Kapitel 7
Quadratische Formen

Eine quadratische Form ist ein Ausdruck der Form

$$F(x,y) = Ax^2 + Bxy + Cy^2 \quad \text{mit} \quad A,B,C \in \mathbb{Z}\,;$$

die Zahlentheorie interessiert sich vor allem dafür, welche Werte $F(x,y)$ für $x,y \in \mathbb{Z}$ annehmen kann.

§ 1: Summen zweier Quadrate

Der einfachste Fall ist die Form $F(x,y) = x^2 + y^2$. Sie 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.

Das Quadrat einer geraden Zahl ist durch vier teilbar, das einer ungeraden Zahl $2k+1$ ist $4k^2+4k+ \equiv 1 \bmod 4$; somit ist jede Summe zweier Quadrate kongruent null, eins oder zwei modulo vier. Eine Zahl kongruent drei modulo vier kann also nicht als Summe zweier Quadratzahlen auftreten.

Auf der Suche nach positiven Ergebnissen können wir uns auf Primzahlen beschränken, denn wie FIBONACCI bereits im dreizehnten Jahrhundert zeigte, gilt:

Lemma: Sind zwei Zahlen $n,m \in \mathbb{N}$ darstellbar als Summen zweier Quadrate, so gilt dasselbe für ihr Produkt nm.

Beweis: Wenn n und m als Summen zweier Quadrate darstellbar sind, gibt es $\alpha, \beta \in \mathbb{Z}[i]$, so daß $n = \mathrm{N}(\alpha)$ und $m = \mathrm{N}(\beta)$ ist. Wegen der Multiplikativität der Norm ist dann $nm = \mathrm{N}(\alpha\beta)$ ebenfalls eine Norm und damit als Summe zweier Quadrate darstellbar. ∎

FIBONACCI bewies dieses Lemma natürlich nicht über den Umweg mit Normen GAUSSscher Zahlen; er fand eine explizite Formel für die Darstellung des Produkts als Summe zweier Quadrate. Es handelt sich dabei um dieselbe Formel, zu der wir auch durch Ausmultiplizieren der Gleichung $\mathrm{N}(\alpha) \cdot \mathrm{N}(\beta) = \mathrm{N}(\alpha\beta)$ für $\alpha = a + ib$ und $\beta = c + id$ kämen:

$$(a^2 + b^2)(c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2 \,.$$

Da $2 = 1^2 + 1^2$ als Summe zweier Quadrate darstellbar ist, müssen wir daher nur die ungeraden Primzahlen untersuchen. Hier wissen wir bereits, daß Zahlen kongruent drei modulo vier keine Summen zweier Quadrate sein können.

Satz: Eine ungerade Primzahl p ist genau dann darstellbar als Summe zweier Quadrate, wenn $p \equiv 1 \bmod 4$. Diese Darstellung ist eindeutig bis auf die Reihenfolge der Summanden.

Beweis: Aus Kapitel I, §8 wissen wir, daß die multiplikative Gruppe des Körper $\mathbb{F}_p$ von einem einzigen Element g erzeugt wird. Für $p = 4k + 1$ ist dann $g^{4k} = 1$, also $g^{2k} = -1$. Somit ist $-1 = p - 1$ in $\mathbb{F}_p$ ein Quadrat.

In $\mathbb{Z}$ gibt es daher Zahlen x, für die $x^2 \equiv -1 \bmod p$ ist oder, anders ausgedrückt, $x^2+1 = \ell p$ für ein $\ell \in \mathbb{N}$. Da jede Restklasse modulo p einen Vertreter mit Betrag kleiner $p/2$ enthält, können wir dabei annehmen, daß $|x| < p/2$ ist; dann ist mit einer geeigneten natürlichen Zahl ℓ

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

Es gibt also ein $\ell < p$, so daß ℓp Summe zweier Quadrate ist. Das kleinste solche ℓ sei m; wir müssen zeigen, daß $m = 1$ ist.

Zunächst ist klar, daß m eine ungerade Zahl sein muß, denn aus der Formel $x^2 + y^2 = mp$ mit geradem m folgt, daß x und y entweder beide

gerade oder beide ungerade sind; $x \pm y$ sind also gerade und

$$\left(\frac{x+y}{2}\right)^2 + \left(\frac{x-y}{2}\right)^2 = \frac{x^2+y^2}{2} = \frac{m}{2}p\,,$$

im Widerspruch zur Minimalität von m.

Falls die Behauptung falsch wäre, müßte somit $m \geq 3$ sein. Wir definieren dann zwei neue Zahlen $u, v \in \mathbb{Z}$ durch die Bedingungen

$$|u| < \frac{m}{2}, \quad |v| < \frac{m}{2}, \quad u \equiv x \bmod m \quad \text{und} \quad v \equiv y \bmod m\,.$$

Offensichtlich können nicht beide dieser Zahlen verschwinden, denn sonst wären x und y beide durch m teilbar, also wäre $x^2 + y^2 = mp$ durch m^2 teilbar. Das kann aber nicht sein, denn p ist prim und $m < p$. Weiter ist

$$u^2 + v^2 \equiv x^2 + y^2 = mp \equiv 0 \bmod m\,,$$

also gibt es eine natürliche Zahl r, so daß $u^2 + v^2 = rm$ ist. Da $u^2 + v^2$ kleiner ist als $\frac{1}{2}m^2$, ist $r < \frac{m}{2}$.

Nach der zu Beginn des Paragraphen zitierten Formel von FIBONACCI, d.h. also durch explizite Berechnung von $(u+iv)(x+iy)$ und Berechnung der Norm davon, erhalten wir die Formel.

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

Dabei ist nach Definition von u und v

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

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

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

Da $r < \frac{m}{2}$, widerspricht dies der Minimalität von m.

Damit haben wir gezeigt, daß $m = 1$ sein muß, d.h. p läßt sich als Summe zweier Quadrate darstellen. Wir müssen uns noch überlegen, daß diese Darstellung bis auf die Reihenfolge der Faktoren eindeutig ist.

Angenommen, es gibt zwei Darstellungen $p = x^2 + y^2 = u^2 + v^2$. In $\mathbb{Z}[i]$ ist dann

$$p = (x + iy)(x - iy) = (u + iv)(u - iv)\,.$$

Alle Faktoren haben Norm p und sind somit irreduzibel, und aus §5 des vorigen Kapitels wissen wir, daß $\mathbb{Z}[i]$ ein EUKLIDischer, insbesondere also faktorieller Ring ist. Daher unterscheiden sich die beiden Zerlegungen nur durch Einheiten von $\mathbb{Z}[i]$. Auch diese kennen wir aus Kapitel 6: Nach dem Lemma aus §6 sind es genau die Elemente ± 1 und $\pm i$. Somit ist entweder $x^2 = u^2$ und $y^2 = v^2$ oder umgekehrt, womit die Eindeutigkeit bewiesen wäre. ∎

Als erste Anwendung davon können wir die Primzahlen im Ring $\mathbb{Z}[i]$ der GAUSSschen Zahlen bestimmen:

Korollar: Eine Primzahl $p \in \mathbb{N}$ ist genau dann irreduzibel in $\mathbb{Z}[i]$, wenn $p \equiv 3 \bmod 4$. Andernfalls zerfällt sie in das Produkt zweier konjugiert komplexer irreduzibler Elemente $r \pm is$ mit $r^2 + s^2 = p$.

Beweis: $p = 2 = (1 + i)(1 - i)$ zerfällt offensichtlich, und dies ist bereits die Primzerlegung, denn $N(1 \pm i) = 2$ hat keine echten Teiler.

Falls eine ungerade Primzahl p einen echten Teiler $r + is$ hat, ist sie auch durch $r - is$ teilbar. Da die Norm von p gleich p^2 ist und $r \pm is$ keine Einheiten, muß $N(r \pm is) = p$ sein. Damit folgt zunächst, daß $r \pm is$ prim sind, denn ein echter Teiler müßte als Norm einen echten Teiler von p haben. Außerdem folgt, daß sich $(r + is)(r - is) = r^2 + s^2$ höchstens durch eine Einheit von p unterscheidet. Da beides positive Zahlen sind, muß diese gleich eins sein, d.h. die Primzerlegung von p in $\mathbb{Z}[i]$ ist

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

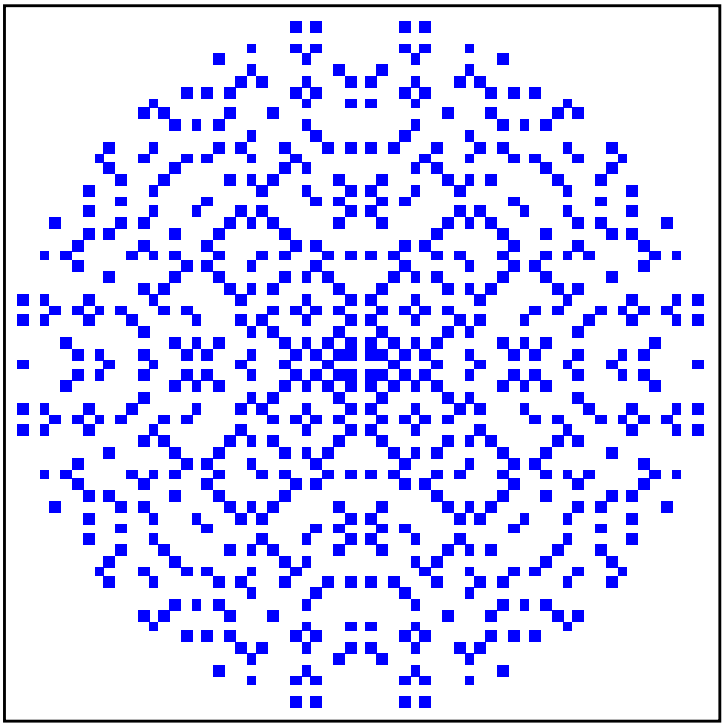
Nach dem Satz ist daher $p \equiv 1 \bmod 4$.

