DIOPHANTINE EQUATIONS IN POLYNOMIALS

PAUL E. GUNNELLS

1. Introduction

These are notes from a talk of the same name given to the PROMYS program on August 6, 2004. The target audience was advanced high-school students, but others also might find the material interesting. The lecture was an elaboration on a chapter in the excellent book *Essays on numbers and figures* by V. V. Prasolov [2], and the interested reader should look there for more information.

2. Diophantine equations

A *Diophantine equation* is a polynomial equation in variables $x, y, z, \dots$ with rational or integral coefficients. What makes such an equation Diophantine is that one puts restrictions on acceptable solutions: given such an equation, one only wants its rational or even integral solutions. Here are some examples.

(1) A standard example is the *Pythagorean equation* $x^2 + y^2 = z^2$. Integral solutions (x, y, z) are called *Pythagorean triples*, because they correspond to right triangles whose sides are whole numbers. For example, we have the standard "SAT" triangles (3,4,5) and (5,12,13). This equation is completely understood: in number theory courses one proves that all solutions have the form $(m^2 - n^2, 2mn, m^2 + n^2)$, where m, n are integers.

(2) The generalization of the Pythagorean equation $x^n + y^n = z^n$, $n > 2$ is called the *Fermat equation*. This equation is a bit more challenging. Indeed, many of the great advances in modern number theory arose out of (failed) attempts to show that there are no solutions to this equation! Today, thanks to the efforts of many mathematicians, we now know that the only integral solutions are the trivial ones, in which at least one of x, y, z is 0.

(3) Those two examples were at the extreme ends of the Diophantine spectrum. How about something in between? Consider the equation $x^2 - 2y^2 = 1$. This is an example of *Pell's equation*[1] The general Pell's equation has the form $x^2 - Dy^2 = 1$, where $D > 0$ is a integer that is not a square. As with the Pythagorean equation, we know how to find all the solutions. In the case

Date: August 13, 2004.

[1] Apparently this is a completely misnamed equation . . . Pell was only marginally involved.

of $D = 2$, one chooses an integer $k \neq 0$ and multiplies out $(1-\sqrt{2})^k$. The answer will have the form $a + b\sqrt{2}$ for a, b integers, and (a, b) is a solution to the equation. What's going on here is more subtle than for the Pythagorean equation, but it's still accessible in a beginning number theory course.

(4) Here's a more interesting problem that's still easier to deal with than Fermat. When is a product of two consecutive integers equal to a product of three consecutive integers? If we translate this into algebra[2] we get $y(y+1) = (x-1)x(x+1)$ or $y^2 + y = x^3 - x$. This is an example of an *elliptic curve*; such objects figure prominently in the proof of Fermat's last theorem. It turns out that there are infinitely many *rational solutions* to this equation, but a theorem of Siegel says that there are at most finitely many *integral* solutions. This is the first example where there is a difference between the qualitative nature of integral and rational solutions, but it's not the only one. For this particular equation, everything about the rational solutions are understood. One starts with the solution $(0, 0)$, and through a geometric procedure (the "group law" of the elliptic curve) one can build all other rational solutions.

3. Polynomials as numbers

Number theorists aren't squeamish about what mathematical techniques they use to solve problems. Analysis, algebra, geometry, topology, representation theory, whatever . . . we don't care, as long as we get results. A guiding force is *analogy*: if you can't solve the problem you want, try to solve an analogous but easier one. Even if the analogy looks completely crazy, and takes the problem into a completely different domain, the solution might shed light on the original problem.

This is the idea we want to develop. We want to discuss the idea that

the set of integers $\mathbf{Z}$

is analogous to

the set $\mathbf{C}[z]$ of complex polynomials in one complex variable z.

Now this might seem kind of strange. After all, a polynomial is not the same thing as a number, right? For example, we can evaluate a polynomial at an integer, so somehow a polynomial is an object that's further up the food chain than an integer. That's certainly true, but this is missing the spirit of our game. Instead we should think of ways that polynomials and integers are *similar*.

(1) For example, we can add and subtract integers and don't leave the world of integers. The same is true for complex polynomials.

[2] It's traditional to write the three consecutive numbers as $x-1, x, x+1$ instead of $x, x+1, x+2$.

(2) We can multiply two integers and obtain an integer. Ditto for complex polynomials.

(3) What about division? Well, we can't always divide one integer into another (e.g. $3/2$ is not an integer). This is also true for polynomials (e.g. $(z+1)/(z-1)$ is not a polynomial), and is the first indication that maybe our analogy isn't so crazy after all. There are some integers n such that $1/n$ is an integer, namely $n = \pm 1$. We call such integers *units*. And the same is true for polynomials: the only polynomials $f(z)$ such that $1/f(z)$ is a polynomial are the nonzero constant polynomials. We'll also call them units.

