Wolfgang K. Seiler

Zahlentheorie

Vorlesung an der Universität Mannheim
im Frühjahrssemester 2007

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

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

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

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

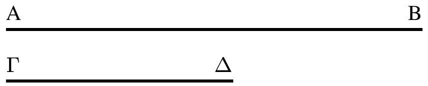
Inhalt

Kapitel 1
Ganze Zahlen und ihre Primzerlegung

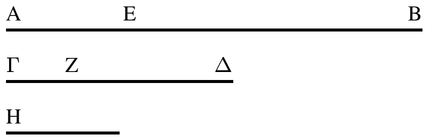
§ 1: Der Euklidische Algorithmus

Bei EUKLID, in Proposition 2 des siebten Buchs seiner *Elemente,* wird er 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 und EUKLID schließlich als *der* Stoichist bekannt wurde 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 Ergebniss 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} diesselben 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.

§2: 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
$$\operatorname{ggT}(a, b) = r_{i-1} = \alpha_{i-1}a + \beta_{i-1}b\,.$$
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
$$\begin{aligned} r_{i+1} = q_i r_i - r_{i-1} &= q_i(\alpha_i a + \beta_i b) - (\alpha_{i-1}a + \beta_{i-1}b) \\ &= (q_i\alpha_i - \alpha_{i-1})a + (q_i\beta_i - \beta_{i-1})b\,; \end{aligned}$$
man setze also
$$\alpha_{i+1} = q_i\alpha_i - \alpha_{i-1} \quad \text{und} \quad \beta_{i+1} = q_i\beta_i - \beta_{i-1}\,.$$

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 ist 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 schon 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. $-\frac{237}{8} \le k \le -\frac{316}{11}$. 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 kann, 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

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

§ 3: 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 $\operatorname{ggT}(a, b)$: 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; wenn 10^{600} eine realistische Aufwandsabschätzung wäre, so wäre es hoffnungslos, an eine Anwendung des EUKLIDischen Algorithmus auf 600-stellige Zahlen auch nur zu denken.

Tatsächlich ist 10^{600} aber natürlich nur eine obere Schranke, von der wir bislang noch nicht wissen, ob sie realistisch 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 Ingenierswissenschaften 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–S^{t}. 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

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

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

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

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

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 also nicht (wie oben bei der naiven Abschätzung) mit b, sondern nur mit $\log b$. 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,* ²1981

§4: 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. ■

§5: 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 $x, y \in \mathbb{Z}$ sind 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 $\times: G \times G \to G$, 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 $\varphi: G \to H$ zwischen zwei Gruppen G und H mit Verknüpfungen $\times$ und $\otimes$ heißt 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 $\varphi: G \to H$ gibt.

c) Ein Ring ist eine Menge R zusammen mit zwei Verknüpfungen $+, \cdot: R \times R \to R$, so daß gilt

1.) Bezüglich + ist R eine abelsche Gruppe.

2.) $(x \cdot y) \cdot z = x \cdot (y \cdot z)$ ür alle $x, y, z \in R$.

3.) Es gibt ein Element $1 \in R$, so daß $1 \cdot x = x \cdot 1$ für alle $x \in R$.

4.) $x(y + z) = xy + yz$ und $(x + y)z = xz + yz$ 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 $\varphi: R \to S$ 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 $\varphi: R \to S$ 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

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

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.

§6: 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 $x+M$ offensichtlich 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 $x \bmod 2$ 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 $0 \le x < m_1 \cdots m_r$. 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 $x \bmod m_i$ 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

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

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 $x \bmod m_1 \cdots m_r$ 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

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

also löst

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

das Problem.

Es 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 somit

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

insbesondere ist die Lösung eindeutig modulo nm.

Als Beispiel betrachten wir die beiden Kongruenzen

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

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 $9 \cdot 17 = 153 \equiv 0 \bmod 17$ und $\equiv 1 \bmod 19$; außerdem ist $-8 \cdot 19 = -152$ durch 19 teilbar und $\equiv 1 \bmod 17$. 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 $x \equiv a_1 \bmod m_1$ und $x \equiv a_2 \bmod m_2$ 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

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

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

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

lösen und die Lösung schreiben als

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

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

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

lösen wir also zunächst nur das System

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

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

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

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 $17 \cdot 13 = 221 \equiv 1 \bmod 110$ und $\equiv 0 \bmod 13$; genauso ist $-2 \cdot 110 = 220 \equiv 1 \bmod 13$ und $\equiv 9 \bmod 110$. 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

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

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.

§7: 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 $\mathrm{ggT}(a, m) = 1$ 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 $\mathrm{ggT}(a, m) = 1$ 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 $x \bmod m$ 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 $\varphi(m) = \prod_{i=1}^{r} \bigl(p_i^{e_i-1}(p_i - 1)\bigr)$.

Beweis: a) Eine Zahl a ist genau dann teilerfremd zum Produkt nm, wenn $a \bmod n$ teilerfremd zu n und $a \bmod m$ 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 $g, gh, \ldots, gh^{r-1}$. Da G die Vereinigung aller Äquivalenzklassen ist, muß die Gruppenordnung somit ein Vielfaches von e 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 Primzahltests werden wir uns näher damit beschäftigen.

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 symmetrischen Verfahrens besteht darin, daß in einem Netzwerk jeder Teilnehmer mit jedem anderen einen Schlüssel vereinbaren muß. In militärischen Netzen war dies traditionellerweise 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 dies 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 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 jeden anderen so zu kommunizieren, und diese 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 zudem noch durch ein sicheres Verfahren wie etwa ein 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. Seine dortige home page hat den URL `http://research.sun.com/people/diffie/` .

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 Schritten 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. DIFFIE und HELLMAN bezeichnen eine Funktion, deren Umkehrfunktion nicht mit vertretbarem Aufwand berechnet werden kann, als *Einwegfunktion* und schlagen als Verschlüsselungsfunktion eine solche Einwegfunktion vor.

Damit hat man aber noch kein praktikables 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 *Goverment 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 Massachussetts 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.

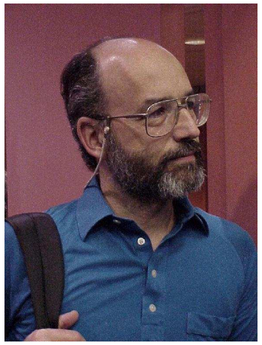
RIVEST, SHAMIR und ADLEMAN gründeten eine Firma namens RSA Computer Security Inc., die 1983 das RSA-Verfahren patentieren ließ und auch nach Auslaufen dieses Patents im September 2000 weiterhin erfolgreich im Kryptobereich tätig ist. 2002 erhielten RIVEST, SHAMIR und ADLEMAN für die Entdeckung des RSA-Systems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

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.

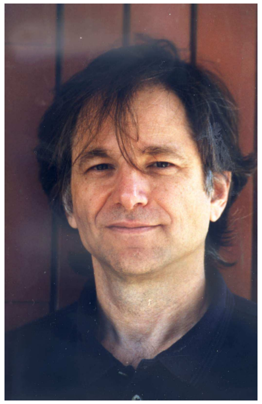
§2: Das RSA-Verfahren

Für eine natürliche Zahl N ist die Funktion $x \mapsto x^e \bmod N$ fast genauso einfach zu berechnen wie wie die Funktion $x \mapsto x^e$, ist aber erheblich chaotischer. Damit ist zumindest vorstellbar, daß diese Funktion Grundlage einer kryptographischen Verschlüsselung sein könnte.

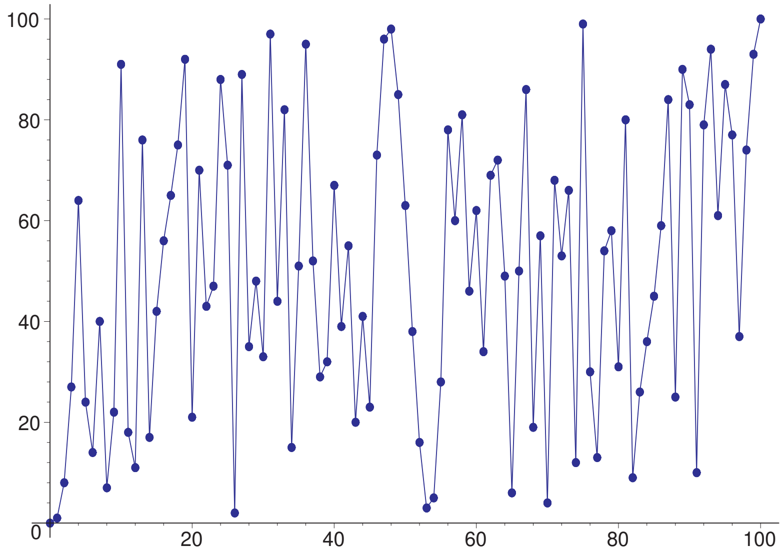
RONALD LINN RIVEST wurde 1947 in Schenectady im US-Bundesstaat New York geboren. Er studierte zunächst Mathematik an der Yale University, wo er 1969 seinen Bachelor bekam; danach studierte er in Stanford Informatik. Nach seiner Promotion 1974 wurde er Assistenzprofessor am Massachussetts Institute of 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 ist arbeitet er an der University of Southern California in Los Angelos. 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`



Die Funktion $x \mapsto x^3$ in $\mathbb{F}_{101}$

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 so 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 $\varphi(p_i)$. 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 Null das einzige Element von $\mathbb{Z}/p$, das keine prime Restklasse hat, und es wird von beiden Funktionen auf sich selbst abgebildet.

■

Jeder, der e und N kennt, kann damit auch d berechnen, allerdings muß er dazu als erstes $\varphi(N)$ bestimmen. Nach der Formel aus Kapitel 1, §6 ist das möglich, wenn er die Primfaktorzerlegung von N kennt. Einfachere alternative Verfahren sind nicht bekannt, und wie wir in Kapitel sechs sehen werden, ist diese Primfaktorisierung für hinreichend große Zahlen N mit den derzeit bekannten Algorithmen nicht mit realistischem Aufwand zu ermitteln.

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, denn wenn man Summe und Produkt zweier Zahlen kennt, kann man auch die Zahlen selbst durch Lösen einer quadratischen Gleichung bestimmen.

Zur praktischen Durchführung des RSA-Verfahrens wählt sich 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 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 den Chiffretext $b \equiv a^e \bmod N$, der als Zahl zwischen null und $N-1$ interpretiert und an den Inhaber des geheimen Schlüssels geschickt wird. Dieser berechnet $b^d \bmod Na^{ed} \bmod N = a$, und da er dazu seinen geheimen Schlüssel braucht, kann dies niemand außer ihm.

In der Praxis wird 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 einfach ist. 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

$$d_k = \frac{1 + \varphi(N)}{e}$$

für diese beiden Werte berechnen; sobald man ein ganzzahliges Ergebnis bekommt, ist d gefunden.

Der private Exponent d wird und sollte fast immer in der Größenordnung von N sein; im nächsten Kapitel werden wir sehen, daß N leicht faktorisiert werden kann, wenn d zu klein ist.

Bei der Berechnung von $x^d \bmod N$ und $x^e \bmod N$ 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 wir können es 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) berechnen.

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 beispielweise 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 Hälte der Ziffernfolge identisch mit der hinteren Hälfte ist. Ohne Kenntnis von d hat man cwpratksich 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 Schlüsseln (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 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. Bislang wurden dazu meist spezielle kryptographische Hash-Funktionen verwendet, die einen 160 Bit langen Wert lieferten, jedoch ist deren Sicherheit inzwischen umstritten, so daß aufwendigere Funktionen, die Hashwerte von ca. 256 Bit liefern, ratsam erscheinen.

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 Nachrichten von A an B lesen, und er kann sich gegenüber B mittels elektronischer Unterschrift als A ausgeben. Daher sind öffentliche Schlüssel meist unterschrieben von einer Zertifizierungsstelle, deren elektronische Unterschrift jeder Teilnehmer kennt (weil sie beispielsweise im Mail- oder Browserprogramm eingebaut ist).

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. Andererseit dürfen die Seriennummern aber auch nicht von der Bank vergeben werden, denn sonst wüßte diese, welcher Kunde Scheine mit welchem 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.

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 authentizieren, 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 Authentizierungsprozedur 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:

$$\begin{aligned} &2135987035920910082395022704999628797051095341 82\backslash\\ &641740644252416500858395774644508840500943086 5999\\ =\ &1113954325148827987925490175477024844070922844843\\ \times\ &1917481702524504439375786268230862180696934189293 \end{aligned}$$

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.

Seit November 1999 haben neu ausgegebene Bankkarten nun noch ein zusätzliches Feld mit einer Unterschrift, die im Gegensatz zum obigen 320-Bit-Modul einen 768-Bit-Modul verwendet. Natürlich kann es 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 Kryptoalgorithmen 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 Signaturgesetztes 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 gütlig sein sollen, sind elektronische Unterschriften also nicht vorgesehen.

Offiziell geht bei den Empfehlungen allgemein um geeignete Algorithmen für elektronische Unterschriften sowie deren Schlüssellängen, aber wie die Entwicklung der letzten Jahre zeigte, drehen sich die Diskussionen, die zu den jeweiligen Empfehlungen führen, tatsächlich fast ausschließlich um die jeweils notwendige Schlüssellänge für RSA.

