negative Werte annehmen kann.

Beispiele von positiv semidefiniten, aber nicht positiv definiten Formen sind etwa $x^2, (x+y)^2$ oder $(3x-4y)^2$ mit zugehörigen Matrizen

$$Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{und} \quad Q_3 = \begin{pmatrix} 3 & -6 \\ -6 & 4 \end{pmatrix}.$$

Für ein allgemeines Kriterium, wann eine Matrix positiv oder negativ (semi-)definit ist, erinnern wir uns an einen Satz aus der linearen Algebra:

Satz: Alle Eigenwerte einer symmetrischen Matrix $Q \in \mathbb{R}^{n\times n}$ sind reell und ihre geometrische Vielfachheit stimmt mit der algebraischen überein. Eigenvektoren zu verschiedenen Eigenwerten stehen senkrecht aufeinander; insbesondere hat $\mathbb{R}^n$ eine Orthonormalbasis aus Eigenvektoren von Q.

Für Leser, die diesen sogenannten *Spektralsatz* nicht kennen, sei kurz ein Beweis für den Spezialfall $n = 2$ skizziert.

Die Matrix $Q = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ hat das charakteristische Polynom

$$\det(Q - \lambda E) = \begin{vmatrix} a-\lambda & b \\ b & c-\lambda \end{vmatrix} = (a-\lambda)(c-\lambda) - b^2$$

$$= \lambda^2 - (a+c)\lambda + ac - b^2 = \left(\lambda - \frac{a+c}{2}\right)^2 - \left(\frac{a+c}{2}\right)^2 + ac - b^2$$

$$= \left(\lambda - \frac{a+c}{2}\right)^2 - \left(\frac{a-c}{2}\right)^2 - b^2\,;$$

die Eigenwerte sind also $\lambda_{1/2} = \dfrac{a+c}{2} \pm \sqrt{\left(\dfrac{a-c}{2}\right)^2 + b^2}$.

Da die Summe zweier Quadrate reeller Zahlen nicht negativ sein kann, ist die Wurzel reell, und damit haben wir auch reelle Eigenwerte.

Für $n = 2$ gibt es genau dann einen Eigenwert mit algebraischer Vielfachheit ungleich eins, wenn die beiden Nullstellen des charakteristischen Polynoms gleich sind, wenn also der Ausdruck unter der Wurzel verschwindet, d.h. $a = c$ und $b = 0$. In diesem Fall ist $Q = \binom{a\ 0}{0\ a}$ eine Diagonalmatrix, der Eigenraum ist also der gesamte $\mathbb{R}^2$, so daß auch die geometrische Vielfachheit des Eigenwerts zwei ist.

Ist $\lambda_1 \neq \lambda_2$ und sind $\vec{a}_1, \vec{a}_2$ Eigenvektoren dazu, so ist

$$(Q\vec{a}_1)\cdot\vec{a}_2 = (\lambda_1\vec{a}_1)\cdot\vec{a}_2 = \lambda_1(\vec{a}_1\cdot\vec{a}_2)\,.$$

Andererseits ist (wie man nötigenfalls leicht nachrechnet) für je zwei Vektoren $\vec{v}, \vec{w}$ und eine Matrix Q das Skalarprodukt $(Q\vec{v})\cdot\vec{w}$ gleich dem Skalarprodukt $\vec{v}\cdot(Q^T\vec{w})$, wobei Q^T die zu Q transponierte Matrix bezeichnet. Für eine symmetrische Matrix ist $Q = Q^T$, also

$$(Q\vec{a}_1)\cdot\vec{a}_2 = \vec{a}_1\cdot(Q\vec{a}_2) = \vec{a}_1\cdot(\lambda_2\vec{a}_2) = \lambda_2(\vec{a}_1\cdot\vec{a}_2)\,.$$

Da $\lambda_1 \neq \lambda_2$, können $\lambda_1(\vec{a}_1\cdot\vec{a}_2)$ und $\lambda_2(\vec{a}_1\cdot\vec{a}_2)$ nur dann gleich sein, wenn $\vec{a}_1\cdot\vec{a}_2$ verschwindet, d.h. $\vec{a}_1$ und $\vec{a}_2$ stehen senkrecht aufeinander. ∎

Auch ein weiteres allgemeines Resultat aus der Linearen Algebra läßt sich hier einfach und direkt beweisen: Das Produkt aller Eigenwerte einer $n \times n$-Matrix ist gleich der Determinante und die Summe gleich der Spur, der Summe der Diagonalelemente also.

Für symmetrische 2×2-Matrizen folgt dies sofort aus den obigen Formeln für die beiden Eigenwerte: Bei der Addition fällt die Wurzel weg,

so daß wir Summe $a + c$ erhalten, und das Produkt ist nach der dritten binomischen Formel gleich

$$\left(\frac{a+c}{2}\right)^2 - \left[\left(\frac{a-c}{2}\right)^2 + b^2\right] = ac - b^2 = \det Q\,.$$

Ist $\vec{a}_1$ ein Eigenvektor zu λ_1 und $\vec{a}_2$ einer zu λ_2, so können wir jeden Vektor $\binom{x}{y}$ aus $\mathbb{R}^2$ als Linearkombination $\binom{x}{y} = u\vec{a}_1 + v\vec{a}_2$ schreiben. Der Zeilenvektor $(x\ y)$ ist dann die entsprechende Linearkombination $u\vec{a}_1^T + v\vec{a}_2^T$ der zu den $\vec{a}_i$ gehörigen Zeilenvektoren $\vec{a}_i^T$, und

$$(x\ y)Q\begin{pmatrix}x\\y\end{pmatrix} = (u\vec{a}_1^T + v\vec{a}_2^T)Q(u\vec{a}_1 + v\vec{a}_2)$$

$$= (u\vec{a}_1^T + v\vec{a}_2^T)\Big(uQ\vec{a}_1 + vQ\vec{a}_2\Big) = (u\vec{a}_1^T + v\vec{a}_2^T)\Big(u\lambda_1\vec{a}_1 + v\lambda_2\vec{a}_2\Big)$$

$$= \lambda_1 u^2\Big(\vec{a}_1^T\vec{a}_1\Big) + \lambda_2 uv\Big(\vec{a}_1^T\vec{a}_2\Big) + \lambda_1 uv\Big(\vec{a}_2^T\vec{a}_1\Big) + \lambda_2 v^2\Big(\vec{a}_2^T\vec{a}_2\Big)\,.$$

Das Matrixprodukt $\vec{a}_i^T\vec{a}_j$ ist gleich dem üblichen Skalarprodukt $\vec{a}_i \cdot \vec{a}_j$; da die a_i aufeinander senkrecht stehen, verschwindet es für $i \neq j$, d.h.

$$(x\ y)Q\begin{pmatrix}x\\y\end{pmatrix} = \lambda_1 u^2\vec{a}_1 \cdot \vec{a}_1 + \lambda_2 v^2\vec{a}_2 \cdot \vec{a}_2\,,$$

wobei die Skalarprodukte der $\vec{a}_i$ mit sich selbst natürlich positiv sind. Falls wir $\vec{a}_1$ und $\vec{a}_2$ als Einheitsvektoren wählen, sind sie eins und können ganz weggelassen werden.

Damit ist klar, daß die quadratische Form genau dann nur nichtnegative Werte annimmt, wenn λ_1 und λ_2 beide positiv sind; genau dann, wenn beide negativ sind, nimmt sie nur nichtpositive Werte an. Die Matrix Q und die zugehörige quadratische Form sind also genau dann positiv *bzw.* negativ semidefinit, wenn beide Eigenwerte ≥ 0 *bzw.* ≤ 0 sind. Sie sind positiv *bzw.* negativ definit, wenn beide echt positiv *bzw.* negativ sind.

Für eine positiv oder negativ semidefinite Matrix muß daher die Determinante ≥ 0 sein; bei einer definiten Matrix muß sie positiv sein. Ob sie positiv oder negativ definit ist, sagt uns dann die Spur, denn da die beiden Eigenwerte (falls $\neq 0$) dasselbe Vorzeichen haben, ist dieses auch das Vorzeichen ihrer Summe, der Spur. Wenn die Determinante $ac - b^2$ positiv ist, müssen a und c dasselbe Vorzeichen haben; da auch $a + c$ gleich der Spur der Matrix ist, folgt

Lemma: *a)* Eine symmetrische 2×2-Matrix ist genau dann positiv oder negativ definit, wenn ihre Determinante positiv ist. Sie ist positiv definit, wenn der Eintrag links oben positiv ist, andernfalls ist sie negativ definit.
b) Die quadratische Form $Ax^2 + Bxy + Cy^2$ ist genau dann definit, wenn ihre Diskriminante $B^2 - 4AC$ negativ ist. Im Falle $A > 0$ ist sie dann positiv, sonst negativ definit.

∎

So nützlich der Wechsel zu einer Basis aus Eigenvektoren in diesem Fall auch war, für die meisten zahlentheoretischen Fragen werden uns nur solche Basiswechsel helfen, die ganzzahlige Punkte wieder in ganzzahlige Punkte überführen. Hier gilt

Lemma: Die lineare Abbildung

$$\varphi\colon\begin{cases}\mathbb{R}^2 \to \mathbb{R}^2\\ \begin{pmatrix}x\\y\end{pmatrix} \mapsto M\begin{pmatrix}x\\y\end{pmatrix}\end{cases}$$

definiert genau dann eine Bijektion $\mathbb{Z}^2 \to \mathbb{Z}^2$, wenn alle Einträge der Matrix A ganzzahlig sind und $\det M = \pm 1$ ist.

Beweis: Da die Spaltenvektoren von M die Bilder der Basisvektoren $\binom{1}{0}$ und $\binom{0}{1}$ sind, ist klar, daß $\varphi(\mathbb{Z}^2)$ genau dann in $\mathbb{Z}^2$ liegt, wenn alle Einträge von M ganzzahlig sind. Soll $\varphi(\mathbb{Z}^2) = \mathbb{Z}^2$ sein, muß zusätzlich $\varphi^{-1}(\mathbb{Z}^2) \subseteq \mathbb{Z}^2$ sein, d.h. auch M^{-1} darf nur ganzzahlige Einträge haben. In diesem Fall sind $\det M$ und $\det M^{-1}$ beide ganzzahlig mit Produkt eins, also ist $\det M = \pm 1$.

Hat umgekehrt eine Matrix M mit ganzzahligen Einträgen Determinante ± 1, so hat auch M^{-1} ganzzahlige Einträge, denn die Spaltenvektoren von M^{-1} sind die Lösungen der linearen Gleichungssysteme $M\binom{x}{y} = \binom{1}{0}$ und $M\binom{x}{y} = \binom{0}{1}$, die wir nach der CRAMERschen Regel ausdrücken können durch Brüche mit ganzzahligen Zählern und $\det M$ im Nenner.

∎

Setzen wir für so eine Matrix M das Bild $M\binom{x}{y}$ an Stelle von $\binom{x}{y}$ in die quadratische Form ein, erhalten wir das Ergebnis

$$\big(M\binom{x}{y}\big)^T \cdot Q \cdot M\binom{x}{y} = (x\ \ y)\big(M^TQM\big)\binom{x}{y},$$

das wir auch erhalten hätten, wenn wir $\binom{x}{y}$ in die quadratische Form zur Matrix M^TQM eingesetzt hätten. Da $\binom{x}{y} \mapsto M\binom{x}{y}$ eine Bijektion von $\mathbb{Z}^2$ nach $\mathbb{Z}^2$ definiert, nehmen die quadratischen Formen zu Q und zu M^TQM also dieselben Werte an. Deshalb definieren wir

Definition: Die quadratischen Formen mit Matrizen Q_1 und Q_2 heißen *äquivalent,* wenn es eine Matrix M mit ganzzahligen Einträgen und $\det M = \pm 1$ gibt, so daß $Q_2 = M^TQ_1M$.

Lemma: Zwei äquivalente quadratische Formen haben dieselbe Diskriminante.

Beweis: Bis auf den Faktor -4 ist die Diskriminante gleich der Determinante der Matrix und $\det Q_2 = \det M^T \cdot \det Q_1 \cdot \det M = \det Q_1$, da $\det M = \det M^T = \pm 1$ ist. ∎

§5: Kettenbruchentwicklung quadratischer Irrationalitäten

Die rationalen Zahlen sind genau diejenigen reellen Zahlen, deren Kettenbruchentwicklung nach endlich vielen Schritten abbricht. Wir wollen sehen, daß wir auch quadratische Irrationalitäten, d.h. Elemente eines quadratischen Zahlkörpers, die nicht in $\mathbb{Q}$ liegen, durch ihre Kettenbruchentwicklung charakterisieren können.

In den Beispielen der Kettenbruchentwicklungen von $\sqrt{2}$ und $\sqrt{3}$ kamen wir in Kapitel 5 auf periodische Folgen. Wie sich zeigen wird, ist dies charakteristisch für quadratische Irrationalitäten.

Nach der Formel am Ende von §2 von Kapitel 5 gilt für die Zahlen α_n aus dem Algorithmus zur Kettenbruchentwicklung die Gleichung

$$\alpha = \frac{\alpha_n p_{n-2} + p_{n-1}}{\alpha_n q_{n-2} + q_{n-1}},$$

wobei p_n und q_n Zähler und Nenner der n-ten Konvergente sind. Mit

$$M = \begin{pmatrix} p_{n-2} & p_{n-1} \\ q_{n-2} & q_{n-1} \end{pmatrix} \quad \text{ist} \quad M\begin{pmatrix} \alpha_n \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha_n p_{n-2} + p_{n-1} \\ \alpha_n q_{n-2} + q_{n-1} \end{pmatrix},$$

die Vektoren $\binom{\alpha}{1}$ und $M\binom{\alpha_n}{1}$ sind also proportional zueinander.

Als quadratische Irrationalität genügt α einer quadratischen Gleichung $A\alpha^2 + B\alpha + C = 0$; der Vektor $\binom{\alpha}{1}$ wird also von der quadratischen Form $Ax^2 + Bxy + Cy^2$ annulliert. Da mit einem Vektor $\binom{x}{y}$ auch alle seine Vielfachen von dieser Form annulliert werden, gilt dasselbe für den zu $\binom{\alpha}{1}$ proportionalen Vektor $M\binom{\alpha_n}{1}$.

Die Matrix zur quadratischen Form $Ax^2 + Bxy + Cy^2$ sei Q. Wie wir aus dem vorigen Paragraphen wissen, erfüllen die Komponenten von $M\binom{\alpha_n}{1}$ dann die quadratische Gleichung zur Form mit Matrix M^TQM. Nach Kapitel 5, §2 ist $\det M = p_{n-2}q_{n-1} - q_{n-2}p_{n-1} = (-1)^{n-1}$; die neue quadratische Form ist also zu der mit Q äquivalent und hat insbesondere dieselbe Diskriminante. Also haben alle α_n dieselbe Diskriminante wie α, denn da die Multiplikation mit M einen Isomorphismus $\mathbb{Z}^2 \to \mathbb{Z}^2$ definiert, sind die Einträge von Q genau dann ganzzahlig und teilerfremd, wenn es die von M^TQM sind.

Dies ist ein wesentlicher Schritt für den Beweis des folgenden Satzes:

Satz (LAGRANGE ∼1766): Die Kettenbruchentwicklung einer irrationalen Zahl α wird genau dann periodisch, wenn α eine quadratische Irrationalzahl ist.

Beweis: Angenommen, α hat eine periodische Kettenbruchentwicklung. Dann gibt es ein n und ein $k > 0$, so daß $\alpha_{n+k} = \alpha_n$ ist. Nach der Formel am Ende von §2 von Kapitel 5 ist daher

$$\alpha = \frac{\alpha_n p_{n-2} + p_{n-1}}{\alpha_n q_{n-2} + q_{n-1}} = \frac{\alpha_{n+k} p_{n+k-2} + p_{n+k-1}}{\alpha_{n+k} q_{n+k-2} + q_{n+k-1}} = \frac{\alpha_n p_{n+k-2} + p_{n+k-1}}{\alpha_n q_{n+k-2} + q_{n+k-1}}.$$

Daraus folgt die Gleichheit von $(\alpha_n p_{n-2} + p_{n-1})(\alpha_n q_{n+k-2} + q_{n+k-1})$ und $(\alpha_n q_{n-2} + q_{n-1})(\alpha_n p_{n+k-2} + p_{n+k-1})$, und ausmultipliziert wird dies zu einer quadratischen Gleichung für α_n. Der Koeffizient von α_n^2 ist $p_{n-2}q_{n+k-2} + q_{n-2}p_{n+k-2}$, was als Summe positiver Zahlen nicht

null sein kann; wir haben also eine echte quadratische Gleichung. Somit läßt sich α_n in der Form $\alpha = r + s\sqrt{D}$ schreiben, und damit auch

$$\alpha = \frac{\alpha_n p_{n-2} + p_{n-1}}{\alpha_n q_{n-2} + q_{n-1}} .$$

Umgekehrt sei $\alpha = r + s\sqrt{D}$ mit quadratfreiem D eine quadratische Irrationalität, die der Gleichung $A_0\alpha^2 + B_0\alpha + C_0 = 0$ genüge. Dann zeigt die Konstruktionsvorschrift für die α_n, daß auch diese Zahlen sowie ihre Inversen in entsprechender Form geschrieben werden können und damit Gleichungen der Form

$$A_n\alpha_n^2 + B_n\alpha_n + C_n = 0$$

genügen. Um diese Gleichung zu bestimmen, betrachten wir die Funktion $f(x) = A_0x^2 + B_0x + C_0$, die für $x = \alpha$ verschwindet, setzen

$$\alpha = \frac{\alpha_n p_{n-2} + p_{n-1}}{\alpha_n q_{n-2} + q_{n-1}}$$

ein und multiplizieren mit dem Hauptnenner. Zumindest für die Koeffizienten A_n und C_n ergeben sich einigermaßen erträgliche Formeln:

$$A_n = A_0p_{n-1}^2 + B_0p_{n-1}q_{n-1} + C_0q_{n-1}^2 = q_{n-1}^2 f\left(\frac{p_{n-1}}{q_{n-1}}\right)$$

und

$$C_n = A_0p_{n-2}^2 + B_0p_{n-2}q_{n-2} + C_0q_{n-2}^2 = q_{n-2}^2 f\left(\frac{p_{n-2}}{q_{n-2}}\right) .$$

Da f eine quadratische Funktion ist, führt die TAYLOR-Entwicklung um α auf die Formel

$$f\left(\frac{p_{n-1}}{q_{n-1}}\right) = f(\alpha) + f'(\alpha)\left(\frac{p_{n-1}}{q_{n-1}} - \alpha\right) + \frac{f''(\alpha)}{2}\left(\frac{p_{n-1}}{q_{n-1}} - \alpha\right)^2 .$$

Hierbei ist $f(\alpha) = 0$, und $\left|\alpha - \frac{p_{n-1}}{q_{n-1}}\right| < 1/q_{n-1}^2 \le 1$. Somit ist

$$|A_n| \le |f'(\alpha)| + |f''(\alpha)| .$$

Genauso zeigt man die Ungleichung $|C_n| \le |f'(\alpha)| + |f''(\alpha)|$. Somit sind die Beträge der Koeffizienten A_n und C_n beschränkt durch eine von n unabhängige Konstante.