(4) How about something fancier, like *prime* numbers? An positive integer p is called prime if $q > 0$ divides p implies $q = p$ or $q = 1$. We can extend this to negative integers by saying that p is a prime if it's only divisible by numbers of the form εp, where $\varepsilon = \pm 1$ is a unit. Now the notion of divisibility makes sense for polynomials, so we can just define a prime polynomial $f(z)$ to be one whose only divisors are $cf(z)$, where c is a unit. Since every polynomial over the complex numbers completely factors into linear terms (i.e. all the roots of the polynomial can be found using the complex numbers), we see that a polynomial is prime if and only if it has degree one. From this we can see that the *fundamental theorem of algebra* becomes in this language the *fundamental theorem of arithmetic*: each polynomial can be factored into a product of prime polynomials, uniquely up to permutation of the factors and multiplication by units.

These facts indicate that our analogy isn't so bad. Since addition, subtraction, and multiplication make sense for polynomials, we can take any polynomial equation (e.g. the Fermat equation) and can plug polynomials into the variables. In other words, we can try to find polynomial solutions to our Diophantine equataions. This is what we'll do in the following sections.

4. Mason's theorem and Fermat's last theorem

The basic tool we'll use is called *Mason's theorem*, or the *ABC theorem* for polynomials. It's a severe restriction on the degrees of polynomials that can appear in linear equations.

Theorem 1. *(Mason) Suppose $a, b, c \in \mathbf{C}[z]$ are pairwise relatively prime complex polynomials*[3]*, and at least one of a, b, c is not constant. Let N_0 be the number of distinct roots of the product abc. Then if $a + b + c = 0$, we have*

$$\deg a, \deg b, \deg c \leq N_0 - 1.$$

[3] This means that any two of them have no common roots.

In other words, the theorem says that if you can add together three polynomials a, b, c and get zero, and if these polynomials have no common roots, then the degrees of each of them can't be too big when compared to the number of distinct roots in the product abc. Note also that N_0 is as big as possible when each polynomial has only simple roots (i.e. in the factorization into distinct factors $a(z) = c(z-\alpha_1)^{e_1} \cdots (z-\alpha_k)^{e_k}$, all exponents are 1), and in this case $N_0 = \deg(abc)$. You can try some simple examples to convince yourself of the truth of the theorem (always a good idea when you learn a new theorem). We won't give the proof here. It's not hard, but is best read and appreciated on one's own. A complete proof is given in [2, Ch. 15].

To see how powerful Theorem 1 is, let's dispense with Fermat's last theorem.

Theorem 2. *Let $n \geq 2$ be an integer, and suppose $a, b, c \in \mathbf{C}[z]$ are pairwise relatively prime polynomials, at least one of which is not a constant, satisfying $a^n + b^n = c^n$. Then $n = 2$.*

Proof. Suppose a, b, c is a solution, and let $A = a^n$, $B = b^n$, $C = c^n$. Then, after multiplying by units, we achieve $A + B + C = 0$. Since $\deg A = n \deg a$, etc., we obtain

$$n \deg a, n \deg b, n \deg c \leq N_0 - 1,$$

where N_0 is the number of distinct zeros of ABC. But $N_0 \leq \deg a + \deg b + \deg c$ (why?), so

$$n \deg a, n \deg b, n \deg c \leq \deg a + \deg b + \deg c - 1.$$

This is really three separate inequalities. If we add them together, we get

$$n(\deg a + \deg b + \deg c) \leq 3(\deg a + \deg b + \deg c) - 3 < 3(\deg a + \deg b + \deg c),$$

and hence $n < 3$. □

Well, that was easy. It almost seems *too* easy, but it's a correct proof. The point is that Mason's theorem, although it doesn't look like much, is very restrictive. Now we can see the power of analogy. Immediately one wonders, what's the analogue of Theorem 1 for integers, and can it be used to attack FLT? We'll talk about this in Section 7; you can skip ahead if you like.

Note that the proof of Theorem 2 doesn't show that there *are* any solutions to the Fermat equation in polynomials for $n = 2$, just that it's only possible to have a solution if $n = 2$. But solutions do exists for $n = 2$: simply take the basic solution in *integers* $(m^2 - n^2, 2mn, m^2 + n^2)$ and replace m, n with any polynomials you like. You can check that the Fermat equation will be satisfied.