Ist umgekehrt $p \equiv 1 \bmod 4$, so gibt es nach dem Satz zwei ganze Zahlen r, s, so daß $p = r^2 + s^2$ ist, d.h. $p = (r + is)(r - is)$ zerfällt in $\mathbb{Z}[i]$, und das Argument aus dem vorigen Abschnitt zeigt, daß dies die Primzerlegung ist.

Somit zerfallen genau die Primzahlen $p \equiv 1 \bmod 4$ und die Zwei, d.h. genau die $p \equiv 3 \bmod 4$ bleiben prim. ∎

In der Abbildung sind die GAUSSschen Primzahlen $a + ib$ der Norm höchstens 1000 durch Quadrate um den Punkt $(a, b) \in \mathbb{R}^2$ dargestellt.

Mancher Leser wird hier ein gelegentlich von Designern verwendetes Muster erkennen.



Kehren wir zurück zur Ausgangsfrage: Wann kann eine vorgegebene natürliche Zahl als Summe zweier Quadrate dargestellt werden?

Satz: Eine natürliche Zahl n läßt sich genau dann als Summe zweier Quadrate schreiben, wenn jeder Primteiler $p \equiv 3 \bmod 4$ in der Primzerlegung von n mit einer geraden Potenz auftritt.

Beweis: Zunächst ist die Bedingung hinreichend, denn da mit n auch jedes Produkt c^2n als Summe zweier Quadrate darstellbar ist, können wir die Primteiler $p \equiv 3 \bmod 4$ ignorieren. Nach dem gerade bewiesenen Satz wissen wir, daß jede Primzahl $p \equiv 1 \bmod 4$ Summe zweier Quadrate ist, und natürlich gilt dies auch für $2 = 1^2 + 1^2$. Damit ist nach dem obigen Lemma auch jedes Produkt solcher Primzahlen als Summe zweier Quadrate darstellbar.

Umgekehrt sei $n = x^2 + y^2$ und $d = \text{ggT}(x, y)$. Mit $x = du$, $y = dv$ und $n = d^2m$ ist dann $m = u^2 + v^2$, und m enthält genau dann einen Primteiler $p \equiv 3 \bmod 4$ in ungerader Potenz, wenn dies für n der Fall ist.

Ein solcher Primteiler p teilt auch $u^2+v^2 = (u+iv)(u-iv)$ im Ring $\mathbb{Z}[i]$ der GAUSSschen Zahlen. Falls p auch dort eine Primzahl ist, muß p mindestens einen der beiden Faktoren teilen; komplexe Konjugation zeigt, daß es dann auch den anderen teilt. Damit teilt es auch deren Summe $2u$ und Differenz $2iv$; da p ungerade ist und i eine Einheit, teilt p also die zueinander teilerfremden Zahlen u und v, ein Widerspruch.

Somit ist p in $\mathbb{Z}[i]$ keine Primzahl; nach obigem Korollar muß daher $p = 2$ oder $p \equiv 1 \bmod 4$ sein. Damit ist jeder Primteiler $p \equiv 3 \bmod 4$ von n zugleich ein Teiler von d und tritt in n daher mit einer geraden Potenz auf. ∎

Für zusammengesetzte Zahlen ist die Darstellung als Summe zweier Quadrate im allgemeinen nicht mehr eindeutig. Über die Primzerlegung in $\mathbb{Z}[i]$ läßt sich die Anzahl verschiedener Darstellungen leicht erkennen: Natürlich entsprechen auch für eine beliebige natürliche Zahl n die Darstellungen als Summe zweier Quadrate den Darstellungen von n als Norm eines Elements von $\mathbb{Z}[i]$, wobei assoziierte Elemente bis auf Reihenfolge auf dieselbe Zerlegung führen.

Aus der Primzerlegung von n in $\mathbb{Z}$ können wir leicht auf die Primzerlegung in $\mathbb{Z}[i]$ schließen: Primzahlen kongruent drei modulo vier bleiben nach obigem Korollar auch in $\mathbb{Z}[i]$ irreduzibel, die kongruent eins modulo vier sind Produkte zweier konjugierter Elemente $x \pm iy$. Die beiden Faktoren sind nicht assoziiert, denn sonst wäre $|x| = |y|$ und $p = x^2 + y^2$ wäre gerade. Die Zwei schließlich ist Produkt der beiden irreduziblen Elemente $1 \pm i$, und die sind assoziiert zueinander, denn $(1-i)\cdot i = 1 + i$.

Wir sortieren daher in der Primzerlegung von n nach den Kongruenzklassen modulo vier der Primfaktoren:

$$n = 2^e \prod_{j=1}^{r} p_j^{f_j} \prod_{k=1}^{s} q_k^{2g_k} \quad \text{mit} \quad p_j \equiv 1 \bmod 4, \quad q_k \equiv 3 \bmod 4\,.$$

Für jedes p_j wählen wir ein $\pi_j \in \mathbb{Z}[i]$ derart, daß $\pi_j \cdot \overline{\pi}_j = p_j$ ist; dann

ist n in $\mathbb{Z}[i]$ assoziiert zu

$$(1+i)^{2e} \prod_{j=1}^{r} \pi_j^{f_j} \prod_{j=1}^{r} \overline{\pi}_j^{f_j} \prod_{k=1}^{s} q_k^{2g_k} \ .$$

Ein Element $\alpha \in \mathbb{Z}[i]$, für das $\mathrm{N}(\alpha) = n$ sein soll, hat daher bis auf eine Einheit die Form

$$\alpha = (1+i)^{e} \prod_{j=1}^{r} \pi_j^{h_j} \prod_{j=1}^{r} \overline{\pi}_j^{f_j - h_j} \prod_{k=1}^{s} q_k^{g_k} \ ,$$

mit $0 \le h_j \le f_j$. Die Anzahl verschiedener Möglichkeiten ist somit gleich dem Produkt der $(f_j + 1)$, wobei hier allerdings die Darstellungen $n = x^2 + y^2$ und $n = y^2 + x^2$ für $x \neq y$ als verschieden gezählt werden.

Die im Vergleich zur Größe von n meisten verschiedenen Darstellungen gibt es offenbar dann, wenn n ein Produkt verschiedener Primzahlen ist, die allesamt kongruent eins modulo vier sind. In diesem Fall ist die Anzahl der Darstellungen gleich zwei hoch Anzahl der Faktoren.

§2: Anwendung auf die Berechnung von π

Aus der Analysis I ist bekannt, daß gilt

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \frac{x^{11}}{11} + \frac{x^{13}}{13} - \frac{x^{15}}{15} + \cdots \ ;$$

falls es jemand nicht mehr weiß: Die Ableitung des Arkustangens ist $1/(1+x^2)$, und nach der Summenformel für die geometrische Reihe ist

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + x^8 - x^{10} + x^{12} - x^{14} + \cdots \ .$$

Durch gliedweise Integration folgt wegen $\arctan 0 = 0$ die obige Formel. Eine bekannte Anwendung davon ist der Spezialfall $x = 1$:

$$\frac{\pi}{4} = \arctan 1 = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \frac{1}{13} - \frac{1}{15} + \cdots \ .$$

Zur praktischen Berechnung von π ist diese Formel allerdings völlig unbrauchbar und der Alptraum eines jeden Numerikers: Zunächst einmal sind alternierende Summen grundsätzlich problematisch, allerdings ist

das hier vergleichsweise harmlos: Wenn wir jeden negativen Summanden von seinem Vorgänger subtrahieren, bekommen wir eine Reihe

$$\frac{\pi}{4} = \frac{2}{1\cdot 3} + \frac{2}{5\cdot 7} + \frac{2}{9\cdot 11} + \frac{2}{13\cdot 15} + \cdots$$

mit lauter positiven Gliedern. Die Summanden sind jedoch immer noch monoton fallend, so daß die Rundungsfehler der ersten Additionen bei hinreichend langer Summation größer sind als die hinteren Summanden. Man muß also, wenn man eine endliche Teilsumme berechnen will, von hinten nach vorne summieren und damit bereits vor Beginn der Rechnung die Anzahl der Terme festlegen. Bei jeder Erhöhung der Anzahl der Summanden muß die gesamte Rechnung von vorne beginnen.

Dazu kommt, daß die Reihe extrem langsam konvergiert: Divierenden wir obige Gleichung durch zwei und berechnen 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 erhalten wir für die ersten Zehnerpotenzen N die folgenden Fehler:

N	10	100	1 000	10 000
$\pi - 8S_N$	$4{,}5\cdot 10^{-2}$	$5{,}0\cdot 10^{-3}$	$5{,}0\cdot 10^{-4}$	$5{,}0\cdot 10^{-5}$
N	100 000	1 000 000	10 000 000	100 000 000
$\pi - 8S_N$	$5{,}0\cdot 10^{-6}$	$5,0\cdot 10^{-7}$	$5{,}0\cdot 10^{-8}$	$5{,}0\cdot 10^{-9}$

Für eine zusätzliche Dezimalstelle muß also der Rechenaufwand ziemlich genau verzehnfacht werden. Angesichts der Tatsache, daß heute mehrere Billionen Ziffern von π bekannt sind ist klar, daß es bessere Wege zur Berechnung von π geben muß.