Natürlich hat in einer Demokratie bei so einer wichtigen Frage auch die Bevölkerung ein Mitspracherecht; deshalb beginnt das BSI jeweils zunächst einen Entwurf, zu dem es um Kommentare bittet; erst einige Monate später wird die endgültige Empfehlung verkündet und im Bundesanzeiger veröffentlicht.

Die interessierte Öffentlichkeit, von der die Kommentare zu den Entwürfen kommen, besteht einerseits aus Anbietern von Hardware und Software für Kryptographie, und als erfahrene Experten für Datensicherheit wissen diese, daß ein Verfahren nur dann wirklich geeignet sein kann, wenn es die eigene Firma im Angebot hat. (Am geeignetsten sind natürlich die Verfahren, die keines der Konkurrenzunternehmen anbietet.)

Andererseits melden sich die Anwender von Kryptoverfahren zu Wort; vor allem sind das Vertreter der Dachverbände des Kreditgewerbes. Diese müssen für eine starke Kryptographie eintreten, denn falls die Kryptographie einer von ihnen ausgegebenen Chipkarte geknackt wird, könnte das für ihre Mitglieder sehr teuer werden. Teuer wird es aber auch, wenn Chipkarten vor Ablauf ihrer Gültigkeit ausgetauscht werden müssen, weil sie nicht mehr den aktuellen Anforderungen entsprechen. Da Chipkarten ein bis zwei Jahre vor Ausgabe in Auftrag gegeben werden müssen und dann im allgemeinen drei Jahre lang gültig sind, versucht dieser Teil der Öffentlichkeit vor allem, die von den Kryptologen für notwendig erachteten Änderungen um ein bis zwei Jahre hinauszuzögern.

Das endgültige Ergebnis ist dann ein Kompromiss zwischen den verschiedenen Positionen.

So ist beispielsweise zu erklären, daß es vieler Anläufe bedurfte, um die Schlüssellänge für RSA auf einen Wert über 1024 Bit zu bringen, denn es gab viele Chips mit Hardware-Implementierungen von RSA für Schlüssellängen von bis zu 1024 Bit, während größere Schlüssellängen zunächst vor allem in *public domain* Software wie PGP zu finden waren.

Bis Ende 2000 galten 768 Bit als ausreichende Größe für das Produkt N der beiden Primzahlen, jener Wert also, den die *neueren* französischen Bankkarten verwenden und den ebenfalls nur die neueren Terminals

lesen können. Schon in den Richtlinien für 1998 wurden 768 Bit jedoch ausdrücklich nur übergangsweise zugelassen; längerfristig, d.h. bei Gültigkeit über 2000 hinaus, waren mindestens 1024 Bit vorgeschrieben.

Die Richtlinien für 2000 erlaubten die 768 Bit ebenfalls noch bis zum Ende des Jahres; für Dokumente mit einer längeren Gültigkeit verlangten sie bis Mitte 2005 eine Mindestgröße von 1024 Bit, danach bis Ende 2005 sogar 2048 Bit.

Anbieterproteste führten dazu, daß nach den Richtlinien von 2001 eine Schlüssellänge von 1024 dann doch noch bis Ende 2006 sicher war; die Schlüssellänge 2048 war nur noch „empfohlen“, also nicht mehr verbindlich.

Im April 2002 erschien der erste Entwurf für die 2002er Richtlinien; darin war für 2006 und 2007 nur eine Mindestlänge von 2048 Bit wirklich sicher. Einsprüche führten im September 2002 zu einem revidierten Entwurf, wonach 2006 doch noch 1024 Bit reichen, 2007 aber mindestens 1536 notwendig werden. Die Mindestlänge von 2048 Bit wurde wieder zur „Empfehlung“ zurückgestuft.

Am 2. Januar 2003 erschienen endlich die offiziellen Richtlinien des Jahres 2002; veröffentlich wurden sie am 11. März 2003 im Bundesanzeiger Nr. 48, S. 4202–4203. Danach reichen 1024 Bit auch noch bis Ende 2007, erst 2008 werden 1280 Bit erforderlich. Die 2048 Bit blieben dringend empohlen.

Nach diesem großen Kraftakt erschienen 2003 keine neuen Richtlinien mehr; erst für 2004 gab es am 2. Januar 2004 neue Empfehlungen (Bundesanzeiger Nr. 30 vom 13. Februar 2004, S. 2537–2538). Für den Zeitraum bis Ende 2008 wurden die alten Empfehlungen beibehalten, bis Ende 2009 aber 1536 Bit gefordert. Die nächsten Richtlinien für 2005 sahen in ihrem ersten Vorentwurf 2048 Bit bis Ende 2010 vor; nach Einsprüchen der Banken, daß das Betriebssystem SECCOS der heute üblichen Chipkarten nur mit maximal 1984 Bit-Schlüsseln umgehen kann, wurde die Länge im zweiten Entwurf auf 1984 gesenkt; in den endgültigen Richtlinien vom 2. Januar 2005 waren es schließlich nur noch 1728.

Die neuesten Richtlinien stammen vom 12. April 2007 (Bundesanzeiger Nr. 69 S. 3759). Sie empfehlen grundsätzlich schon heute 2048 Bit, aber wirklich verbindlich sind

bis Ende	2007	2008	2009	2010	2012	
Minimallänge	1024	1280	1536	1728	1976	*Bit.*

(1976 unterscheidet sich nicht wesentlich von 2048; der minimal kleinere Wert wurde in Hinblick auf die oben erwähnten Probleme mit SECCOS gewählt.)

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

$$\varepsilon_1 = 0{,}1 \quad \text{und} \quad \varepsilon_2 = 30$$

vorgeschlagen; ist p die kleinere der beiden Primzahlen, soll also

$$2^{-10}p < q < 2^{30}p \approx 10^9 p$$

gelten, d.h. die beiden Primzahlen sollten 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 sechs noch genauer damit befassen.

§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.

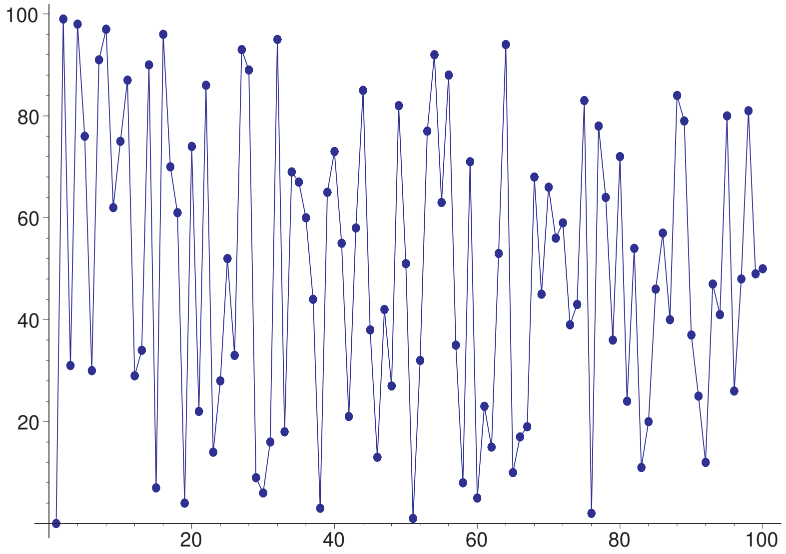
Ausgangspunkt ist wieder das Potenzieren im Körper $\mathbb{F}_p$; hier betrachten wir aber die Exponentialfunktion $x \mapsto a^x$ zu einer geeigneten Basis a. Ihre Umkehrfunktion bezeichnet man als *Index* oder *diskreten Logarithmus* zur Basis a:

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

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 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 Zweierpo-



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

tenzen, und jeder Fall erfordert einen neue 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.

Falls sich allerdings die sogenannten *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.

§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.

Aus Sicht der amerikanischen Behörden hat DSA gegenüber RSA und Verfahren wie DIFFIE-HELLMAN vor allem einen großen Vorteil: Es läßt sich *nur* für elektronische Unterschriften benutzen, nicht zur Verschlüsselung.

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, die im ursprünglichen Standard eine Länge von mindestens 160 Bit haben sollte. Laut Bundesnetzagentur sollte diese Länge auch noch bis Ende 2009 ausreichen, bis Ende 2012 sind allerdings nach dem Entwurf für 2007 mindestens 224 Bit vorgeschrieben, was wahrscheinlich mehr mit den verwendeten Hashfunktionen als mit der Sicherheit der Unterschrift zu tun hat.

Zu dieser Primzahl q sucht man eine Primzahl $p \equiv 1 \bmod q$, für deren Länge die Bundesnetzagentur bis Ende 2007 mindestens 1024 Bit vorschreibt, bis Ende 2008 mindestens 1280, bis Ende 2009 mindestens 1536 und bis Ende 2012 mindestens 2048. „Empfohlen“ sind auch hier 2048 Bit.

Daß diese Zahlen (bis auf die unwesentliche Differenz zwischen 2048 und 1976) mit den RSA-Modullängen für die entsprechenden Jahre übereinstimmen, 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 die auch 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 160 *bzw.* 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 160 Bit lagen Zahlen r und s. 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 Übetragung 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 Browserrprogramme 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, frägt 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; gelegentlich werden wir auch im Laufe dieser Vorlesung darauf zurückkommen.

Mit Ausnahme von Verfahren wie dem *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

$$\begin{gathered}114381625757888867669235779976146612010218296721242362562561842935\backslash\\ 706935245733897830597123563958705058989075147599290026879543541\end{gathered}$$

(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

$$
\begin{array}{r}
490529510847650949147849619903898133417764638493387843990820577 \\
\times 32769132993266709549961988190834461413177642967992942539798288533\ .
\end{array}
$$

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

Auch bei heute 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 – sicher kann aber auch das niemand vorhersagen.

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

BUCHMANN: Einführung in die Kryptographie, *Springer,* ^{3}2004

oder natürlich auch in der Kryptologie-Vorlesung des nächsten Semesters.

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
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:

$$m : r_1 = q_1 \ \text{ Rest } \ r_2 \Longrightarrow \frac{m}{r_1} = q_1 + \frac{r_2}{r_1} \Longrightarrow \alpha = q_0 + \cfrac{1}{q_1 + \cfrac{r_2}{r_1}}\ .$$

Ist auch noch r_2 von Null verschieden, wird sodann r_1 durch r_2 dividiert:

$$r_1 : r_2 = q_2 \ \text{ Rest } \ r_3 \Longrightarrow \frac{r_1}{r_2} = q_2 + \frac{r_3}{r_2} \Longrightarrow \alpha = q_0 + \cfrac{1}{q_1 + \cfrac{1}{q_2 + \cfrac{r_3}{r_2}}}\ ,$$

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

$$\alpha = c_0 + \alpha_0 = c_0 + \cfrac{1}{c_1 + \alpha_1} = c_0 + \cfrac{1}{c_1 + \cfrac{1}{c_2 + \alpha_2}}$$

$$= \cdots = c_0 + \cfrac{1}{c_1 + \cfrac{1}{c_2 + \cfrac{1}{\ddots\, c_{r-1} + \cfrac{1}{c_r + \alpha_r}}}} .$$

Falls der Algorithmus mit $\alpha_r = 0$ abbricht, steht im untersten Bruch natürlich nur c_r im Nenner.

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 $c_0 = \left[\sqrt{2}\right] = 1$ 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. $c_1 = \left[1 + \sqrt{2}\right] = 2$ und $\alpha_2 = 1 + \sqrt{2} - 2 = \sqrt{2} - 1 = \alpha_1$. Damit wiederholt sich ab jetzt alles, d.h.

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cdots}}}}} .$$

Die ersten Partialbrüche sind

$$P_0 = 1, \quad P_1 = 1 + \frac{1}{2} = 1{,}5, \quad P_2 = 1 + \cfrac{1}{2 + \cfrac{1}{2}} = \frac{7}{5} = 1{,}4 ,$$

$$P_3 = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2}}} = \frac{17}{12} = 1{,}41\overline{6} \quad \text{und} \quad P_4 = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2}}}} = \frac{41}{29} ,$$

was ungefähr gleich 1,4137931 ist. Die Fehler $\sqrt{2} - P_n$ sind, gerundet auf sechs Nachkommastellen, die Zahlen

$$0{,}414214, \quad -0{,}085786, \quad 0{,}014214, \quad -0{,}002453 \quad \text{und} \quad 0{,}000420 ;$$

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 beginnt ist somit

$$\pi = 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{292 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cdots}}}}}}}}}$$

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

Tatsächlich werden wir sehen, daß die Konvergenten der Kettenbruchentwicklung einer irrationalen Zahl stets die bei vorgegebener Größenordnung des Nenners bestmögliche rationale Approximation der 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 (p, q) mit ganzzahligen Koordinaten liegt. Rationale Approximationen erhalten wir durch Punkte $(p, q) \in \mathbb{Z} \times \mathbb{Z}$, die in der Nähe der Geraden liegen.

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.

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.

Der gerichtete vertikale Abstand des Punktes $P_n = (q_n, p_n)$ von der Geraden $y = \alpha x$ ist $d_n = p_n - \alpha q_n$; damit ausgedrückt ist

$$a_n = \left[\left|\frac{d_{n-2}}{d_{n-1}}\right|\right].$$

Somit verschwindet a_n genau dann, wenn $|d_{n-2}| < |d_{n-1}|$ ist.

