

August 2, 2007

CHAPTER SEVEN

DIOPHANTINE EQUATIONS

§1. NORM FORMS

Let $p(x)$ be a monic irreducible polynomial of degree n with integer coefficients, and suppose that θ is a root of the polynomial. This polynomial has n distinct roots, $\theta_1 = \theta, \theta_2, \dots, \theta_n$; suppose that s of the roots, θ_i with $1 \leq i \leq s$ are real and that $2t$, θ_i with $s+1 \leq i \leq s+2t$ of them are nonreal, the nonreal roots consisting of t complex conjugate pairs, $(\theta_{s+j}, \theta_{s+t+j})$, with $1 \leq j \leq t$. Thus, $n = s + 2t$.

The function

$$f(\mathbf{x}) = \prod_{i=1}^n (x_0 + x_1\theta_i + \dots + x_{n-1}\theta_i^{n-1})$$

is a polynomial in n variables x_0, \dots, x_{n-1} with real coefficients known as a *norm form*, its value being the norm $N(\xi)$ of the element $\xi = x_0 + x_1\theta + \dots + x_{n-1}\theta^{n-1}$ in the field extension $\mathbf{Q}(\theta)$ of \mathbf{Q} . We wish to study the Diophantine equation $f(\mathbf{x}) = \pm 1$.

A simple example of an equation of this type is *Pell's equation* $x^2 - dy^2 = \pm 1$, where d is a nonsquare integer, as the left side can be factored as $(x + \sqrt{d}y)(x - \sqrt{d}y)$, \sqrt{d} being a root of the irreducible polynomial $x^2 - d$.

It turns out that the set of solutions of this Diophantine equation enjoys the algebraic structure of a group whose character can be described. If \mathbf{x} is a solution of this equation, then we say that the element $x_0 + x_1\theta + \dots + x_{n-1}\theta^{n-1}$ is a *unit* in the ring $\mathbf{Z}(\theta)$; these are the elements of the ring whose multiplicative inverses also lie in the ring.

The key result is the

Dirichlet Unit Theorem; *There exists a set of units $\{\epsilon_1, \dots, \epsilon_r\}$, where $r = s + t - 1$ such that every unit ϵ in $\mathbf{Z}(\theta)$ can be written uniquely in the form*

$$\epsilon = \zeta \epsilon_1^{k_1} \epsilon_2^{k_2} \dots \epsilon_r^{k_r},$$

where ζ is a root of unity and each ϵ_i is a unit of infinite order.

A sketch of the proof of this result will be given. We begin by describing a *vector lattice*. Let $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$ be a linearly independent set of vectors in a real vector space of dimension $n \geq m$. The set of vectors of the form $\sum_i a_i \mathbf{e}_i$, where $a_i \in \mathbf{Z}$ is a vector lattice. The lattice is *full* when $m = n$. Two linearly independent sets $\{\mathbf{e}_i\}$ and $\{\mathbf{f}_i\}$ give the same vector lattice if and only if they are related by a linear transformation $\mathbf{F} = \mathbf{C}\mathbf{E}$ where the determinant of the transformation matrix has absolute value 1. When the vector space is given a topology defined by the inner product with respect to the basis, the lattice generated by the basis is a discrete set. Indeed, any discrete subgroup of the vector space is a lattice. The *fundamental parallelepiped* of the lattice is the set $T \equiv \{\sum_i u_i \mathbf{e}_i : 0 \leq u_i < 1\}$. The translates $T + \mathbf{z}$, where \mathbf{z} belongs to the lattice are pairwise disjoint sets and there are only finitely many of them that intersect any ball $B(\mathbf{0}, r)$.

Let M be the \mathbf{Z} -module in \mathbf{C} consisting of numbers of the form $\xi = x_0 + x_1\theta + \dots + x_{n-1}\theta^{n-1}$, where each x_i is an integer, and let $\xi_i = x_0 + x_1\theta_1 + \dots + x_{n-1}\theta_i^{n-1}$. This module can be coordinatized through its associates as $(\xi_1, \xi_2, \dots, \xi_s, \xi_{s+1}, \dots, \xi_{s+t})$ for $\xi \in M$, where each ξ_i ($1 \leq i \leq s$) is real and each ξ_j is nonreal (with two real coordinates) for $s+1 \leq j \leq t$. This induces on M the inner product norm given by

$$\|\xi\|^2 = \xi_1^2 + \dots + \xi_s^2 + |\xi_{s+1}|^2 + \dots + |\xi_t|^2.$$

For each $\eta \in M$, the mapping $\alpha \rightarrow \eta\alpha$ maps M linearly into itself with determinant $N(\eta)$.