Einer davon benutzt Zahlen mit einer großen Anzahl verschiedener Darstellungen als Summen zweier Quadrate. Die Reihe für den Arkustangens konvergiert sicherlich umso besser, je kleiner der Wert von x ist. Wenn wir also den Winkel $\frac{\pi}{4}$ aufteilen können in mehrere kleine Winkel,

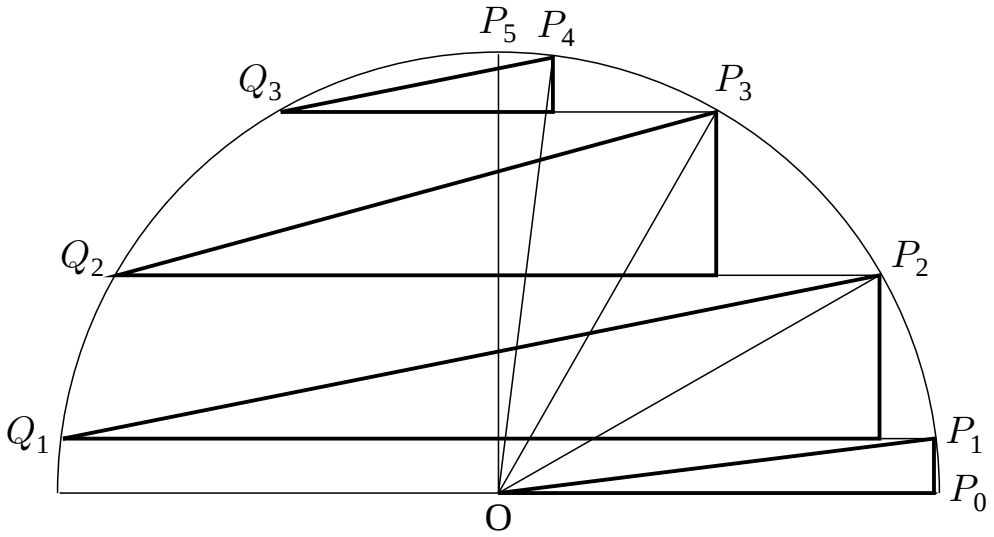
deren Tangens wir kennen, sollten bessere Ergebnisse zu erwarten sein. Genau das können wir mit solchen Zahlen erreichen.

Angenommen, wir haben für eine Zahl n die r verschiedenen Darstellungen

$$n = x_1^2 + y_1^2 = \cdots = x_r^2 + y_r^2$$

als Summen von Quadraten, wobei $y_1 < \cdots < y_r$ sei. Dann ist $x_i = y_{r-i}$, denn wir können ja in jeder Darstellung die Reihenfolge der Faktoren vertauschen. Wir wollen außerdem voraussetzen, daß n nicht das Doppelte eines Quadrats ist, so daß stets $x_i \neq y_i$ und somit r eine gerade Zahl ist.

Die Punkte $P_i = (x_i, y_i)$ und $Q_i = (-x_i, y_i)$ für $i = 1, \ldots, r$ liegen auf der Kreislinie $x^2 + y^2 = n$ um den Nullpunkt O, genauso die drei Punkte $P_0 = \left(\sqrt{n}, 0\right)$, $Q_0 = \left(-\sqrt{n}, 0\right)$ und $P_{r+1} = \left(0, \sqrt{n}\right)$.



Da die y-Koordinaten y_i der P_i der Größe nach geordnet sind, ist

$$\frac{\pi}{2} = \sum_{i=0}^{r} \measuredangle OP_iP_{i+1} = 2 \sum_{i=0}^{r/2-1} \measuredangle OP_iP_{i+1} + \measuredangle OP_{r/2}P_{r/2+1}\,.$$

Leider ist keines der Dreiecke $\triangle OP_iP_{i+1}$ rechtwinklig, so daß uns die ganzzahligen Koordinaten der (meisten) P_i bei der Berechnung der Winkel $\measuredangle OP_iP_{i+1}$ nichts nützen.

Nun lehrt uns aber ein Satz der Elementargeometrie, der (im Anhang zu diesem Paragraphen bewiesene) Satz vom Innenwinkel, daß der Winkel $\measuredangle OP_iP_{i+1}$ doppelt so groß ist wie der Winkels $\measuredangle Q_iP_iP_{i+1}$. Letzterer gehört zu einem rechtwinkligen Dreieck, denn natürlich ändert sich nichts am Winkel, wenn wir den Punkt P_i ersetzen durch die senkrechte Projektion $P_i' = (x_{i+1}, y_i)$ von P_{i+1} auf die Gerade Q_iP_i. Somit ist

$$\frac{\pi}{2} = 2\measuredangle OP_0'P_1 + 4\sum_{i=1}^{r/2-1} \measuredangle Q_iP_i'P_{i+1} + 2\measuredangle Q_{r/2}P_{r/2}'P_{r/2+1}\,.$$

Division durch zwei macht daraus

$$\frac{\pi}{4} = \measuredangle OP_0'P_1 + 2\sum_{i=1}^{r/2-1} \measuredangle Q_iP_i'P_{i+1} + \measuredangle Q_{r/2}P_{r/2}'P_{r/2+1}\,.$$

In dieser Darstellung sind die drei Punkte, die den Winkel bestimmen, in allen Fällen die Eckpunkte eines rechtwinkligen Dreiecks, sie haben allesamt ganzzahlige Koordinaten, und zumindest die Katheten der Dreiecke haben ganzzahlige Längen. Somit können wir alle auftretenden Winkel ausdrücken durch Arkustangenswerte rationaler Zahlen.

Als Beispiel betrachten wir das kleinste Produkt dreier verschiedener Primzahlen kongruent eins modulo vier, also $n = 5\cdot 13\cdot 17 = 1105$. Aus den Darstellungen

$$5 = 1^2 + 2^2, \quad 13 = 2^2 + 3^2 \quad \text{und} \quad 17 = 1^2 + 4^2$$

verschafft man sich leicht die vier Darstellungen

$$1105 = 4^2 + 33^2 = 9^2 + 32^2 = 12^2 + 31^2 = 23^2 + 24^2\,,$$

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

$$P_1 = (33,4)\,,\ P_2 = (32,9)\,,\ P_3 = (31,12)\,,\ P_4 = (24,23)\,,$$
$$P_8 = (4,33)\,,\ P_7 = (9,32)\,,\ P_6 = (12,31)\,,\ P_5 = (23,24)\,;$$

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

Die Q_i für $1 \leq i \leq 8$ unterscheiden sich von den P_i nur durch das Vorzeichen der Abszisse. Damit können wir die Tangenten aller Winkel bei O berechnen:

$$\tan \measuredangle OP_0P_1 = \tan \measuredangle OP_8P_9 = \frac{y_1}{x_1} = \frac{4}{33}$$

$$\tan \measuredangle OP_1P_2 = \tan \measuredangle OP_7P_8 = \tan 2\measuredangle Q_1P_1P_2 = \frac{y_2 - y_1}{x_1 + x_2} = \frac{5}{65} = \frac{1}{13}$$

$$\tan \measuredangle OP_2P_3 = \tan \measuredangle OP_6P_7 = \tan 2\measuredangle Q_2P_2P_3 = \frac{y_3 - y_2}{x_2 + x_3} = \frac{3}{63} = \frac{1}{21}$$

$$\tan \measuredangle OP_3P_4 = \tan \measuredangle OP_5P_6 = \tan 2\measuredangle Q_3P_3P_4 = \frac{y_4 - y_3}{x_3 + x_4} = \frac{11}{55} = \frac{1}{5}$$

$$\tan \measuredangle OP_4P_5 = \tan 2\measuredangle Q_4P_4P_5 = \frac{y_5 - y_4}{x_4 + x_5} = \frac{1}{47}$$

Die Summe aller dieser Winkel ist

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

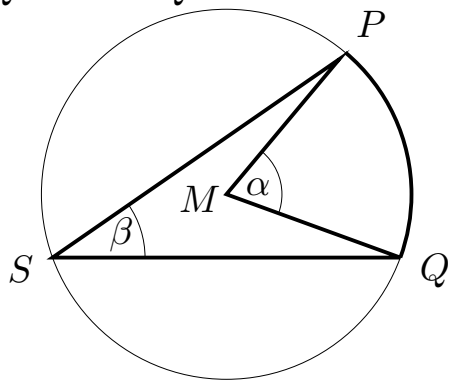
Approximieren wir dies, indem wir jede der TAYLOR-Reihen durch das TAYLOR-Polynom vom Grad n ersetzen, erhalten wir die folgenden betragsmäßigen Abweichungen Δ_n zwischen π und dem Vierfachen dieser Summe:

n	1	3	5	7	9
Δ_n	$2{,}5 \cdot 10^{-2}$	$5{,}2 \cdot 10^{-4}$	$1{,}4 \cdot 10^{-5}$	$4{,}4 \cdot 10^{-7}$	$1{,}4 \cdot 10^{-8}$
n	11	13	15	17	19
Δ_n	$5 \cdot 10^{-10}$	$4{,}9 \cdot 10^{-10}$	$6 \cdot 10^{-13}$	$2{,}1 \cdot 10^{-14}$	$7{,}7 \cdot 10^{-16}$