5. A GENERALIZATION OF FERMAT

How about a generalization of Fermat's last theorem, in which the three exponents are allowed to be arbitrary positive integers?

Theorem 3. *Suppose $a, b, c \in \mathbf{C}[z]$ are pairwise relatively prime polynomials, at least one of which is not a constant, satisfying $a^p + b^q = c^r$, where $2 \leq p \leq q \leq r$. Then (p, q, r) must be one of*

(1) $(2, 2, r)$, $r \geq 2$,
(2) $(2, 3, 3)$,
(3) $(2, 3, 4)$,
(4) $(2, 3, 5)$.

This sounds much more difficult than Fermat, but it too can be handled easily with Mason's theorem.

Proof. (Sketch) We only prove that $p = 2$ and $q \leq 3$. Let $\alpha = \deg a$, $\beta = \deg b$, $\gamma = \deg c$. Again $N_0 \leq \alpha + \beta + \gamma$, so we have the three inequalities

$$p\alpha \leq \alpha + \beta + \gamma - 1, \tag{1}$$

$$q\beta \leq \alpha + \beta + \gamma - 1, \tag{2}$$

$$r\gamma \leq \alpha + \beta + \gamma - 1, \tag{3}$$

Now $p(\alpha + \beta + \gamma) \leq p\alpha + q\beta + r\gamma$, so if we add all three inequalities together we get

$$p(\alpha + \beta + \gamma) \leq 3(\alpha + \beta + \gamma - 1) \implies p = 2.$$

Since $p = 2$, (1) becomes

$$\alpha \leq \beta + \gamma - 1. \tag{4}$$

Now (2),(3),(4), together with $q \leq r$, give

$$q(\beta + \gamma) \leq \alpha + 3\beta + 3\gamma - 3.$$

If we plug in (4) again, we get

$$q < 4 \implies q = 2, 3.$$

We have to show $r \leq 5$ if $q = 3$. This is very similar to the argument used to show $q \leq 4$. Simply put $q = 3$ in (2), and then use the result with (4) a few times. □

Just like with Theorem 2, the proof of Theorem 3 doesn't show that solutions actually exist for the indicated (p, q, r), but they do. For example, we have the identity

$$\left(\frac{x^r + 1}{2}\right)^2 - \left(\frac{x^r - 1}{2}\right)^2 = x^r,$$

which shows that $(2, 2, r)$ has solutions. There are solutions as well for the three exceptional cases $(2, 3, 3)$, $(2, 3, 4)$, $(2, 3, 5)$, but they are more involved (cf. [2]).

Here is a geometric interpretations of Theorem 3. Suppose we fix (p, q, r), and consider the inequalites (1)–(3). Consider α, β, γ to be *variables* instead of integers. Then we can draw the graphs of (1)–(3) in three-dimensional space $\mathbf{R}^3$, where the coordinates are labelled α, β, γ instead of x, y, z. The result is three "half-spaces," that is, three solid regions in $\mathbf{R}^3$ bounded on one side by a plane. The intersection of these three half-spaces determines a region, which may in fact be empty. A point (α, β, γ) in this region corresponds to a triple of degrees if and only if it lies in the integer lattice $\mathbf{Z}^3 \subset \mathbf{R}^3$ and its coordinates are nonnegative. The nonnegativity conditions give us three more half-spaces, so altogher we have *six* inequalties

$$\begin{aligned} p\alpha &\leq \alpha + \beta + \gamma - 1, \\ q\beta &\leq \alpha + \beta + \gamma - 1, \\ r\gamma &\leq \alpha + \beta + \gamma - 1, \\ \alpha &\geq 0, \\ \beta &\geq 0, \\ \gamma &\geq 0, \end{aligned}$$

in the variables α, β, γ. These six inequalities determine a region $R(p, q, r)$. For example, suppose $p = q = r = 2$. Then $R(2, 2, 2)$ is a triangular cone with vertex at $(1, 1, 1)$, and with face angles $\pi/3$ (Figure 1).

We can have solutions to our original diophantine equation only if $R(p, q, r)$ contains nontrivial lattice points, so our question becomes, for which (p, q, r) does $R(p, q, r)$ contain at least one integral point? From Figure 1, it's clear that $R(2, 2, 2)$ contains infinitely many lattice points, since it's an infinite region, the apex is an integral point, and the defining rays are rational.[4] It turns out that $R(p, q, r)$ is infinite if and only if (p, q, r) is taken from the list in Theorem 3, and in all other cases is empty! And certainly an empty region contains no lattice points.