Consider the vector space \mathbf{R}^{s+t} . We define a mapping L from M into \mathbf{R}^{s+t} by

$$L(\xi) = (\log |\xi_1|, \log |\xi_2|, \dots, \log |\xi_s|, \log |\xi_{s+1}|^2, \log |\xi_{s+2}|^2, \dots, \log |\xi_{s+t}|^2) .$$

Observe that $L(\xi\eta) = L(\xi) + L(\eta)$, so that the mapping L is a homomorphism of the nonzero values of M , considered as a group under multiplication, to an additive subgroup of \mathbf{R}^{s+t} . The kernel of this homomorphism consists of those elements of absolute value 1.

Let E be the set of units contained in M and U be the subset of those units that belong to the kernel of L . U contains all the roots of unity that happen to be in M , in particular ± 1 , so that U is nontrivial. In fact, U contains only roots of unity. To see this, we can assign to the points ξ in V a norm equal to

$$\sqrt{\xi_1^2 + \dots + \xi_s^2 + |\xi_{s+1}|^2 + \dots + |\xi_{s+t}|^2} ,$$

so that the set U is bounded in M . If $\alpha \in U$, then all powers of α belong to the bounded set U ; accordingly, there can be only finitely many of them, from which we deduce that some power of α must equal 1. Thus, U is a finite cyclic group of even order and consists only of the roots of unity in M .

Since the norm of any number of E is equal to ± 1 , the sum of the entries of $L(\xi)$ for any ξ satisfies $\sum_i l_i = 0$, so that $L(E)$ lies in a subspace of dimension $s+t-1$. It remains to show that $L(E)$ is full in the subspace of \mathbf{R}^{s+t} of vectors for which the sum of the entries is 0.

To show that the lattice is full, we use the criterion that *a lattice N contained in an inner product linear space V is full if and only if there is a bounded set S such that V is contained in the union $S + N$ of its translates by elements of N* . If the lattice is full, S can be the fundamental parallelepiped. If the lattice is not full, let S be any bounded set. Then there is a number r for which $\|x\| < r$ for all $x \in S$. Since the subspace W generated by N is proper in V , we can find $y \in W$ orthogonal to V for which $\|y\| > r$. Suppose if possible that $y = u + z$ with $u \in S$ and $z \in N$. Then

$$\|y\|r < \|y\|^2 = \langle y, r \rangle = \langle y, u \rangle \leq \|y\|\|u\| < \|y\|r ,$$

which is a contradiction.

Minkowski Theorem on Convex Bodies. *Let J be a full lattice in \mathbf{R}^n whose fundamental parallelepiped has volume Δ , and let X be a bounded, centrally symmetric convex set with volume Γ . If $\Gamma > 2^n \Delta$, then X contains at least one nonzero point of J .*

Proof. Note that, if a bounded set Y is such that its translates $Y_z = Y + z$ for $z \in J$ are pairwise nonintersecting, then the volume of Y is less than Δ . To see this, note that, where T is the fundamental parallelepiped,

$$\text{Vol}(Y) = \sum \{ \text{Vol}(Y \cap T_{-z}) : z \in J \} = \sum \{ \text{Vol}(Y_z \cap T) : z \in J \} ,$$

where the right sum computes the volume of finitely many disjoint subsets of T and therefore must be less than $\text{Vol}(T) = \Delta$.