Wie wir oben gesehen haben, hat α_n dieselbe Diskriminante wie α; die Diskriminante $\Delta = B_n^2 - 4A_nC_n$ hängt also nicht ab von n. Daher folgen aus der obigen Schranken für A_n und C_n auch Schranken für $B_n^2 = \Delta + 4A_nC_n$, so daß auch der Betrag von B_n beschränkt ist.

Somit gibt es nur endlich viele Tripel (A_n, B_n, C_n), also auch nur endlich viele verschiedene Werte für α_n. Es muß daher zwei Zahlen n, k mit $k \ge 1$ geben derart, daß $\alpha_n = \alpha_{n+k}$ ist, und die Kettenbruchentwicklung wird spätestens ab der n-ten Stelle periodisch. ∎

Der gerade bewiesene Satz charakterisiert Zahlen, deren Kettenbruchentwicklung periodisch *wird;* er besagt nicht, daß die Kettenbruchentwicklung einer quadratischen Irrationalität von Anfang an periodisch ist, und in der Tat kennen wir Beispiele wie $\sqrt{2} = [1, \overline{2}]$ oder $\sqrt{3} = [1, \overline{1, 2}]$, bei denen das nicht der Fall ist. Für eine rein periodische Kettenbruchentwicklung brauchen wir also noch zusätzliche Bedingungen:

Satz: Die Kettenbruchentwicklung einer quadratischen Irrationalzahl α ist genau dann rein periodisch, wenn $\alpha > 1$ ist und ihr konjugiertes Element $\overline{\alpha}$ zwischen -1 und 0 liegt.

Beweis: Sei zunächst $\alpha > 1$ und $-1 < \overline{\alpha} < 0$. Der Trick zum Beweis der reinen Periodizität der Folge der c_i besteht darin, die $c_i = [1/\alpha_i]$ durch die konjugierten Elemente $\overline{\alpha}_{i+1}$ auszudrücken und so von der Gleichheit zweier Koeffizienten auch auf die ihrer Vorgänger zu schließen.

Die Gleichung $\alpha = c_0 + \alpha_1$ wird, da c_0 eine rationale Zahl ist, durch Konjugation zu $\overline{\alpha} = c_0 + \overline{\alpha}_1$. Da $\overline{\alpha}$ nach Voraussetzung zwischen -1 und 0 liegt, ist somit $0 < -\overline{\alpha}_1 - c_0 < 1$ und $c_0 = [-\overline{\alpha}_1]$. Wegen $c_0 = [\alpha] \ge 1$ folgt außerdem $-1 < \dfrac{1}{\overline{\alpha}_1} < 0$.

Wir wollen induktiv zeigen, daß auch für alle $i > 0$ gilt

$$c_i = [-\overline{\alpha}_{i+1}] \quad \text{und} \quad -1 < \frac{1}{\overline{\alpha}_{i+1}} < 0 .$$

Dazu nehmen wir an, dies gelte für $i - 1$. Aus

$$\frac{1}{\alpha_i} = c_i + \alpha_{i+1} \quad \text{und} \quad -1 < \frac{1}{\overline{\alpha}_i} < 0$$

folgt wie im Fall $i = 0$, daß $c_i = [-\overline{\alpha}_{i+1}]$ ist, und da die Koeffizienten c_i für $i > 0$ bei jeder Kettenbruchentwicklung mindestens gleich eins sind, folgt auch die Ungleichung für $1/\overline{\alpha}_{i+1}$ genau wie dort.

Daraus folgt nun leicht die Periodizität der Kettenbruchentwicklung von α: Wir wissen bereits, daß sie periodisch *wird;* es gibt also irgendeinen Index $m \geq 0$ und eine Periode k, so daß $\alpha_{n+k} = \alpha_n$ für alle $n \geq m$. Wir betrachten das minimale m mit dieser Eigenschaft. Die Kettenbruchentwicklung von α ist genau dann rein periodisch, wenn $m = 0$ ist. Für $m \geq 1$ können wir aber aus $\alpha_{m+k} = \alpha_m$ und $c_{m+k} = c_m$ folgern, daß auch $c_{m+k-1} = [-\overline{\alpha}_{m+k}] = [-\overline{\alpha}_m] = c_{m-1}$ ist. Aus den Gleichungen

$$\frac{1}{\alpha_{m+k-1}} = c_{m+k-1} + \alpha_{m+k} \quad \text{und} \quad \frac{1}{\alpha_{m-1}} = c_{m-1} + \alpha_m$$

folgt dann aber, daß auch $\alpha_{m-1+k} = \alpha_{m-1}$ ist, im Widerspruch zur Minimalität von m. Somit ist $m = 0$, die Kettenbruchentwicklung von α also rein periodisch.

Umgekehrt habe α eine rein periodische Kettenbruchentwicklung der Periode k mit Koeffizienten $c_0, c_1, \ldots$. Wegen $c_k = c_0$ ist dabei auch c_0 positiv, denn alle c_n mit $n > 0$ müssen ja positiv sein. Somit ist insbesondere $\alpha > 1$.

Um zu sehen, daß $\overline{\alpha}$ zwischen -1 und 0 liegt, beachten wir, daß $\overline{\alpha}$ dieselbe quadratische Gleichung erfüllt wie α. Da diese Gleichung genau zwei Nullstellen hat und α größer als eins ist, genügt es, wenn wir zeigen, daß diese Gleichung im Intervall $(-1,\ 0)$ eine Nullstelle hat. Das wiederum folgt aus dem Zwischenwertsatz, wenn wir zeigen können, daß die quadratische Funktion auf der linken Seite an den Stellen 0 und -1 Werte mit entgegengesetzten Vorzeichen annimmt.

Für $k = 1$ ist $\alpha = c_0 + \alpha_1 = c_0 + \frac{1}{\alpha} \Longrightarrow \alpha^2 - c_0\alpha - 1 = 0$. Die quadratische Funktion $x^2 - c_0 x - 1$ nimmt an der Stelle 0 den Wert -1 an, und bei $x = 1$ den Wert $c_0 > 0$; somit gibt es eine Nullstelle zwischen diesen beiden Punkten.

Für $k \geq 2$ verwenden wir die bereits im vorigen Satz benutzte Formel aus Kapitel 5, §2, und beachten, daß $\alpha_k = 1/\alpha$ ist. Dies führt auf die

Gleichung

$$\alpha = \frac{\alpha_k p_{k-2} + p_{k-1}}{\alpha_k q_{k-2} + q_{k-1}} = \frac{p_{k-2} + p_{k-1}\alpha}{q_{k-2} + q_{k-1}\alpha} .$$

Überkreuzmultiplikation macht daraus die quadratische Gleichung

$$q_{k-1}\alpha^2 + (q_{k-2} - p_{k-2})\alpha - p_{k-1} = 0 .$$

Hier nimmt die quadratische Funktion bei 0 den Wert $-p_{k-1} < 0$ an, und an der Stelle -1 den Wert

$$q_{k-1} - (q_{k-2} - p_{k-2}) - p_{k-1} = (q_{k-1} - q_{k-2}) + (p_{k-2} - p_{k-1}) .$$

Dieser ist positiv, da sowohl die Folge der Zähler als auch die der Nenner der Konvergenten von α monoton steigt. ∎

§6: Die Pellsche Gleichung

Im letzten Kapitel hatten wir gesehen, daß eine Einheit $x + y\sqrt{D}$ von $\mathcal{O}_D$ die Gleichung $x^2 - Dy^2 = \pm 1$ erfüllen muß. Hauptziel dieses Paragraphen ist die Lösung der PELLschen Gleichung $x^2 - Dy^2 = 1$ für $(x, y) \in \mathbb{Z}^2$ oder – da es auf das Vorzeichen von x und y nicht ankommt – $(x, y) \in \mathbb{N}^2$.

Faktorisierung der linken Seite der PELLschen Gleichung führt auf

$$(x + y\sqrt{D})(x - y\sqrt{D}) = 1 ,$$

und damit ist

$$x - y\sqrt{D} = \frac{1}{x + y\sqrt{D}} \Longrightarrow \frac{x}{y} - \sqrt{D} = \frac{1}{y^2\left(\frac{x}{y} + \sqrt{D}\right)} .$$

Wegen der Positivität der rechten Seite ist $\frac{x}{y} > \sqrt{D}$, also folgt

$$\left|x - y\sqrt{D}\right| = x - y\sqrt{D} = \frac{1}{y^2\left(\frac{x}{y} + \sqrt{D}\right)} < \frac{1}{2y^2\sqrt{D}} < \frac{1}{2y^2} .$$

Nach dem Satz aus Kapitel 5, §3 muß $\frac{x}{y}$ somit eine Konvergente der Kettenbruchentwicklung von $\sqrt{D}$ sein.

Umgekehrt liefert aber nicht jede Konvergente der Kettenbruchentwicklung von $\sqrt{D}$ eine Lösung der PELLschen Gleichung: Beispielsweise hat $\sqrt{13} = [3, 1, 1, 1, 1, 6, \dots]$ die Brüche

$$\frac{4}{1}, \quad \frac{7}{2}, \quad \frac{11}{3}, \quad \frac{18}{5}, \quad \frac{119}{33}$$

als seine ersten Konvergenten, aber

$$4^2 - 13 = 3, \quad 7^2 - 13 \cdot 2^2 = -3, \quad 11^2 - 13 \cdot 3^2 = 4$$
$$18^2 - 13 \cdot 5^2 = -1 \quad \text{und} \quad 119^2 - 13 \cdot 33^2 = 4\,.$$

Zumindest *a priori* ist nicht klar, ob es überhaupt eine Konvergente gibt, die auf eine Lösung der PELLschen Gleichung führt.

Um hier mehr zu erfahren, müssen wir uns die Kettenbruchentwicklung von $\sqrt{D}$ genauer ansehen. Dabei sei D im folgenden stets eine quadratfreie natürliche Zahl.

Das konjugierte Element zu $\sqrt{D}$ ist $-\sqrt{D}$ und somit kleiner als -1; die Kettenbruchentwicklung von $\sqrt{D}$ ist also nicht rein periodisch. Betrachten wir aber $\alpha = \lfloor\sqrt{D}\rfloor + \sqrt{D}$, so ist natürlich $\alpha > 1$. und $\overline{\alpha} = \lfloor\sqrt{D}\rfloor - \sqrt{D}$ liegt zwischen -1 und 0. Somit hat α eine rein periodische Kettenbruchentwicklung. Die Periode sei k und die Koeffizienten seien $c_0, c_1, \dots$.

Die Kettenbruchentwicklung von $\sqrt{D} = \alpha - \lfloor\sqrt{D}\rfloor$ unterscheidet sich von der von α nur im ganzzahligen Anteil. Dieser ist im Falle von α gleich $2\lfloor\sqrt{D}\rfloor$, im Falle von $\sqrt{D}$ nur $\lfloor\sqrt{D}\rfloor$. Danach folgen in beiden Fällen die c_i mit $i \geq 1$. Wegen $c_k = c_0 = 2\lfloor\sqrt{D}\rfloor$ gilt daher

Satz: Ist D eine quadratfreie natürliche Zahl, so ist die Folge $c_0, c_1, \dots$ der Koeffizienten der Kettenbruchentwicklung von $\sqrt{D}$ ab c_1 periodisch. Bezeichnet k die Periode, so ist $c_k = 2c_0 = 2\lfloor\sqrt{D}\rfloor$. ∎

Bezeichnet p_n/q_n wieder die n-te Konvergente dieser Kettenbruchentwicklung, so ist nach der schon oft benutzten Formel

$$\sqrt{D} = \frac{\alpha_n p_{n-2} + p_{n-1}}{\alpha_n q_{n-2} + q_{n-1}}\,.$$

Ist speziell $n = rk$ ein Vielfaches einer Periode, hat $1/\alpha_n$ eine Kettenbruchentwicklung mit Koeffizienten $c_{rk}, c_{rk+1}, c_{rk+2}, \dots$; nach dem gerade bewiesenen Satz stimmt das überein mit der Folge $2c_0, c_1, c_2, \dots$, d.h.

$$\frac{1}{\alpha_{rk}} = c_0 + \sqrt{D} = \lfloor\sqrt{D}\rfloor + \sqrt{D}\,.$$

Einsetzen in die obige Formel führt auf

$$\sqrt{D} = \frac{\alpha_{rk} p_{rk-2} + p_{rk-1}}{\alpha_{rk} q_{rk-2} + q_{rk-1}} = \frac{p_{rk-2} + p_{rk-1}\left(c_0 + \sqrt{D}\right)}{q_{rk-2} + q_{rk-1}\left(c_0 + \sqrt{D}\right)}$$

oder

$$(q_{rk-2} + q_{rk-1}c_0)\sqrt{D} + q_{rk-1}D = (p_{rk-2} + p_{rk-1}c_0) + p_{rk-1}\sqrt{D}\,.$$

Durch Koeffizientenvergleich folgt:

$$p_{rk-2} = q_{rk-1}D - p_{rk-1}c_0 \quad \text{und} \quad q_{rk-2} = p_{rk-1} - q_{rk-1}c_0\,.$$

Setzen wir dies ein in die aus Kapitel 5, §2, bekannte Formel

$$p_m q_{m-1} - q_m p_{m-1} = (-1)^{m-1}$$

mit $m = rk - 1$, erhalten wir die Gleichung

$$p_{rk-1}^2 - p_{rk-1}q_{rk-1}c_0 - q_{rk-1}^2 D + q_{rk-1}p_{rk-1}c_0$$
$$= p_{rk-1}^2 - Dq_{rk-1}^2 = (-1)^{rk-2}\,.$$

Im Falle einer geraden Periode k ist somit (p_{kr-1}, q_{kr-1}) für jedes $r \in \mathbb{N}$ eine Lösung der PELLschen Gleichung ; für ungerade Perioden liefern nur die geradzahligen Vielfachen von k Lösungen, während die ungeradzahligen zu Lösungen der Gleichung $x^2 - Dy^2 = -1$ führen.

Im Eingangsbeispiel $D = 13$ zeigt eine genauere Rechnung, daß sich die Koeffizienten $1, 1, 1, 1, 6$ periodisch wiederholen, wir haben also die ungerade Periode fünf. Damit liefern die vierte, vierzehnte, vierundzwanzigste Konvergente der Kettenbruchentwicklung Lösungen der Gleichung $x^2 - Dy^2 = -1$, was wir für die vierte bereits nachgerechnet haben. Lösungen der PELLschen Gleichung liefern die neunte, neunzehnte *usw.* Konvergente. Die neunte Konvergente ist

$$[3, 1, 1, 1, 1, 6, 1, 1, 1] = \frac{649}{180}\,,$$

und in der Tat ist

$$649^2 - 13 \cdot 180^2 = 421\,201 - 13 \cdot 32\,400 = 421\,201 - 421\,200 = 1\,.$$

Allgemein haben wir gezeigt, daß die PELLsche Gleichung für jedes quadratfreie D eine Lösung hat; zusammen mit dem Satz aus Kapitel 6, §6 folgt, daß die Einheitengruppe eines jeden reellquadratischen Zahlkörpers unendlich ist und daß es speziell für die Gruppe der Einheiten mit Norm eins (der sogenannten Einseinheiten) ein Element $\alpha \in \mathcal{O}_D$ gibt, so daß jede Einseinheit in der Form $\pm\alpha^r$ mit einem $r \in \mathbb{Z}$ geschrieben werden kann. α ist die kleinste Einseinheit größer eins.

Natürlich kann auch α in der Form $p_n + q_n\sqrt{D}$ geschrieben werden, wobei p_n/q_n eine Konvergente der Kettenbruchentwicklung von $\sqrt{D}$ ist. Da Zähler und Nenner der Konvergenten strikt monoton ansteigen mit n, handelt es sich hier um die *erste* Konvergente p_n/q_n, für die $p_n^2 - Dq_n^2 = 1$ ist.

Mit Rechnungen, die sehr ähnlich zu den obigen sind, kann man zeigen, daß die oben gefundenen Indizes m mit $p_m^2 - Dq_m^2 = \pm 1$ tatsächlich die einzigen sind mit dieser Eigenschaft. Da wir schon viel mit Kettenbrüchen gerechnet haben und es noch viele andere interessante Teilgebiete der Zahlentheorie zu entdecken gilt, möchte ich auf diese Rechnungen verzichten.

Wer sich für diese Rechnungen interessiert, findet sie zum Beispiel in

WINFRIED SCHARLAU, HANS OPOLKA: Von Fermat bis Minkowski – Eine Vorlesung über Zahlentheorie und ihre Entwicklung, *Springer,* 1980

im Kapitel über LAGRANGE im (nur im Inhaltsverzeichnis benannten) Paragraphen *Lösung der Fermatschen (Pellschen) Gleichung* ab Seite 64. Es gibt zwar Rückverweise, aber wer den obigen Beweis verstanden hat, muß nur einem wirklich folgen. Zu beachten sind die unterschiedlichen Bezeichnungen: Was hier α heißt, ist dort θ, aber das dortige θ_n ist hier $1/\alpha_n$. Die hiesigen c_n werden dort mit a_n bezeichnet.