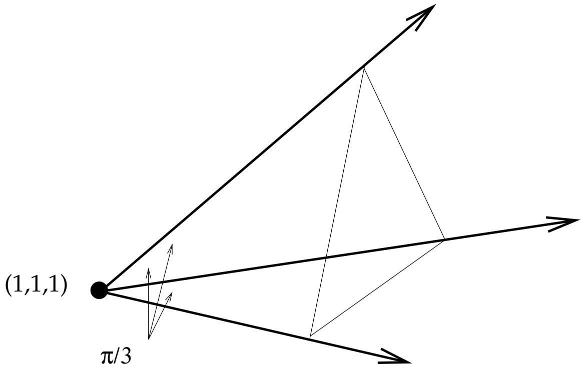

FIGURE 1.

[4] Conditions like these are essential. Consider for instance the ray ρ with tail at the origin in $\mathbf{R}^2$ and going through the point $(\pi, 1)$. The origin is the only nontrivial lattice point on ρ, precisely because π is irrational.

6. The ADE pattern

There is another geometric interpretation of the list of triples in Theorem 3 that involves the *Platonic solids*, otherwise known as the regular polyhedra in three dimensions. They are the tetrahedron, cube, octahedron, dodecahedron, and icosahedron (Figure 2).

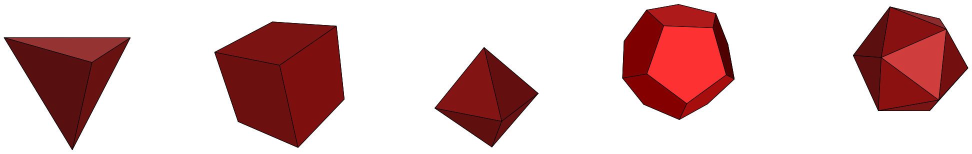

Figure 2. Platonic solids

The connection is very simple. Take a regular polyhedron P, and add a vertex at the midpoint of every edge and the center of each face. Join these new vertices together so that the center of each face gets joined radially to the original vertices and the midpoints of the neighboring edges. Hence each face will become subdivided into triangles; if the original face had n sides, it will now contain $2n$ triangles. This operation is called *barycentric subdivision*.

Now imagine that each polyhedron is made of airtight flexible material, like rubber or latex. Inflate each polyhedron until it becomes round. The result is a sphere tiled with congruent spherical triangles.[5] Something remarkable happens:

- Even though there were five solids to begin with, we only obtain three different tiled spheres: the cube/octahedron and dodecahedron/icosahedron each become the same tiled sphere.
- The angles of the triangular tiles are $(\pi/p, \pi/q, \pi/r)$, where (p, q, r) is one of $(2, 3, 3)$, $(2, 3, 4)$, $(2, 3, 5)$.

Hence the "exceptional triples" (the ones not contained in the infinite family $(2, 2, r)$) can be understood in terms of the Platonic solids! There is also a geometric explanation of $(2, 2, r)$ in the same spirit. Take a regular polygon with r sides, subdivide each of its edges, then build a "double pyramid" over it. In other words, attach two pyramids, one on the top and one on the bottom. The result

[5] There is a notion of geometry on a sphere very similar to that of the usual Euclidean geometry. A "line" is defined to be a great circle on the sphere, i.e. a circle on the surface with center the same as that of the sphere. On the earth, for instance, the equator and lines of longitude are examples of great circles. Hence there is a well defined notion of spherical triangles. In fact, there is even a version of trigonometry for the sphere, called (what else?) spherical trigonometry. Courses in spherical trigonometry are taught to future navigators at maritime academies.

is a solid with $4r$ triangular faces, each with angles $(\pi/2, \pi/2, \pi/r)$. This isn't a regular polyhedron (it's not symmetric enough to earn that title), but in some sense it's a three-dimensional solid generated by a regular polygon, which is the two-dimensional "Platonic polygon." So the solutions $(2, 2, r)$ can be seen to be a part of the same geometric picture.

It's very bizarre that there's any connection at all between our original equation and regular polyhedra. Is there any explanation? Well, there's not really an explanation that most people would accept as one, i.e. some kind of justification that explains why this *should* happen. All we can say is that we set out to perform a classification of some objects, and the answers ended up being the same. What's even more bizarre is that this phenomenon occurs in mathematics *over and over again*, with essentially the same objects as answers.