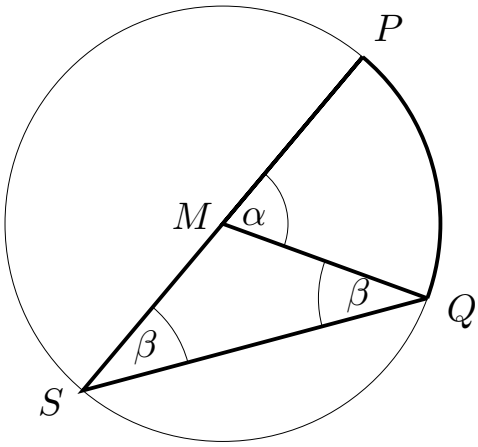
Die Verbesserung gegenüber der Berechnung via $\frac{\pi}{4} = \arctan 1$ ist dramatisch: Die dort betrachtete Teilsumme S_N entspricht der Auswertung des TAYLOR-Polynoms vom Grad $n = 4N + 3$, und selbst wenn wir N auf hundert Millionen setzen, haben wir noch einen Fehler von $5 \cdot 10^{-7}$. Mit dem neuen Ansatz kommen wir bereits mit TAYLOR-Polynomen vom Grad neun auf einen Fehler, der gerade mal ein Zehntel davon beträgt. An Stelle von hundert Millionen Summanden mußten wir dazu nur fünf TAYLOR-Polynome mit jeweils fünf Summanden auswerten.

Anhang: Der Satz vom Innenwinkel

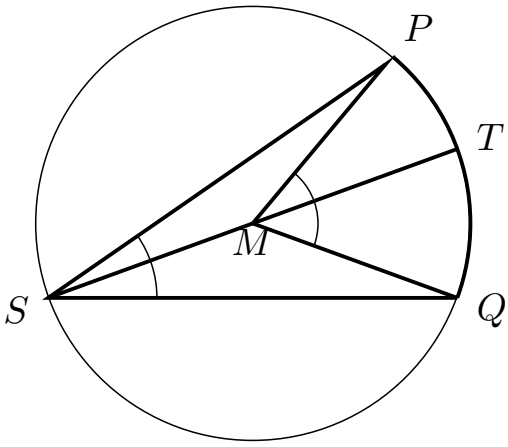
Satz: P, Q, S seien Punkte auf einer Kreislinie mit Mittelpunkt M. Dann ist $\angle MPQ = 2\angle SPQ$.



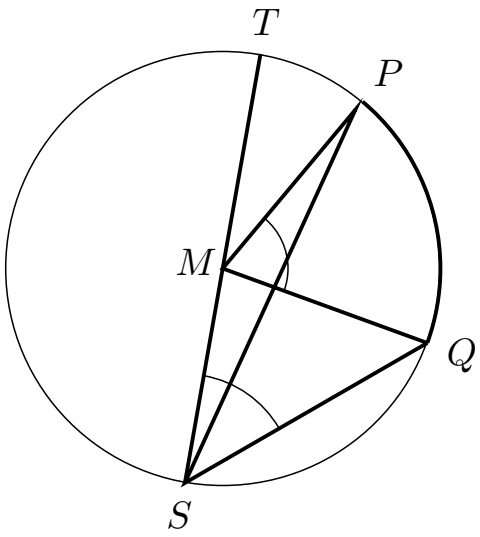
Beweis: Am einfachsten ist der Fall, daß M auf der Verbindungsstrecke von S mit einem der beiden Punkte P und Q liegt; wir nehmen an, er liege auf $\overline{SP}$. (Der andere Fall ist spiegelsymmetrisch dazu und geht genauso.) Dann ist das Dreieck $\triangle MSQ$ gleichschenklig, d.h. wir haben bei S und bei Q denselben Winkel β. Der verbleibende Dreieckswinkel bei M ist somit $\pi - 2\beta$. Andererseits ist dies aber der Komplementärwinkel zu $\alpha = \angle MPQ$, also ist $\alpha = 2\beta$, wie behauptet.



Der allgemeine Fall kann auf diesen Spezialfall zurückgeführt werden: Liegen P und Q auf verschiedenen Seiten des Durchmessers durch S, dessen anderer Endpunkt T sei, so erfüllen auch die Punkte S, P, T, M sowie die Punkte S, Q, T, M die Voraussetzung des Satzes, und in beiden Fällen sind wir in der Situation des bereits bewiesenen Spezialfalls. Addition der Ergebnisse für diese beiden Fälle liefert das Ergebnis für die Punkte S, P, Q, M.



Bleibt noch der Fall, daß P und A auf derselben Seite des Durchmessers $\overline{ST}$ liegen. Auch in diesem Fall erfüllen wieder sowohl die Punkte S, P, T, M als auch die Punkte S, Q, T, M die Voraussetzungen des Satzes, und beides Mal sind wir in der Situation des eingangs bewiesenen Spezialfalls. Dieses Mal führt die Subtraktion dieser beiden Ergebnisse zum gewünschten Resultat für die Ausgangssituation mit den Punkten S, P, Q, M.



Damit ist der Satz vollständig bewiesen. ∎

§3: Der Satz von Lagrange

Es ist nicht möglich, eine beliebige natürliche Zahl als Summe von höchstens drei Quadratzahlen zu schreiben; das kleinste Gegenbeispiel ist die Sieben. Wie EULER vermutete und LAGRANGE bewies, kommt man aber immer mit höchstens vier Quadratzahlen aus.

Einer der vielen Beweise dieses Satzes ist recht ähnlich zu dem des Zweiquadratesatzes aus §1; statt mit dem Ring $\mathbb{Z}[i]$ der GAUSSschen Zahlen arbeiten wir aber mit dem Ring

$$\mathbb{Z} \oplus \mathbb{Z}\mathbf{i} \oplus \mathbb{Z}\mathbf{j} \oplus \mathbb{Z}\mathbf{k}$$

der ganzen Quaternionen. Auch hier haben wir eine Normabbildung, und eine ganze Zahl n ist offensichtlich genau dann als Summe von vier Quadraten darstellbar, wenn sie Norm einer ganzen Quaternion ist. Wegen der Multiplikativität der Norm reicht es also wieder, wenn wir Primzahlen p betrachten.

Zur Vorbereitung zeigen wir zunächst

Lemma: Zu jeder Primzahl p gibt es ganze Zahlen $x, y, z \in \mathbb{Z}$ und eine natürliche Zahl $m < p$, so daß gilt: $mp = x^2 + y^2 + z^2$

Beweis: Für $p = 2$ ist $2 = 1^2 + 1^2 + 0^2$; sei also p ungerade.

Von den Zahlen a^2 mit $0 \le a \le \frac{1}{2}(p-1)$ sind keine zwei kongruent modulo p, denn $a^2 - b^2 = (a+b)(a-b)$, und falls $0 \le a, b < \frac{1}{2}(p-1)$ sind beide Faktoren kleiner als p. Damit gibt es auch in den Mengen

$$\mathcal{M}_1 = \left\{ -a^2 \;\middle|\; 0 \le a \le \tfrac{1}{2}(p-1) \right\}$$

und

$$\mathcal{M}_2 = \left\{ 1 + a^2 \;\middle|\; 0 \le a \le \tfrac{1}{2}(p-1) \right\}$$

keine zwei Elemente, die modulo p kongruent sind. Da die beiden Mengen disjunkt sind und jede davon $\frac{1}{2}(p+1)$ Elemente hat, enthält ihre Vereinigung $p+1$ Elemente; hier muß es also mindestens zwei Elemente geben, die modulo p kongruent sind. Es gibt also Zahlen $x, y \in \mathbb{Z}$ mit $-x^2 \equiv 1 + y^2 \bmod p$, d.h. $x^2 + y^2 + 1^2 = mp$ ist durch p teilbar. Da $x, y \le \frac{1}{2}(p-1)$, ist dabei $m < p$ und das Lemma ist bewiesen. ∎

Lemma: Jede Primzahl p läßt sich als Summe von höchstens vier Quadraten schreiben.

Beweis: Für $p = 2$ wissen wir das; sei also p wieder ungerade. Nach dem vorigen Lemma gibt es eine natürliche Zahl $m < p$ derart, daß mp als Summe von sogar höchstens drei Quadraten darstellbar ist; ℓ sei die kleinste natürliche Zahl, für die ℓp als Summe von höchstens vier Quadraten darstellbar ist. Natürlich ist dann auch $\ell < p$.

Wäre ℓ 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

$$\left(\frac{w+x}{2}\right)^2 + \left(\frac{w-x}{2}\right)^2 + \left(\frac{y+z}{2}\right)^2 + \left(\frac{y-z}{2}\right)^2 = \frac{w^2+x^2+y^2+z^2}{2} = \frac{\ell}{2}p\,,$$

im Widerspruch zur Minimalität von ℓ. Also ist ℓ ungerade, und falls das Lemma falsch wäre, müßte $\ell \ge 3$ sein.