The volume of the set $\frac{1}{2}X$ obtained from X by a dilatation of factor $1/2$ exceeds Δ . Hence, two of its translates by elements of J must intersect, so that there exist elements $x_1, x_2 \in X$ and $z_1, z_2 \in J$ so that $z_1 \neq z_2$ and $\frac{1}{2}x_1 + z_1 = \frac{1}{2}x_2 + z_2$. Hence

$$z_1 - z_2 = \frac{1}{2}(-x_1) + \frac{1}{2}(x_2) \in X .$$

and the result follows. ♠

To relate all of this to M , consider the coordinatization of M with respect to its associates, where $\xi_j = \eta_j + \zeta_j i$ for $s+1 \leq j \leq t$ where η_j and ζ_j are real. If X is the bounded set of points ξ for which $|\xi_i| < c_i$ ($1 \leq i \leq s$) and $|\xi_j|^2 < c_j$ ($s+1 \leq j \leq t$), then the volume of X is

$$\int_{-c_1}^{c_1} d\xi_1 \cdots \int_{-c_s}^{c_s} d\xi_s \int_{\eta_{s+1}^2 + \zeta_{s+1}^2 < c_{s+1}} d\eta_{s+1} d\zeta_{s+1} \cdots \int_{\eta_t^2 + \zeta_t^2 < c_t} d\eta_t d\zeta_{s+1} = 2^s \pi^t c_1 \cdots c_t .$$

Noting that $n = s + 2t$, we obtain from Minkowski's theorem that if the fundamental parallelepiped of the full lattice M has volume Δ and if $c_1 c_2 \cdots c_{s+2t} > (4/\pi)^t \Delta$, then there is a nonzero element ξ of M for which

$$|\xi_1| < c_1, \dots, |\xi_s| < c_s, |\xi_{s+1}|^2 < c_{s+1}, \dots, |\xi_t|^2 < c_t .$$

§2. PELL'S EQUATION

Pell's equations arise from the norm form for the real n th primitive root of unity. In the quadratic case, it is the familiar $x^2 - dy^2 = \pm 1$, where d is a nonsquare integer. When $d < 0$, then this has finitely many solutions, which all arise from roots of unity in $\mathbf{Q}(\sqrt{-d})$. When $d > 0$, then it has infinitely many solutions (x_m, y_m) arising from $x_m + y_m \sqrt{d} = (u + v\sqrt{d})^m$, where $u + v\sqrt{d}$ is a "fundamental" unit in $\mathbf{Z}(\sqrt{d})$.

In fact, $(x_m, y_m) = (T_m(u), U_m(u)v)$ where T_m and U_m are Chebyshev polynomials of the first and second types. To see this, observe that $dv^2 = u^2 - 1$ and that

$$\begin{aligned} (T_m(u) + U_m(u)v\sqrt{d})(u + v\sqrt{d}) &= (uT_m(u) + (u^2 - 1)U_m(u)) + (T_m(u) + uU_m(u))v\sqrt{d} \\ &= T_{m+1}(u) + U_{m+1}(u)v\sqrt{d} . \end{aligned}$$

The cubic version of Pell's equation is

$$x^3 + cy^3 + c^2z^3 - 3cxyz = \pm 1 ,$$

where c is an integer not equal to a perfect cube. By Dirichlet's theorem, the set of solutions is, up to roots of unity, a cyclic group; they can be found by taking powers of a fundamental unit $u + v\theta + w\theta^2$, where θ is the real cube root of c .

The sequence of solutions generated by (u, v, w) is defined by the recursion $(x_{m+1}, y_{m+1}, z_{m+1})^t = M(x_m, y_m, z_m)^t$, where the transition matrix

$$M = \begin{pmatrix} u & cw & cv \\ v & u & cw \\ w & v & u \end{pmatrix}$$

has characteristic polynomial $\lambda^3 - 3u\lambda^2 + 3(u^2 - cvw)\lambda - 1 = 0$. Thus, $\{x_m\}$, $\{y_m\}$ and $\{z_m\}$ each satisfy the recursion

$$t_{m+1} = 3ut_m - 3(u^2 - cvw)t_{m-1} + t_{m-2} .$$

Example 1. When $c = 2$, the fundamental solution is $(1, 1, 1)$, the recursion is $t_{m+1} = 3t_m + 3t_{m-1} + t_{m-2}$ and the sequence of solutions is

$$\{\dots, (-1, 1, 0), (1, 0, 0), (1, 1, 1), (5, 4, 3), (19, 15, 12), (73, 58, 46), \dots\} .$$

Example 2. When $c = 3$, the fundamental solution is $(4, 3, 2)$, the recursion is $t_{m+1} = 12t_m - m + 6t_{m-1} + t_{m-2}$ and the sequence of solutions is

$$\{\dots, (4, 3, -4), (-2, 0, 1), (1, 0, 0), (4, 3, 2), (52, 36, 25), \dots\} .$$