Let's try to be more specific, while being as vague as possible. In various contexts certain mathematical objects are defined. These contexts are very diverse, and usually have no apparent relationship to each other. Properities and structures of the basic objects are explored and organized in lists (theorems, propositions, lemmas, etc.). Sometimes one attempts to classify the objects, i.e. to use the accumulated knowledge to understand completely the set of possible objects. The surprising discovery has been that if this is possible, i.e. if the set of objects is tractable, then usually it ends up being "the same" as the set of Platonic solids.[6] This is called the *ADE pattern*, because after the work of Cartan and Killing on simple complex Lie[7] algebras, a certain notational scheme was standardized. Why does this happen? To the best of our knowledge no one knows.

7. THE ABC CONJECTURE

Let's return to number theory after this detour through geometry. Recall that we were originally interested in Diophantine equations over the integers and rational numbers, and that we decided to look at the same equations over complex polynomials to gain insight. Did we get any?

Perhaps the main insight is that Mason's theorem is extremely powerful. With it in hand, we were able to derive Fermat's last theorem with essentially no work. This makes us ask, what is the analogue of Mason's theorem for integers, and is it true? To answer this, we have to reverse-engineer our integers → polynomials analogy. In other words, what appears in the statement of Mason's theorem, and what is the analogous concept for integers?

The main ingredient we have to worry about is the degree of a polynomial. What properties does it have? Well, in some sense the degree is a measure of size: $\deg f > \deg g$ intuitively means that f is bigger than g. Also, $\deg fg =$

[6] Including of course their higher dimensional analogues, cf. [1].

[7] rhymes with *free*, not *die*.

$\deg f + \deg g$, which fits well with this. The degree is also the number of linear factors of a polynomial, and since a linear polynomial is prime, this suggests that the "degree" of an integer should have something to do with the number of its prime divisors. But "number of prime divisors" isn't the right definition: different primes are different sizes, whereas degree one polynomials all have the same size.

To work our way out of this, we take the analogue of degree to be the absolute value $|n|$. Clearly this is a measure of the size of n. The additive formula $\deg fg = \deg f + \deg g$ gets replaced by a multiplicative one $|mn| = |m||n|$. For an analogue of $N_0(f)$ we take the radical $\operatorname{rad} n$ of n, defined to be the product over all distinct primes that divide n. For example, $\operatorname{rad} 2 = \operatorname{rad} 8 = 2$, and $\operatorname{rad} 6 = \operatorname{rad} 12 = 6$.

Conjecture 1 (ABC conjecture)**.** *Suppose a, b, c are three pairwise relatively prime integers with $a + b + c = 0$. Then for every $\varepsilon > 0$ there exists a constant C_ε such that*

$$|a|, |b|, |c| \leq C_\varepsilon |\operatorname{rad}(abc)|^{1+\varepsilon}.$$

Here's how to interpret the $\varepsilon, C_\varepsilon$ business. The cleanest statement possible would be

$$|a|, |b|, |c| \leq |\operatorname{rad}(abc)|, \tag{5}$$

but in practice this appears to be too strong. We'd like to adjust (5) so that the right hand is a little bigger that the left. This can be done by scaling the right by a constant and raising the radical to a power > 1. It turns out that the best way to do this is to first choose the power you want, and then modify the constant accordingly. So if one picks a big ε, say $\varepsilon = 1$. then the left will be much smaller in general then $\operatorname{rad}(abc)^2$, so we can take a small C_ε. But if one takes ε very very tiny, like 10^{-23}, then this is ok, but one is forced to choose C_ε to be very big. In other words, we can get as close to the pure statement (5) as we want, at the expense of multiplying the right by bigger and bigger constants. What's amazing about the statement of Conjecture 1 is that once one picks ε, there is supposed to be a C_ε that will work for *every a, b, c*.

Given the ABC conjecture, one can prove many theorems in number theory that look very difficult without it, for example Fermat's last theorem for sufficiently large exponents. Unfortunately, we don't know if the ABC conjecture is true or not.[8] Opinions vary about whether or not it's likely to be true. The appearance of $\varepsilon, C_\varepsilon$ means that it's impossible to check computationally: if one were to find a counterexample, couldn't it be that C_ε was chosen poorly?

Finally, we should mention something about the proof of Mason's theorem. We didn't describe it in detail here, but it makes sense to contemplate why we can prove Theorem 1 but can't prove Conjecture 1. What is so special about polynomials? The answer: you can differentiate polynomials, and this is an essential

[8] (Summer 2004) There are rumors that it has been proved.

ingredient in the proof of Theorem 1. Hence ultimately our analogy breaks down. Polynomials really are different from integers.

References

1. H. S. M. Coxeter, *Regular polytopes*, Dover, 1980.
2. V. V. Prasolov, *Essays on numbers and figures*, Mathematical world, vol. 16, American Mathematical Society, 2000.

Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003

E-mail address: gunnells@math.umass.edu