Wir betrachten die modulo ℓ zu w, x, y, z kongruenten ganzen Zahlen W, X, Y, Z vom Betrag kleiner $\ell/2$. Wie schon beim Zwei-Quadrate-Satz können diese nicht allesamt verschwinden, denn sonst wären w, x, y, z durch ℓ teilbar, also ihre Quadratsumme ℓp durch ℓ^2, was wegen $\ell < p$ für eine Primzahl p nicht möglich ist.

Somit ist $0 < W^2 + X^2 + Y^2 + Z^2 < 4 \cdot \left(\frac{\ell}{2}\right)^2 = \ell^2$. Andererseits ist aber

$$W^2 + X^2 + Y^2 + Z^2 \equiv w^2 + x^2 + y^2 + z^2 \equiv 0 \bmod \ell\,;$$

also ist

$$W^2 + X^2 + Y^2 + Z^2 = \ell m \quad \text{mit} \quad 1 \le m < \ell\,.$$

Damit haben die Quaternionen

$$q = w + \mathbf{i}x + \mathbf{j}y + \mathbf{k}z \quad \text{und} \quad Q = W + \mathbf{i}X + \mathbf{j}Y + \mathbf{k}Z$$

die Normen $\mathrm{N}(q) = \ell p$ und $\mathrm{N}(Q) = \ell m$, ihr Produkt hat also die Norm $\ell^2 mp$. Zumindest von der Norm her spricht also nichts dagegen, daß dieses Produkt durch ℓ teilbar sein könnte.

Tatsächlich ist $q\overline{Q}$ durch ℓ teilbar, und das sieht man am schnellsten durch brutales Nachrechnen: In

$$\begin{aligned} q\overline{Q} = \quad & (wW + xX + yY + zZ) \\ & + (-wX + xW - yZ + zY)\mathbf{i} \\ & + (-wY + yW - zX + xZ)\mathbf{j} + (-wZ + zW - xY + yX)\mathbf{k} \end{aligned}$$

sind alle vier Klammern durch ℓ teilbar, denn modulo ℓ sind alle Großbuchstaben gleich den entsprechenden Kleinbuchstaben, so daß die Koeffizienten von $\mathbf{i}, \mathbf{j}, \mathbf{k}$ trivialerweise modulo ℓ verschwinden, und für den Realteil haben wir

$$wW + xX + yY + zZ \equiv w^2 + x^2 + y^2 + z^2 = \ell p \equiv 0 \bmod \ell\,.$$

Somit ist

$$\frac{q\overline{Q}}{\ell} = A + B\mathbf{i} + C\mathbf{j} + D\mathbf{k}$$

eine Quaternion mit ganzzahligen Koeffizienten, und

$$A^2 + B^2 + C^2 + D^2 = \mathrm{N}\left(\frac{q\overline{Q}}{\ell}\right) = \frac{\mathrm{N}(q\overline{Q})}{\ell^2} = \frac{\mathrm{N}(q)\mathrm{N}(Q)}{\ell^2} = mp\,.$$

Dies widerspricht aber der Minimalität von ℓ.

Somit muß $\ell = 1$ sein, und der Satz ist bewiesen. ∎

Satz (LAGRANGE)**:** Jede natürliche Zahl läßt sich als Summe von höchstens vier Quadraten schreiben.

Beweis: Wie wir in Kapitel 6, §7 gesehen haben, läßt sich eine Zahl n genau dann als Summe von höchstens vier Quadraten schreiben, wenn sie Norm einer ganzen Quaternion ist. Da wir gerade gesehen haben, daß sich jede Primzahl als Summe von höchstens vier Quadraten schreiben läßt (und die Eins natürlich auch), folgt die Behauptung aus der Multiplikativität der Norm. ∎

§4: Quadratische Formen und Matrizen

Nachdem wir in den vorigen Paragraphen gesehen haben, daß die spezielle quadratische Form x^2+y^2 vielfältige Beziehungen sowohl zum Zahlkörper $\mathbb{Q}[i]$ als auch zu Anwendungen außerhalb der Zahlentheorie haben, wollen wir uns nun etwas mit der allgemeinen Theorie solcher Formen beschäftigen. In den nächsten Paragraphen werden wir sie dann auf quadratische Zahlkörper und die PELLsche Gleichung anwenden.

Viele abstrakte Aussagen über quadratische Formen werden einfacher, wenn wir sie in lineare Algebra übersetzen. In Matrixschreibweise ist

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

die quadratische Form kann also auch durch die symmetrische Matrix Q beschrieben werden.

Die Determinante von Q ist $AC - \frac{1}{4}B^2$; bis auf einen Faktor -4 ist das die Zahl $B^2 - 4AC$, die wir in Kapitel 6, §2 als Diskriminante eines Elements eines quadratisches Zahlkörpers definiert haben. Wir können also hoffen, daß uns die lineare Algebra via Determinantentheorie Aussagen über die Werte einer quadratischen Form sowie über Zusammenhänge zwischen den Diskriminanten verschiedener Elemente eines quadratischen Zahlkörpers gibt.

Die reellen Werte, die eine quadratische Form annehmen kann, hängen nicht davon ab, in welcher Basis des $\mathbb{R}^2$ wir das Argument $\binom{x}{y}$ darstellen; wir können die Basis daher bei Bedarf beliebig ändern.

Das ist zum Beispiel nützlich bei der Frage, wann eine quadratische Form nur positive oder nur negative Werte annimmt:

Definition: Eine symmetrische Matrix $Q \in \mathbb{R}^{2\times 2}$ und die dadurch definierte quadratische Form $f_Q(x,y) = (x\ y)Q\binom{x}{y}$ heißen $\left\{\begin{matrix}\text{positiv}\\ \text{negativ}\end{matrix}\right\}$ semidefinit, wenn $f_Q(x,y)$ $\left\{\begin{matrix}\geq\\ \leq\end{matrix}\right\}$ 0 für alle $x, y \in \mathbb{R}$. Sie heißen $\left\{\begin{matrix}\text{positiv}\\ \text{negativ}\end{matrix}\right\}$ definit, wenn zusätzlich $f(x,y)$ nur für $x = y = 0$ verschwindet.

Ein typisches Beispiel einer positiv definiten quadratischen Form ist x^2+y^2; hier ist Q einfach die Einheitsmatrix. Die negative Einheitsmatrix führt auf die negativ definite Form $-x^2 - y^2$, die Diagonalmatrix mit Einträgen +1 und -1 auf $x^2 - y^2$, was sowohl positive als auch negative Werte annehmen kann.

Beispiele von positiv semidefiniten, aber nicht positiv definiten Formen sind etwa x^2, $(x + y)^2$ oder $(3x - 4y)^2$ mit zugehörigen Matrizen

$$Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{und} \quad Q_3 = \begin{pmatrix} 3 & -6 \\ -6 & 4 \end{pmatrix}.$$

Für ein allgemeines Kriterium, wann eine Matrix positiv oder negativ (semi-)definit ist, erinnern wir uns an einen Satz aus der linearen Algebra:

Satz: Alle Eigenwerte einer symmetrischen Matrix $Q \in \mathbb{R}^{n\times n}$ sind reell und ihre geometrische Vielfachheit stimmt mit der algebraischen überein. Eigenvektoren zu verschiedenen Eigenwerten stehen senkrecht aufeinander; insbesondere hat $\mathbb{R}^n$ eine Orthonormalbasis aus Eigenvektoren von Q.

Für Leser, die diesen sogenannten *Spektralsatz* nicht kennen, sei kurz ein Beweis für den Spezialfall $n = 2$ skizziert.

Die Matrix $Q = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ hat das charakteristische Polynom

$$\det(Q - \lambda E) = \begin{vmatrix} a-\lambda & b \\ b & c-\lambda \end{vmatrix} = (a-\lambda)(c-\lambda) - b^2$$

$$= \lambda^2 - (a+c)\lambda + ac - b^2 = \left(\lambda - \frac{a+c}{2}\right)^2 - \left(\frac{a+c}{2}\right)^2 + ac - b^2$$

$$= \left(\lambda - \frac{a+c}{2}\right)^2 - \left(\frac{a-c}{2}\right)^2 - b^2 ;$$

die Eigenwerte sind also $\lambda_{1/2} = \frac{a+c}{2} \pm \sqrt{\left(\frac{a-c}{2}\right)^2 + b^2}$.

Da die Summe zweier Quadrate reeller Zahlen nicht negativ sein kann, ist die Wurzel reell, und damit haben wir auch reelle Eigenwerte.

Für $n = 2$ gibt es genau dann einen Eigenwert mit algebraischer Vielfachheit ungleich eins, wenn die beiden Nullstellen des charakteristischen Polynoms gleich sind, wenn also der Ausdruck unter der Wurzel verschwindet, d.h. $a = c$ und $b = 0$. In diesem Fall ist $Q = \left(\begin{smallmatrix} a & 0 \\ 0 & a \end{smallmatrix}\right)$ eine Diagonalmatrix, der Eigenraum ist also der gesamte $\mathbb{R}^2$, so daß auch die geometrische Vielfachheit des Eigenwerts zwei ist.