Wenn wir dieses Ergebnis akzeptieren, können wir die Einheitengruppe eines jeden reellquadratischen Zahlkörpers $\mathbb{Q}[\sqrt{D}]$ explizit berechnen, zumindest für $D \not\equiv 1 \bmod 4$: Dann ist $\mathcal{O}_D = \mathbb{Z} \oplus \mathbb{Z}\sqrt{D}$, so daß die Einheiten genau den ganzzahligen Lösungen der beiden Gleichungen

$x^2 - Dy^2 = \pm 1$ entsprechen. Ist k die Periode der Kettenbruchentwicklung von $\sqrt{D}$ und p/q die $(k-1)$-te Konvergente, so ist $\alpha = p + q\sqrt{D}$ die Grundeinheit, und jede andere Einheit läßt sich als $\pm\alpha^r$ mit einem $r \in \mathbb{Z}$ schreiben. Für gerades k sind dies alles Einseinheiten, für ungerades k bekommen wir für gerade r Einseinheiten und sonst Einheiten der Norm -1.

Bleibt die Frage, für welche D die Periode k gerade *bzw.* ungerade ist. Diese Frage muß nicht nur in dieser Vorlesung unbeantwortet bleiben: Es handelt sich hier um eines der vielen zahlentheoretischen Probleme, die trotz jahrhundertelanger Bemühungen auch heute noch offen sind.

Die zweite Frage ist: Was passiert für $D \equiv 1 \bmod 4$? Wie wir wissen, sind dann auch die Zahlen $\frac{1}{2}(x + y\sqrt{D})$ für ungerade ganze Zahlen x, y ganz, es kann also auch Einheiten dieser Form geben. In der Tat haben wir beim Eingangsbeispiel $D = 13$ bereits solche Fälle kennengelernt: Für die dritte Konvergente $\frac{11}{4}$ ist $11^2 - 13 \cdot 3^2 = 4$, d.h. $\mathrm{N}\left(\frac{1}{2}(11 + 3\sqrt{13}\right) = 1$. Wie eine genauere Untersuchung zeigt, ist dies genau dann möglich, wenn $D \equiv 5 \bmod 8$, jedoch nicht für alle solche D. Wenn es eine Grundeinheit dieser Form gibt, liegt ihre dritte Potenz in $\mathbb{Z} \oplus \mathbb{Z}\sqrt{D}$, der Kettenbruchalgorithmus gibt dann also nur die dritte Potenz der Grundeinheit. Einzelheiten findet man beispielsweise in §16, 5D des Buchs

HELMUT HASSE: Vorlesungen über Zahlentheorie, *Springer,* ^{2}1964.

Kapitel 8
Quadratische Reste

§1: Das Legendre-Symbol

Definition: Für eine Primzahl p und eine nicht durch p teilbare natürliche Zahl a ist das LEGENDRE-Symbol

$$\left(\frac{a}{p}\right) = \begin{cases} +1 & \text{falls es ein } x \in \mathbb{N} \text{ gibt mit } x^2 \equiv a \bmod p \\ -1 & \text{sonst} \end{cases}$$

Im ersten Fall bezeichnen wir a als *quadratischen Rest* modulo p, andernfalls als quadratischen Nichtrest. Für eine durch p teilbare Zahl a setzen wir $\left(\frac{a}{p}\right) = 0$.

Sind a, b zwei modulo p kongruente natürliche Zahlen, so ist offensichtlich $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$. Wir haben daher auch für $a \in \mathbb{F}_p^\times$ ein wohldefiniertes LEGENDRE-Symbol $\left(\frac{a}{p}\right)$, das durch die Vorschrift $\left(\frac{0}{p}\right) = 0$ auf ganz $\mathbb{F}_p$ fortgesetzt wird.

ADRIEN-MARIE LEGENDRE (1752–1833) wurde in Toulouse oder Paris geboren; jedenfalls ging er in Paris zur Schule und studierte Mathematik und Physik am dortigen Collège Mazarin. Ab 1775 lehrte er an der Ecole Militaire und gewann einen Preis der Berliner Akademie für eine Arbeit über die Bahn von Kanonenkugeln. Andere Arbeiten befaßten sich mit der Anziehung von Ellipsoiden und der Himmelsmechanik. Ab etwa 1785 publizierte er auch Arbeiten über Zahlentheorie, in denen er z.B. das quadratische Reziprozitätsgesetz bewies sowie die Irrationalität von π und π^2.

Lemma: Das LEGENDRE-Symbol definiert einen Gruppenhomomorphismus

$$\left(\frac{\cdot}{p}\right) : \begin{cases} \mathbb{F}_p^\times \to \{+1, -1\} \\ a \mapsto \left(\frac{a}{p}\right) \end{cases} .$$

Für $p = 2$ ist dies der triviale Homomorphismus, für ungerade p ist er surjektiv. Insbesondere gibt es dann jeweils $\frac{p-1}{2}$ quadratische Reste und Nichtreste.

Beweis: Für $p = 2$ ist $\mathbb{F}_2^\times = \{1\}$, und $1 = 1^2$ ist ein quadratischer Rest.

Sei nun p ungerade. Der Homomorphismus

$$\begin{cases} \mathbb{F}_p^\times \to \mathbb{F}_p^\times \\ x \mapsto x^2 \end{cases}$$

hat den Kern $\{+1, -1\}$, also besteht das Bild aus $\frac{p-1}{2}$ Elementen, den quadratischen Resten.

Trivialerweise ist das Produkt zweier quadratischer Reste wieder ein quadratischer Rest. Ist $a = x^2$ ein quadratischer Rest und b ein Nichtrest, so ist auch ab ein quadratischer Nichtrest, denn wäre $ab = y^2$, wäre $b = (yx^{-1})^2$ ein quadratischer Rest. Da Multiplikation mit b injektiv ist, folgt, daß sich jeder quadratische Nichtrest in der Form bc darstellen läßt, wobei c ein quadratischer Rest ist. Damit folgt, daß das Produkt zweier quadratischer Nichtreste ein quadratischer Rest ist, denn $bc \cdot bd = b^2cd$, wobei c und d Quadrate in $\mathbb{F}_p^\times$ sind. ∎

Lemma (EULER): Ist p eine ungerade Primzahl und kein Teiler von a, so ist $\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \bmod p$.

Beweis: g sei ein erzeugendes Element von $\mathbb{F}_p^\times$. Dann ist offensichtlich jede Potenz g^r mit geradem r ein quadratischer Rest, und da es genau $\frac{p-1}{2}$ verschiedene solcher Potenzen gibt, sind das auch *alle* quadratischen Reste. Somit ist g^r genau dann ein quadratischer Rest, wenn r gerade ist.

Da g ein erzeugendes Element ist, kann $g^{(p-1)/2}$ nicht gleich eins sein; da nach dem kleinen Satz von FERMAT aber sein Quadrat $g^{p-1} = 1$ ist, folgt $g^{(p-1)/2} = -1$. Für $a = g^r$ ist somit

$$a^{\frac{p-1}{2}} = (g^r)^{\frac{p-1}{2}} = \left(g^{\frac{p-1}{2}}\right)^r = (-1)^r$$

genau dann gleich eins, wenn a ein quadratischer Rest ist, und -1 sonst. ∎

Korollar: Für ungerades p ist

$$\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}} = \begin{cases} +1 & \text{falls } p \equiv 1 \bmod 4 \\ -1 & \text{falls } p \equiv 3 \bmod 4 \end{cases}.$$

∎

§2: Das quadratische Reziprozitätsgesetz

Quadratisches Reziprozitätsgesetz: Für zwei verschiedene ungerade Primzahlen p, q ist

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{(p-1)(q-1)}{4}} \quad \text{und} \quad \left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}.$$

Zum *Beweis* betrachten wir ein zum Nullpunkt symmetrisches Vertretersystem von $\mathbb{F}_p^\times$ in $\mathbb{Z}$, nämlich

$$R = \{-h, \ldots, -1, 1, \ldots, h\} \quad \text{mit} \quad h = \frac{p-1}{2}.$$

Weiter sei $S = \{q, 2q, \ldots, hq\}$. Da p und q teilerfremd sind, haben zwei verschiedene Elemente von S verschiedene Restklassen modulo p.

1. Schritt (GAUSS): q sei eine beliebige Primzahl und $p \neq q$ eine ungerade Primzahl. Dann ist $\left(\frac{q}{p}\right) = (-1)^m$, wobei m die Anzahl jener Elemente von S bezeichnet, die modulo p kongruent sind zu einem negativen Element von R.

Beweis: $a_1, \ldots, a_m$ seien die negativen Elemente von R, die zu Elementen aus S kongruent sind, $b_1, \ldots, b_n$ die positiven. Dann ist

$$a_1 \cdots a_m b_1 \cdots b_n \equiv \prod_{i=1}^{h} (iq) = h!\, q^h \bmod p.$$

Natürlich sind a_i und a_j für $i \neq j$ zwei verschiedene Zahlen, genauso auch b_i und b_j. Außerdem kann auch nie $|a_i| = |b_j|$ sein, denn sonst wäre einerseits $a_i + b_j = 0$, andererseits gäbe es aber Zahlen $1 \leq k, \ell \leq h$, so daß $a_i \equiv kq$ und $b_j \equiv \ell q \bmod p$. Also wäre $(k + \ell)q$ durch p teilbar, was nicht möglich ist, denn $k + \ell \leq 2h = p - 1$. Damit sind die Beträge der a_i und der b_j genau die Zahlen von 1 bis h, d.h.

$$a_1 \cdots a_m b_1 \cdots b_n = (-1)^m h!\,.$$

Vergleich mit der obigen Kongruenz zeigt, daß dann $q^h \equiv (-1)^m \bmod p$ ist, also nach dem vorigen Lemma $\left(\frac{q}{p}\right) = (-1)^m$. ∎

2. Schritt (GAUSS): Für zwei ungerade Primzahlen $p \neq q$ ist

$$\left(\frac{q}{p}\right) = (-1)^M \quad \text{mit} \quad M = \sum_{i=1}^{h} \left[\frac{iq}{p}\right] \quad \text{und} \quad \left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}.$$

Im *Beweis* sei zunächst auch noch der Fall $q = 2$ zugelassen. Für $i \leq h$ sei $r_i = iq - p \cdot \left[\frac{iq}{p}\right]$; dann ist $0 \leq r_i < p$ und $iq = p \cdot \left[\frac{iq}{p}\right] + r_i$. Falls iq modulo p kongruent ist zu einem negativen Element $a_j \in R$, ist also $r_i = p + a_j$; falls $r_i \equiv b_j > 0$ ist dagegen $r_i = b_j$. Somit ist

$$\sum_{i=1}^{h} iq = p \sum_{i=1}^{h} \left[\frac{iq}{p}\right] + \sum_{i=1}^{m} (a_i + p) + \sum_{i=1}^{n} b_i = pM + mp + \sum_{i=1}^{m} a_i + \sum_{i=1}^{n} b_i.$$

Andererseits ist

$$\sum_{i=1}^{h} iq = \frac{h(h+1)}{2} \cdot q = \frac{1}{2} \cdot \frac{p-1}{2} \cdot \frac{p+1}{2} \cdot q = \frac{p^2-1}{8} \cdot q.$$

Außerdem wissen wir aus dem ersten Schritt, daß

$$\{-a_1, \ldots, -a_m, b_1, \ldots, b_n\} = \{1, \ldots, h\}$$

ist, d.h.

$$-\sum_{i=1}^{m} a_i + \sum_{i=1}^{n} b_i = \sum_{i=1}^{h} i = \frac{h(h+1)}{2} = \frac{p^2-1}{8}$$

und damit ist $\sum_{i=1}^{n} b_i = \frac{p^2-1}{8} + \sum_{i=1}^{m} a_i$. Setzen wir das alles in die obige Formel ein, erhalten wir die Beziehung

$$\frac{p^2-1}{8} \cdot q = (M+m)p + \frac{p^2-1}{8} + 2\sum_{i=1}^{m} a_i$$

oder

$$\frac{p^2-1}{8} \cdot (q-1) = (M+m)p + 2\sum_{i=1}^{m} a_i \,.$$

Im Falle einer ungeraden Primzahl q steht rechts eine gerade Zahl; damit muß auch $M+m$ gerade sein, d.h. $(-1)^M = (-1)^m$, und die Behauptung folgt aus dem ersten Schritt.

Für $q = 2$ ist $M = 0$, da $\left[\frac{2i}{p}\right]$ für alle $i \le h$ verschwindet. Modulo zwei wird die obige Beziehung daher zu

$$\frac{p^2-1}{8} \equiv mp \equiv m \bmod 2\,,$$

so daß die Behauptung auch hier aus dem ersten Schritt folgt. ∎

3. Schritt (EISENSTEIN): p und q seien ungerade Primzahlen,

$$h = \frac{p-1}{2}, \quad k = \frac{q-1}{2}, \quad M = \sum_{i=1}^{h} \left[\frac{iq}{p}\right] \quad \text{und} \quad N = \sum_{i=1}^{k} \left[\frac{ip}{q}\right] .$$

Dann ist $M + N = hk$.

Beweis: Im Innern des Rechtecks mit Ecken $(0,0), (\frac{p}{2}, 0), (0, \frac{q}{2})$ und $(\frac{p}{2}, \frac{q}{2})$ liegen hk Gitterpunkte, nämlich die Punkte (i, j) mit $1 \le i \le h$ und $1 \le j \le k$.

Die Diagonale des Rechtecks liegt auf der Geraden $y = \frac{q}{p}x$ und enthält keine Gitterpunkte. Unterhalb der Diagonalen liegen $\left[\frac{iq}{p}\right]$ Punkte mit Abszisse i, insgesamt also M Punkte. Darüber liegen $\left[\frac{ip}{q}\right]$ Punkte mit Ordinate i, insgesamt also N Punkte. Somit ist $hk = M + N$. ∎

Zum *Beweis* des quadratischen Reziprozitätsgesetzes müssen wir nun nur noch alles kombinieren: Nach dem zweiten und dem dritten Schritt ist

$$\left(\frac{q}{p}\right)\left(\frac{p}{q}\right) = (-1)^M \cdot (-1)^N = (-1)^{M+N} = (-1)^{hk} = (-1)^{\frac{p-1}{2}\frac{q-1}{2}} .$$

∎

CARL FRIEDRICH GAUSS (1777–1855) leistete wesentliche Beiträge zur Zahlentheorie, zur nichteuklidischen Geometrie, zur Funktionentheorie, zur Differentialgeometrie und Kartographie, zur Fehlerrechnung und Statistik, zur Astronomie und Geophysik *usw.* Als Direktor der Göttinger Sternwarte baute er zusammen mit dem Physiker Weber den ersten Telegraphen. Er leitete die erste Vermessung und Kartierung des Königreichs Hannover und zeitweise auch den Witwenfond der Universität Göttingen; seine hierbei gewonnene Erfahrung benutzte er für erfolgreiche Spekulationen mit Aktien. Seine 1801 veröffentlichten *Disquisitiones arithmeticae* sind auch noch heute fundamental für die Zahlentheorie.

FERDINAND GOTTHOLD MAX EISENSTEIN (1823–1852), genannt Gotthold, wurde in Berlin geboren. Als einziges seiner sechs Geschwister starb er nicht bereits während der Kindheit an Meningitis. Im Alter von 17 Jahren, noch als Schüler, besuchte er Mathematikvorlesungen der Universität, unter anderem bei DIRICHLET. Ab 1842 las er die *Disquisitiones arithmeticae* von GAUSS, den er 1844 in Göttingen besuchte. Trotz zahlreicher wichtiger Arbeiten erhielt er nie eine gut bezahlte Position und überlebte vor allem dank der Unterstützung durch ALEXANDER VON HUMBOLDT. 1847 habilitierte er sich in Berlin und hatte dort unter anderem RIEMANN als Studenten. Er starb 29-jährig an Tuberkulose.

Bemerkung: Die rechten Seiten der Gleichungen im quadratischen Reziprozitätsgesetz lassen sich auch durch Kongruenzbedingungen ausdrücken: $(p-1)/2$ ist genau dann gerade, wenn $p \equiv 1 \bmod 4$, entsprechend für q. Somit ist

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = \begin{cases} +1 & \text{falls } p \equiv 1 \bmod 4 \text{ oder } q \equiv 1 \bmod 4 \\ -1 & \text{falls } p \equiv q \equiv 3 \bmod 4 \end{cases} .$$

Ist $p = 8r + k$, so ist $p^2 = 64r^2 + 16r + k^2 \equiv k^2 \bmod 16$, also ist $p^2 - 1 \equiv k^2 - 1 \bmod 16$. Für $k = \pm 1$ ist dies null, für $k = \pm 3$ acht. Somit ist

$$\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}} = \begin{cases} +1 & \text{falls } p \equiv \pm 1 \bmod 8 \\ -1 & \text{falls } p \equiv \pm 3 \bmod 8 \end{cases}.$$

Das quadratische Reziprozitätsgesetz läßt sich gelegentlich dazu verwenden, um ein LEGENDRE-Symbol einfach zu berechnen. Wenn wir beispielsweise entscheiden wollen, ob sieben ein quadratischer Rest modulo 17 ist, sagt es uns (da $17 \equiv 1 \bmod 4$), daß $\left(\frac{7}{17}\right) = \left(\frac{17}{7}\right)$ ist. Letzteres ist gleich $\left(\frac{3}{7}\right)$, da $17 \equiv 3 \bmod 7$. Hier haben wir zwei Primzahlen, die beide kongruent drei modulo vier sind, also ist $\left(\frac{3}{7}\right) = -\left(\frac{7}{3}\right) = -\left(\frac{1}{3}\right) = -1$, denn die Eins ist natürlich modulo jeder Primzahl ein quadratischer Rest. Also ist sieben modulo 17 ein quadratischer Nichtrest.

Genauso können wir auch leicht feststellen, ob 13 quadratischer Rest modulo 1 000 003 ist: Da $13 \equiv 1 \bmod 4$, ist $\left(\frac{13}{1\,000\,003}\right) = \left(\frac{1\,000\,003}{13}\right)$. Da $1\,000\,003 \equiv 4 \bmod 13$, ist dies gleich $\left(\frac{4}{13}\right)$, und das ist natürlich eins, da $4 = 2^2$ modulo jeder Primzahl ein Quadrat ist. Somit ist auch 13 ein Quadrat modulo 1 000 003.

Das Problem bei dieser Vorgehensweise besteht darin, daß wir normalerweise nicht soviel Glück haben wie hier und als Reduktionen stets Primzahlen erhalten. Wir sollten daher ein quadratisches Reziprozitätsgesetz haben, das auch funktioniert, wenn die beteiligten Zahlen nicht prim sind.