Ist dagegen $|d_{n-1}| < |d_{n-2}|$, so ist $a_n \geq 1$, und da $P_n = P_{n-2} + a_n P_{n-1}$ auf derselben Seite der Geraden liegt wie P_{n-2}, ist auch

$$d_n = d_{n-2} + a_n d_{n-1} = d_{n-2} + \left[\left|\frac{d_{n-2}}{d_{n-1}}\right|\right] d_{n-1}$$

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 $a_i \geq 1$.

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 $P_0 = (1, a_0)$ mit $a_0 = [\alpha]$, 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 a_n von Null verschieden und $|d_n| < |d_{n-1}|$.

Aus den Beziehungen

$$p_n = p_{n-2} + a_n p_{n-1} \quad \text{und} \quad q_n = q_{n-2} + a_n q_{n-1}$$

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

$$\begin{pmatrix} p_n & q_n \\ p_{n-1} & q_{n-1} \end{pmatrix} = \begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} p_{n-1} & q_{n-1} \\ p_{n-2} & q_{n-2} \end{pmatrix};$$

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

$$p_{n-1} q_{n-2} - q_{n-1} p_{n-2} = p_{-1} q_{-2} - q_{-1} p_{-2} = 0 \cdot 0 - 1 \cdot 1 = -1;$$

daraus folgt induktiv, daß

$$p_n q_{n-1} - q_n p_{n-1} = (-1)^{n-1}$$

ist. 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 $n \geq 0$

$$\begin{aligned} \left|\alpha - \frac{p_n}{q_n}\right| &\leq \left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| = \left|\frac{p_{n+1}q_n - q_{n+1}p_n}{q_n q_{n+1}}\right| \\ &= \frac{1}{q_n q_{n+1}} = \frac{1}{q_n(q_{n-1} + a_{n+1}q_n)} \leq \frac{1}{q_n^2}. \end{aligned}$$

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 a_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 $a_n = \left[\left|d_{n-2}/d_{n-1}\right|\right]$ ist dann $a_n = [1/\alpha_n]$. Division der Beziehung $d_n = d_{n-2} + a_n d_{d-1}$ durch d_{n-1} führt auf

$$\frac{d_n}{d_{n-1}} = \frac{d_{n-2}}{d_{n-1}} + a_n \quad \text{oder} \quad -\alpha_{n+1} = -\frac{1}{\alpha_n} + a_n \quad \text{oder} \quad \frac{1}{\alpha_n} = a_n + \alpha_{n+1},$$

was zusammen mit $a_n = [1/\alpha_n]$ und dem Induktionsanfang $\alpha = a_0 + \alpha_1$ genau auf die zu Beginn des Abschnitts konstruierten Folgen der a_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

Als nächstes wollen wir uns überlegen, daß Kettenbrüche in der Tat bestmögliche Approximationen geben: Ist $\frac{r}{s}$ 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 zur Geraden $y = \alpha x$ einen kleineren Betrag hat als der von R.

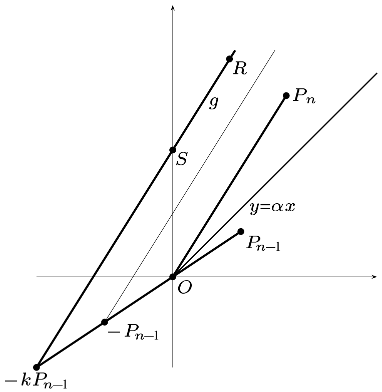
Wir schreiben R als ganzzahlige Linearkombination $R = kP_{n-1} + \ell P_n$ der Punkte P_{n-1} und P_n. Das ist möglich, denn die Determinante des linearen Gleichungssystems

$$\begin{pmatrix} p_{n-1} & p_n \\ q_{n-1} & q_n \end{pmatrix} \begin{pmatrix} k \\ \ell \end{pmatrix} = \begin{pmatrix} r \\ s \end{pmatrix}$$

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 Vorraussetzung $R \neq P_{n-1}$ 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 $a_0 = [\sqrt{3}] = 1$ und $\alpha_1 = \sqrt{3} - 1$. Der Kehrwert davon ist

$$\frac{1}{\sqrt{3}-1} = \frac{\sqrt{3}+1}{2} \Longrightarrow a_1 = 1 \quad \text{und} \quad \alpha_2 = \frac{\sqrt{3}-1}{2}\,.$$

Der Kehrwert davon ist

$$\frac{2}{\sqrt{3}-1} = \sqrt{3}+1 \Longrightarrow a_2 = 2 \quad \text{und} \quad \alpha_3 = \sqrt{3}-1 = \alpha_1\,.$$

Ab hier wiederholt sich also alles periodisch, d.h.

$$\sqrt{3} = 1 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{2 + \cdots}}}}\,.$$

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)* Für eine irrationale Zahl α und jedes $n \geq 2$ ist mindestens eine der beiden Relationen

$$\left|\alpha - \frac{p_{n-1}}{q_{n-1}}\right| < \frac{1}{2q_{n-1}^2} \quad \text{und} \quad \left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{2q_n^2}$$

erfüllt.
b) Ist für eine rationale Zahl $\left|\alpha - \frac{p}{q}\right| < \frac{1}{2q^2}$, so ist 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

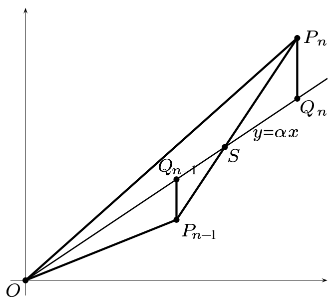
$$|q_{n-1}\alpha - p_{n-1}| \geq \frac{1}{2q_{n-1}} \quad \text{und} \quad |q_n\alpha - p_n| \geq \frac{1}{2q_n}\,.$$

Wir nehmen für den Beweis wieder an, daß $p_{n-1}/q_{n-1} < \alpha < p_n/q_n$ ist; der andere Fall geht völlig analog.

Nach unserer Annahnme 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 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 $p_nq_{n-1} - q_np_{n-1} = (-1)^{n-1}$ 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 $y = \alpha x$. Nach unserer Annahme 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 des Dreiecks mindestens gleich $1/4$.



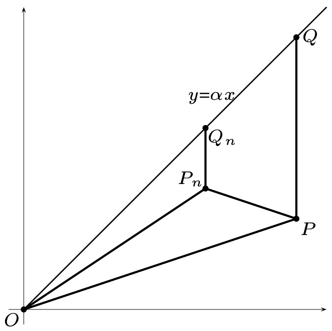
Ist S der Schnittpunkt der Geraden $y = \alpha x$ mit der Verbindungsstrecke von P_{n-1} und P_n, so ist das Dreieck $\triangle OP_{n-1}P_n$ die Vereinigung der Dreiecke $\triangle OP_{n-1}Q_{n-1}$, $\triangle OP_nQ_n$ und $\triangle P_{n-1}Q_{n-1}S$, minus dem Dreieck $\triangle SP_nQ_n$. Die Dreiecke 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.

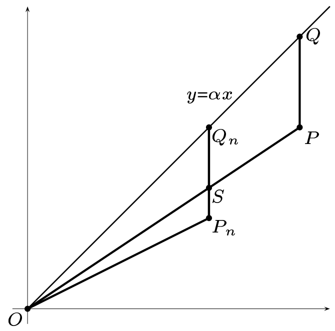
Da die Folge der Nenner q_n strikt monoton ansteigt, gibt es genau ein n, so daß $q_n \le q < q_{n+1}$ ist; wir müssen zeigen, daß $p/q = p_n/q_n$ ist. Andernfalls ist $pq_n - qp_n \neq 0$, also – da dies eine ganze Zahl ist – $|pq_n - qp_n| \ge 1$. Setzen wir $P = (q, p)$, so ist also die Fläche des 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.



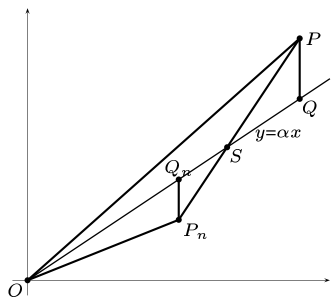
Betrachten wir als erstes den Fall, daß p_n/q_n zwischen α und p/q liegt. Dann liegt der Punkt P_n im Innern des Dreiecks $\triangle OPQ$, also ist das gesamte Dreieck $\triangle OPP_n$ imd Dreieck $\triangle OPQ$ enthalten. Da ersteres mindestens die Fläche $1/2$ hat, letzteres aber weniger als $1/4$, 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 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 Fläche des Dreiecks $\triangle OPQ$, von der wir wissen, daß sie höchstens gleich $1/4$ 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 Dreiecke 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. ■

§4: 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 möglicherweise 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.

Ausgangspunkt ist die Gleichung $ed - k\varphi(N) = 1$, die wir umschreiben können 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 auch 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 4
Quadratische Zahlkörper

Ein Zahlkörper ist ein Körper K, der den Körper $\mathbb{Q}$ der rationalen Zahlen enthält und als $\mathbb{Q}$-Vektorraum betrachtet endlichdimensional ist. Im zweidimensionalen Fall reden wir von quadratischen Zahlkörpern. Die algebraische Zahlentheorie untersucht die (noch zu definierenden) ganzen Zahlen eines solchen Zahlkörpers.

§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 sogenannten *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. Zur Bequemlichkeit des Lesers sei hier auch die Definition von Ringen noch einmal wiederholt:

Definition: *a)* Ein Ring ist eine Menge R zusammen mit zwei Rechenoperationen „+“ und „·“, so daß gilt:

1.) R bildet bezüglich „+“ eine abelsche Gruppe.

2.) Die Verknüpfung „·“: $R \times R \to R$ erfüllt das Assoziativgesetz $x(yz) = (xy)z$, und es gibt ein Element $1 \in R$, so daß $1x = x1 = x$.

3.) „+“ und „·“ erfüllen die Distributivgesetze $x(y + z) = xy + xz$ und $(x + y)z = xz + yz$.

b) Ein Ring heißt *kommutativ,* falls zusätzlich noch das Kommutativgesetz $xy = yx$ der Multiplikation erfüllt ist.

c) Ein Ring heißt *nullteilerfrei* wenn gilt: Falls ein Produkt xy verschwindet, muß mindestens einer der beiden Faktoren x, y verschwinden. Ein nullteilerfreier kommutativer Ring heißt *Integritätsbereich.*

d) 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 $x = qu$.

e) $u \in R$ 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: $v|u$.

f) Ein Element $e \in R$ heißt *Einheit,* falls es ein $e' \in R$ gibt mit $ee' = 1$. Die Menge aller Einheiten von R bezeichnen wir mit $R^\times$.

g) Zwei Elemente $x, y \in R$ heißen assoziiert, wenn es eine Einheit $e \in R$ gibt, so daß $y = ex$.

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 eine Primzahl 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 $a, b \neq 0$.

Der 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

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

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 gilt: Ist $xz = yz$ für ein Element $z \neq 0$ und zwei beliebige Elemente x, y, so ist $x = y$.
c) Zwei Elemente x, y eines Integritätsbereich R sind genau dann assoziert, 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 $g_i \le e_i$ für alle i, und Teiler von y, wenn $g_i \le f_i$. ■

§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 Zahlen A, B, C, so daß $A\alpha^2 + B\alpha + C = 0$ verschwindet. Nach der Lösungsformel für quadratische Gleichungen folgt

$$\alpha = -\frac{B}{2A} \pm \frac{\sqrt{B^2 - 4AC}}{2A}\,.$$

Wegen der Irrationalität von α muß auch $\sqrt{B^2 - 4AC}$ irrational sein, d.h. $B^2 - 4AC$ ist kein Quadrat einer rationalen Zahl. Wegen der Eindeutigkeit der Primzerlegung in $\mathbb{Z}$ können wir aber ganze Zahlen p, q und D finden, so daß

$$B^2 - 4AC = \frac{p^2 D}{q^2} \quad \text{und} \quad \sqrt{B^2 - 4AC} = \frac{p}{q}\sqrt{D}$$

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 quadratfreie 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:

$$\frac{r + s\sqrt{D}}{u + v\sqrt{D}} = \frac{(r + s\sqrt{D})(u - v\sqrt{D})}{(u + v\sqrt{D})(u - v\sqrt{D})} = \frac{ru + svD}{u^2 - v^2 D} + \frac{rv + su}{u^2 - v^2 D}\,.$$

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.

Jede Zahl aus $\alpha = r + s\sqrt{D} \in K$ genügt einer quadratischen Gleichung, zum Beispiel der Gleichung $(\alpha - r)^2 = s^2 D$. Durch Multiplikation mit dem Hauptnenner der Koeffizienten und gegebenenfalls noch Kürzen mit dem ggT erhalten wir eine Gleichung

$$A\alpha^2 + B\alpha + C = 0 \quad \text{mit} \quad A, B, C \in \mathbb{Z} \quad \text{und} \quad \mathrm{ggT}(A, B, C) = 1\,.$$

Nach der Lösungsformel für quadratische Gleichungen ist

$$\alpha = -\frac{B}{2A} \pm \frac{\sqrt{B^2 - 4AC}}{2A}\,.$$

Die Zahl $\Delta = B^2 - 4AC$ bezeichnen wir als die *Diskriminante* von α.