Ist $\lambda_1 \neq \lambda_2$ und sind $\vec{a}_1, \vec{a}_2$ Eigenvektoren dazu, so ist

$$(Q\vec{a}_1) \cdot \vec{a}_2 = (\lambda_1 \vec{a}_1) \cdot \vec{a}_2 = \lambda_1 (\vec{a}_1 \cdot \vec{a}_2) .$$

Andererseits ist (wie man nötigenfalls leicht nachrechnet) für je zwei Vektoren $\vec{v}, \vec{w}$ und eine Matrix Q das Skalarprodukt $(Q\vec{v}) \cdot \vec{w}$ gleich dem Skalarprodukt $\vec{v} \cdot (Q^T \vec{w})$, wobei Q^T die zu Q transponierte Matrix bezeichnet. Für eine symmetrische Matrix ist $Q = Q^T$, also

$$(Q\vec{a}_1) \cdot \vec{a}_2 = \vec{a}_1 \cdot (Q\vec{a}_2) = \vec{a}_1 \cdot (\lambda_2 \vec{a}_2) = \lambda_2 (\vec{a}_1 \cdot \vec{a}_2) .$$

Da $\lambda_1 \neq \lambda_2$, können $\lambda_1(\vec{a}_1 \cdot \vec{a}_2)$ und $\lambda_2(\vec{a}_1 \cdot \vec{a}_2)$ nur dann gleich sein, wenn $\vec{a}_1 \cdot \vec{a}_2$ verschwindet, d.h. $\vec{a}_1$ und $\vec{a}_2$ stehen senkrecht aufeinander. ■

Auch ein weiteres allgemeines Resultat aus der Linearen Algebra läßt sich hier einfach und direkt beweisen: Das Produkt aller Eigenwerte

einer $n \times n$-Matrix ist gleich der Determinante und die Summe gleich der Spur, der Summe der Diagonalelemente also.

Für symmetrische 2×2-Matrizen folgt dies sofort aus den obigen Formeln für die beiden Eigenwerte: Bei der Addition fällt die Wurzel weg, so daß wir Summe $a + c$ erhalten, und das Produkt ist nach der dritten binomischen Formel gleich

$$\left(\frac{a+c}{2}\right)^2 - \left[\left(\frac{a-c}{2}\right)^2 + b^2\right] = ac - b^2 = \det Q\,.$$

Ist $\vec{a}_1$ ein Eigenvektor zu λ_1 und $\vec{a}_2$ einer zu λ_2, so können wir jeden Vektor $\binom{x}{y}$ aus $\mathbb{R}^2$ als Linearkombination $\binom{x}{y} = u\vec{a}_1 + v\vec{a}_2$ schreiben. Der Zeilenvektor $(x\ y)$ ist dann die entsprechende Linearkombination $u\vec{a}_1^T + v\vec{a}_2^T$ der zu den $\vec{a}_i$ gehörigen Zeilenvektoren $\vec{a}_i^T$, und

$$(x\ y)Q\begin{pmatrix}x\\y\end{pmatrix} = (u\vec{a}_1^T + v\vec{a}_2^T)Q(u\vec{a}_1 + v\vec{a}_2)$$

$$= (u\vec{a}_1^T + v\vec{a}_2^T)\Big(uQ\vec{a}_1 + vQ\vec{a}_2\Big) = (u\vec{a}_1^T + v\vec{a}_2^T)\Big(u\lambda_1\vec{a}_1 + v\lambda_2\vec{a}_2\Big)$$

$$= \lambda_1 u^2\Big(\vec{a}_1^T\vec{a}_1\Big) + \lambda_2 uv\Big(\vec{a}_1^T\vec{a}_2\Big) + \lambda_1 uv\Big(\vec{a}_2^T\vec{a}_1\Big) + \lambda_2 v^2\Big(\vec{a}_2^T\vec{a}_2\Big)\,.$$

Das Matrixprodukt $\vec{a}_i^T\vec{a}_j$ ist gleich dem üblichen Skalarprodukt $\vec{a}_i \cdot \vec{a}_j$; da die a_i aufeinander senkrecht stehen, verschwindet es für $i \neq j$, d.h.

$$(x\ y)Q\begin{pmatrix}x\\y\end{pmatrix} = \lambda_1 u^2\vec{a}_1 \cdot \vec{a}_1 + \lambda_2 v^2\vec{a}_2 \cdot \vec{a}_2\,,$$

wobei die Skalarprodukte der $\vec{a}_i$ mit sich selbst natürlich positiv sind. Falls wir $\vec{a}_1$ und $\vec{a}_2$ als Einheitsvektoren wählen, sind sie eins und können ganz weggelassen werden.

Damit ist klar, daß die quadratische Form genau dann nur nichtnegative Werte annimmt, wenn λ_1 und λ_2 beide positiv sind; genau dann, wenn beide negativ sind, nimmt sie nur nichtpositive Werte an. Die Matrix Q und die zugehörige quadratische Form sind also genau dann positiv *bzw.* negativ semidefinit, wenn beide Eigenwerte ≥ 0 *bzw.* ≤ 0 sind. Sie sind positiv *bzw.* negativ definit, wenn beide echt positiv *bzw.* negativ sind.

Für eine positiv oder negativ semidefinite Matrix muß daher die Determinante ≥ 0 sein; bei einer definiten Matrix muß sie positiv sein. Ob sie positiv oder negativ definit ist, sagt uns dann die Spur, denn da die beiden Eigenwerte (falls $\neq 0$) dasselbe Vorzeichen haben, ist dieses auch das Vorzeichen ihrer Summe, der Spur. Wenn die Determinante $ac - b^2$ positiv ist, müssen a und c dasselbe Vorzeichen haben; da auch $a + c$ gleich der Spur der Matrix ist, folgt

Lemma: *a)* Eine symmetrische 2×2-Matrix ist genau dann positiv oder negativ definit, wenn ihre Determinante positiv ist. Sie ist positiv definit, wenn der Eintrag links oben positiv ist, andernfalls ist sie negativ definit.
b) Die quadratische Form $Ax^2 + Bxy + Cy^2$ ist genau dann definit, wenn ihre Diskriminante $B^2 - 4AC$ negativ ist. Im Falle $A > 0$ ist sie dann positiv, sonst negativ definit.
∎

So nützlich der Wechsel zu einer Basis aus Eigenvektoren in diesem Fall auch war, für die meisten zahlentheoretischen Fragen werden uns nur solche Basiswechsel helfen, die ganzzahlige Punkte wieder in ganzzahlige Punkte überführen. Hier gilt

Lemma: Die lineare Abbildung

$$\varphi\colon \begin{cases} \mathbb{R}^2 \to \mathbb{R}^2 \\ \begin{pmatrix} x \\ y \end{pmatrix} \mapsto M \begin{pmatrix} x \\ y \end{pmatrix} \end{cases}$$

definiert genau dann eine Bijektion $\mathbb{Z}^2 \to \mathbb{Z}^2$, wenn alle Einträge der Matrix A ganzzahlig sind und $\det M = \pm 1$ ist.

Beweis: Da die Spaltenvektoren von M die Bilder der Basisvektoren $\binom{1}{0}$ und $\binom{0}{1}$ sind, ist klar, daß $\varphi(\mathbb{Z}^2)$ genau dann in $\mathbb{Z}^2$ liegt, wenn alle Einträge von M ganzzahlig sind. Soll $\varphi(\mathbb{Z}^2) = \mathbb{Z}^2$ sein, muß zusätzlich $\varphi^{-1}(\mathbb{Z}^2) \subseteq \mathbb{Z}^2$ sein, d.h. auch M^{-1} darf nur ganzzahlige Einträge haben. In diesem Fall sind $\det M$ und $\det M^{-1}$ beide ganzzahlig mit Produkt eins, also ist $\det M = \pm 1$.

Hat umgekehrt eine Matrix M mit ganzzahligen Einträgen Determinante ± 1, so hat auch M^{-1} ganzzahlige Einträge, denn die Spaltenvektoren von M^{-1} sind die Lösungen der linearen Gleichungssysteme $M\binom{x}{y} = \binom{1}{0}$ und $M\binom{x}{y} = \binom{0}{1}$, die wir nach der CRAMERschen Regel ausdrücken können durch Brüche mit ganzzahligen Zählern und det M im Nenner. ∎

Setzen wir für so eine Matrix M das Bild $M\binom{x}{y}$ an Stelle von $\binom{x}{y}$ in die quadratische Form ein, erhalten wir das Ergebnis

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

das wir auch erhalten hätten, wenn wir $\binom{x}{y}$ in die quadratische Form zur Matrix $M^T Q M$ eingesetzt hätten. Da $\binom{x}{y} \mapsto M\binom{x}{y}$ eine Bijektion von $\mathbb{Z}^2$ nach $\mathbb{Z}^2$ definiert, nehmen die quadratischen Formen zu Q und zu $M^T Q M$ also dieselben Werte an. Deshalb definieren wir

Definition: Die quadratischen Formen mit Matrizen Q_1 und Q_2 heißen *äquivalent*, wenn es eine Matrix M mit ganzzahligen Einträgen und $\det M = \pm 1$ gibt, so daß $Q_2 = M^T Q_1 M$.

Lemma: Zwei äquivalente quadratische Formen haben dieselbe Diskriminante.