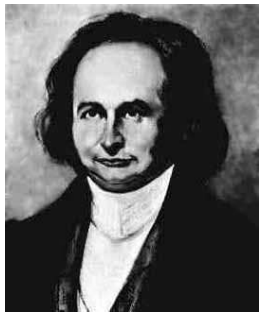
§3: Das Jacobi-Symbol

Wie wir in §1 gesehen haben, definiert das LEGENDRE-Symbol in Bezug auf seinen „Zähler“ einen Homomorphismus; wir können versuchen, es zu erweitern, indem wir dasselbe auch für den „Nenner“ postulieren:

Definition: Ist $n = \prod_{i=1}^{r} p_i^{e_i}$ eine ungerade Zahl und m eine zu n teilerfremde Zahl, so ist das JACOBI-Symbol definiert als

$$\left(\frac{m}{n}\right) = \prod_{i=1}^{r} \left(\frac{m}{p_i}\right)^{e_i}.$$

Falls m und n nicht teilerfremd sind, setzen wir $\left(\frac{m}{n}\right) = 0$.



CARL GUSTAV JACOB JACOBI (1804–1851) wurde in Potsdam als Sohn eines jüdischen Bankiers geboren und erhielt den Vornamen Jacques Simon. Im Alter von zwölf Jahren bestand er sein Abitur, mußte aber noch vier Jahre in der Abschlußklasse des Gymnasiums bleiben, da die Berliner Universität nur Studenten mit mindestens 16 Jahren aufnahm. 1824 beendete er seine Studien mit dem Staatsexamen für Mathematik, Griechisch und Latein und wurde Lehrer. Außerdem promovierte er 1825 und begann mit seiner Habilitation. Etwa gleichzeitig konvertierte er zum Christentum, so daß er ab 1825 an der Universität Berlin und ab 1826 in Königsberg lehren konnte. 1832 wurde er dort Professor. Zehn Jahre später mußte er aus gesundheitlichen Gründen das rauhe Klima Königsbergs verlassen und lebte zunächst in Italien, danach für den Rest seines Lebens in Berlin. Er ist vor allem berühmt durch seine Arbeiten zur Zahlentheorie und über elliptische Integrale.

Für eine Primzahl n und ein nicht dadurch teilbares m stimmt das JACOBI-Symbol natürlich mit dem LEGENDRE-Symbol überein, und man kann sich fragen, ob man hier wirklich einen neuen Namen braucht. Dieser ist gerechtfertigt, weil es einen ganz wesentlichen Unterschied zwischen den beiden Symbolen gibt: Beispielsweise ist

$$\left(\frac{2}{15}\right) = \left(\frac{2}{3}\right)\left(\frac{2}{5}\right) = (-1)^{\frac{3^2-1}{8}} \cdot (-1)^{\frac{5^2-1}{8}} = (-1)\cdot(-1) = 1\,,$$

aber zwei ist offensichtlich kein quadratischer Rest modulo 15: Sonst müßte es schließlich erst recht quadratischer Rest modulo drei und modulo fünf sein, aber die entsprechenden LEGENDRE-Symbole sind -1. In der Tat gibt es modulo 15 nur vier quadratische Reste: $1, 4, 6$ und 10.

Das JACOBI-Symbol gibt daher keine Auskunft darüber, ob eine Zahl quadratischer Rest ist oder nicht; lediglich wenn es gleich -1 ist, können wir sicher sein, daß wir es mit einem quadratischen Nichtrest zu tun haben, denn dann muß ja auch schon für mindestens einen Primteiler

des „Nenners“ das LEGENDRE-Symbol gleich -1 sein, während ein quadratischer Rest modulo einer Zahl n erst recht quadratischer Rest modulo eines jeden Teilers von n sein muß.

Die Nützlichkeit des JACOBI-Symbols kommt in erster Linie daher, daß auch dafür das quadratische Reziprozitätsgesetz gilt und es somit zur Berechnung von LEGENDRE-Symbolen verwendet werden kann:

Satz: Für zwei ungerade Zahlen m, n mit ggT$(m, n) = 1$ ist

$$\left(\frac{n}{m}\right)\left(\frac{m}{n}\right) = (-1)^{\frac{n-1}{2}\frac{m-1}{2}} \quad \text{und} \quad \left(\frac{2}{m}\right) = (-1)^{\frac{m^2-1}{8}} .$$

Beweis: Sei $n = \prod_{i=1}^{r} p_i^{e_i}$ und $m = \prod_{j=1}^{s} q_j^{f_j}$. Nach Definition des JACOBI-Symbols und weil das LEGENDRE-Symbol bei festgehaltenem „Nenner“ einen Homomorphismus definiert, ist dann

$$\left(\frac{n}{m}\right) = \prod_{j=1}^{s}\left(\frac{n}{q_j}\right)^{f_j} = \prod_{j=1}^{s}\prod_{i=1}^{r}\left(\frac{p_i}{q_j}\right)^{e_i f_j} \text{ und } \left(\frac{m}{n}\right) = \prod_{j=1}^{s}\prod_{i=1}^{r}\left(\frac{q_j}{p_i}\right)^{e_i f_j}.$$

Nach dem quadratischen Reziprozitätsgesetz aus §2 ist daher

$$\left(\frac{n}{m}\right)\left(\frac{m}{n}\right) = \prod_{i=1}^{r}\prod_{j=1}^{s}\left((-1)^{\frac{p_i-1}{2}\frac{q_j-1}{2}}\right)^{e_i f_j} = (-1)^{\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{p_i-1}{2}\frac{q_j-1}{2}e_i f_j}$$

$$= (-1)^{\left(\sum_{i=1}^{r}\frac{p_i-1}{2}e_i\right)\left(\sum_{j=1}^{s}\frac{q_j-1}{2}f_j\right)} = \left((-1)^{\sum_{i=1}^{r}\frac{p_i-1}{2}e_i}\right)^{\sum_{j=1}^{s}\frac{q_j-1}{2}f_j} .$$

Dies ist genau dann gleich +1, wenn mindestens einer der beiden Exponenten gerade ist; andernfalls ist es gleich -1.

Im Produkt

$$(-1)^{\sum_{i=1}^{r}\frac{p_i-1}{2}e_i} = \prod_{i=1}^{r}(-1)^{\frac{p_i-1}{2}e_i}$$

können wir alle Faktoren weglassen, für die e_i gerade ist oder aber $p_i \equiv 1 \bmod 4$. Das Produkt ist also gleich $(-1)^N$ mit

$$N = \text{Anzahl der Indizes } i \text{ mit } p_i \equiv 3 \bmod 4 \text{ und } e_i \text{ ungerade}.$$

Die Faktoren $p_i^{e_i}$ sind genau dann kongruent eins modulo vier, wenn $p_i \equiv 1 \bmod 4$ oder e_i gerade ist, denn $3^2 \equiv 1 \bmod 4$. Andernfalls ist $p_i^{e_i} \equiv 3 \equiv -1 \bmod 4$. Somit ist auch $n \equiv (-1)^N \bmod 4$, also

$$(-1)^{\sum_{i=1}^{r}\frac{p_i-1}{2}e_i} = (-1)^N = (-1)^{\frac{n-1}{2}} .$$

Ist dies gleich +1, so ist die rechte Seite der Gleichung für $\left(\frac{n}{m}\right)\left(\frac{m}{n}\right)$ ebenfalls +1, andernfalls zeigt das gleiche Argument für m, daß sie gleich $(-1)^{(m-1)/2}$ ist. In jedem Fall erhalten wir daher die gewünschte Formel

$$\left(\frac{m}{n}\right)\left(\frac{n}{m}\right) = (-1)^{\frac{n-1}{2}\frac{m-1}{2}} .$$

Genauso folgt auch, daß $\left(\frac{2}{m}\right) = (-1)^{(m^2-1)/8}$ ist, denn dies ist +1 für $m \equiv \pm 1 \bmod 8$ und -1 für $m \equiv \pm 3 \bmod 8$. Das Produkt zweier Primzahlen kongruent ± 1 modulo acht ist wieder kongruent ± 1, genauso das zweier Primzahlen kongruent ± 3 modulo acht. Damit führt dieselbe Argumentation wie oben zum Ziel. ∎

Als Anwendung können wir uns überlegen, modulo welcher Primzahlen eine vorgegebene Zahl a quadratischer Rest ist. Modulo seiner Primteiler verschwindet a und ist somit ein Quadrat. Sei also p kein Teiler von a.

Für $a = 2$ haben wir gesehen, daß $\left(\frac{2}{p}\right)$ nur von der Kongruenzklasse p mod 8 abhängt; wegen der Multiplikativität des JACOBI-Symbols reicht es also, wenn wir ungerade a betrachten. Nach dem gerade bewiesenen Gesetz ist dann

$$\left(\frac{a}{p}\right) = (-1)^{\frac{a-1}{2}\frac{p-1}{2}}\left(\frac{p}{a}\right) .$$

Für festes a ist $(a-1)/2$ ein konstanter Wert, $(p-1)/2$ hängt nur ab von p mod 4, und $\left(\frac{p}{a}\right)$ hängt ab von p mod a. Insgesamt hängt es also nur ab von p mod $4a$, ob a ein quadratischer Rest oder Nichtrest modulo p ist.

Betrachten wir als Beispiel den Fall $a = 3$. Hier ist $(a-1)/2 = 1$, also

$$(-1)^{\frac{(a-1)}{2}\frac{p-1}{2}} = (-1)^{\frac{p-1}{2}} = \begin{cases} +1 & \text{falls } p \equiv 1 \bmod 4 \\ -1 & \text{falls } p \equiv 3 \bmod 4 \end{cases},$$

und

$$\left(\frac{p}{3}\right) = \begin{cases} +1 & \text{falls } p \equiv 1 \bmod 3 \\ -1 & \text{falls } p \equiv 2 \bmod 3 \end{cases} .$$

Somit ist für eine Primzahl $p > 3$

$$\left(\frac{3}{p}\right) = \begin{cases} +1 & \text{falls } p \bmod 12 \in \{1, 11\} \\ -1 & \text{falls } p \bmod 12 \in \{5, 7\} \end{cases} .$$

Für $a = 5$ ist $(a-1)/2 = 2$ gerade, also

$$\left(\frac{5}{p}\right) = \left(\frac{p}{5}\right) = \begin{cases} +1 & \text{falls } p \bmod 5 \in \{1, 4\} \\ -1 & \text{falls } p \bmod 5 \in \{2, 3\} \end{cases} ,$$

§4: Berechnung der modularen Quadratwurzel

Das quadratische Reziprozitätsgesetz zeigt uns schnell, ob die Gleichung $x^2 \equiv a \bmod p$ für eine gegebene Primzahl p und eine ganze Zahl a lösbar ist; es gibt uns aber keinen Hinweis darauf, wie wir diese Lösung finden können. Darum soll es in diesem Paragraphen gehen.

Am einfachsten lassen sich Quadratwurzeln modulo zwei ziehen, denn modulo zwei ist jede Zahl ihre eigene Quadratwurzel. Im folgenden sei daher p eine ungerade Primzahl; wir zerlegen

$$p - 1 = 2^e \cdot q$$

in eine Zweierpotenz 2^e und eine ungerade Zahl q.

Da die multiplikative Gruppe $\mathbb{F}_p^\times$ modulo p zyklisch ist, gibt es ein Element $g \in \mathbb{F}_p^\times$, so daß

$$\mathbb{F}_p^\times = \{g, g^2, \ldots, g^{p-1} = 1\}$$

genau aus den Potenzen von g besteht.

Die Ordnung eines Elements g^r läßt sich leicht berechnen: Da g^{rn} genau dann zu eins wird, wenn $p-1$ den Exponenten rn teilt, ist rn für die Ordnung n das kleinste gemeinsame Vielfache von r und $p-1$. Das kleinste gemeinsame Vielfache ist bekanntlich gleich dem Produkt, dividiert durch den größten gemeinsamen Teiler. Damit ist die Ordnung

$$n = \frac{p-1}{\mathrm{ggT}(r, p-1)} .$$

Speziell für die beiden Elemente

$$y = g^{2^e} \quad \text{und} \quad z = g^q$$

folgt, daß y die Ordnung q hat und z die Ordnung 2^e.

Da q und 2^e teilerfremd sind, gibt es nach dem erweiterten EUKLIDischen Algorithmus ganze Zahlen u, v, so daß

$$2^e u + qv = 1$$

ist. Hierbei muß v offensichtlich eine ungerade Zahl sein, denn sonst wäre die Summe auf der linken Seite eine gerade Zahl.

Damit ist

$$g = g^{2^e u + qv} = g^{2^e u} g^{qv} = y^u z^v$$

und entsprechend für jedes r

$$g^r = y^{ur} z^{vr} .$$

Da es bei Potenzen von y nur auf den Exponenten modulo q ankommt und bei solchen von z nur auf den Exponenten modulo 2^e, heißt dies, daß sich jedes Element von $\mathbb{F}_p^\times$ in der Form

$$a = y^\alpha z^\beta \quad \text{mit} \quad 0 \le \alpha < q \quad \text{und} \quad 0 \le \beta < 2^e$$

schreiben läßt.

$a = g^r$ ist genau dann ein quadratischer Rest, wenn r eine gerade Zahl ist; die beiden Quadratwurzeln sind dann $\pm g^{r/2}$. Falls wir nur a kennen, ist dies allerdings nicht sonderlich nützlich zur Berechnung dieser Wurzeln, denn erstens kennen wir im allgemeinen das Element g nicht explizit, und zweitens kennen wir auch den Exponenten r nicht. Ersteres wäre kein großes Problem, denn es gibt effiziente probabilistische Algorithmen, um sich mögliche Werte für g zu verschaffen, letzteres aber ist ein diskretes Logarithmenproblem modulo p, also nur dann effizient lösbar, wenn die Primzahl p ziemlich klein ist.

Auch der etwas komplizierteren (und bislang ebenfalls nicht explizit bekannten) Form $a = y^\alpha z^\beta$ können wir ansehen, wann a quadratischer Rest ist: Da v eine ungerade Zahl ist, ist r genau dann gerade, wenn auch

vr gerade ist, und das wiederum ist äquivalent dazu, daß $\beta = vr \bmod 2^e$ eine gerade Zahl ist.

$a^q = y^{\alpha q} z^{\beta q} = z^{\beta q}$ ist eine Potenz von z; es gibt also eine ganze Zahl k zwischen null und $2^e - 1$, so daß $a^q z^k = 1$ ist, nämlich

$$k = -\beta q \bmod 2^e \in \{0, 1, 2, \ldots, 2^e - 1\} ,$$

und auch diese Zahl ist genau dann gerade, wenn a quadratischer Rest ist. Mit dieser Zahl k sind dann

$$x = \pm a^{(q+1)/2} z^{k/2}$$

die beiden Quadratwurzeln des quadratischen Rests a, denn

$$x^2 = a^{q+1} z^k = a(a^q z^k) = a .$$

Sobald wir den Wert $z^{k/2}$ kennen, können wir also die Quadratwurzeln von a bestimmen, und wir wir gleich sehen werden, läßt sich dieser Wert erheblich schneller berechnen als ein diskreter Logarithmus.

Zumindest für die „Hälfte“ aller Primzahlen haben wir damit überhaupt kein Problem: Eine ungerade Zahl hat bei Division durch vier offensichtlich entweder Rest eins oder Rest drei; wir betrachten zunächst den Fall, daß $p \equiv 3 \bmod 4$. (Solche Primzahlen werden in der Kryptologie gelegentlich als BLUMsche Primzahlen bezeichnet.)

Ist $p \equiv 3 \bmod 4$, so ist $p - 1 \equiv 2 \bmod 4$, d.h. $p - 1$ ist zwar durch zwei, nicht aber durch vier teilbar. Mithin ist $p - 1 = 2q$ mit einer ungeraden Zahl q, d.h. der oben definierte Exponent e ist eins. Damit kommen für k nur die Werte $k = 0$ und $k = 1$ in Frage, und für einen quadratischen Rest a muß $k = 0$ sein. Dann sind sowohl z^k als auch $z^{k/2}$ gleich eins, also sind die beiden Wurzeln aus a einfach

$$x_{1/2} = \pm a^{(q+1)/2} = \pm a^{(\frac{p-1}{2}+1)/2} = \pm a^{(p+1)/4} ,$$

was sich leicht berechnen läßt. Die Richtigkeit dieser Formel läßt sich auch direkt nachprüfen, denn nach dem Lemma von EULER ist

$$x_{1/2}^2 = a^{(p+1)/2} = a \cdot a^{(p-1)/2} = a \cdot \left(\frac{a}{p}\right) .$$

Ist also a ein quadratischer Rest, so ist $x_{1/2}^2 = a$; wendet man die Formel fälschlicherweise auf einen quadratischen Nichtrest an, ist $x_{1/2}^2 = -a$.

Für $p \equiv 1 \bmod 4$ ist entweder $p \equiv 1 \bmod 8$ oder $p \equiv 5 \bmod 8$. Zumindest im letzteren Fall können wir wieder eine explizite Formel für die Quadratwurzeln aufstellen: In diesem Fall ist $p - 1 \equiv 4 \bmod 8$, d.h. $p - 1$ ist zwar durch vier, nicht aber durch acht teilbar. Somit ist

$$p - 1 = 4q = 2^2 q \quad \text{mit einer ungeraden Zahl } q ,$$

d.h. $e = 2$. Daher kann k nun die vier Werte $0, 1, 2, 3$ annehmen; für quadratische Reste gibt es die beiden Möglichkeiten $k = 0$ und $k = 2$.

Im Fall $k = 0$ können wir wie oben argumentieren: Dann ist

$$x_{1/2} = \pm a^{(q+1)/2} = \pm a^{(\frac{p-1}{4}+1)/2} = \pm a^{(p+3)/8} .$$

Für $k = 2$ sind die Wurzeln $\pm a^{(q+1)/2} z$, wobei z wie oben definiert und nicht explizit bekannt ist. z ist nach Definition q-te Potenz eines primitiven Elements, und das gilt offenbar für jedes Element, dessen Ordnung gleich 2^e ist. (Hier ist natürlich $e = 2$, aber das folgende Argument gilt für jedes e.)