Für $\alpha = \sqrt{D}$ beispielsweise haben wir die quadratische Gleichung $\alpha^2 - D = 0$ mit ganzzahligen, teilerfremden Koeffizienten; somit ist die Diskriminante von $\sqrt{D}$ gleich $4D$. 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 $\Delta = 150^2 - 4 \cdot 225 \cdot 7 = 16200$.

§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: Eine 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.

Als (in dieser Vorlesung einziges) Beispiel betrachten wir den 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 r. 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 $u + \beta_D$ 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 unserere 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 assozierter 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 des 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 $\nu(x)\le\nu(y)$, 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 $\nu: R \setminus \{0\} \to \mathbb{N}_0$ 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 $\nu(x) \le n$ 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 $e_i, f_j \ge 1$ 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 $s \ne 0$ 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 zeig 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 $23 + 5i$ 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 $2 + 3i$. 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

$$\mathrm{N}(r+is\sqrt{D}) = (r+is\sqrt{D})(r-is\sqrt{D}) = r^2+s^2D = \left|r+is\sqrt{D}\right|^2$$

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 weden, 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.

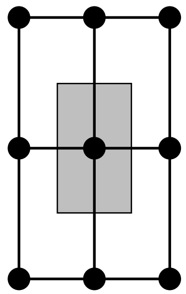
Der Wirkungsbereich eines Gitterpunkts z unterscheidet sich von dem des Nullpunkts nur durch eine Verschiebung um z; entsprechendes gilt auch für die Kreise mit Radius eins um die beiden Punkte. Daher reicht es, zu untersuchen, wann der Wirkungsbereich des Nullpunkts ganz im Innern des Einheitskreises liegt.

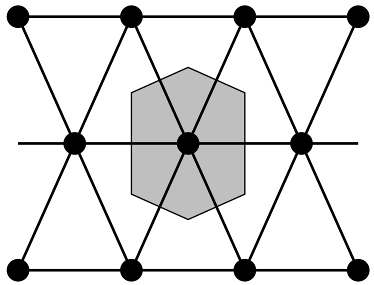
Die Struktur des Wirkungsbereichs hängt ab von $-D \bmod 4$: Falls $D \not\equiv -1 \bmod 4$, d.h. $D \not\equiv 3 \bmod 4$, ist $\mathcal{O}_{-D} = \mathbb{Z} \oplus \mathbb{Z}[\sqrt{-D}]$. In der komplexen Zahlenebene bilden diese Punkte ein Rechteckgitter mit den Gitterpunkten $q = r + is\sqrt{D}$ 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 $D \equiv 1 \bmod 4$ 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 die 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

$$\left\{\left(\pm\frac{1}{2}\pm\frac{i}{2}\sqrt{D}\right)+\left(\mp\frac{\sqrt{D}}{2}\pm\frac{i}{2}\right)t \;\Big|\; t\in\mathbb{R}\right\}.$$

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 $\nu\colon\mathcal{O}_D\setminus\{0\}\to\mathbb{N}_0$ 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

$$\left|(x-r)^2-(y-v)^2D\right|<1\,.$$

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 $\nu\colon R\setminus\{0\}\to\mathbb{N}_0$ 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 imaginärquadratische 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 $(x,y)=(\pm1,0)$, sowie im Fall $D=-1$ der GAUSSschen Zahlen $(x,y)=(0,\pm1)$. 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

$$\lambda\colon \begin{cases} \mathcal{O}_D^\times \to \mathbb{R}^2 \\ \alpha \mapsto (\log|\alpha|\,, \log|\overline{\alpha}|) \end{cases}$$

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 $\log|\alpha| \leq M$ und $\log|\overline{\alpha}| \leq M$. Ist $\log R = M$, so ist also $|\alpha| \leq R$ und $|\overline{\alpha}| \leq R$. Damit ist $|\mathrm{Sp}(\alpha)| \leq 2R$ und $|\mathrm{N}(\alpha)| \leq R^2$.

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 $s \geq 0$ beschränken.

Für einen solchen Punkt $\lambda(\beta) = (s, -s)$ gibt es jedenfalls ein größtes $n \in \mathbb{N}_0$, so daß $nr \leq s$ 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 $0 \leq s - nr < r$; 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 Elementen der Form $\pm\alpha^n$ mit $n \in \mathbb{Z}$ und eine geeignete Einheit $\alpha \in \mathcal{O}_D^\times$.

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 wurde, 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.

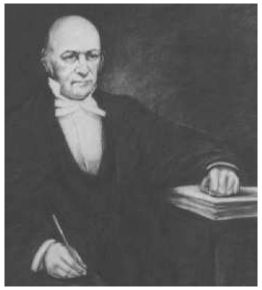
Die damaligen Mathematiker waren jedoch bescheidener: Ihnen genügte es, einfach irgendeine Art von Multiplikation zu finden, die nicht unbedingt den Körperaxiomen genügte – von Körpern sprach damals ohnehin noch niemand.

Erst 1940 konnte HEINZ HOPF (1894–1971) (auf dem Umweg über Vektorfelder auf Sphären) zeigen werden, daß das nicht möglich ist: Selbst eine bilineare Abbidlung $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ 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 uber 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 für 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 5
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 $x, y \in \mathbb{Z}$ annimmt.

§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 mit Normen GAUSSsche 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 durch Ausmultiplizieren der Gleichun $\mathrm{N}(\alpha) \cdot \mathrm{N}(\beta) = \mathrm{N}(\alpha\beta)$ 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 die Primzahlen 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, §7 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 $\mathbb{F}_p$ das Quadrat von g^k.

In $\mathbb{Z}$ gibt es daher Zahlen x, für die $x^2 \equiv -1 \bmod p$ ist oder, anders ausgedrückt, $x^2 + 1 = kp$ für ein $k \in \mathbb{N}$. 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 k

$$x^2 + 1^2 = kp < \frac{p^2}{4} + 1 < \frac{p^2}{2} \Longrightarrow k < p\,.$$

Es gibt also eine natürliche Zahl $1 \le k < p$, so daß kp darstellbar ist als Summe zweier Quadrate. Die kleinste solche Zahl sei m; wir müssen zeigen, daß sie gleich eins 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 $m \ge 3$ sein. Wir definieren zwei neue Zahlen u, v 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 ℓ, so daß $u^2 + v^2 = \ell m$ ist. Da $u^2 + v^2$ kleiner ist als $\frac{1}{2}m^2$, ist $\ell < \frac{m}{2}$.

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.

$$(\ell m)(mp) = (u^2 + v^2)(x^2 + y^2) = (xu + yv)^2 + (xv - yu)^2\,.$$

Dabei ist

$$xu + yv \equiv x^2 + y^2 \equiv 0 \bmod m \text{ und } xv - yu \equiv xy - yx \equiv 0 \bmod m\,,$$

beide Zahlen sind also durch m teilbar. Somit gibt es natürliche Zahlen X, Y mit

$$(\ell m)(mp) = m^2\ell p = (mX)^2 + (mY)^2 \quad \text{oder} \quad \ell p = X^2 + Y^2\,.$$

Da $\ell < \frac{m}{2}$, 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 Kapitel IV, §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 IV: 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 aber gleich eins sein, d.h. die Primzerlegung von p in $\mathbb{Z}[i]$ ist

$$p = (r+is)(r-is) = r^2 + s^2 .$$

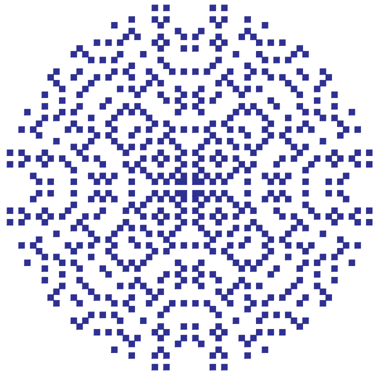
Nach dem Satz ist daher $p \equiv 1 \bmod 4$.

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 eine beliebige natürliche Zahl als Summe zweier Quadrate dargestellt werden kann:

Satz: Eine natürliche Zahl n läßt sich genau dann als Summe zweier Quadrate schreiben, wenn jede Primteiler $p \equiv 3 \bmod 4$ mit einer gerader 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

$$n = x^2 + y^2 \quad \text{und} \quad d = \mathrm{ggT}(x, y) .$$

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 m der Fall ist. Sei p ein solcher Primteiler. Dann ist p ein Teiler

von
$$u^2 + v^2 = (u + iv)(u - iv)$$
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, tritt in n also 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 assoziert, 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 Arcustangens 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
$\frac{\pi}{8}-S_N$	$5{,}68\cdot 10^{-3}$	$6{,}19\cdot 10^{-4}$	$6{,}24\cdot 10^{-5}$	$6{,}25\cdot 10^{-6}$
N	100 000	1 000 000	10 000 000	
$\frac{\pi}{8}-S_N$	$6{,}25\cdot 10^{-7}$	$6,26\cdot 10^{-7}$	$6,4\cdot 10^{-8}$	

Man muß also für jede weitere Dezimalstelle den Rechenaufwand ungefähr verzehnfachen. Angesichts der Tatsache, daß heute mehrere Billionen Ziffern von π bekannt sind, kann das wohl kaum der beste Weg zur Berechnung von π sein.

Zahlen mit einer großen Anzahl verschiedener Darstellungen als Summen von Quadraten können uns hier zu besseren Ergebnissen helfen: Die Reihe für den Arcustangens konvergiert sicherlich umso besser, je kleiner der Wert von x ist. Wenn wir also den Winkel $\frac{\pi}{4}$ aufteilen können

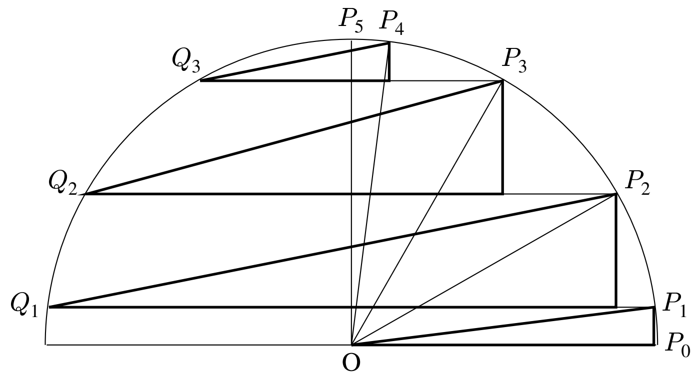
in mehrere kleine Winkel, deren Tangens wir kennen, sollten bessere Ergebnisse zu erwarten sein. Genau das können wir mit solchen Zahlen erreichen.

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 $i=1,\dots,r$ liegen auf der Kreislinie $x^2+y^2=N^2$ um den Nullpunkt O, genauso die drei Punkte $P_0=(\sqrt{n},0)$, $Q_0=(-\sqrt{n},0)$ und $P_{r+1}=(0,\sqrt{n})$.



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 Koordiaten 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 Arcustangenswerte 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

$$1105 = 4^2 + 3^2 = 9^2 + 32^2 = 12^2 + 31^2 = 23^2 + 24^2\,,$$

zu denen natürlich auch noch vier mit vertauschten Faktoren kommen. Wir haben also

$$\begin{aligned} &P_1 = (33,4)\,,\ P_2 = (32,9)\,,\ P_3 = (31,12)\,,\ P_4 = (24,23)\,,\\ &P_8 = (4,33)\,,\ P_7 = (9,32)\,,\ P_6 = (12,31)\,,\ P_5 = (23,24)\,; \end{aligned}$$

dazu kommen noch die beiden Randpunkte $P_0 = (\sqrt{1105}, 0)$ sowie $P_9 = (0, \sqrt{1105})$.

Die Q_i für $1 \le i \le 8$ 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

$$\frac{\pi}{4} = \arctan\frac{4}{33} + 2\arctan\frac{1}{13} + 2\arctan\frac{1}{21} + 2\arctan\frac{1}{5} + \arctan\frac{1}{47}\,.$$

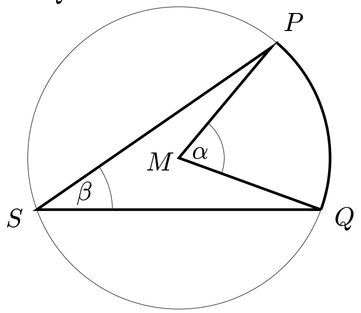
Berechnen wir das Summation von 0 bis n in der TAYLOR-Reihe, erhalten wir folgende (auf eine geltende Ziffer gerundeten) Abweichungen Δ_n von π:

n	1	2	3	4	5
Δ_n	$5\cdot 10^{-4}$	$1\cdot 10^{-5}$	$4\cdot 10^{-7}$	$1\cdot 10^{-8}$	$5\cdot 10^{-10}$
n	6	7	8	9	10
Δ_n	$2\cdot 10^{-11}$	$6\cdot 10^{-13}$	$2\cdot 10^{-14}$	$8\cdot 10^{-16}$	$3\cdot 10^{-17}$

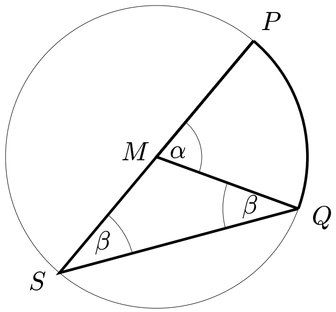
Anhang: Der Satz vom Innenwinkel

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:

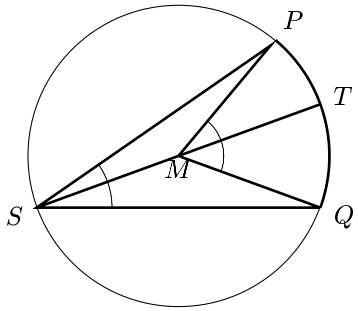
Satz: P, Q, S seien Punkte auf einer Kreisline mit Mittelpunkt M. Dann ist $\angle MPQ = 2\angle SPQ$.