The quartic version of Pell's equation, obtained from calculating the norm of $x + y\theta + z\theta^2 + w\theta^3$, where θ is a fourth root of c , is given by

$$(x^2 + cz^2 - 2cyw)^2 - c(2xz - y^2 - cw^2)^2 = \pm 1 .$$

When $c < 0$, then the equation $t^4 - c = 0$ has two pairs of complex conjugates roots, and so the set of solutions of this equation is essentially a cyclic group. When $c > 0$ and c is not a square, then the set of solutions is essentially a free group on two generators (up to roots of unity). If we write $X = x^2 + cz^2 - 2cyw$ and $Y = 2xz - y^2 - cw^2$, then it appears as though the two generators will yield $(X, Y) = (1, 0)$ and (X, Y) a fundamental solution of $X^2 - cY^2 = \pm 1$. It would be nice to find a nice representation for the set of solutions.

The quintic Pell's equation is the rather formidable looking

$$\begin{aligned} (x^5 + cy^5 + c^2z^5 + c^3u^5 + c^4v^5) - 5c(x^yv + x^3zu + xy^3z) - 5c^2(y^3uv + xz^3v + yz^3u + xyu^3) \\ - 5c^3(zu^3v + xuv^3 + yvv^3) + 5c(x^2y^2u + x^2yz^2) \\ + 5c^2(x^2u^2v + x^2zv^2 + xy^2v^2 + xz^2u^2 + y^2z^2v + y^2zu^2) \\ + 5c^3(yu^2v^2 + z^2uv^2) - 5c^2(xyzuv) = \pm 1 . \end{aligned}$$

In this case, $s = 1$, $t = 2$, so that there the group of solutions is essentially free with two generators.

The sixth degree Pell's equation has the form

$$p^2 - cq^2 = r^3 + cs^3 + c^2t^3 - 3crst = \pm 1$$

where

$$\begin{aligned} p = x^3 + (3xu^2 + 3y^2v + z^3 - 3xyw - 3xvz - 3uyz)c \\ + (v^3 + 3zw^2 - 3uvw)c^2 , \end{aligned}$$

a

$$\begin{aligned} q = (3x^2u + y^3 - 3xyz) \\ + (u^3 + 3yv^2 + 3z^2w - 3xvw - 3uyw - 3uvz)c + wc^2 , \\ r = x^2 + 2czv - cu^2 - 2cyw , \\ s = 2xz + cv^2 - y^2 - 2cuw , \end{aligned}$$

and

$$t = z^2 + 2xv - 2yu - cw^2 .$$

When c is positive and not a cube, we have $(s, t) = (2, 2)$ so that the group of solutions is essentially free with three generators. There are four types of solutions, those for which

$$(p, q) = (\pm 1, 0) \quad (r, s, t) = (\pm 1, 0, 0)$$

$$(p, q) = (\pm 1, 0) \quad (r, s, t) \text{ nontrivial}$$

$$(p, q) \text{ nontrivial} \quad (r, s, t) = (\pm 1, 0, 0)$$

and both (p, q) and (r, s, t) nontrivial.

For example, when $c = 2$, we have solutions $[(x, y, z, u, v, w), (p, q), (r, s, t), \pm 1]$ given by

$$\begin{aligned} [(1, 1, 0, 0, 0, 0), (1, 1), (1, -1, 0), -1] \\ [(1, 0, 0, 1, 0, 0), (7, 5), (-1, 0, 0), -1] \\ [(1, 0, 1, 0, 1, 0), (1, 0), (5, 4, 3), +1] \\ [(3, 2, 2, 2, 2, 2), (3, 2), (1, 0, 0), +1] \\ [(11, 10, 9, 8, 7, 6), (3, 2), (5, 4, 3), +1] \\ [(145, 138, 126, 108, 90, 78), (1, 0), (1, 0, 0), +1] \end{aligned}$$

§3. Other norm forms

If θ is a root of the equation $t^3 = t + 1$, we get the norm form equation

$$x^3 + y^3 + z^3 + 2x^2z + xz^2 - xy^2 - yz^2 - 3xyz$$

for which solutions can be readily found.

Reference

Edward J. Barbeau, *Pell's equation*. Springer, 2003