Wenn wir die q-te Potenz *irgendeines* Elements aus $\mathbb{F}_p^\times$ berechnen, erhalten wir ein Element, dessen Ordnung eine Zweierpotenz ist, die 2^e teilt. Falls sie nicht gleich 2^e ist, muß das Element Quadrat eines anderen sein; die Elemente von Zweipotenzordnung, die genaue Ordnung 2^e haben, sind daher genau die quadratischen Nichtreste. Einen solchen können wir uns verschaffen, wenn wir irgendeinen quadratischen Nichtrest modulo p kennen: Da q ungerade ist, ist dann auch dessen q-te Potenz quadratischer Nichtrest, und wegen $p - 1 = 2^e q$ muß diese Potenz Zweipotenzordnung haben. Um z zu finden, reicht es also, *irgendeinen* quadratischen Nichtrest modulo p zu finden.

Letzteres ist im Fall $p \equiv 5 \bmod 8$ einfach: Hier ist zwei nach dem quadratischen Reziprozitätsgesetz quadratischer Nichtrest; daher können wir

$$z = 2^q = 2^{(p-1)/4}$$

setzen. Für $k = 2$ ist somit

$$x_{1/2} = \pm a^{(p+3)/8} \cdot 2^{(p-1)/4} .$$

Um das Ganze wirklich explizit zu machen, müssen wir nun nur noch die beiden Fälle $k = 0$ und $k = 2$ algorithmisch voneinander unterscheiden, aber das ist einfach: Wir berechnen zunächst

$$w = a^{(p+3)/8};$$

falls $u^2 = a$ ist, sind $x_{1/2} = \pm w$ die beiden Quadratwurzeln. Andernfalls multiplizieren wir w mit $2^{(p-1)/4}$; falls das Quadrat dieses Elements von $\mathbb{F}_p$ gleich a ist, sind die Quadratwurzeln $\pm 2^{(p-1)/4} w$; andernfalls ist a quadratischer Nichtrest.

Auch dies läßt sich wieder direkt nachrechnen:

$$w^4 = a^{(p+3)/4} = a^2 \cdot a^{(p-1)/2} = a^2 \left(\frac{a}{p}\right) = a^2$$

nach EULER, falls a quadratischer Rest modulo p ist. Damit ist $w^2 = \pm a$.

Falls $w^2 = a$, sind die beiden Wurzeln $\pm w$; andernfalls ist $w^2 = -a$. Da zwei quadratischer Nichtrest ist, ist nach EULER $2^{(p-1)/2} = -1$, also hat $w \cdot 2^{(p-1)/4}$ das Quadrat a.

Bleibt noch der Fall, daß $p \equiv 1 \bmod 8$. Dies ist der schwerste und allgemeinste Fall, denn während $p \equiv 3 \bmod 4$ äquivalent ist zu $e = 1$ und $p \equiv 5 \bmod 8$ zu $e = 2$ kann e hier jeden Wert größer oder gleich drei annehmen. Damit wird auch die Anzahl 2^e der möglichen Werte von k deutlich größer; die Suche nach dem richtigen Exponenten k wird also aufwendiger.

Auch z selbst ist im allgemeinen Fall schwerer zu finden: Während wir für $p \equiv 5 \bmod 8$ wissen, daß zwei ein quadratischer Nichtrest ist, gibt es für $p \equiv 1 \bmod 8$ keine entsprechende Wahl; hier ist $\left(\frac{2}{p}\right) = 1$, und man weiß nur, daß der kleinste quadratische Nichtrest *irgendeine* Zahl zwischen eins und $1 + \sqrt{p}$ ist, was für große Werte von p viel zu viele Möglichkeiten offenläßt.

Leider gibt es keinen effizienten deterministischen Algorithmus zur Bestimmung eines quadratischen Nichtrests modulo einer beliebigen Primzahl; da wir aber wissen, daß die Hälfte aller Restklassen modulo p quadratische Nichtreste sind, kann man in der Praxis leicht welche finden,

indem man einfach Zufallszahlen erzeugt und nach der EULERschen Formel oder dem quadratischen Reziprozitätsgesetz das LEGENDRE-Symbol berechnet. Die Wahrscheinlichkeit, mehr als zehn Versuche zu brauchen, liegt dann bei $1 : 1024$, die für mehr als zwanzig Versuche ist kleiner als eins zu einer Million, *usw.*

Der folgende Algorithmus von SHANKS funktioniert für beliebige ungerade Primzahlen p; für $p \equiv 3 \bmod 4$ und $p \equiv 5 \bmod 8$ ist es aber natürlich effizienter, die obigen Verfahren zu benutzen.

p sei eine beliebige ungerade Primzahl, und $a \in \mathbb{F}_p^\times$ sei ein quadratischer Rest; der Algorithmus bestimmt unter dieser Voraussetzung ein Element $x \in \mathbb{F}_p^\times$ mit der Eigenschaft, daß x und $-x$ Quadrat a haben.

Indem wir $p - 1$ so lange wie möglich durch zwei dividieren (oder die hinteren Bits betrachten) bestimmen wir zunächst die Zerlegung $p - 1 = 2^e q$ mit einer ungeraden Zahl q.

Im *ersten Schritt* wird dann ein quadratischer Nichtrest modulo p bestimmt, indem wir so lange Zufallszahlen $n \bmod p$ erzeugen, bis $\left(\frac{n}{p}\right) = -1$ ist. Dann berechnen wir $z = n^q \bmod p$ und wissen, daß diese Restklasse genau die Ordnung 2^e hat.

Im *zweiten Schritt* werden einige Variablen initialisiert; wir setzen

$$y \leftarrow z, \quad r \leftarrow e, \quad u \leftarrow a^{(q-1)/2}, \quad b \leftarrow au^2, \quad x \leftarrow au\,,$$

wobei alle Berechnungen modulo p durchgeführt werden. Danach ist

$$ab = x^2, \quad y^{2^{r-1}} = -1 \quad \text{und} \quad b^{2^{r-1}} = 1\,,$$

denn $ab = a^2u^2 = x^2$, und die beiden hinteren Gleichungen bedeuten nach EULER einfach, daß $y = z$ quadratischer Nichtrest ist, a und damit auch $b = au^2$ aber (nach Voraussetzung) quadratischer Rest.

Diese drei Gleichungen werden als Schleifeninvarianten durch den gesamten Algorithmus beibehalten; für $b = 1$ besagen sie, daß x eine Quadratwurzel aus a ist.

Im *dritten Schritt* testen wir daher, ob $b = 1$ ist; falls ja, endet der Algorithmus mit den Lösungen $\pm x$. Andernfalls suchen wir die kleinste

natürliche Zahl m, für die $b^{2^m} = 1$ ist; die dritte Schleifeninvariante zeigt, daß es ein solches m gibt und $m \le r-1$ ist.

Im *vierten Schritt* setzen wir

$$t \leftarrow y^{2^{r-m-1}}, \quad y \leftarrow t^2, \quad r \leftarrow m, \quad x \leftarrow xt, \quad b \leftarrow by$$

und gehen zurück zum dritten Schritt.

Die obigen Schleifeninvarianten gelten auch wieder nach den Zuweisungen im vierten Schritt: Für $ab = x^2$ kommt das einfach daher, daß das neue x gleich dem alten mal t ist, wohingegen das neue b aus dem alten durch Multiplikation mit $y = t^2$ entsteht. Die Gleichung $y^{2^r} = -1$ gilt weiterhin, weil das neue y die 2^{r-m}-te Potenz des alten ist, so daß seine m-te Potenz gleich dem alten Wert von y^{2^r} ist; da das neue r gleich m ist, folgt die Behauptung. Daß auch die dritte Schleifeninvariante erhalten bleibt, folgt aus der Gültigkeit der zweiten, und der Algorithmus endet nach endlich vielen Iterationen, da r bei jeder Wiederholung des vierten Schritt um mindestens eins kleiner wird und $r = 1$ äquivalent ist zu $b = 1$.

DANIEL SHANKS (1917–1996) wurde in Chicago geboren, wo er zur Schule ging und 1937 einen Bachelorgrad in Physik der University of Chicago erwarb. Er arbeitete bis 1950 in verschiedenen Positionen als Physiker, danach als Mathematiker. 1949 begann er ein *graduate* Studium der Mathematik an der University of Maryland, zu dessen Beginn er der erstaunten Fakultät als erstes eine fertige Doktorarbeit vorlegte. Da zu einem *graduate* Studium auch Vorlesungen und Prüfungen gehören, wurde diese noch nicht angenommen; da er während seines Studiums Vollzeit arbeitete, dauerte es noch bis 1954, bevor er alle Voraussetzungen erfüllte; dann wurde die Arbeit in praktisch unveränderter Form akzeptiert. Erst 1977 entschloß er sich, eine Stelle an einer Universität anzunehmen; ab dann bis zu seinem Tod war er Professor an der University of Maryland.

SHANKS schrieb außer seinem Buch *Solved and unsolved problems in number theory* über achtzig Arbeiten, vor allem auf dem Gebiet der algorithmischen Zahlentheorie und der Primzahlverteilung, aber auch der Numerik. 1962 berechnete er π mit der für damals sensationellen Genauigkeit von 100 000 Dezimalstellen. Näheres findet man in seinem Nachruf in den *Notices of the AMS* vom August 1997, der unter `www.ams.org/notices/199707/comm-shanks.pdf` auch im Netz zu finden ist.

§5: Anwendungen quadratischer Reste

Zum Abschluß dieses Kapitels sollen kurz noch einige Anwendungen quadratischer Reste vorgestellt werden:

a) Quadratische Formen und quadratische Reste

Im ersten Paragraphen des Kapitels über quadratische Formen haben wir uns überlegt, welche natürliche Zahlen sich als Summe zweier Quadrate darstellen lassen. Allgemeiner kann man sich fragen, welche ganzen Zahlen sich durch irgendeine vorgegebene Quadratische Form darstellen lassen. GAUSS untersucht diese Frage für die quadratische Formen

$$Q(x,y) = ax^2 + 2bxy + cy^2 ;$$

im Gegensatz zu unserer bisherigen (auf LAGRANGE zurückgehenden) Praxis beschränkt er sich also auf Formen, bei denen der Koeffizient des gemischten Terms gerade ist. Die Matrix $M = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ zu einer solchen Form hat nur ganzzahlige Einträge, und auf Grund früherer Resultate von LAGRANGE wußte er, daß er für solche Formen bessere Ergebnisse erwarten konnte. In Abschnitt 154 seines zahlentheoretischen Hauptwerks *Disquisitiones Arithmeticae* zeigt er den folgenden

Satz: Ist $n = ax^2 + 2bxy + cy^2$ für zwei zueinander teilerfremde ganze Zahlen x, y, so ist $b^2 - ac$ ein quadratischer Rest modulo n.

Sein *Beweis* startet mit der für beliebige Zahlen u, v gültigen Formel

$$\begin{aligned}&(ax^2 + 2bxy + cy^2)(av^2 - 2buv + cu^2)\\ &= \big((u(bx+cy) - v(ax+by)\big)^2 - (b^2-ac)(ux+vy)^2 ,\end{aligned}$$

die der Leser durch Ausmultiplizieren nachrechnen soll. Da es wohl kaum falsch sein kann, einem Ratschlag von GAUSS zu folgen, möchte ich diese Aufforderung auch an die Leser dieses Skriptums weitergeben.

Da x und y teilerfremd sind, lassen sich nach dem erweiterten EUKLIDischen Algorithmus ganze Zahlen u, v finden, für die $ux + vy = 1$ ist. Setzt man diese in obige Formel ein, vereinfacht sich diese zu

$$n \cdot (av^2 - 2buv + cu^2) = \big((u(bx+cy) - v(ax+by)\big)^2 - (b^2-ac) ,$$

modulo n ist also $b^2 - ac$ ein Quadrat, wie behauptet. ∎

Als Beispiel können wir nochmals die quadratische Form $Q(x, y) = x^2 + y^2$ betrachten. Hier ist $b^2 - ac = -1$, falls sich eine natürliche Zahl n als Summe zweier teilerfremder Quadrate darstellen läßt, muß also -1 modulo n ein Quadrat sein.

Ist eine Primzahl $p = x^2 + y^2$ als Summe zweier Quadrate darstellbar, so sind x und y automatisch teilerfremd, also muß $\left(\frac{-1}{p}\right) = 1$ sein. Nach dem Korollar am Ende von §1 ist dies für ungerades p genau dann der Fall, wenn $p \equiv 1 \bmod 4$ ist; für $p = 2$ ist $\left(\frac{-1}{2}\right) = \left(\frac{1}{2}\right) = 1$. Somit kann eine Primzahl $p \equiv 3 \bmod p$ nicht als Summe zweier Quadrate geschrieben werden.

Das wissen natürlich bereits aus §1 des vorigen Paragraphen; für beliebige quadratische Formen aber ist obiger Satz von GAUSS der Ausgangspunkt für eine genauere Untersuchung. Details dazu führen sehr schnell weit in die algebraische Zahlentheorie und können daher im Rahmen dieser Vorlesung nicht behandelt werden.

b) Münzwurf per Telephon

A und B können sich nicht einigen, wer von ihnen eine dringend notwendige aber unangenehme Arbeit übernehmen soll. Also werfen sie eine Münze. Vorher entscheidet sich etwa A für „Wappen“, B für „Zahl“, dann wirft A die Münze in die Luft. Wenn sie mit Wappen nach oben auf den Boden fällt, hat er gewonnen, andernfalls B.

Stellen wir uns nun aber vor, A und B stehen nicht nebeneinander, sondern befinden sich an verschiedenen Orten und diskutieren per Telephon, wer was machen soll. Auch hier könnte A wieder eine Münze werfen, allerdings sieht jetzt nur A, wie sie zu Boden fällt; wenn er gewinnt, muß B sehr viel Vertrauen in ihn haben, um das zu glauben.

Mit Hilfe von quadratischen Resten läßt sich der Münzwurf so simulieren, daß *beide* den Ausgang überprüfen können und jeder mit der gleichen Wahrscheinlichkeit gewinnt.

Dazu wählt sich A zwei Primzahlen p und q die so groß sind, daß B das Produkt $N = pq$ nicht mit einem Aufwand von nur wenigen Minuten faktorisieren kann. (p und q können also deutlich kleiner sein als bei

RSA, wo man mit Gegnern rechnen muß, die monatelang rechnen.) Dieses N schickt er an B.

B wählt sich nun eine zufällige Zahl x zwischen eins und N und schickt deren Quadrat $y = x^2 \bmod N$ an A.

A kennt die Faktorisierung von N; nach dem Algorithmus aus dem vorigen Paragraphen kann er also die Gleichungen

$$z^2 \equiv y \bmod p \quad \text{und} \quad z^2 \equiv y \bmod q$$

lösen; die jeweiligen Lösungsmengen seien $\{a_1, a_2\}$ und $\{b_1, b_2\}$. Nach dem chinesischen Restesatz kann er sich vier Elemente

$$u_{ij} \equiv \begin{cases} a_i \bmod p \\ b_j \bmod q \end{cases} \quad \text{mit} \quad i, j \in \{1, 2\}$$

zwischen null und $N - 1$ konstruieren, die allesamt die Kongruenz $u_{ij}^2 \equiv y \bmod N$ erfüllen. Er entscheidet sich zufällig für eine dieser vier Möglichkeiten (dies entspricht dem Münzwurf) und schickt das entsprechende $u = u_{ij}$ an B.

B kennt nun zwei Zahlen x und u, die beide das Quadrat y haben. Möglicherweise ist $u = x$; in diesem Fall hat er keine neue Information bekommen, und er hat verloren. Das gleiche gilt im Fall $u \equiv -x \bmod N$, d.h. $u = N - x$.

Ist aber $u \neq \pm x$, was mit 50%-iger Wahrscheinlichkeit eintritt, hat B gewonnen und muß das nun gegenüber A beweisen.

Aus der Kongruenz $x^2 \equiv y \bmod N = pq$ folgen natürlich die entsprechenden Kongruenzen modulo p und modulo q; es gibt daher Indizes $\mu, \nu \in \{1, 2\}$, so daß $x \equiv a_\mu \bmod p$ und $x \equiv b_\nu \bmod q$. Falls sich A genau für dieses Indexpaar entschieden hat, ist $x = u$; falls er sich für das Paar (i, j) mit $\mu \neq i$ und $\nu \neq j$ entschieden hat, ist $u \equiv -x \bmod N$, denn $a_1 \equiv -a_2 \bmod p$ und $b_1 \equiv -b_2 \bmod q$.

Falls sich A aber für ein Paar (i, j) entschieden hat, in dem genau einer der beiden Indizes gleich dem entsprechenden Index in (μ, ν) ist, sind x und u modulo einer der beiden Primzahlen p, q kongruent, modulo der anderen ist x kongruent zu $-u$. Um zu beweisen, daß dieser für ihn günstige Fall eingetreten ist, kann B daher ggT$(x - u, N)$ berechnen

und erhält einen der beiden Primfaktoren p oder q. (Der andere ist ggT$(x + u, N)$.) Damit hat er N faktorisiert und schickt das Ergebnis an A.

Wenn B sich nicht an die Regeln hält und ein y an A schickt, das kein Quadrat modulo N ist, merkt A dies bei der Berechnung der modularen Quadratwurzeln; falls A ein u schickt, dessen Quadrat von y verschieden ist, kann B dies leicht feststellen, denn wenn er verloren hat, muß $u = x$ oder $u = N - x$ sein. (Er kann natürlich auch $u^2 \bmod N$ berechnen.)

c) Akustik von Konzerthallen

Alte Konzerthallen waren zwangsläufig sehr hoch: Andernfalls wäre die Luft während eines längeren Konzert bei voll besetztem Saal zu schnell verbraucht gewesen. Mit den Fortschritten der Lüftungstechnik verschwand diese Notwendigkeit; dafür sorgten steigende Bau- und Heizungskosten für immer niedrigere Säle. Auf die Luftqualität hatte das keinen nennenswerten Einfluß; die Akustik der Hallen allerdings wurde deutlich schlechter.

Der Grund dafür ist intuitiv recht klar und wurde auch durch Messungen und Hörerbefragungen in einer Reihe von Konzertsälen experimentell bestätigt: Die Hörer bevorzugen Schall, der von den Seitenwänden kommt und daher mit verschiedener Stärke bei den beiden Ohren eintrifft gegenüber Schall von oben, der beide Ohren mit gleicher Stärke erreicht und somit keinen räumlichen Eindruck hinterläßt.