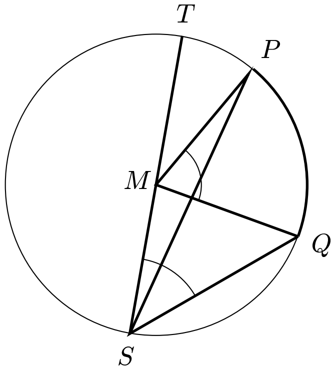
Beweis: Am einfachsten ist der Fall, daß M auf der Verbindungsstrecke von S mit einem der beiden Punkte P und Q liegt; wir nehmen an, er liege auf $\overline{SP}$. (Der andere Fall ist spiegelsymmetrisch dazu und geht genauso.) Dann ist das Dreieck $\triangle MSQ$ gleichschenklig, d.h. wir haben bei 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.

Der Beweis 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. $x^2 + y^2 + 1 = mp$ ist durch p teilbar. Da $x, y \le \frac{1}{2}(p-1)$, ist dabei $m < p$. Da $1 = 1^2$ ein Quadrat ist, ist damit das Lemma 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 IV, §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 dieser 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

$$Ax^2 + Bxy + Cy^2 = (\,x \quad y\,)\, Q \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{mit} \quad Q = \begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix},$$

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 IV, §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 Werte, die eine quadratische Form annehmen kann, hängen nicht davon ab, in welcher Basis 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 $Q \in \mathbb{R}^{2\times 2}$ sowie 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 heißt $\left\{\begin{matrix}\text{positiv}\\ \text{negativ}\end{matrix}\right\}$ definit, wenn zusätzlich $f(x,y) = 0$ nur gilt für $x = y = 0$.

Wie aus der linearen Algebra bekannt ist, gibt es zu einer symmetrischen reellen Matrix stets eine Basis aus Eigenvektoren; bezüglich derer hat die Matrix Diagonalgestalt, wobei in der Diagonale die beiden (reellen) Eigenwerte stehen. Das Produkt dieser Eigenwerte ist die Determinante der Matrix, ihre Summe die Spur.

Offensichtlich ist eine Diagonalmatrix genau dann positiv semidefinit, wenn beide Eigenwerte ≥ 0 sind und genau dann positiv definit, wenn sie sogar echt positiv sind. Entsprechendes gilt für negativ (semi-)definite Matrizen. 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 $AC - \frac{1}{4}B^2$ 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. Das Gleichheitszeichen gilt genau dann, wenn auch $\varphi^{-1}(\mathbb{Z}^2) \subseteq \mathbb{Z}^2$ ist, d.h. wenn auch M^{-1} lauter ganzzahlige Einträge hat. 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

$$^t\!\left(M\begin{pmatrix} x \\ y \end{pmatrix}\right) \cdot Q \cdot M\begin{pmatrix} x \\ y \end{pmatrix} = (x \;\; y)\left({}^tMQM\right)\begin{pmatrix} x \\ y \end{pmatrix},$$

das wir auch erhalten hätten, wenn wir $\binom{x}{y}$ in die quadratische Form zur Matrix ${}^tMQM$ 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 ${}^tMQM$ 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 $Q_2 = {}^tMQ_1M$.

Lemma: Zwei äquivalente quadratische Formen haben dieselbe Diskriminante.

Beweis: Bis auf den Faktor -4 ist die Diskriminante gleich der Determinante der Matrix und

$$\det Q_2 = \det {}^tM \cdot \det Q_1 \cdot \det M = \det Q_1\,,$$

da $\det M = \det {}^tM = \pm 1$ 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 III auf periodische Folgen. Wie sich zeigen wird, ist dies charakteristisch für quadratische Irrationalitäten.

Nach der Formel am Ende von §2 von Kapitel III gilt für die Zahlen α_n aus dem Algorithmus zur Kettenbruchentwickung 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. Zähler und Nenner des rechtsstehenden Bruchs sind die beiden Komponenten des Vektors

$$M\begin{pmatrix} \alpha_n \\ 1 \end{pmatrix} \quad \text{mit} \quad M = \begin{pmatrix} p_{n-2} & p_{n-1} \\ q_{n-2} & q_{n-1} \end{pmatrix},$$

und wie wir ebenfalls aus Kapitel III wissen, ist

$$\det M = p_{n-2}q_{n-1} - q_{n-2}p_{n-1} = (-1)^{n-1}\,.$$

Somit sind die Vektoren $\binom{\alpha}{1}$ und $M\binom{\alpha_n}{1}$ 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 $\binom{x}{y}$ auch alle Vielfachen dieses Vektors dieselbe Gleichung erfüllen, gilt dasselbe für den dazu proportionalen Vektor $M\binom{\alpha_n}{1}$.

Die Matrix zu dieser quadratischen Form sei Q. Wie wir aus dem vorigen Paragraphen wissen, erfüllen die Komponenten von $M\binom{\alpha_m}{1}$ dann die quadratische Gleichung zur Form mit Matrix ${}^tMQM$, die zu der mit Q äquivalent ist und somit dieselbe Diskriminante hat. Also haben alle α_n dieselbe Diskriminante wie α, denn da 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 ${}^tMQM$ sind.

Dies ist ein wesentlicher Schritt für den Beweis des folgenden Satzes:

Satz (LAGRANGE $\sim$1766): Die Kettenbruchentwicklung einer irrationalen Zahl α ist 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 III 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. Diese Gleichung können wir uns explizit verschaffen, indem wir

$$\alpha = \frac{\alpha_n p_{n-2} + p_{n-1}}{\alpha_n q_{n-2} + q_{n-1}}$$

in die Gleichung $f(\alpha) = A_0\alpha^2 + B_0\alpha + C_0 = 0$ einsetzen und mit dem Nenner multiplizieren. 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 $k \geq 1$ 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 ja Beispiele wie

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cdots}}}}}$$

oder

$$\sqrt{3} = 1 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{2 + \cdots}}}},$$

bei denen das nicht der Fall ist. Für eine rein periodische Kettenbruchentwicklung brauchen wir also noch zusätzliche Bedingungen:

Satz: Die Kettenbruchentwicklung von α ist genau dann rein periodisch, wenn $\alpha > 1$ ist und sein 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 $\overline{\alpha}_i$ auszudrücken.

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

$$0 < -\overline{\alpha}_1 - c_0 < 1 \quad \text{und} \quad c_0 = [-\overline{\alpha}_1]\,.$$

Wegen $c_0 = [\alpha] \geq 1$ 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 $c_i = [-\alpha_{i+1}]$ 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

$$\alpha = a_0 + \alpha_1 = c_0 + \frac{1}{\alpha} \Longrightarrow \alpha^2 - c_0\alpha - 1 = 0\,.$$

Die quadratische Funktion $x^2 - c_0x - 1$ 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 III, §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 $-p_{k-2} < 0$ 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.

Damit ist der Satz vollständig bewiesen. ■

§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 III, §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

$$\sqrt{13} = 3 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{6 + \cdots}}}}}\,,$$

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 $\alpha = \left[\sqrt{D}\right] + \sqrt{D}$, so ist natürlich $\alpha > 1$. und $\overline{\alpha} = \left[\sqrt{D}\right] - \sqrt{D}$ 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 $\sqrt{D} = \alpha - \left[\sqrt{D}\right]$ unterscheidet sich von der von α nur im ganzzahligen Anteil. Dieser ist im Falle von α gleich $2\left[\sqrt{D}\right]$, im Falle von $\sqrt{D}$ nur $\left[\sqrt{D}\right]$. Danach folgen in beiden Fällen die c_i mit $i \geq 1$. Wegen $c_k = c_0 = 2\left[\sqrt{D}\right]$ 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 $c_k = 2c_0 = 2\left[\sqrt{D}\right]$. ∎

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.

$$\frac{1}{\alpha_{rk}} = c_0 + \sqrt{D} = \left[\sqrt{D}\right] + \sqrt{D}\,.$$

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 III, §2, bekannte Formel

$$p_m q_{m-1} - q_m p_{m-1} = (-1)^{m-1}$$

mit $m = rk - 1$, erhalten wir die Gleichung

$$\begin{aligned} p_{rk-1}^2 - p_{rk-1}q_{rk-1}c_0 - q_{rk-1}^2 D + q_{rk-1}p_{rk-1}c_0 \\ = p_{rk-1}^2 - Dq_{rk-1}^2 = (-1)^{rk-2}\,. \end{aligned}$$

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

$$\sqrt{13} = 3 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{6 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1}}}}}}}} = \frac{649}{180}\,,$$

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 IV, §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 hießigen 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 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 $11/4$ ist $11^2 - 13 \cdot 3^2 = 4$, d.h.

$$\mathrm{N}\left(\tfrac{1}{2}(11 + 3\sqrt{13})\right) = 1\,.$$

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 6
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 Irratinalität 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): Für ungerades p ist und $a \in \mathbb{F}_p^\times$ ist $\left(\frac{a}{p}\right) = a^{\frac{p-1}{2}}$.

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

$$R = \{-h, \dots, -1, 1, \dots, h\} \quad \text{mit} \quad h = \frac{p-1}{2}.$$

Weiter sei $S = \{q, 2q, \dots, hq\}$. 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: $a_1, \dots, a_m$ seien die negativen Elemente von R, die zu Elementen aus S kongruent sind, $b_1, \dots, b_n$ 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 Konguenz 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

$$\{-a_1, \dots, -a_m, b_1, \dots, b_n\} = \{1, \dots, h\}$$

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 auchnoch 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.

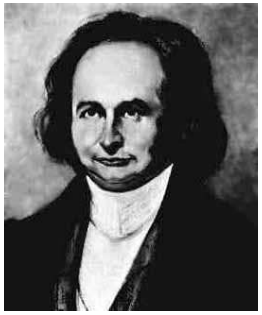
§3: Das Jacobi-Symbol

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 $\mathrm{ggT}(m,n) = 1$ 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

$$\left(\frac{n}{m}\right) = \prod_{j=1}^{s}\left(\frac{n}{q_j}\right)^{f_j} = \prod_{j=1}^{s}\prod_{i=1}^{r}\left(\frac{p_i}{q_j}\right)^{e_i f_j} .$$

Damit ist

$$\left(\frac{n}{m}\right)\left(\frac{m}{n}\right) = \prod_{i=1}^{r}\prod_{j=1}^{s}\left((-1)^{\frac{p_i-1}{2}\frac{q_j-1}{2}}\right)^{e_i f_j} = (-1)^{\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{s}\frac{p_i-1}{2}\frac{q_j-1}{2}e_i f_j}$$

$$= (-1)^{\left(\sum\limits_{i=1}^{r}\frac{p_i-1}{2}e_i\right)\left(\sum\limits_{j=1}^{s}\frac{q_j-1}{2}f_j\right)} = \left((-1)^{\sum\limits_{i=1}^{r}\frac{p_i-1}{2}e_i}\right)^{\sum\limits_{j=1}^{s}\frac{q_j-1}{2}f_j} .$$

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 $p \bmod 8$ abhängt; wegen der Multiplikativität des JACOBI-Symbols reicht es also, wenn wir ungerade a betrachten. Nach dem gerade bewiesenen Gesetz ist dann

$$\left(\frac{a}{p}\right) = (-1)^{\frac{a-1}{2}}(-1)^{\frac{p-1}{2}}\left(\frac{p}{a}\right) .$$

Für festes a ist $(-1)^{(a-1)/2}$ ein konstanter Wert, $(-1)^{(p-1)/2}$ hängt nur ab von $p \bmod 4$, und $\left(\frac{p}{a}\right)$ hängt ab von $p \bmod a$. Insgesamt hängt es also nur ab von $p \bmod 4a$, ob a ein quadratischer Rest oder Nichtrest modulo p ist.