Beweis: Bis auf den Faktor -4 ist die Diskriminante gleich der Determinante der Matrix und $\det Q_2 = \det M^T \cdot \det Q_1 \cdot \det M = \det Q_1$, da $\det M = \det M^T = \pm 1$ ist. ∎

§5: Kettenbruchentwicklung quadratischer Irrationalitäten

Die rationalen Zahlen sind genau diejenigen reellen Zahlen, deren Kettenbruchentwicklung nach endlich vielen Schritten abbricht. Wir wollen sehen, daß wir auch quadratische Irrationalitäten, d.h. Elemente eines quadratischen Zahlkörpers, die nicht in $\mathbb{Q}$ liegen, durch ihre Kettenbruchentwicklung charakterisieren können.

In den Beispielen der Kettenbruchentwicklungen von $\sqrt{2}$ und $\sqrt{3}$ kamen wir in Kapitel 5 auf periodische Folgen. Wie sich zeigen wird, ist dies charakteristisch für quadratische Irrationalitäten.

Nach der Formel am Ende von §2 von Kapitel 5 gilt für die Zahlen α_n aus dem Algorithmus zur Kettenbruchentwicklung die Gleichung

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

wobei p_n und q_n Zähler und Nenner der n-ten Konvergente sind. Mit

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

die Vektoren $\binom{\alpha}{1}$ und $M\binom{\alpha_n}{1}$ sind also proportional zueinander.

Als quadratische Irrationalität genügt α einer quadratischen Gleichung $A\alpha^2 + B\alpha + C = 0$; der Vektor $\binom{\alpha}{1}$ wird also von der quadratischen Form $Ax^2 + Bxy + Cy^2$ annulliert. Da mit einem Vektor $\binom{x}{y}$ auch alle seine Vielfachen von dieser Form annulliert werden, gilt dasselbe für den zu $\binom{\alpha}{1}$ proportionalen Vektor $M\binom{\alpha_n}{1}$.

Die Matrix zur quadratischen Form $Ax^2 + Bxy + Cy^2$ sei Q. Wie wir aus dem vorigen Paragraphen wissen, erfüllen die Komponenten von $M\binom{\alpha_n}{1}$ dann die quadratische Gleichung zur Form mit Matrix $M^T Q M$. Nach Kapitel 5, §2 ist $\det M = p_{n-2}q_{n-1} - q_{n-2}p_{n-1} = (-1)^{n-1}$; die neue quadratische Form ist also zu der mit Q äquivalent und hat insbesondere dieselbe Diskriminante. Also haben alle α_n dieselbe Diskriminante wie α, denn da die Multiplikation mit M einen Isomorphismus $\mathbb{Z}^2 \to \mathbb{Z}^2$ definiert, sind die Einträge von Q genau dann ganzzahlig und teilerfremd, wenn es die von $M^T Q M$ sind.

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

Satz (LAGRANGE ∼1766): Die Kettenbruchentwicklung einer irrationalen Zahl α wird genau dann periodisch, wenn α eine quadratische Irrationalzahl ist.

Beweis: Angenommen, α hat eine periodische Kettenbruchentwicklung. Dann gibt es ein n und ein $k > 0$, so daß $\alpha_{n+k} = \alpha_n$ ist. Nach der Formel

am Ende von §2 von Kapitel 5 ist daher

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

Daraus folgt die Gleichheit von $(\alpha_n p_{n-2} + p_{n-1})(\alpha_n q_{n+k-2} + q_{n+k-1})$ und $(\alpha_n q_{n-2} + q_{n-1})(\alpha_n p_{n+k-2} + p_{n+k-1})$, und ausmultipliziert wird dies zu einer quadratischen Gleichung für α_n. Der Koeffizient von α_n^2 ist $p_{n-2}q_{n+k-2} + q_{n-2}p_{n+k-2}$, was als Summe positiver Zahlen nicht null sein kann; wir haben also eine echte quadratische Gleichung. Somit läßt sich α_n in der Form $\alpha = r + s\sqrt{D}$ schreiben, und damit auch

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

Umgekehrt sei $\alpha = r + s\sqrt{D}$ mit quadratfreiem D eine quadratische Irrationalität, die der Gleichung $A_0\alpha^2 + B_0\alpha + C_0 = 0$ genüge. Dann zeigt die Konstruktionsvorschrift für die α_n, daß auch diese Zahlen sowie ihre Inversen in entsprechender Form geschrieben werden können und damit Gleichungen der Form

$$A_n\alpha_n^2 + B_n\alpha_n + C_n = 0$$

genügen. Um diese Gleichung zu bestimmen, betrachten wir die Funktion $f(x) = A_0x^2 + B_0x + C_0$, die für $x = \alpha$ verschwindet, setzen

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

ein und multiplizieren mit dem Hauptnenner. Zumindest für die Koeffizienten A_n und C_n ergeben sich einigermaßen erträgliche Formeln:

$$A_n = A_0p_{n-1}^2 + B_0p_{n-1}q_{n-1} + C_0q_{n-1}^2 = q_{n-1}^2 f\left(\frac{p_{n-1}}{q_{n-1}}\right)$$

und

$$C_n = A_0p_{n-2}^2 + B_0p_{n-2}q_{n-2} + C_0q_{n-2}^2 = q_{n-2}^2 f\left(\frac{p_{n-2}}{q_{n-2}}\right).$$

Da f eine quadratische Funktion ist, führt die TAYLOR-Entwicklung um α auf die Formel

$$f\left(\frac{p_{n-1}}{q_{n-1}}\right) = f(\alpha) + f'(\alpha)\left(\frac{p_{n-1}}{q_{n-1}} - \alpha\right) + \frac{f''(\alpha)}{2}\left(\frac{p_{n-1}}{q_{n-1}} - \alpha\right)^2.$$

Hierbei ist $f(\alpha) = 0$, und $\left|\alpha - \frac{p_{n-1}}{q_{n-1}}\right| < 1/q_{n-1}^2 \leq 1$. Somit ist

$$|A_n| \leq |f'(\alpha)| + |f''(\alpha)| \ .$$

Genauso zeigt man die Ungleichung $|C_n| \leq |f'(\alpha)| + |f''(\alpha)|$. 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 Beispiele wie $\sqrt{2} = [1, \overline{2}]$ oder $\sqrt{3} = [1, \overline{1, 2}]$, bei denen das nicht der Fall ist. Für eine rein periodische Kettenbruchentwicklung brauchen wir also noch zusätzliche Bedingungen:

Satz: Die Kettenbruchentwicklung einer quadratischen Irrationalzahl α ist genau dann rein periodisch, wenn $\alpha > 1$ ist und ihr konjugiertes Element $\overline{\alpha}$ zwischen -1 und 0 liegt.

Beweis: Sei zunächst $\alpha > 1$ und $-1 < \overline{\alpha} < 0$. Der Trick zum Beweis der reinen Periodizität der Folge der c_i besteht darin, die $c_i = [1/\alpha_i]$ durch die konjugierten Elemente $\overline{\alpha}_{i+1}$ auszudrücken und so von der Gleichheit zweier Koeffizienten auch auf die ihrer Vorgänger zu schließen.

Die Gleichung $\alpha = c_0 + \alpha_1$ wird, da c_0 eine rationale Zahl ist, durch Konjugation zu $\overline{\alpha} = c_0 + \overline{\alpha}_1$. Da $\overline{\alpha}$ nach Voraussetzung zwischen -1 und 0 liegt, ist somit $0 < -\overline{\alpha}_1 - c_0 < 1$ und $c_0 = [-\overline{\alpha}_1]$. Wegen $c_0 = [\alpha] \geq 1$ 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 $c_i = [-\overline{\alpha}_{i+1}]$ ist, und da die Koeffizienten c_i für $i > 0$ bei jeder Kettenbruchentwicklung mindestens gleich eins sind, folgt auch die Ungleichung für $1/\overline{\alpha}_{i+1}$ genau wie dort.

Daraus folgt nun leicht die Periodizität der Kettenbruchentwicklung von α: Wir wissen bereits, daß sie periodisch *wird;* es gibt also irgendeinen Index $m \geq 0$ und eine Periode k, so daß $\alpha_{n+k} = \alpha_n$ für alle $n \geq m$. Wir betrachten das minimale m mit dieser Eigenschaft. Die Kettenbruchentwicklung von α ist genau dann rein periodisch, wenn $m = 0$ ist. Für $m \geq 1$ können wir aber aus $\alpha_{m+k} = \alpha_m$ und $c_{m+k} = c_m$ folgern, daß auch $c_{m+k-1} = [-\overline{\alpha}_{m+k}] = [-\overline{\alpha}_m] = c_{m-1}$ ist. Aus den Gleichungen

$$\frac{1}{\alpha_{m+k-1}} = c_{m+k-1} + \alpha_{m+k} \quad \text{und} \quad \frac{1}{\alpha_{m-1}} = c_{m-1} + \alpha_m$$

folgt dann aber, daß auch $\alpha_{m-1+k} = \alpha_{m-1}$ ist, im Widerspruch zur Minimalität von m. Somit ist $m = 0$, die Kettenbruchentwicklung von α also rein periodisch.

Umgekehrt habe α eine rein periodische Kettenbruchentwicklung der Periode k mit Koeffizienten $c_0, c_1, \ldots$. Wegen $c_k = c_0$ ist dabei auch c_0 positiv, denn alle c_n mit $n > 0$ müssen ja positiv sein. Somit ist insbesondere $\alpha > 1$.