Eine mögliche Abhilfe bestünde darin, die Decken aus absorbierendem Material zu bauen. Dem steht entgegen, daß in einem großen Konzertsaal aller Schall, der von der Bühne kommt, den Hörer auch wirklich erreichen sollte: Ansonsten müßte der Schall aus Lautsprechern kommen und man könnte sich das Konzert genauso gut daheim per Radio oder CD anhören.

Der Schall muß daher von der Decke reflektiert werden, darf die Ohren der Zuhörer aber nicht von oben erreichen. Er sollte daher beispielsweise möglichst diffus zu den Seitenwänden hin gestreut werden, so daß der größte Teil der Energie die Zuhörer über die Seitenwände erreicht.

Der Einfachheit halber wollen wir uns auf eindimensionale Wellen beschränken und damit auch nur diffuse Reflektion in einer Richtung betrachten, der Querrichtung des Konzertsaals.

Eine Welle hat eine räumliche wie auch zeitliche Periodizität. Zeitlich periodische Funktionen sind beispielsweise Sinus und Kosinus; wie die FOURIER-Analysis lehrt, läßt sich jede stückweise stetige zeitlich periodische Funktion (bis auf sogenannte Nullfunktionen) aus Sinus- und Kosinusfunktionen zusammensetzen, so daß es reicht, solche Funktionen zu betrachten.

Da der Umgang mit den Additionstheoremen für trigonometrische Funktionen recht umständlich ist, schreibt man Wellen allerdings meist komplex in der Form $f(t) = Ae^{i\omega t}$ mit der Maßgabe, daß nur der Realteil dieser Funktion physikalische Realität beschreibt. Aufgrund der EULERschen Formel $e^{i\varphi} = \cos\varphi + i\sin\varphi$ lassen sich so, falls man für A beliebige komplexe Konstanten zuläßt, alle Funktionen der Art $a\cos\omega t + b\sin\omega t$ als Realteile erhalten, und da beispielsweise

$$\cos(\alpha+\beta) = \mathfrak{Re}\, e^{i(\alpha+\beta)} = \mathfrak{Re}(e^{i\alpha}e^{i\beta}) = \cos\alpha\,\cos\beta - \sin\alpha\,\sin\beta$$

ist, lassen sich auf diese Weise auch die Additionstheoreme auf einfache Multiplikationen von Exponentialfunktionen zurückführen.

Auch die räumliche Periodizität läßt sich mit trigonometrischen oder – besser – Exponentialfunktionen ausdrücken; hier schreiben wir entsprechend $g(x) = Be^{ikx}$.

Um einen räumlich und zeitlich periodischen Vorgang zu beschreiben, kombinieren wir die beiden Ansätze und betrachten beispielsweise die Funktion

$$\psi(x,t) = Ae^{i(\omega t - kx)} = Ae^{ik(\frac{\omega}{k}t - x)}\,.$$

Wie man der zweiten Form ansieht, hängt $\psi(x,t)$ nur ab von $x - \frac{\omega}{t}t$, was wir auch so interpretieren können, daß

$$v = \frac{\omega}{k} = \frac{\lambda}{T} = \frac{\lambda\omega}{2\pi}$$

die Ausbreitungsgeschwindigkeit der Welle ist; denn eine Änderung der Zeit um Δt hat denselben Effekt wie eine Änderung des Orts um $v\cdot\Delta t$.

Da Sinus und Kosinus die Periode 2π haben, müssen wir für eine Schwingung der Frequenz ν den Parameter ω gleich $2\pi\nu$ wählen, denn dann fallen $1/\nu$ Perioden in das Intervall $0 \le t \le 1$. Aus diesem Grund wird $\omega = 2\pi\nu$ als die *Kreisfrequenz* der Schwingung bezeichnet. In der räumlichen Dimension nimmt die Wellenlänge λ die Rolle der zeitlichen Periode ein; dementsprechend muß hier $k = 2\pi/\lambda$ gesetzt werden. Diese Konstante wird als *Wellenzahl* bezeichnet.

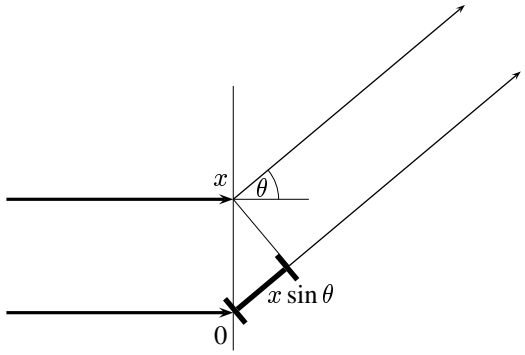
Schallwellen breiten sich bei 20° C in Luft mit einer Geschwindigkeit von etwa $v = 343\,\text{m/s}$ aus; der hörbare Frequenzbereich beginnt bei $\nu = 16\,\text{Hz}$ und kann bis zu etwa $\nu 20\,\text{kHz}$ gehen. Die Wellenlängen, mit denen wir es zu tun haben, variieren also zwischen etwa $\lambda = 21{,}5\,\text{m}$ und $\lambda = 1{,}75\,\text{cm}$. Der Kammerton a' mit 440 Hz hat eine Wellenlänge von knapp 78 cm.

Bei einer Reflektion können wir nach HUYGENS annehmen, daß von jedem Punkt der reflektierenden Fläche eine neue Welle ausgeht; ihre Amplitude ist gleich der Amplitude der dort eintreffenden Welle mal einem Reflektionsfaktor $\rho(x)$, der im Idealfall gleich eins ist.

CHRISTIAAN HUYGENS (1629–1695) kam aus einer niederländischen Diplomatenfamilie. Dadurch und später auch durch seine Arbeit hatte er Kontakte zu führenden europäischen Wissenschaftlern wie DESCARTES und PASCAL. Nach seinem Studium der Mathematik und Jura arbeitete er teilweise auch selbst als Diplomat, interessierte sich aber bald vor allem für Astronomie und den Bau der dazu notwendigen Instrumente. Er entwickelte eine neue Methode zum Schleifen von Linsen und erhielt ein Patent für die erste Pendeluhr. Trotz des französisch-niederländischen Kriegs arbeitete er einen großen Teil seines Lebens an der *Académie Royale des Sciences* in Paris, wo beispielsweise LEIBNIZ viel Mathematik bei ihm lernte. HUYGENS war ein scharfer Kritiker sowohl von NEWTONS Theorie des Lichts als auch seiner Gravitationstheorie, die er für absurd und nutzlos hielt. Gegen Ende seines Lebens beschäftigte er sich mit der Möglichkeit außerirdischen Lebens.

Da es uns nur um den mittleren Schalldruck, nicht aber um seine Variation geht, können wir den ωt-Term ignorieren und einfach mit der Funktion Ae^{-ikx} arbeiten. Wir interessieren uns, wieviel Schall unter welchem Winkel reflektiert wird.

Die Schallwellen die von zwei verschiedenen Punkten unter einem Winkel θ ausgehen haben, wie die Zeichnung zeigt, einen Laufwegunterschied von $x\sin\theta$, wobei x den Abstand der beiden Punkte bezeichnet.



Der Laufwegunterschied von $x\sin\theta$ entspricht einem Phasenfaktor $e^{-ikx\sin\theta}$. Wählen wir also die Phase im Nullpunkt als Referenz (die wir in den zu ignorierenden Phasenfaktor der einfallenden Welle hineinziehen können), ist die Summe aller unter dem Winkel θ abgehenden Strahlen gleich

$$\int_{-\infty}^{\infty} \rho(x) e^{-ikx\sin\theta}\,dx\,;$$

das ist die sogenannte FOURIER-Transformierte von $\rho(x)$, ausgewertet im Punkt $u = k\sin\theta$. Wenn wir den Schall möglichst gleichmäßig verteilen wollen, müssen wir die Funktion ρ daher so wählen, daß ihre FOURIER-Transformierte möglichst konstant ist.

Eine Möglichkeit dazu sind das, was Physiker als *Reflektions-Phasengitter* bezeichnen: Die Decke besteht aus einem Material mit konstantem, möglichst großem Reflektionsgrad, aber die Höhe der Decke variiert stufenförmig mit dem Querschnitt. Wenn die Höhe der einer festen Stelle um den Betrag h über der Nullinie liegt, muß der dort reflektierte Schall gegenüber dem an der Nullinie reflektierten den zusätzlichen Weg $2h$ zurücklegen; dies kann man formal so ausdrücken, daß man in der Reflektionsfunktion $r(x)$ den zusätzlichen Faktor $e^{2i\omega h}$ einfügt.

Bei den sogenannten SCHROEDER-Reflektoren werden die Abstände zur Nullinie so gewählt, daß die Längen $2\omega h$ gleich den quadratischen Resten modulo einer ungeraden Primzahl sind, die Decke ist also treppenförmig aufgebaut, wobei die n-te Stufe eine Höhe proportional zu $n^2 \bmod p$ hat. Das obige FOURIER-Integral läßt sich dann approximieren durch die diskrete FOURIER-Transformierte

$$\widehat{r}(m) = \frac{1}{\sqrt{p}} \sum_{n=0}^{p-1} e^{2\pi i n^2/p} e^{-2\pi i n m t} = \frac{1}{\sqrt{p}} \sum_{n=0}^{p-1} e^{2\pi i n(n-m)/p} \,.$$

Ihr Betragsquadrat ist

$$\begin{aligned} |\widehat{r}(m)|^2 &= \frac{1}{\sqrt{p}} \sum_{n=0}^{p-1} e^{2\pi i n(n-m)/p} \cdot \frac{1}{\sqrt{p}} \sum_{n=0}^{p-1} e^{-2\pi i n(n-m)/p} \\ &= \frac{1}{p} \sum_{n=0}^{p-1} \sum_{k=0}^{p-1} e^{2\pi i n(n-m)/p} e^{-2\pi i k(k-m)/p} \\ &= \frac{1}{p} \sum_{n=0}^{p-1} \sum_{k=0}^{p-1} e^{2\pi i \left(n^2-k^2-(n-k)m\right)/p} \end{aligned}$$

Die Summanden hängen nur ab von den Restklassen modulo p der Indizes k und n, und für festes n durchläuft mit k auch $n-k$ alle diese Restklassen. Daher können wir dies weiter ausrechnen als

$$\begin{aligned} |\widehat{r}(m)|^2 &= \frac{1}{p} \sum_{n=0}^{p-1} \sum_{k=0}^{p-1} e^{2\pi i \left(n^2-(n-k)^2-km\right)/p} \\ &= \frac{1}{p} \sum_{k=0}^{p-1} e^{-2\pi i k m/p} \sum_{n=0}^{p-1} e^{2\pi i \left((n^2-(n-k)^2\right)/p} \\ &= \frac{1}{p} \sum_{k=0}^{p-1} e^{-2\pi i k m/p} \sum_{n=0}^{p-1} e^{2\pi i (2kn-k^2)/p} \,. \end{aligned}$$

Die zweite Summe können wir schreiben als

$$e^{-2\pi i k^2} \sum_{n=0}^{p-1} e^{4\pi i k n/p} \,.$$

Für $k = 0$ ist sowohl der Vorfaktor wie auch jeder der Summanden gleich eins, wir erhalten also insgesamt p. Für $k \neq 0$ und $k < p$ ist die Summe aber gleich null, denn

$$e^{4\pi i k/p} \sum_{n=0}^{p-1} e^{4\pi i k n/p} = \sum_{n=0}^{p-1} e^{4\pi i (k+1) n/p} = \sum_{n=1}^{p} e^{4\pi i k n/p} = \sum_{n=0}^{p-1} e^{4\pi i k n/p} \,,$$

die Summe ändert also ihren Wert nicht, wenn man sie mit der von eins verschiedenen Zahl $e^{4\pi i k/p}$ multipliziert, und damit muß sie verschwinden. Somit ist

$$|\widehat{r}(m)|^2 = \frac{1}{p} e^0 \cdot p = 1$$

für alle m, wir haben also die gewünschte Diffusionseigenschaft.

Die obige Abbildung zeigt den Querschnitt über ein solches Phasengitter, hier für $p = 23$. Entsprechende SCHROEDER-Reflektoren zu den verschiedensten Primzahlen gibt es in vielen Konzertsälen und Opernhäusern, oft allerdings verborgen hinter schalldurchlässigem Material.

MANFRED ROBERT SCHROEDER wurde 1926 in Deutschland geboren. Er studierte Physik an der Universität Göttingen, wo er 1952 promovierte. Danach arbeitete er bei den AT & T Bell Laboratories in Murray Hill, New Jersey auf dem Gebiet der Akustik; diese Arbeit führte unter anderem zu 45 Patenten. 1969 wechselte er als Professor für Akustik an die Universität Göttingen, wo er bis zu seiner Emeritierung lehrte. Er schrieb mehrere Bücher, unter anderem *Number theory in Science and Communication* und *Fractals, Chaos, Power Laws*. Der Inhalt dieses Abschnitts ist kurz im ersten dieser Bücher dargestellt sowie ausführlich in M.R. SCHROEDER: Binaural dissimilarity and optimum ceilings for concert halls: More lateral sound diffusion, *J. Acoust. Soc. Am.* **65** (4), 1979

www.physik3.gwdg.de/~mrs

Kapitel 9
Die Fermat-Vermutung für Zahlen und für Polynome

§1: Zahlen und Funktionen

Zum Beweis der eindeutigen Primzerlegung in der Hauptordnung eines Zahlkörpers mußten wir nur nachweisen, daß diese ein EUKLIDischer Ring ist, denn wie wir aus Kapitel 6, §5 wissen, ist jeder EUKLIDische Ring faktoriell. Da der Polynomring in einer Veränderlichen über einem Körper EUKLIDisch ist, gilt also auch dort das Gesetz von der eindeutigen Primzerlegung, wobei die irreduziblen Polynome die Rolle der Primzahlen einnehmen.

Besonders einfach ist die Situation, wenn der Grundkörper k algebraisch abgeschlossen ist: Dann hat jedes nichtkonstante Polynom $f \in k[x]$ eine Nullstelle a und ist damit durch $x - a$ teilbar. In diesem Fall sind also alle irreduziblen Polynome linear.

Die Zerlegung in irreduzible Elemente ist bekanntlich nur eindeutig bis auf Einheiten, und die Einheiten eines Polynomrings sind nach §1 aus Kapitel 6 gerade die des Koeffizientenrings, hier also die von Null verschiedenen Elemente von k. Durch Multiplikation mit einem solchen Element kann man den höchsten Koeffizienten eines jeden Polynoms zu eins machen; die irreduziblen Polynome über einem algebraisch abgeschlossenen Körper sind also bis auf Assoziiertheit genau die Polynome der Form $x - a$ mit $a \in k$ und sie entsprechen eindeutig den Elementen von k.

Den Primzahlen von $\mathbb{Z}$ entsprechen daher im Polynomring über einem algebraisch abgeschlossenen Körper die Punkte der affinen Geraden über k, wir können also geometrisch argumentieren.

Natürlich gibt es – zum Teil beträchtliche – Unterschiede zwischen $\mathbb{Z}$ und dem Polynomring über einem Körper, aber gerade das macht die Analogie so interessant: Da es für jeden der beiden Ringe ein eigenes Instrumentarium gibt, kann man versuchen die damit bewiesenen Resultate auf den jeweils anderen Fall zu übertragen, was idealerweise zu neuen Sätzen und sonst zumindest zu interessanten Vermutungen führt.

Als Beispiel für Parallelen und Unterschiede zwischen den beiden Situationen wollen wir die FERMAT-Vermutung betrachten. FERMAT schrieb bekanntlich um 1637 an den Rand seiner Arithmetik des DIOPHANTOS von Alexandrien, daß die Gleichung

$$x^n + y^n = z^n$$

für $n \geq 3$ keine Lösung in ganzen Zahlen habe – außer natürlich den trivialen Lösungen, bei denen eine der beiden Variablen verschwindet. (Die französische Übersetzung der Arithmetik, die er dabei benutzte, stammt übrigens von BACHET DE MÉZIRIAC, denn wir bereits als Entdecker des erweiterten EUKLIDischen Algorithmus kennen. Bekannt wurde FERMATs Randbemerkung erst, als dessen Sohn CLÉMENT-SAMUEL DE FERMAT 1670 die Arithmetik mit den Randbemerkungen seines fünf Jahre zuvor gestorbenen Vaters veröffentlichte.)

Es ist nicht möglich, einen Kubus in zwei Kuben oder ein Biquadrat in zwei Biquadrate und ganz allgemein irgendeine der unendlich vielen Potenzen jenseits des Quadrats in zwei ebensolche zu teilen. Ich habe einen wunderbaren Beweis hierfür gefunden, aber der Rand ist zu schmal, um ihn zu fassen.

Die direkte Verallgemeinerung auf Polynomringe ist sicherlich falsch: Die Gleichung $f^n + g^n = h^n$ ist zumindest für *konstante* Polynome über einem algebraisch abgeschlossenen Körper immer lösbar: Für beliebig vorgegebene Konstanten $f, g \in k$ muß man einfach $h = \sqrt[n]{f^n + g^n}$ setzen. Das sind allerdings, wenn wir uns wirklich für Polynome interessieren, uninteressante Lösungen, vergleichbar den Lösungen $x^n + 0^n = x^n$ der klassischen FERMAT-Gleichung.

Auch wenn wir verlangen, daß die Grade aller beteiligter Polynome positiv sein sollen, gibt es triviale Lösungen: Ist f irgendein beliebiges Polynom und sind $a, b, c \in k$ so, daß gilt $a^n + b^n = c^n$, ist natürlich auch $(af)^n + (bf)^n = (cf)^n$. Was wir höchstens erwarten können ist also das folgende Analogon zur klassischen FERMAT-Vermutung:

Für $n \geq 3$ gibt es keine paarweise teilerfremden Polynome f, g, h mit positivem Grad, so daß $f^n + g^n = h^n$ ist.

Für Körper positiver Charakteristik ist selbst das noch falsch: Über einen Körper der Charakteristik p ist schließlich $f^p + g^p = (f + g)^p$ für beliebige Polynome f und g, und dasselbe gilt auch wenn man den Exponenten p durch eine seiner Potenzen ersetzt. Wir können also höchstens für Körper der Charakteristik null erwarten, daß diese Vermutung für alle Exponenten $n \geq 3$ richtig ist, und genau das werden wir im übernächsten Paragraphen (zumindest für den Körper der komplexen Zahlen) auch beweisen. Zunächst aber wollen wir schauen, was im bei FERMAT ausgeschlossenen Fall des Exponenten zwei passiert.