Betrachten wir als Beispiel den Fall $a = 3$. Hier ist $(-1)^{(a-1)/2} = -1$,

$$\left(\frac{p}{3}\right) = \begin{cases} +1 & \text{falls } p \equiv 1 \bmod 3 \\ -1 & \text{falls } p \equiv 2 \bmod 3 \end{cases} ,$$

und wie wir bereits wissen ist

$$(-1)^{\frac{p-1}{2}} = \begin{cases} +1 & \text{falls } p \equiv 1 \bmod 4 \\ -1 & \text{falls } p \equiv 3 \bmod 4 \end{cases}.$$

Somit ist

$$\left(\frac{3}{p}\right) = \begin{cases} +1 & \text{falls } p \bmod 12 \in \{5,7\} \\ -1 & \text{falls } p \bmod 12 \in \{1,11\} \end{cases}.$$

Für $a = 5$ ist $(-1)^{(a-1)/2} = +1$ und

$$\left(\frac{p}{5}\right) = \begin{cases} +1 & \text{falls } p \bmod 5 \in \{1,4\} \\ -1 & \text{falls } p \bmod 5 \in \{2,3\} \end{cases},$$

also

$$\left(\frac{5}{p}\right) = \begin{cases} +1 & \text{falls } p \bmod 20 \in \{1,3,7,9\} \\ -1 & \text{falls } p \bmod 20 \in \{11,13,17,19\} \end{cases}.$$

§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 $\pm r^{r/2}$. 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

$$k = -\beta q \bmod 2^e \in \{0, 1, 2, \dots, 2^e - 1\}\,,$$

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

$$p - 1 = 2q \quad \text{mit einer ungeraden Zahl } q\,,$$

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 leicht direkt nachprüfen, denn

$$x_{1/2}^2 = a^{(p+1)/2} = a \cdot a^{(p-1)/2} = a \cdot \left(\frac{a}{p}\right)$$

nach dem Lemma von EULER. 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. Damit ist

$$p - 1 = 4q = 2^2 q \quad \text{mit einer ungeraden Zahl } q\,,$$

d.h. $e = 2$. Damit 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 ist auch die Anzahl 2^e der möglichen Werte von k nicht mehr beschränkt; 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 wir wissen nur, daß der kleinste quadratische Rest *irgendeine* Zahl

zwischen eins und $1 + \sqrt{p}$ ist, was für große Werte von p viel zu viele Möglichkeiten läßt.

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 beispielsweise nach der EULERschen Formel das LEGENDRE-Symbol ausrechnet. 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 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.

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 unverändertet 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 zwei Anwendungen quadratischer Reste vorgestellt werden:

a) 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. Sie mit Wappen nach oben auf den Boden fällt, also hat A gewonnen.

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 $u \not\equiv \pm x$, 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 $\mathrm{ggT}(x - u, N)$ berechnen und erhält einen der beiden Primfaktoren p oder q. (Der andere ist $\mathrm{ggT}(x + u, N)$.) 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.)

b) 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 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. Zeitliche 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 $v = 343$ m/s aus; der hörbare Frequenzbereich beginnt bei

$\nu = 16$ Hz und kann bis zu etwa $\nu 20$ kHz gehen. Die Wellenlängen, mit denen wir es zu tun haben, variieren also zwischen etwa $\lambda = 21{,}5$ m und $\lambda = 1{,}75$ cm. Der Kammerton a' mit 440 Hz hat eine Wellenlänge von knapp 78 cm.

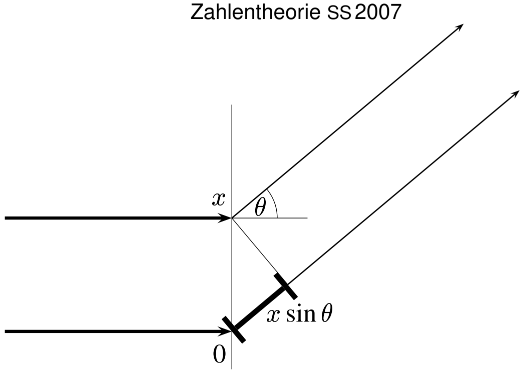
Bei einer Reflektion können wir nach HUYGENS annehmen, daß von jedem Punkt der reflektierenden Fläche eine neue Welle ausgeht; ihre Amplitude ist gleich der Amplitude der dort eintreffenden Welle mal einem Reflektionsfaktor $\rho(x)$, der im Idealfall gleich eins ist.

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 die 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

$$\begin{aligned}
\left|\widehat{r}(m)\right|^2 &= \frac{1}{\sqrt{p}} \sum_{n=0}^{p-1} e^{2\pi i n(n-m)/p} \cdot \frac{1}{\sqrt{p}} \sum_{n=0}^{p-1} e^{-2\pi i n(n-m)/p} \\
&= \frac{1}{p} \sum_{n=0}^{p-1} \sum_{k=0}^{p-1} e^{2\pi i n(n-m)/p} e^{-2\pi i k(k-m)/p} \\
&= \frac{1}{p} \sum_{n=0}^{p-1} \sum_{k=0}^{p-1} e^{2\pi i \left(n^2-k^2-(n-k)m\right)/p}
\end{aligned}$$

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

$$\begin{aligned}
\left|\widehat{r}(m)\right|^2 &= \frac{1}{p} \sum_{n=0}^{p-1} \sum_{k=0}^{p-1} e^{2\pi i \left(n^2-(n-k)^2-km\right)/p} \\
&= \frac{1}{p} \sum_{k=0}^{p-1} e^{-2\pi i k m/p} \sum_{n=0}^{p-1} e^{2\pi i \left((n^2-(n-k)^2)\right)/p} \\
&= \frac{1}{p} \sum_{k=0}^{p-1} e^{-2\pi i k m/p} \sum_{n=0}^{p-1} e^{2\pi i (2kn-k^2)/p}\,.
\end{aligned}$$

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 ingesamt 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 $e^{4\pi ik/p}$ multipliziert, und damit muß sie verschwinden. Somit ist

$$\left|\widehat{r}(m)\right|^2 = \frac{1}{p}e^0 \cdot p = 1$$

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 7
Primzahlen

§1: 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:

Wenn man alle Primzahlen kleiner oder gleich einer Zahl N finden möchte, schreibt 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.

Ansonsten ist eine Primzahl eine Zahl, die außer der Eins und sich selbst keine Teiler hat. Damit muß zwei eine Primzahl sein.

Die echten Vielfachen von zwei sind natürlich keine Primzahlen; also streichen wir sie durch. 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 hat sie natürlich auch eine Primzahl als echten Teiler.

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 $p^2 \leq x$.

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; denn 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.

ERATOSTHENES (Ερατοσθένες) wurde 276 v.Chr. in Cyrene im heutigen Lybien 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.

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 ims 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 übrigbleibt, 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.

§2: Der Fermat-Test

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 $1 \leq a \leq p-1$ gilt $a^{p-1} \not\equiv 1 \bmod p$, kann p keine Primzahl sein.

Beispiel: Ist $F_{20} = 2^{2^{20}} + 1$ eine Primzahl? Falls ja, ist nach dem kleinen Satz von FERMAT insbesondere

$$3^{F_{20}-1} = 1 \bmod F_{20} .$$

Nachrechnen zeigt, daß

$$3^{(F_{20}-1)/2} \not\equiv \pm 1 \bmod F_{20} ,$$

die Zahl ist also nicht prim. (Das „Nachrechnen“ ist bei dieser 315 653-stelligen Zahl natürlich keine Übungsaufgabe für Taschenrechner: 1988 brauchte eine Cray X-MP dazu 82 Stunden, eine Cray-2 immerhin noch zehn; siehe *Math. Comp.* **50** (1988), 261–263. Die anscheinend etwas weltabgewandt lebenden Autoren meinen, dies sei die teuerste bislang produzierte 1-Bit-Information.)

Ein anderes Beispiel, daß sich leicht mit einem Computeralgebrasystem nachrechnen läßt, wäre $M_{67} = 2^{67} - 1$. Hier ist

$$13^{M_{67}-1} \equiv 81\,868\,480\,399\,682\,966\,751 \bmod M_{67} ,$$

also ist auch M_{67} keine Primzahl.

Zur Not auch mit dem Taschenrechner läßt sich

$$F_5 = 2^{2^5} + 1 = 2^{32} + 1 = 2\,863\,311\,531$$

überprüfen: Hier ist zum Beispiel

$$3^{F_5-1} = 3^{2^{32}} \equiv 2\,863\,311\,531 \bmod F_5 ,$$

wie man durch 32-faches Quadrieren modulo F_5 feststellt.

(Allgemein bezeichnet man $F_n = 2^{2^n} + 1$ als die n-te FERMAT-Zahl, da FERMAT 1650 behauptet hatte, er könne beweisen, daß alle diese Zahlen prim seien. $F_0 = 3, F_1 = 5, F_2 = 17, F_3 = 257$ und $F_4 = 65\,537$ sind in der Tat prim, aber 1732 zeigte EULER, daß F_5 durch 641 teilbar ist. Inzwischen ist bekannt, daß F_n für $5 \le n \le 32$ sowie viele andere Werte von n zusammengesetzt ist; FERMATsche Primzahlen mit $n > 4$ sind nicht bekannt.

So einfach es ist, auf diese Weise eine Zahl als zusammengesetzt zu erkennen, so unmöglich ist es, umgekehrt so zu beweisen, daß sie prim ist. So ist beispielsweise

$$18^{322} \equiv 1 \bmod 323 \quad \text{aber} \quad 323 = 17 \cdot 19 .$$

Immerhin gibt es nicht viele $a \le 323$ mit $a^{322} \equiv 1 \bmod 323$: Die einzigen Möglichkeiten sind $a = \pm 1$ und $a = \pm 18$.

Zumindest theoretisch läßt sich der FERMAT-Test allerdings auch ausbauen zu einem echten Primzahlbeweis:

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, §7 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. ■

Beispiel: Ist $p = 2^{16} + 1$ prim? Hier hat $p - 1 = 2^{16}$ nur die Zwei als Primteiler; nach dem Satz ist p also prim, falls wir eine Zahl a finden können, so daß $a^{p-1} \equiv 1 \bmod p$, aber $a^{(p-1)/2} \not\equiv 1 \bmod p$. Für $a = 2$ sind beide Potenzen eins, für $a = 3$ aber ist die zweite gleich -1. Somit ist p eine Primzahl.

Für FERMAT-Zahlen läßt sich also recht einfach entscheiden, ob sie prim sind oder nicht; bei sonstigen Zahlen hat man das Problem, daß zunächst $p-1$ faktorisiert werden muß. Falls p Primfaktor eines sicheren RSA-Moduls werden soll, im Idealfall also über dreihundert Dezimalstellen haben sollte, ist dies jedoch unrealistisch; hier braucht man alternative Verfahren.

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, allerdings 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 $\operatorname{ggT}(a,n)=1$ 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.

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

Beweis: Die Primzerlegung von n sei $\prod_{i=1}^{r} p_i^{e_i}$. Falls $a^{n-1} \equiv 1 \bmod n$ ist, haben wir erst recht $a^{n-1} \equiv 1 \bmod p_i^{e_i}$ für jedes i. Andererseits wissen wir, daß die Ordnung von a in $\mathbb{Z}/p_i^{e_i}\mathbb{Z}$ die Gruppenordnung $\varphi(p_i^{e_i}) = p_i^{e_i-1}(p_i-1)$ teilt, außerdem muß sie $n-1$ teilen. Damit kann sie auf keinen Fall ein Teiler von $p_i^{e_i-1}$ sein, denn das ist ein Teiler von n. Also muß sie ein Teiler von $p_i - 1$ sein.

Für $e_i > 1$ kann dies aber nicht für alle zu p_i primen a der Fall sein, denn beispielsweise ist $(p_i+1)^{p_i} \equiv 1 \bmod p_i^2$, so daß nicht gleichzeitig $(p_i+1)^{p_i-1} \equiv 1 \bmod p_i^2$ sein kann, denn die beiden Potenzen unterscheiden sich um den Faktor $p_i+1 \not\equiv 1 \bmod p_i^2$. Daher muß $e_i = 1$ sein, und da es in $(\mathbb{Z}/p_i\mathbb{Z})^\times$ Elemente der Ordnung $p_i - 1$ gibt, ist $p_i - 1$ ein Teiler von $n-1$.

Schließlich müssen wir uns noch überlegen, daß $r \geq 3$ ist. Wäre $n = pq$ nur das Produkt zweier Primzahlen, 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 augeschlossen 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.

Die kleinste CARMICHAEL-Zahl ist $561 = 3 \cdot 11 \cdot 17$; wie man inzwischen weiß, gibt es unendlich viele CARMICHAEL-Zahlen, auch wenn sie ziemlich selten sind.

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 Fehlerwahrscheinlichkeit ε:

$p \approx 10^{60}$	10^{70}	10^{80}	10^{90}	10^{100}
$\varepsilon \leq 7{,}16 \cdot 10^{-2}$	$2{,}87 \cdot 10^{-3}$	$8{,}46 \cdot 10^{-5}$	$1{,}70 \cdot 10^{-6}$	$2{,}77 \cdot 10^{-8}$
$p \approx 10^{120}$	10^{140}	10^{160}	10^{180}	10^{200}
$\varepsilon \leq 5{,}28 \cdot 10^{-12}$	$1{,}08 \cdot 10^{-15}$	$1{,}81 \cdot 10^{-19}$	$2{,}76 \cdot 10^{-23}$	$3{,}85 \cdot 10^{-27}$

(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 gelegent zu höherende 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 industrielle Anwendungen gut genug ist.

§3: 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.