Um zu sehen, daß $\overline{\alpha}$ zwischen -1 und 0 liegt, beachten wir, daß $\overline{\alpha}$ dieselbe quadratische Gleichung erfüllt wie α. Da diese Gleichung genau zwei Nullstellen hat und α größer als eins ist, genügt es, wenn wir zeigen, daß diese Gleichung im Intervall $(-1,\ 0)$ eine Nullstelle hat. Das wiederum folgt aus dem Zwischenwertsatz, wenn wir zeigen können, daß die quadratische Funktion auf der linken Seite an den Stellen 0 und -1 Werte mit entgegengesetzten Vorzeichen annimmt.

Für $k = 1$ ist $\alpha = c_0 + \alpha_1 = c_0 + \frac{1}{\alpha} \Longrightarrow \alpha^2 - c_0\alpha - 1 = 0$. Die quadratische Funktion $x^2 - c_0 x - 1$ nimmt an der Stelle 0 den Wert -1 an, und bei $x = 1$ den Wert $c_0 > 0$; somit gibt es eine Nullstelle zwischen diesen beiden Punkten.

Für $k \geq 2$ verwenden wir die bereits im vorigen Satz benutzte Formel aus Kapitel 5, §2, und beachten, daß $\alpha_k = 1/\alpha$ ist. Dies führt auf die Gleichung

$$\alpha = \frac{\alpha_k p_{k-2} + p_{k-1}}{\alpha_k q_{k-2} + q_{k-1}} = \frac{p_{k-2} + p_{k-1}\alpha}{q_{k-2} + q_{k-1}\alpha} .$$

Überkreuzmultiplikation macht daraus die quadratische Gleichung

$$q_{k-1}\alpha^2 + (q_{k-2} - p_{k-2})\alpha - p_{k-1} = 0 .$$

Hier nimmt die quadratische Funktion bei 0 den Wert $-p_{k-1} < 0$ an, und an der Stelle -1 den Wert

$$q_{k-1} - (q_{k-2} - p_{k-2}) - p_{k-1} = (q_{k-1} - q_{k-2}) + (p_{k-2} - p_{k-1}) .$$

Dieser ist positiv, da sowohl die Folge der Zähler als auch die der Nenner der Konvergenten von α monoton steigt.

∎

§6: Die Pellsche Gleichung

Im letzten Kapitel hatten wir gesehen, daß eine Einheit $x + y\sqrt{D}$ von $\mathcal{O}_D$ die Gleichung $x^2 - Dy^2 = \pm 1$ erfüllen muß. Hauptziel dieses Paragraphen ist die Lösung der PELLschen Gleichung $x^2 - Dy^2 = 1$ für $(x, y) \in \mathbb{Z}^2$ oder – da es auf das Vorzeichen von x und y nicht ankommt – $(x, y) \in \mathbb{N}^2$.

Faktorisierung der linken Seite der PELLschen Gleichung führt auf

$$(x + y\sqrt{D})(x - y\sqrt{D}) = 1 ,$$

und damit ist

$$x - y\sqrt{D} = \frac{1}{x + y\sqrt{D}} \Longrightarrow \frac{x}{y} - \sqrt{D} = \frac{1}{y^2\left(\frac{x}{y} + \sqrt{D}\right)} .$$

Wegen der Positivität der rechten Seite ist $\frac{x}{y} > \sqrt{D}$, also folgt

$$\left|x - y\sqrt{D}\right| = x - y\sqrt{D} = \frac{1}{y^2\left(\frac{x}{y} + \sqrt{D}\right)} < \frac{1}{2y^2\sqrt{D}} < \frac{1}{2y^2}\,.$$

Nach dem Satz aus Kapitel 5, §3 muß $\frac{x}{y}$ somit eine Konvergente der Kettenbruchentwicklung von $\sqrt{D}$ sein.

Umgekehrt liefert aber nicht jede Konvergente der Kettenbruchentwicklung von $\sqrt{D}$ eine Lösung der PELLschen Gleichung: Beispielsweise hat $\sqrt{13} = [3, 1, 1, 1, 1, 6, \ldots]$ 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 $c_0, c_1, \ldots$.

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 $c_0, c_1, \ldots$ 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 $c_{rk}, c_{rk+1}, c_{rk+2}, \ldots$; nach dem gerade bewiesenen Satz stimmt das überein mit der Folge $2c_0, c_1, c_2, \ldots$, 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 5, §2, bekannte Formel

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

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

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

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

$$[3, 1, 1, 1, 1, 6, 1, 1, 1] = \frac{649}{180},$$

und in der Tat ist

$$649^2 - 13 \cdot 180^2 = 421\,201 - 13 \cdot 32\,400 = 421\,201 - 421\,200 = 1\,.$$

Allgemein haben wir gezeigt, daß die PELLsche Gleichung für jedes quadratfreie D eine Lösung hat; zusammen mit dem Satz aus Kapitel 6, §6 folgt, daß die Einheitengruppe eines jeden reellquadratischen Zahlkörpers unendlich ist und daß es speziell für die Gruppe der Einheiten mit Norm eins (der sogenannten Einseinheiten) ein Element $\alpha \in \mathcal{O}_D$ gibt, so daß jede Einseinheit in der Form $\pm\alpha^r$ mit einem $r \in \mathbb{Z}$ geschrieben werden kann. α ist die kleinste Einseinheit größer eins.

Natürlich kann auch α in der Form $p_n + q_n\sqrt{D}$ geschrieben werden, wobei p_n/q_n eine Konvergente der Kettenbruchentwicklung von $\sqrt{D}$ ist. Da Zähler und Nenner der Konvergenten strikt monoton ansteigen mit n, handelt es sich hier um die *erste* Konvergente p_n/q_n, für die $p_n^2 - Dq_n^2 = 1$ ist.

Mit Rechnungen, die sehr ähnlich zu den obigen sind, kann man zeigen, daß die oben gefundenen Indizes m mit $p_m^2 - Dq_m^2 = \pm 1$ tatsächlich die einzigen sind mit dieser Eigenschaft. Da wir schon viel mit Kettenbrüchen gerechnet haben und es noch viele andere interessante Teilgebiete der Zahlentheorie zu entdecken gilt, möchte ich auf diese Rechnungen verzichten.

Wer sich für diese Rechnungen interessiert, findet sie zum Beispiel in

WINFRIED SCHARLAU, HANS OPOLKA: Von Fermat bis Minkowski – Eine Vorlesung über Zahlentheorie und ihre Entwicklung, *Springer,* 1980

im Kapitel über LAGRANGE im (nur im Inhaltsverzeichnis benannten) Paragraphen *Lösung der Fermatschen (Pellschen) Gleichung* ab Seite 64. Es gibt zwar Rückverweise, aber wer den obigen Beweis verstanden hat, muß nur einem wirklich folgen. Zu beachten sind die unterschiedlichen Bezeichnungen: Was hier α heißt, ist dort θ, aber das dortige θ_n ist hier $1/\alpha_n$. Die hiesigen c_n werden dort mit a_n bezeichnet.

Wenn wir dieses Ergebnis akzeptieren, können wir die Einheitengruppe eines jeden reellquadratischen Zahlkörpers $\mathbb{Q}[\sqrt{D}]$ explizit berechnen, zumindest für $D \not\equiv 1 \bmod 4$: Dann ist $\mathcal{O}_D = \mathbb{Z} \oplus \mathbb{Z}\sqrt{D}$, so daß die Einheiten genau den ganzzahligen Lösungen der beiden Gleichungen $x^2 - Dy^2 = \pm 1$ entsprechen. Ist k die Periode der Kettenbruchentwicklung von $\sqrt{D}$ und p/q die $(k-1)$-te Konvergente, so ist $\alpha = p + q\sqrt{D}$ die Grundeinheit, und jede andere Einheit läßt sich als $\pm\alpha^r$ mit einem $r \in \mathbb{Z}$ schreiben. Für gerades k sind dies alles Einseinheiten, für ungerades k bekommen wir für gerade r Einseinheiten und sonst Einheiten der Norm -1.

Bleibt die Frage, für welche D die Periode k gerade *bzw.* ungerade ist. Diese Frage muß nicht nur in dieser Vorlesung unbeantwortet bleiben: Es handelt sich hier um eines der vielen zahlentheoretischen Probleme, die trotz jahrhundertelanger Bemühungen auch heute noch offen sind.

Die zweite Frage ist: Was passiert für $D \equiv 1 \bmod 4$? Wie wir wissen, sind dann auch die Zahlen $\frac{1}{2}(x + y\sqrt{D})$ für ungerade ganze Zahlen x, y ganz, es kann also auch Einheiten dieser Form geben. In der Tat haben wir beim Eingangsbeispiel $D = 13$ bereits solche Fälle kennengelernt: Für die dritte Konvergente $\frac{11}{4}$ ist $11^2 - 13 \cdot 3^2 = 4$, d.h. $\mathrm{N}\left(\frac{1}{2}(11 + 3\sqrt{13}\right) = 1$. Wie eine genauere Untersuchung zeigt, ist dies genau dann möglich, wenn $D \equiv 5 \bmod 8$, jedoch nicht für alle solche D. Wenn es eine Grundeinheit dieser Form gibt, liegt ihre dritte Potenz in $\mathbb{Z} \oplus \mathbb{Z}\sqrt{D}$, der Kettenbruchalgorithmus gibt dann also nur die dritte Potenz der Grundeinheit. Einzelheiten findet man beispielsweise in §16, 5D des Buchs

HELMUT HASSE: Vorlesungen über Zahlentheorie, *Springer,* ^{2}1964.