§2: Pythagoräische Tripel

Betrachten wir zunächst den Fall der Polynome, wobei wir uns der Einfachheit halber gleich auf Polynome mit komplexen Koeffizienten beschränken wollen. Ist $f^2 + g^2 = h^2$, so ist

$$f^2 = h^2 - g^2 = (h + g)(h - g)\,.$$

Wenn wir f und g als teilerfremd voraussetzen, sind auch g und h teilerfremd und somit auch $h + g$ und $h - g$, denn jeder gemeinsame Teiler dieser beiden Polynome wäre auch ein Teiler ihrer Summe $2h$ sowie ihrer Differenz $2g$.

Wenn wir die Zerlegung von f^2 in irreduzible Faktoren vergleichen mit der von $h + g$ und $h - g$ folgt somit, daß jeder irreduzible Faktor von f entweder in $h + g$ oder in $h - g$ in gerader Potenz auftreten muß. Da jede komplexe Zahl ein Quadrat ist, können wir auch eine eventuell auftretende Einheit als Quadrat schreiben; somit gibt es zwei Polynome $u, v \in \mathbb{C}[x]$ derart, daß

$$u^2 = h + g, \quad v^2 = h - g \quad \text{und} \quad uv = f$$

ist, d.h.

$$f = uv, \quad g = \frac{u^2 - v^2}{2} \quad \text{und} \quad h = \frac{u^2 + v^2}{2}\,.$$

Starten wir umgekehrt mit zwei beliebigen teilerfremden Polynomen $u, v \in \mathbb{C}[x]$, erhalten mit Hilfe dieser Formeln Lösungen der Gleichung $f^2 + g^2 = h^2$. Damit kennen wir alle teilerfremden Lösungen, und die restlichen erhalten wir, indem wir alle drei Polynome mit einem gemeinsamen Faktor multiplizieren.

Nehmen wir als ein einfaches Beispiel $u = 2x$ und $v = 2$, ist also

$$f = 4x, \quad g = x^2 - 1 \quad \text{und} \quad h = x^2 + 1\,;$$

in der Tat ist

$$(2x)^2 + \left(x^2 - 1\right)^2 = \left(x^2 + 1\right)^2\,.$$

Hier erhalten wir also mit geringem Aufwand eine vollständige Übersicht über alle Lösungen.

Versuchen wir das gleiche auch für den klassischen Fall! Wegen des Satzes von PYTHAGORAS bezeichnet man ein Tripel (x, y, z) ganzer Zahlen mit $x^2 + y^2 = z^2$ als pythagoräisches Tripel; für jedes solche Tripel gibt es ein rechtwinkliges Dreieck mit Seitenlängen $|x|$, $|y|$ und $|z|$.

Wir bezeichnen das Tripel (x, y, z) als *primitiv,* wenn sie sich nicht als Vielfaches eines anderen schreiben läßt, wenn also die Zahlen x, y, z keinen gemeinsamen Teiler haben. Sobald wir alle primitiven Lösungen kennen, können wir daraus die gesamte Lösungsmenge konstruieren, denn jede nichtprimitive Lösung ist Vielfaches einer primitiven.

Wie im Fall der Polynome gehen wir aus von einem primitiven Tripel (x, y, z) und wenden die dritte binomische Formel an:

$$x^2 = z^2 - y^2 = (z + y)(z - y)\,.$$

Hier können wir leider nicht mehr ohne weiteres folgern, daß $z + y$ und $z - y$ teilerfremd und damit Quadrate sind: Sind y und z beide ungerade, so sind ihre Summe und ihre Differenz beide gerade, also durch zwei teilbar. Wir müssen uns also zunächst über die Paritäten von y und z klarwerden.

Für eine primitive Lösung (x, y, z) müssen bereits x und y teilerfremd sein, denn ist d ein gemeinsamer Teiler von x und y, so sind x^2 und y^2 beide durch d^2 teilbar, also auch ihre Summe z^2. Wegen der eindeutigen

Zerlegbarkeit einer natürlichen Zahl in Primfaktoren ist dann auch z durch d teilbar, d.h. d ist ein gemeinsamer Teiler von x, y und z.

Insbesondere können daher x und y nicht beide gerade sein; mindestens eine der beiden Zahlen muß ungerade sein. Andererseits können aber auch nicht beide Zahlen ungerade sein: Wäre nämlich $x = 2u + 1$ und $y = 2v + 1$, so wäre

$$z^2 = (2u+1)^2 + (2v+1)^2 = 4u^2 + 4u + 1 + 4v^2 + 4v + 1 \equiv 2 \bmod 4\,,$$

was unmöglich ist, da modulo vier nur null und eins Quadrate sind.

Somit muß in einem primitiven pythagoräischen Tripel (x, y, z) eine der beiden Zahlen x, y gerade sein und die andere ungerade. Da mit (x, y, z) auch (y, x, z) ein primitives pythagoräisches Tripel ist, genügt es, wenn wir diejenigen Tripel betrachten, in denen x gerade ist und y ungerade. Offensichtlich ist dann auch z ungerade.

Für so ein Tripel steht in der Gleichung

$$x^2 = z^2 - y^2 = (z+y)(z-y)$$

rechts das Produkt zweier gerader Zahlen. Im Gegensatz zur Situation bei den Polynomen haben wir hier also keine teilerfremden Faktoren.

Dividieren wir aber durch zwei, so können wir wie oben argumentieren, daß $\frac{1}{2}(z+y)$ und $\frac{1}{2}(z-y)$ teilerfremd sind, denn jeder gemeinsame Teiler wäre Teiler ihrer Summe z und ihrer Differenz y.

Jetzt können wir wieder die eindeutige Primzerlegung anwenden: Da

$$x^2 = 2^2 \cdot \left(\frac{z+y}{2}\right) \cdot \left(\frac{z-y}{2}\right)\,,$$

wobei die beiden Klammern teilerfremd sind, gibt es ganze Zahlen u, v, so daß

$$u^2 = \left(\frac{z+y}{2}\right), \quad v^2 = \left(\frac{z+y}{2}\right) \quad \text{und} \quad 2uv = x$$

ist, also

$$x = 2uv, \quad y = u^2 - v^2 \quad \text{und} \quad z = u^2 + v^2\,.$$

Umgekehrt definieren diese Formeln für beliebige ganze Zahlen u und v ein primitives pythagoräisches Tripel, und bis auf die Reihenfolge von x und y erhalten wir so auch jedes dieser Tripel, für $u = 2$ und $v = 1$ etwa das seit Jahrtausenden bekannte Tripel $(4, 3, 5)$. Da sich in alten Sakralbauten und -anlagen auch deutlich kompliziertere pythagoräische Tripel nachweisen lassen, steht zu vermuten, daß die obige Konstruktion möglicherweise bereits in einigen steinzeitlichen Kulturen zumindest teilweise bekannt waren; entsprechende Thesen vertritt beispielsweise bekannte algebraische Geometer B.L. VAN DER WAERDEN (1903–1996), der nach seiner Emeritierung auch mehrere Bücher über die Geschichte der Mathematik veröffentlichte. Mit pythagoräischen Tripel beschäftigt er sich ausführlich in

B.L. VAN DER WAERDEN: Geometry and algebra in ancient civilizations, *Springer,* 1983

§3: Der Satz von Mason

In diesem Abschnitt wollen wir, wie bereits angekündigt, sehen, daß es für $n \geq 3$ keine zueinander teilerfremden Polynome positiven Grades $f, g, h \in \mathbb{C}[x]$ gibt mit $f^n + g^n = h^n$.

Der *Beweis* beruht darauf, daß die Polynome f^n und g^n dieselben Nullstellen wie f und g haben, aber mit n-facher Vielfachheit. Ist $f^n + g^n = h^n$, so hat auch die Summe dieser beiden Potenzen im Vergleich zum Grad relativ wenige Nullstellen, diese aber mit mindestens n-facher Vielfachheit. Nach einem 1983 von R.C. MASON bewiesenen Satz können in einer solchen Situation aber f^n, g^n und h^n nicht zu wenige verschiedene Nullstellen haben:

Satz: Bezeichnet $n_0(f)$ die Anzahl verschiedener (komplexer) Nullstellen eines Polynoms f, so gilt für drei nichtkonstante, teilerfremde Polynome f, g, h mit $f + g = h$

$$n_0(fgh) \geq \max(\deg f, \deg g, \deg h) + 1\,.$$

Bevor wir diesen Satz beweisen, wollen wir uns zunächst überlegen, daß daraus wirklich die FERMAT-Vermutung für Polynome folgt:

Für drei nichtkonstante teilerfremde Polynome f, g, h mit $f^n + g^n = h^n$ ist nach dem Satz von MASON

$$\begin{aligned} n_0\big(f^n g^n h^n\big) &\geq \max\big(\deg f^n, \deg g^n, \deg h^n\big) + 1 \\ &= n \max\big(\deg f, \deg g, \deg h\big) + 1 \,. \end{aligned}$$

Andererseits ist aber

$$\begin{aligned} n_0\big(f^n g^n h^n\big) &= n_0\big(fgh\big) \\ &\leq \deg f + \deg g + \deg h \\ &\leq 3 \max\big(\deg u(x), \deg v(x), \deg w(x)\big) \,, \end{aligned}$$

denn die Anzahl *verschiedener* Nullstellen einer Potenz eines Polynoms ist gleich der Anzahl verschiedener Nullstellen des Polynoms selbst, und die Nullstellenanzahl eines Polynom kann nicht größer sein als der Grad.

Damit haben wir insgesamt die Ungleichung

$$\begin{aligned} 3 \max\big(\deg f, \deg g, \deg h\big) &\geq n_0\big(f^n g^n h^n\big) \\ &\geq n \max\big(\deg f, \deg g, \deg h\big) + 1 \,, \end{aligned}$$

die bei nichtkonstanten Polynomen nur für $n \leq 2$ gelten kann. Somit gibt es für $n \geq 3$ keine nichtkonstanten teilerfremden Polynome, für die $f^n + g^n = h^n$ ist.

Zu einem vollständigen Beweis der FERMAT-Vermutung für Polynome fehlt nun nur noch der Beweis des Satzes von MASON. Die Idee dazu ist folgende: Ist $f + g = h$, so betrachten wir den Quotienten g/f im rationalen Funktionenkörper $\mathbb{C}(x)$. Da f und g teilerfremd sind, ist das ein gekürzter Bruch. Falls wir diesen auch in der Form $g/f = G/F$ schreiben können mit Polynomen F, G vom Grad höchstens $n_0(fgh) - 1$ schreiben können, so haben auch f und g höchstens den Grad $n_0(fgh) - 1$. Wegen $f + g = h$ gilt dasselbe auch für h, und damit wäre der Satz bewiesen.

Um g/f als Quotienten zweier neuer Polynome auszudrücken, schreiben wir zunächst

$$\frac{g}{f} = \frac{S}{R} \quad \text{mit} \quad R = \frac{f}{h} \;\text{ und }\; S = \frac{g}{h} \,.$$

Dabei ist $R + S = 1$, die Summe $R' + S'$ der Ableitungen verschwindet also. Aus der Gleichung

$$R' + S' = \frac{R'}{R} R + \frac{S'}{S} S$$

folgt die neue Darstellung

$$\frac{g}{f} = \frac{S}{R} = -\frac{R'/R}{S'/S} \,.$$

Rechts stehen die logarithmischen Ableitungen von R und S im Zähler und Nenner, und damit lassen sich gut die Nullstellen von f, g und h ins Spiel bringen: Nach der LEIBNIZ-Regel ist bekanntlich

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

die logarithmische Ableitung eines Produkts ist also einfach die Summe der logarithmischen Ableitungen der Faktoren. Daraus folgt sofort, daß die logarithmische Ableitung eines Quotienten gleich der Differenz aus logarithmischer Ableitung des Zählers und logarithmischer Ableitung des Nenners ist. Schreiben wir

$$f = f_0 \prod_{i=1}^{r} (x - a_i)^{n_i}, \quad g = g_0 \prod_{j=1}^{s} (x - b_j)^{s_j} \quad \text{und} \quad h = h_0 \prod_{k=1}^{t} (x - c_k)^{p_k} \,,$$

so ist also

$$\frac{R'}{R} = \frac{f'}{f} - \frac{h'}{h} = \sum_{i=1}^{r} \frac{n_i}{x - a_i} - \sum_{k=1}^{t} \frac{p_k}{x - c_k} \,,$$

$$\frac{S'}{S} = \frac{g'}{g} - \frac{h'}{h} = \sum_{j=1}^{r} \frac{m_j}{x - b_j} - \sum_{k=1}^{t} \frac{p_k}{x - c_k}$$

$$\text{und} \quad \frac{g}{f} = \frac{R'/R}{S'/S} = - \frac{\displaystyle\sum_{i=1}^{r} \frac{n_i}{x - a_i} - \sum_{k=1}^{t} \frac{p_k}{x - c_k}}{\displaystyle\sum_{j=1}^{s} \frac{m_j}{x - b_j} - \sum_{k=1}^{t} \frac{p_k}{x - c_k}} \,.$$

Erweitern wir Zähler und Nenner mit dem Hauptnenner aller Summanden erweitern, d.h. mit dem Polynom vom Grad $r+s+t = n_0(fgh)$

$$H = \prod_{i=1}^{r}(x-a_i)\cdot\prod_{j=1}^{s}(x-b_j)\cdot\prod_{k=1}^{t}(x-c_k)\,,$$

so erhalten wir im Zähler wie auch im Nenner Summen von Polynomen vom Grad $n_0(fgh)-1$, als Polynome vom Grad höchstens $n_0(fgh)-1$, wie gewünscht. Damit ist der Satz von MASON bewiesen.

§4: Die abc-Vermutung

Der Erfolg des Satzes von MASON beim Beweis der FERMAT-Vermutung für Polynome legt es nahe, etwas ähnliches auch im klassischen Fall zu versuchen.

Da natürliche Zahlen weder Grade noch Nullstellen haben, müssen wir dazu den Satz von MASON zunächst einmal so umformulieren, daß wir eine Aussage bekommen, die ein sinnvolles Analogon für natürliche Zahlen hat.

Dazu ordnen wir einem Polynom f anstelle der Anzahl $n_0(f)$ seiner (verschiedenen) Nullstellen ein *Polynom* $N_0(f)$ dazu, das genau diese Nullstellen mit jeweils der Vielfachheit eins haben soll: Für

$$f = f_0\prod_{i=1}^{r}(x-a_i)^{n_i} \quad \text{sei} \quad N_0(f) \underset{\text{def}}{=} \prod_{i=1}^{r}(x-a_i)\,,$$

so daß der Grad von $N_0(f)$ gerade die im vorigen Paragraphen definierte Zahl $n_0(f)$ ist.

Der Vorteil des Polynoms $N_0(f)$ besteht darin, daß wir eine analoge Definition leicht auch für natürliche Zahlen hinschreiben können: Für

$$n = \prod_{i=1}^{r}p_i^{e_i} \quad \text{setzen wir} \quad N_0(n) \underset{\text{def}}{=} \prod_{i=1}^{r}p_i\,.$$

Mit Hilfe der Polynome $N_0(f)$ läßt sich der Satz von MASON folgendermaßen umformulieren:

Gilt für drei teilerfremde Polynome f, g und h die Gleichung $f+g=h$, so hat jedes der drei Polynome einen kleineren Grad als das Polynom $N_0(fgh)$.

In dieser Formulierung kommt immer noch der Grad vor, für den wir bei natürlichen Zahlen keine Verwendung haben. Betrachten wir aber den Grad lediglich als eine Methode, einem Polynom eine Zahl aus $\mathbb{N}_0$ zuzuordnen, so können wir, wenn wir bereits natürliche Zahlen haben, einfach ganz auf ihn verzichten; falls wir ganze Zahlen betrachten, liegt es nahe, ihn durch den Betrag zu ersetzen.

Gemäß dieser Philosophie können wir nun probeweise die folgende Aussage formulieren:

A1: *Ist $c = a+b$ für drei zueinander teilerfremde natürliche Zahlen a, b, c, so ist jede der drei Zahlen kleiner als $N_0(abc)$.*

Damit haben wir eine sinnvolle Aussage über natürliche Zahlen gefunden, die – falls sie korrekt ist – sofort die FERMAT-Vermutung impliziert: Gibt es nämlich drei natürliche Zahlen x, y, z mit der Eigenschaft, daß $x^n+y^n=z^n$ für ein $n \geq 3$, so gibt es auch drei zueinander teilerfremde Zahlen x, y, z mit dieser Eigenschaft: Wir müssen einfach die drei Zahlen durch ihren größten gemeinsamen Teiler kürzen. Alsdann muß, falls obige Aussage richtig ist, jede der drei Potenzen x^n, y^n, z^n kleiner sein als $N_0(x^n y^n z^n)$. Nun ist aber

$$N_0(x^n y^n z^n) = N_0(xyz) \leq xyz\,,$$

d.h. jede der drei Zahlen x^n, y^n, z^n wäre kleiner als xyz. Damit wäre

$$(xyz)^n = x^n y^n z^n < (xyz)^3\,,$$

was für $n \geq 3$ offensichtlich nicht möglich ist.

Angesichts der Komplexität des WILESschen Beweises fällt es schwer, an einen so einfachen Beweis zu glauben, und in der Tat ist die obige Aussage in dieser Form falsch:

Betrachten wir etwa die Gleichung $8+1=9$. Offensichtlich sind die drei Summanden teilerfremd zueinander, aber sowohl 8 als auch 9 sind größer als $N_0(8\cdot 1\cdot 9) = 2\cdot 3 = 6$. Ganz so einfach geht es also nicht.

Da der Grad eines Polynoms nicht durch konstante Faktoren beeinflußt wird, könnte man versuchen, als „richtiges“ Analogon zum Satz von MASON eine abgeschwächte Aussage zu nehmen, die nur eine Abschätzung bis auf einen konstanten Faktor enthält, etwa

A2: *Ist $c = a+b$ für drei zueinander teilerfremde natürliche Zahlen a, b, c, so gibt es eine Konstante K derart, daß jede der drei Zahlen kleiner ist als $K \cdot N_0(abc)$.*

Diese Aussage ist trivialerweise richtig: Wir müssen nur eine Konstante K wählen, die größer ist als das Maximum von a, b und c. Leider ist sie auch völlig nutzlos, denn solange die Konstante von a, b und c abhängen darf, haben wir keine Chance, damit die FERMAT-Vermutung zu beweisen.