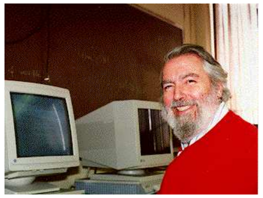
Beispiel: Ist 247 eine Primzahl? Wir wählen $a = 77$, und da $77^{246} \bmod 247 = 1$ ist, können wir mit FERMAT nicht ausschließen, daß 247 prim ist. Da aber $77^{123} \bmod 247 = 77$ ist, sagt uns der Algorithmus von MILLER und RABIN im zweiten Schritt, wenn wir $77^2 \equiv 1 \bmod 247$ betrachten,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

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



Der amerikanische Mathematiker JOHN L. SELFRIDGE promivierte 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`

§4: 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,* 2005,

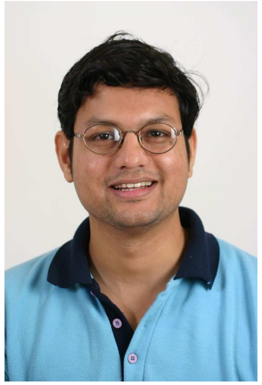
dem dieser Paragraph im wesentlichen folgt, argumentiert SHOUP, daß alternative Algorithmen, so man sich auf Zahlen von weniger als 2^{256} Bit

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. Derzeit arbeitet er am Institute for Advances Study in Princeton, wo er bereits im akademischen Jahr 2003/2004 als *visiting student research collaborator* war. Neuere Arbeiten beschäftigen sich mit der Komplexität des Isomorphieproblems bei endlichen Ringen sowie der Lösbarkeit von bivariaten Polynomgleichungen über endlichen Körpern.

`http://www.math.ias.edu/∼kayaln/`



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. Derzeit arbeitet er als Postdoc in der Gruppe *Quantum Computing and Advanced Systems Research* am Centrum voor Wiskunde en Informatica in Amsterdam. Sein Interesse gilt für algorithmischen Verfahren der Algebra und Zahlentheorie sowie Fragen der Komplexitätstheorie.

`http://homepages.cwi.nl/∼ns/`

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 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 $1 \le i \le n-1$ 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

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

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 $\sigma_k\colon R \to R$, 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 $j = 1, \ldots, \ell$ ist andererseits auch

$$\sigma_k(X + j) = (X + j)^k\,,$$

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

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

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

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

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

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 $S = \tau\bigl(D(n)\bigr)$:

Lemma: $S = \tau\bigl(D(n)\bigr)$ 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 q 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

$$I = \bigl\{ n^u p^v \bigm| 0 \le u, v \le [\sqrt{t}] \bigr\}$$

mindestens $\bigl([\sqrt{t}]+1\bigr)^2$ 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 Polynomes $X^k - X^{k'}$. 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 $p > r > t \ge m$ ist, sind die Zahlen von 1 bis m auch modulo p paarweise verschieden. Die Teilmenge

$$P = \Bigl\{ \prod_{j=1}^{m} (X+j)^{e_j} \Bigm| e_j \in \{0,1\} \text{ und } \sum_{j=1}^{m} e_j < m \Bigr\}$$

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

$$P(\eta) = \bigl\{ f(\eta) \bigm| f \in P \bigr\} \subseteq R \quad \text{und} \quad P(\zeta) = \bigl\{ f(\zeta) \bigm| f \in P \bigr\} \subseteq K .$$

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 $\tau\bigl(P(\eta)\bigr) = P(\zeta)$ 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 mod 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 Untergrupe 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.

§5: Die Verteilung der Primzahlen

Wenn wir Primzahlen einer vorgegebenen Größenordnung suchen (z.B. für einen RSA-Schlüssel), sollten wir zumindest ungefähr wissen, wie die Primzahlen verteilt sind. Damit können wird dann beispielsweise abschätzen, wie groß ein Intervall sein muß, damit wir eine einigermaßen gute Chance haben, dort mindestens eine Primzahl zu finden.

Natürlich sind die Abstände zwischen aufeinanderfolgenden Primzahlen sehr ungleichmäßig verteilt: 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 fünf und drei, aber auch der zwischen den Primzahlen $10^{50} + 18307$ und $10^{50} + 18309$. 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 beweisen, daß es wirklich unendlich viele gibt.

Eine obere Grenze für den Abstand zwischen zwei aufeinanderfolgenden Primzahlen gibt es nicht: 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

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

Die Abbildungen auf der folgenden 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 differenzerbaren Funktion; auf den ersten Blick sieht sie sogar 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:

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

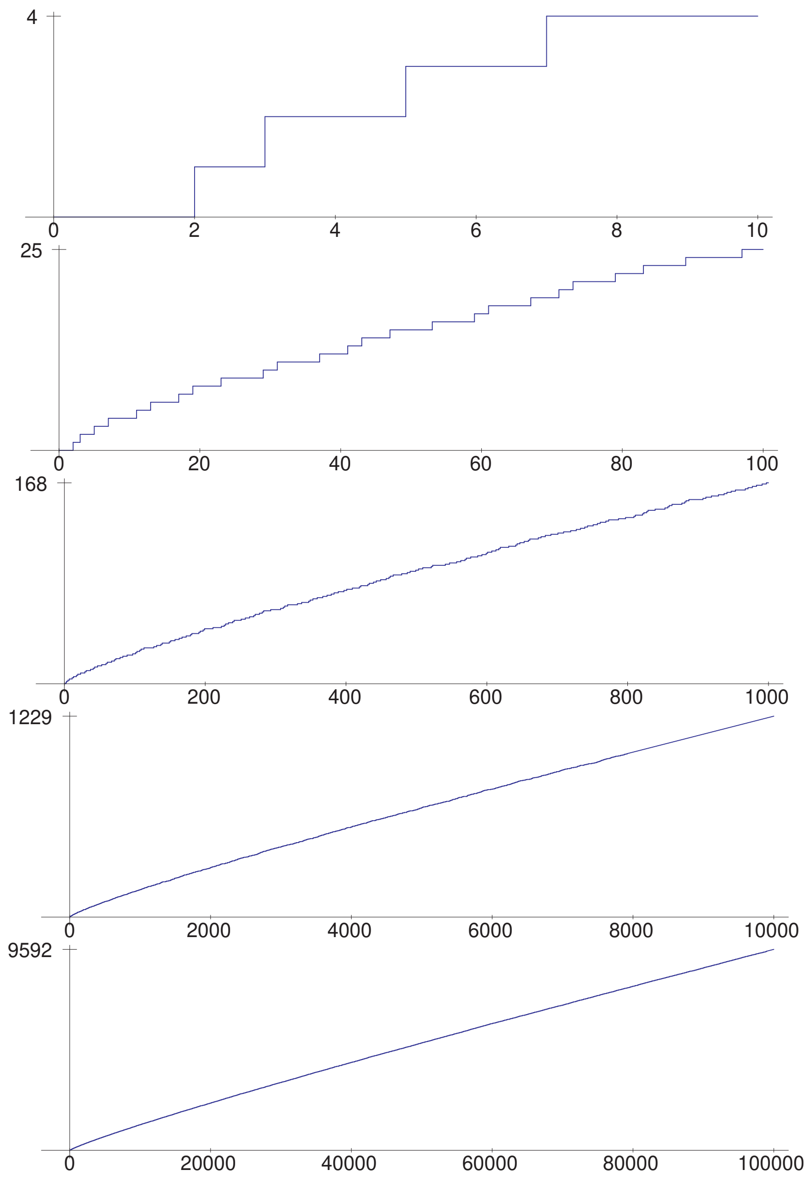
Beweis: Wir betrachten die neue Funktion

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

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

$$\pi(x) = \sum_{p \leq x} \frac{\log p}{\log p} \geq \sum_{p \leq x} \frac{\log p}{\log x} = \frac{\vartheta(x)}{\log x} ,$$



andererseits ist

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

und damit auch

$$\pi(x) < \frac{2\vartheta(x)}{\log x} + \pi\left(\sqrt{x}\right) < \frac{2\vartheta(x)}{\log x} + \sqrt{x}\,.$$

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

$$e_p = \sum_{k\ge1}\left[\frac{n}{p^k}\right] \quad\text{und}\quad \log n! = \sum_{p\le n} e_p \log p = \sum_{p\le n}\sum_{k\ge1}\left[\frac{n}{p^k}\right]\log p\,.$$

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

$$\sum_{p\le n}\sum_{k\ge2}\left[\frac{n}{p^k}\right]\log p \le \sum_{p\le n}\left(\log p\cdot\sum_{k\ge2}\frac{n}{p^k}\right) = n\sum_{p\le n}\frac{\log p}{p(p-1)}$$

nach der Summenformel für die geometrische Reihe:

$$\sum_{k\ge2}\frac{1}{p^k} = \frac{1}{p^2}\cdot\frac{1}{1-\frac{1}{p}} = \frac{1}{p^2-p} = \frac{1}{p(p-1)}\,.$$

Zur weiteren Abschätzung ersetzen wir die Summe über alle Primzahlen kleiner oder gleich n durch die Summe aller natürlicher Zahlen bis n und beachten, daß für alle reellen Zahlen $x \ge 2$ gilt

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

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

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

$$\sum_{k\ge2}\frac{1}{p^k} = O(n)\,.$$

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 sie noch nicht kennen, im Anhang zu diesem Paragraphen skizziert ist. Kombinieren wir dies mit der gerade bewiesenen Formel, ist also

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

Damit ist

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

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

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

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

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

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

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

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 Figurs 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 ČEBYŠËV (1821–1894) zeigte 1851: *Falls*

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

existiert, dann muß er den Wert eins haben.

1852 zeigte er dann ein deutlich schärferes Resultat als den oben bewiesenen Satz: 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 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 RIEMANNschen Zetafunktion

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

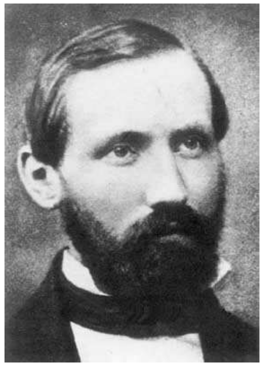
Es ist im wesentlichen eine Analysis I Übungsaufgabe zu zeigen, daß diese Summe für reelle $s > 1$ konvergiert; wer mit komplexen Zahlen umgehen kann, folgert daraus dann leicht, daß sie auch für alle komplexen s mit Realteil größer eins konvergiert.

Etwas trickreicher, aber durchaus noch im Rahmen einer Vorlesung *Funktionentheorie I* durchführbar, ist der Beweis, daß $\zeta(s)$ zu einer analytischen Funktion auf $\mathbb{C} \setminus \{1\}$ forgesetzt werden kann. (Für $s = 1$ haben wir eine harmonische Reihe, und die divergiert bekanntlich, so daß der von rechts kommende Limes von $\zeta(s)$ für $s \to 1$ unendlich sein muß.) Wie RIEMANN erkannte, hängt die Primzahlverteilung eng mit der Frage zusammen, welche Nullstellen $\zeta(s)$ für jene Argumente s hat, deren Realteil zwischen null und eins liegt.

Nach einer berühmten Vermutung von RIEMANN haben alle diese Nullstellen 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. einen Preis von jeweils einer Million Dollar ausgesetzt hat.



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.

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

$$\int_k^{k+1} \left(\{x\} - \tfrac12\right) f'(x)\,dx = \left(x - k - \tfrac12\right) f(x)\,\Big|_k^{k+1} - \int_k^{k+1} f(x)\,dx$$

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

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

$$\int_1^n \left(\{x\} - \tfrac12\right) f'(x)\,dx = \frac{f(1)}{2} + \sum_{k=2}^{n-1} f(k) + \frac{f(n)}{2} - \int_1^n f(x)\,dx\,,$$

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

$$\sum_{k=1}^n f(k) = \int_1^n f(x)\,dx + \frac{f(1) + f(n)}{2} + \int_1^n \left(\{x\} - \tfrac12\right) f'(x)\,dx\,.$$

■

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

$$\log n! = \int_1^n \log x\,dx + \frac{\log n}{2} + \int_1^n \frac{\{x\} - \frac12}{x}\,dx$$

$$= x(\log x - 1)\,\Big|_1^n + \frac{\log n}{2} + \int_1^n \frac{\{x\} - \frac12}{x}\,dx$$

$$= n(\log n - 1) + 1 + \frac{\log n}{2} + \int_1^n \frac{\{x\} - \frac12}{x}\,dx\,.$$

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

$$\int_k^{k+1} \frac{\{x\} - \frac12}{x}\,dx = \int_{-\frac12}^{\frac12} \frac{x}{k + \frac12 + x}\,dx$$

$$= \int_0^{\frac12} \left(\frac{x}{k+\frac12+x} - \frac{x}{k+\frac12-x} \right) dx = \int_0^{\frac12} \frac{-2x^2}{\left(k+\frac12\right)^2 - x^2}\, dx\,.$$

Im Intervall von 0 bis $\frac12$ ist der Integrand monoton fallend, d.h.

$$0 \ge \frac{-2x^2}{\left(k+\frac12\right)^2 - x^2} \ge \frac{-\frac12}{\left(k+\frac12\right)^2 - \frac14} = \frac{-2}{(2k+1)^2-1} \ge -\frac{1}{2k^2}\,,$$

und damit ist

$$0 \ge \int_k^{k+1} \frac{\{x\}-\frac12}{x}\, dx = \int_0^{\frac12} \frac{-2x^2}{\left(k+\frac12\right)^2 - x^2}\, dx \ge -\frac{1}{4k^2}\,,$$

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

$$0 \ge \int_1^n \frac{\{x\}-\frac12}{x}\, dx \ge -\sum_{k=1}^{n-1} \frac{1}{4k^2}$$

für das störende Integral aus der obigen Formel.