Wir müssen die Aussage also noch einmal umformulieren:

A3: *Es gibt eine Konstante K, so daß gilt: Ist $c = a+b$ für drei zueinander teilerfremde natürliche Zahlen a, b, c, so ist jede der drei Zahlen kleiner als $K \cdot N_0(abc)$.*

Wie wir gleich sehen werden, würde hieraus die FERMAT-Vermutung zumindest für alle hinreichend großen Exponenten n folgen, allerdings ist die Aussage, so wie sie dasteht, leider immer noch falsch:

Betrachten wir die Gleichung

$$a_n + b_n = c_n \quad \text{mit} \quad a_n = 3^{2^n} - 1, \quad b_n = 1 \quad \text{und} \quad c_n = 3^{2^n}. \quad (*)$$

Wäre sie richtig, müßte für jedes n gelten:

$$3^{2^n} \le K N_0\big((3^{2^n} - 1)3^{2^n}\big) = K \cdot 3 \cdot N_0(3^{2^n} - 1)\,.$$

Um $N_0(3^{2^n} - 1)$ abzuschätzen, beachten wir, daß gilt

$$3^{2^n} = \big(3^{2^{n-1}}\big)^2 \quad \text{und} \quad 3^{2^n} - 1 = \big(3^{2^{n-1}} + 1\big)\big(3^{2^{n-1}} - 1\big)$$

nach der dritten binomischen Formel. Wenden wir dies mehrfach an, erhalten wir

$$\begin{aligned} 3^{2^n} - 1 &= \big(3^{2^{n-1}} + 1\big)\big(3^{2^{n-1}} - 1\big) \\ &= \big(3^{2^{n-1}} + 1\big)\big(3^{2^{n-2}} + 1\big)\big(3^{2^{n-2}} - 1\big) \\ &= \big(3^{2^{n-1}} + 1\big)\big(3^{2^{n-2}} + 1\big)\big(3^{2^{n-3}} + 1\big)\big(3^{2^{n-3}} - 1\big) \\ &= \cdots \\ &= \big(3^{2^{n-1}} + 1\big)\big(3^{2^{n-2}} + 1\big) \ldots \big(3^2 + 1\big)\big(3^1 + 1\big)\big(3^1 - 1\big)\,. \end{aligned}$$

In der letzten Zeile steht ein Produkt aus $n+1$ geraden Zahlen; somit ist $3^{2^n} - 1$ durch 2^{n+1} teilbar. Das Produkt $N_0(3^{2^n} - 1)$ aller *verschiedener* Primteiler von $3^{2^n} - 1$ erfüllt daher die Ungleichung

$$N_0(3^{2^n} - 1) \le 2 \cdot \frac{3^{2^n} - 1}{2^{n+1}} = \frac{3^{2^n} - 1}{2^n}\,,$$

denn das Produkt aller ungerader Primteiler kann höchstens gleich $(3^{2^n} - 1)/2^n$ sein.

Falls **A3** korrekt wäre, müßte nach Gleichung $(*)$ also gelten

$$3^{2^n} \le \frac{3K}{2^n}(3^{2^n} - 1) \quad \text{für alle } n\,.$$

Das kann aber unmöglich der Fall sein, denn für hinreichend große n ist der Faktor $\frac{3K}{2^n}$ kleiner als eins, so daß 3^{2^n} echt kleiner als sich selbst sein müßte.

Auf der Suche nach einem Analogon für den Satz von MASON müssen wir daher noch weiter abschwächen. *Eine* Möglichkeit dazu ist die 1986 aufgestellte

abc-Vermutung von MASSER und OESTERLÉ: Zu jedem $\varepsilon > 0$ gibt es eine Konstante $K(\varepsilon)$, so daß für alle teilerfremden natürlichen Zahlen a, b, c mit $a+b = c$ gilt: Jede der drei Zahlen a, b, c ist kleiner oder gleich $K(\varepsilon) \cdot N_0(abc)^{1+\varepsilon}$.

Diese Vermutung ist, im Gegensatz zur FERMAT-Vermutung, bis heute offen.

Wir wollen uns überlegen, daß sie zumindest für große Exponenten n die FERMAT-Vermutung impliziert.

Dazu betrachten wir eine Lösung $x^n + y^n = z^n$ mit o.B.d.A. teilerfremden natürlichen Zahlen x, y, z und wählen uns irgendein $\varepsilon > 0$. Nach der abc-Vermutung gibt es dazu eine Konstante $K(\varepsilon)$, so daß x^n, y^n und z^n allesamt höchstens gleich

$$K(\varepsilon)N_0(x^n y^n z^n)^{1+\varepsilon} = K(\varepsilon)N_0(xyz)^{1+\varepsilon} \leq K(\varepsilon)(xyz)^{1+\varepsilon}$$

sind. Für ihr Produkt gilt daher

$$x^n y^n z^n \leq K(\varepsilon)^3 (xyz)^{3(1+\varepsilon)} \quad \text{oder} \quad (xyz)^{n-3-3\varepsilon} \leq K(\varepsilon)^3 .$$

$K(\varepsilon)^3$ ist eine feste Zahl; es gibt daher einen Exponenten m derart, daß $2^m > K(\varepsilon)^3$ ist. Da das Produkt xyz auf jeden Fall nicht kleiner als zwei sein kann, ist dann für $n - 3 - 3\varepsilon > m$ oder $n > m + 3 + 3\varepsilon$ insbesondere

$$(xyz)^{n-3-3\varepsilon} > K(\varepsilon)^3 .$$

Für Exponenten $n > m + 3 + 3\varepsilon$ kann daher die FERMAT-Gleichung keine Lösung in natürlichen Zahlen haben.

Ob und gegebenenfalls welche konkreten Schranken für n man damit erreichen kann, hängt natürlich davon ab, wie $K(\varepsilon)$ von ε abhängt. Dazu gibt es im Augenblick nicht einmal Vermutungen.

Für weitere Informationen zu §3 und S4 sei auf einen Vortrag verwiesen, den SERGE LANG in Zürich vor einem „allgemeinen“ Publikum hielt und dem ich hier im wesentlichen gefolgt bin:

SERGE LANG: Die abc-Vermutung, *Elemente der Mathematik* **48** (1993), 89-99

Der Artikel ist (wie die gesamte Zeitschrift *Elemente der Mathematik)* unter `http://www.bibliothek.uni-regensburg.de/ezeit/?2135837` frei zugänglich.

§5: Die Frey-Kurve

Da die FERMAT-Vermutung seit 1994 bewiesen ist, die abc-Vermutung aber immer noch offen, mußte der Beweis der FERMAT-Vermutung

natürlich andere Wege gehen. Die meisten dieser Wege führen in Gebiete, die weit jenseits dessen liegen, was selbst ein guter auf Zahlentheorie spezialisierter Diplom-Mathematiker im Laufe seines Studiums lernen kann, aber zumindest die Grundidee der abc-Vermutung, daß man nämlich Summenbeziehungen zwischen großen Zahlen nicht ohne ein gewisses Minimum an verschiedenen Primfaktoren realisieren kann, spielt in modifizierter Weise in der Tat eine große Rolle.

Der Anstoß kam 1984 von GERHARD FREY, damals Professor an der Universität Saarbrücken, wo er auf dem Gebiet der Arithmetik elliptischer Kurven arbeitete. Heute leitet er die Arbeitsgruppe Zahlentheorie am Institut für experimentelle Mathematik der (inzwischen mit Duisburg vereinigten) Universität Essen und beschäftigt sich mit der Anwendung elliptischer (und anderer) Kurven in der Kryptologie.

Elliptische Kurven sind ebene Kurven, die durch eine Gleichung der Form $y^2 = f_3(x)$ beschrieben werden mit einem Polynom $f_3(x)$ vom Grad drei mit drei verschiedenen Nullstellen. Da das Quadrat einer reellen Zahl nicht negativ sein kann, gibt es im Reellen nur Punkte mit x-Koordinaten, für die $f_3(x) \geq 0$ ist. Im Falle $f_3(x) > 0$ erfüllt mit y auch $-y$ die obige Gleichung, die Kurve ist also symmetrisch zur x-Achse.

Falls $f_3(x)$ nur zwei verschiedene Nullstellen hat, muß eine der Nullstellen doppelt sein, und bei diesem x-Wert überkreuzt sich die Kurve; wir reden dann von einer Knotenkurve.

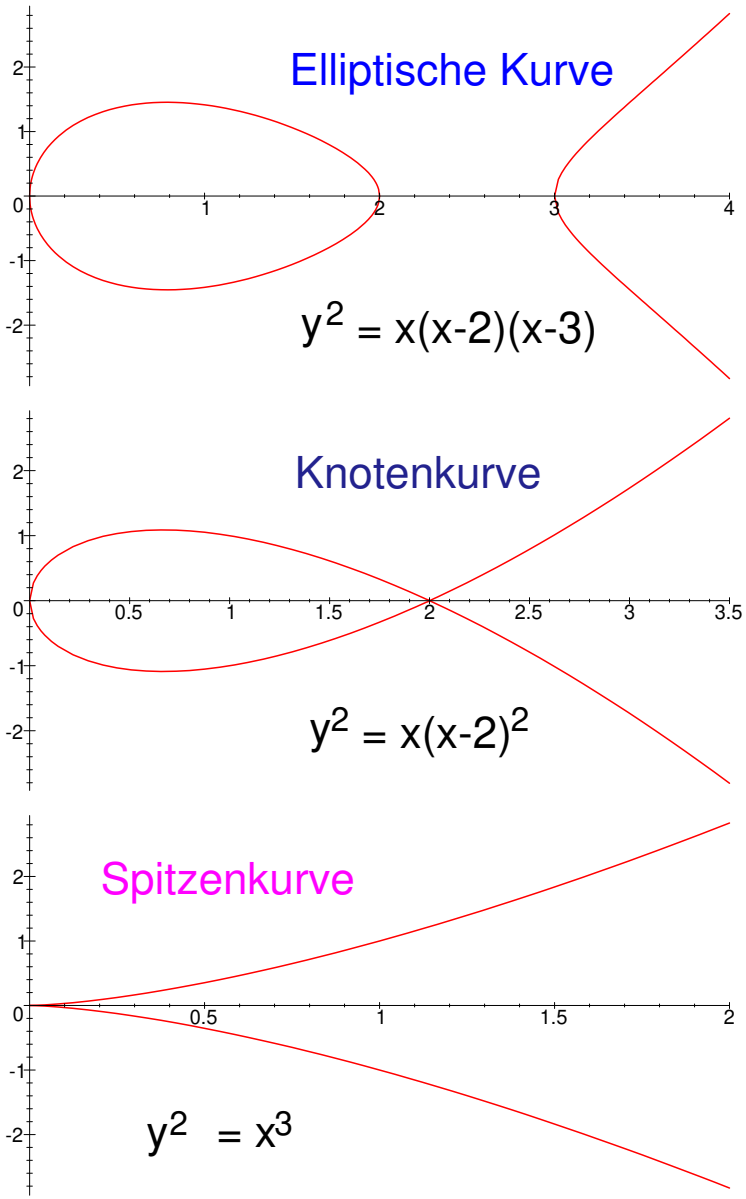
Hat schließlich $f_3(x)$ nur eine, dafür aber dreifache Nullstelle, entsteht eine Spitzenkurve.

FREY betrachtete eine (hypothetische) Lösung

$$x^n + y^n = z^n$$

der FERMAT-Gleichung mit teilerfremden natürlichen Zahlen x, y, z und $n \geq 5$. (Den Fall $n = 4$ hat, wie wir gesehen haben, bereits FERMAT selbst gelöst, den Fall $n = 3$ nicht viel später EULER.) Wenn es eine solche Lösung gibt, dann gibt es auch eine Lösung für einen Primzahlexponenten ℓ, denn ist ℓ ein Primteiler von n und $n = \ell m$, so ist auch

$$a^\ell + b^\ell = c^\ell \quad \text{mit} \quad a = x^m, \quad b = y^m \quad \text{und} \quad c = z^m$$



eine Lösung, und auch a, b, c sind teilerfremd. Auch für ℓ genügt es, den Fall $\ell \geq 5$ zu betrachten, denn wenn wir für ℓ den größten Primteiler von n nehmen, bedeutet $\ell = 2$, daß n eine Zweierpotenz sein muß, was für $n = 2$ kein Widerspruch zur FERMAT-Vermutung ist und für $n = 4$ und damit auch jede höhere Zweierpotenz nach FERMATs Beweis ausgeschlossen ist. Für den Fall $\ell = 3$ kann wieder auf EULER verwiesen werden.

Zur obigen Lösung definiert FREY die elliptische Kurve

$$y^2 = x(x - a^\ell)(x + b^\ell)$$

zu, die er aber nicht nur über den reellen oder komplexen Zahlen betrachtet, sondern auch über den ganzen Zahlen modulo einer Primzahl p:

Ist allgemein

$$y^2 = (x - x_1)(x - x_2)(x - x_3)$$

eine Kurvengleichung mit ganzen Zahlen x_1, x_2, x_3, so können wir für x und y auch ganze Zahlen einsetzen und diese Gleichung modulo p betrachten. Wir sprechen wieder von einer elliptischen Kurve, einer Knotenkurve oder einer Spitzenkurve je nachdem wie viele der Nullstellen x_1, x_2 und x_3 modulo p noch verschieden sind.

Die obige Gleichung definiert genau dann eine elliptische Kurve, wenn alle drei Nullstellen verschieden sind, wenn also die sogenannte Diskriminante

$$\Delta = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$$

von Null verschieden ist. Modulo p definiert sie eine elliptische Kurve, wenn Δ auch modulo p noch von Null verschieden ist, wenn also p kein Teiler von Δ ist.

Speziell für die FREY-Kurve $y^2 = x(x - a^\ell)(x + b^\ell)$ ist die Diskriminante

$$\Delta = (0 - a^\ell)(0 - b^\ell)\left(a^\ell - (-b)^\ell\right) = a^\ell b^\ell (a^\ell + b^\ell) = a^\ell b^\ell c^\ell = (abc)^\ell$$

stets von Null verschieden; modulo p verschwindet sie genau dann, wenn p ein Teiler von Δ ist, d.h., wenn p eine der drei Zahlen a, b, c teilt.

Da die Diskriminante als ℓ-te Potenz von abc verglichen mit a, b, c ziemlich groß ist, heißt das, daß es im Verhältnis zur Größe der Diskriminante

erstaunlich wenige Primzahlen gibt, modulo derer wir *keine* elliptische Kurve erhalten; wir sind also wieder einer ähnlichen Situation wie bei der abc-Vermutung. Die FREYsche Kurve sieht damit so aus, als sei sie fast zu schön, um wirklich zu existieren.

Einen Anhaltspunkt zum Beweis dieser Nichtexistenz liefert eine Vermutung, die auf um 1955 durchgeführte Rechnungen und Spekulationen des japanischen Mathematikers TANIYAMA zurückgeht und heute je nach Autor mit irgendeiner Kombination der drei Namen TANIYAMA, SHIMURA und WEIL bezeichnet wird. Danach sollte es zu einer elliptischen Kurve E mit ganzzahligen Koeffizienten eine surjektive Abbildung $X_0(N) \to E$ von einer sogenannten Modulkurve $X_0(N)$ auf E geben, wobei N im wesentlichen das Produkt aller Primzahlen p ist, modulo derer E keine elliptische Kurve mehr ist. Wie FREYs Rechnungen zeigen, hat seine Kurve vor diesem Hintergrund sehr seltsame Eigenschaften.

Als er damals hier in Mannheim über seine Resultate vortrug, meinte er noch, er glaube nicht, daß die FERMAT-Vermutung so bewiesen werde; er veröffentlichte sein Ergebnis auch nicht in einer der großen internationalen Fachzeitschriften, sondern als Band 1, Heft 1 einer gerade neu gestarteten Schriftenreihe der Universität Saarbrücken, in einfachster Aufmachung xerographiert mit einem nur schwarz-weiß gestalteten Karton als Umschlag:

GERHARD FREY: Links between stable elliptic curves and certain diophantine equations, *Annales Universitatis Saraviensis, Series Mathematicae,* **1** *(1), 1986*

1987 verschärfte der französische Mathematiker JEAN-PIERRE SERRE die TANIYAMA-Vermutung, und aus dieser stärkeren Vermutung folgt in der Tat, daß die FREY-Kurve nicht existieren kann. Leider ist die SERREsche Vermutung bis heute noch nicht bewiesen.

SERRE erhielt übrigens 2002 den ersten der vom norwegischen Parlament gestifteten ABEL-Preise, die seither zur Erinnerung an den norwegischen Mathematiker NIELS HENRIK ABEL (1802–1829) jedes Jahr in gleicher Weise und gleicher Ausstattung wie die Nobel-Preise für hervorragende Leistungen auf dem Gebiet der Mathematik vergeben werden.

SERRE stellte jedoch noch zusätzlich seine sogenannte ε-Vermutung auf,

und auch aus der TANIYAMA-Vermutung zusammen mit der ε-Vermutung folgt die Nichtexistenz der FREY-Kurve und damit die FERMAT-Vermutung. Diese ε-Vermutung bewies KENNETH RIBET von der Universität Berkeley 1990. Die Grundidee seines Beweises läßt sich interpretieren als eine Art zweidimensionale Version eines Beweises von ERNST EDUARD KUMMER (1810–1893), der die FERMAT-Vermutung 1846 für sogenannte reguläre Primzahlen als Exponenten bewies. (Eine Primzahl p heißt regulär, wenn die Hauptordnung des Körpers $\mathbb{Q}[\zeta_p]$ der p-ten Einheitswurzeln faktoriell ist.) Der Beweis von RIBET ist allerdings erheblich aufwendiger.

Damit war also die FERMAT-Vermutung zurückgeführt auf die TANIYAMA-Vermutung. Diese Vermutung schließlich (die für die weitere mathematische Forschung erheblich wichtiger ist als die FERMAT-Vermutung) bewies WILES 1994.