Wie wohl jeder schon einmal in einer Analysis I Übungsaufgabe zeigen mußte, konvergiert die rechtsstehende Summe (egal ob mit oder ohne vier im Nenner) für $n \to \infty$; wer mit FOURIER-Reihen vertraut ist, weiß wahrscheinlich auch, daß der Grenzwert $\pi^2/24$ ist. Auf jeden Fall können wir folgern, daß das uneigentliche Integral

$$\int_1^\infty \frac{\{x\}-\frac12}{x}\, dx$$

konvergiert; den Grenzwert wollen wir mit I bezeichnen. Dann ist

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

also folgt insbesondere die Abschätzung

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

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

Kapitel 8
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 Sekunden; 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

FRANK NELSON COLE (1861–1926) wurde in Massachusetts geboren; 1878 ging er dort an die Harvard University, wo er 1882 seinen Bachelor erhielt. Mit einem Stipendium konnte er dann drei Jahre lang nach Deutschland gehen, wo er bei FELIX KLEIN in Leipzig studierte. Mit einer von KLEIN betreuten Arbeit über Gleichungen sechsten Grades wurde er 1886 in Harvard promoviert. Nach verschiedenen Positionen in Harvard und Michigan ging er 1895 als Professor an die Columbia University in New York, wo er bis zu seinem Tod lehrte. Seine Arbeiten befassen sich hauptsächlich mit Primzahlen und mit 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.

In diesem Kapitel soll es um zumindest einige der Verfahren gehen, mit denen man heute das Problem der Faktorisierung von Zahlen wie $2^{67} - 1$ und auch erheblich größeren Zahlen behandelt.

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 Größenordung ist, wird man anders vorgehen als etwa bei einer Zahl der Form $a^n \pm 1$. 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. 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

Falls eine Zahl n zusammengesetzt ist, hat sie mindestens einen Primteiler $p \leq [\sqrt{n}]$. Bei kleinen Zahlen n besteht die effizienteste Art der Faktorisierung im allgemeinen darin, einfach alle diese Primzahlen durchzuprobieren, indem man sie der Reihe nach so lange abdividiert, wie es geht.

Für große Werte von n ist $[\sqrt{n}]$ zu groß; trotzdem sollte man auch da zumindest alle Primteiler bis zu einer gewissen Schranke N eliminieren. Ein typischer Wert für PCs wäre etwa $N = 2^{15}$ oder $N = 2^{16}$. Die Vorgehensweise ist folgende:

1. Schritt: Bestimme nach ERATOSTHENES die Folge $p_1, \ldots, p_r$ aller Primzahlen $p_i \leq N$ und setze $m = n$ sowie $e_1 = \cdots = e_r = 0$.

2. Schritt: Führe für $i = 1, \ldots, r$ die folgenden Anweisungen aus:
Falls m nicht durch p_i teilbar, geht es weiter mit dem nächsten i; andernfalls wird so lange m durch m/p_i und e_i durch $e_i + 1$ ersetzt, bis p_i kein Teiler von m mehr ist. Falls $m = 1$, geht es weiter zu Schritt drei, andernfalls geht es weiter mit dem nächsten i.

3. Schritt: Falls $m = 1$, ist $n = \prod_{i=1}^{r} p_i^{e_i}$ faktorisiert; andernfalls ist $n = \prod_{i=1}^{r} p_i^{e_i} \cdot m$ mit einer Zahl m, die keinen Primteiler $p \leq N$ hat. Falls $m < N^2$, ist m eine Primzahl, und n ist ebenfalls komplett faktorisiert. Andernfalls teste man, ob m nicht eventuell doch prim ist, womit die Faktorisierung ebenfalls beendet ist. Im Falle eines zusammengesetzten m muß dieses mit einem anderen Verfahren weiter untersucht werden.

§2: Die Verfahren von Pollard und ihre Varianten

a) Die Monte-Carlo-Methode

Als etwas weniger systematische Alternative zum Abdividieren könnte man auch eine Folge von Zufallszahlen x_i erzeugen und jeweils den ggT von x_i mit der zu faktorisierenden Zahl N bilden. Dies hat zwar den Nachteil, daß ein EUKLIDischer Algorithmus aufwendiger ist als eine bloße Division mit Rest und daß der ggT möglicherweise eine zusammengesetzte Zahl ist, daür testet man aber in vielen Schritten mehrere Primzahlen auf einmal, und selbst ein zusammengesetzter Faktor ist nützlich, denn je kleiner eine Zahl ist, desto einfacher ist sie zu faktorisieren. Eine weitere Optimierung wird dadurch erreicht, daß wir mehrere x_i modulo N miteinander multiplizieren können und dann erst den ggT des Produkts modulo N mit N bilden. Offensichtlich ist dieser ggT genau dann durch eine Primzahl p teilbar, wenn diese Teiler von N

und von mindestens einem der Faktoren ist. Die Anzahl der Faktoren darf natürlich nicht zu groß sein, denn sonst besteht die Gefahr, daß der ggT einfach gleich N ist. Wenn man aber die kleinen Primzahlen bereits durch Abdividieren eliminiert hat, kann man i.a. relativ gefahrlos mit der Zusammenfassung von etwa hundert Zufallszahlen arbeiten.

Ist p ein Primteiler von N, so sollte bei echten Zufallszahlen etwa jede p-te durch p teilbar sein; ist also p der kleinste Primteiler von N, so kann man erwarten, daß nach p Versuchen ein nichttrivialer Faktor gefunden wird, der p enthält. Dies ist kein Problem für vierstellige Faktoren (die wir allerdings mindestens genauso schnell auch durch Abdividieren bestimmen können), ist aber schon für achtstellige Faktoren viel zu aufwendig.

POLLARDs Idee zur Beschleunigung beruht auf dem 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 $\sqrt{p}$ Folgegliedern 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.

In dieser Form ist das Verfahren allerdings 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 Folgeglieder 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.)

Wegen der speziellen Form der Rekursion hängt die Restklasse $x_{i+1} \bmod p$ 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 $Q\bigl(Q(y)\bigr)$; berechne dann $\mathrm{ggT}(x-y, N)$. 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 und die der x_{2i} simultan, ohne Zwischenergebnisse zu speichern.

Natürlich kann man auch bei dieser Form des Algorithmus mehrere ggT-Berechnungen zusammenfassen: Sollen etwa jeweils s Berechnungen zusammengefaßt weden, so führt man eine neue Variable P ein mit Anfangswert ein ersetzt P im i-ten Schritt durch $P \cdot (x-y) \bmod N$. Nur falls i durch r 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.

JOHN M. POLLARD ist ein britischer Mathematiker, der hauptsächlich bei British Telecom arbeitete. Er publizierte rund zwanzig mathematische Arbeiten, größtenteils auf dem Gebiet der algorithmischen Zahlentheorie. Bekannt sind auch 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 schnellten Faktorisierungsalgorithmen für große Zahlen sind.

b) Die $(p-1)$-Methode

POLLARDs zweite Methode beruht auf dem kleinen Satz von FERMAT: Ist p ein Primteiler von N und r ein Vielfaches von $p-1$, so ist $a^r \equiv 1 \bmod$

p für jedes zu p teilerfremde a; der ggT von a^r mod 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 mod 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} mod N.

Schritt 3: Berechne $\text{ggT}(a, N)$. 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 eine nichttriviale Faktorisierung gefunden, und ein Primzahltest zeigt, daß sowohl der gefundene Faktor als auch sein Komplement prim sind.

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 $N(p) = p-1$. Ein anderer Vorschlag von POLLARD war $G_p = \mathbb{F}_{p^2}^\times / \mathbb{F}_p^\times$; hier ist

$$N(p) = \frac{p^2-1}{p-1} = p+1\,,$$

daher der Name $(p+1)$-Methode. Zur Konstruktion vom G_n brauchen wir zunächst eine Gruppe, die modulo p auf $\mathbb{F}_p^\times$ reduziert; dazu können wir eine geeignete quadratische Erweiterung von $(\mathbb{Z}/n)^\times$ nehmen. G_n ist dann die Faktorgruppe dieser Gruppe nach $(\mathbb{Z}/n)^\times$.

Derzeit am populärsten ist aber 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 $y = mx+c$ ein 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

$$(x_1,y_1) \oplus (x_2,y_2) = \bigl(x_3, -(mx_3+c)\bigr)\,.$$

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.

§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

$$x^2 - y^2 = (x+y)(x-y)\,.$$

Ist $N = pq$ Produkt zweier ungerader Primzahlen, so ist

$$N = (x+y)(x-y) \quad \text{mit} \quad x = \frac{p+q}{2} \quad \text{und} \quad y = \frac{p-q}{2}\,;$$

zusammen mit obiger Formel folgt $N + y^2 = x^2$.

FERMAT berechnet für $y = 0, 1, 2, \dots$ die Zahlen $N + y^2$; falls er auf ein Quadrat x^2 stößt, hat er zwei Faktoren $x \pm y$ gefunden.

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 dies die beiden Zahlen eins und N, so daß dies auf den ersten Blick keine Vorteile gegenüber dem klassischen FERMAT-Verfahren hat, insbesondere da es keine offensichtliche Wahl für k gibt.

Es gibt allerdings eine ganze Reihe von Algorithmen, die ohne Rücksicht auf einen konkreten Wert von k einfach Zahlen $x, y \in \mathbb{Z}$ suchen, für die

$$x^2 \equiv y^2 \bmod N$$

ist, und unter diesen Verfahren sind die besten derzeit bekannten zur Faktorisierung großer Zahlen ohne kleine Primteiler.

Auch hier sind mit etwas Glück $\mathrm{ggT}(x \pm y, N)$ echte Faktoren von N, wenn man Pech hat sind es einfach wieder eins und N. In der Praxis wird man daher von vornherein gleich mehrere Paare (x, y) mit $x^2 \equiv y^2 \bmod N$ suchen um die Chance zu erhöhen, daß zumindest ein Paar auf eine nichttriviale Faktorisierung führt.

Hier soll nur kurz der Grundalgorithmus, das sogenannte quadratische Sieb, beschrieben werden; die wirklich für Rekordfaktorisierungen benutzen Modifikationen sind teilweise mathematisch recht anspruchsvolle Varianten davon.

Im einfachsten Fall arbeiten wir ausschließlich mit dem 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 $x_1, x_2, \dots, x_r$ 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.

Um die 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 etwa 100–120 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

$$\prod_{i=1}^{r} f(x_i)^{\varepsilon_i} = \prod_{p \in \mathcal{B}} p^{\sum_{i=1}^{r} \varepsilon_i e_{ip}}$$

genau dann ein Quadrat, wenn

$$\sum_{i=1}^{r} \varepsilon_i e_{ip}$$

für alle $p \in \mathcal{B}$ gerade ist. Dies hängt natürlich nur ab von den $\varepsilon_i \bmod 2$ und den $e_{ip} \bmod 2$; 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

$$\sum_{i=1}^{r} \varepsilon_i e_{ip} = 0 \quad \text{für alle } p \in \mathcal{B}\,.$$

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

$$\prod_{i=1}^{r} f(x_i)^{\varepsilon_i} = \prod_{i=1}^{r} \left(x + \left[\sqrt{N}\right]\right)^{2\varepsilon_i} \bmod N$$

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 $\mathbb{N}_0$)

$$x = \prod_{p\in\mathcal{B}} p^{\frac{1}{2}\sum_{i=1}^{r}\varepsilon_i e_{ip}} \bmod N \quad \text{und} \quad y = \prod_{i=1}^{r} \left(x + \left[\sqrt{N}\right]\right)^{\varepsilon_i} \bmod N\,.$$

Zum besseren Verständnis des Verfahrens 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	$x + \left[\sqrt{N}\right]$	$f(x)$	*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

$$\text{ggT}(4+1, 15) = 5 \quad \text{und} \quad \text{ggT}(4-1, 15) = 3\,.$$

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:

$$\begin{aligned} 6^2 &\equiv 3 \cdot 7 && \bmod 15\\ 9^2 &\equiv 2 \cdot 3 \cdot 11 && \bmod 15\\ 13^2 &\equiv 2 \cdot 7 \cdot 11 && \bmod 15\\ 54^2 &\equiv 2 \cdot 3 \cdot 7^2 \cdot 11 && \bmod 15 \end{aligned}$$

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

$$\text{ggT}(702 - 462, 15) = \text{ggT}(240, 15) = 15$$

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
$$\text{ggT}(4446 - 3234, 15) = \text{ggT}(1212, 15) = 3\,,$$
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
$$f(x) = \left(x - \left[\sqrt{N}\right]\right)^2 - N = 0 \Longleftrightarrow \left(x - \left[\sqrt{N}\right]\right)^2 = N\,,$$
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
$$x = \left[\sqrt{N}\right] \pm w\,;$$
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, die Faktorbasis sollte also nur Primzahlen enthalten, für die N modulo p ein Quadrat ist. Mit Hilfe des quadratischen Reziprozitätsgesetzes läßt sich leicht bestimmen, für welche Primzahlen dies der Fall ist. Für solche p kann man dann (z.B. mit dem Algorithmus von SHANKS) die beiden Lösungen der Gleichung $f(x) = 0$ in $\mathbb{F}_p$ berechnen.

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 ganzzahlige 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 ganzzahlige 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 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.