ELLIPTIC CURVES

BJORN POONEN

1. Introduction

The theme of this lecture is to show how geometry can be used to understand the rational number solutions to a polynomial equation. We will illustrate this first in the context of rational points on the circle, and then in the more advanced context of rational points on elliptic curves.

2. Plane curves

A plane curve is the set of the form $\{(x, y) : f(x, y) = 0\}$ where $f(x, y)$ is a polynomial in two variables[1]. There are many familiar examples of plane curves: for example, the circle $(x-3)^2+(y-2)^2 = 4$ is a plane curve, as one sees by taking $f(x, y)$ to be $(x-3)^2+(y-2)^2-4$.

The *degree* of the curve is the total degree of f; this is defined as the maximum of $i + j$ such that there is a monomial ax^iy^j occurring in f with $a \neq 0$. For example, the plane curve

$$x^3 - 10x^2y^2 + 9y^3 + 20 = 0$$

has degree 4 because of the monomial of largest degree in it is $-10x^2y^2$, which has degree $2 + 2 = 4$.

3. Plane curves of low degree

Plane curves of degree 1 are called *lines.* They are defined by equations of the form $ax + by + c = 0$, where a, b are not both zero.

Plane curves of degree 2 are called *conic sections* or simply *conics*[2]. These have the form $ax^2 + bxy + cy^2 + dx + ey + f = 0$ for some numbers a, b, c, d, e, f. The conics include ellipses (including the special case of circles), parabolas, hyperbolas, as well as some "degenerate" cases such as $xy = 0$ (two lines), $x^2 - 1 = 0$, or $x^2 = 0$. (Many people would exclude some or all of the last three examples from the definition of a conic.)

Plane curves of degree 3 are called *cubic curves.* The general form of such a curve is

$$a_1x^3 + a_2x^2y + a_3xy^2 + a_4y^3 + a_5x^2 + a_6xy + a_7y^2 + a_8x + a_9y + a_{10} = 0,$$

Date: January 3, 2012.

[1] There are a couple of subtleties here. Usually we will insist that $f(x, y)$ be non-constant, since if $f(x, y)$ is a constant, then the set of solutions to $f(x, y) = 0$ is either empty or the entire plane, depending on whether the constant is nonzero or zero. Also, although we will usually draw the set of solutions to $f(x, y) = 0$ where x and y are real numbers, the theory actually works better when one allows complex number solutions as well. For example, the "curve" $x^2 + y^2 + 1 = 0$ looks empty if one only takes real number solutions, but acquires many solutions if x and y are permitted to be complex numbers.

[2] This is because they arise by slicing a double cone in space such as $x^2 + y^2 = z^2$ with a plane.

where $a_1, \ldots, a_{10}$ are numbers. *Elliptic curves* are certain cubic curves; namely they are the curves defined by equations of the form

$$y^2 = f(x)$$

or equivalently

$$y^2 - f(x) = 0,$$

where $f(x)$ is a squarefree polynomial of degree 3. "Squarefree" means that $f(x)$ has no multiple roots. For instance

$$y^2 = x^3 - 3x + 2$$

does not define an elliptic curve, because

$$x^3 - 3x + 2 = (x-1)^2(x+2)$$

has 1 as a multiple root. Similarly $y^2 = x^3$ is not an elliptic curve, but $y^2 = x^3 + 1$ is an elliptic curve.

By scaling the coordinates and translating, one can convert any elliptic curve into one of the form $y^2 = x^3 + Ax + B$ where A and B are numbers. It turns out that a general curve of the form $y^2 = x^3 + Ax + B$ is an elliptic curve (i.e., $x^3 + Ax + B$ is squarefree) if and only if $-(4A^3 + 27B^2) \neq 0$. From now on, we will always assume that our elliptic curves are *defined over* $\mathbb{Q}$; this means that the coefficients of the polynomial defining an elliptic curve are *rational* numbers.

4. Rational points on the unit circle

A *rational point* on a plane curve is a point on the curve with rational coordinates. For example, $(3/5, 4/5)$ is a rational point on the circle C with equation $x^2 + y^2 = 1$.

As one can guess from the example just given, rational points on C are closely related to Pythagorean triples, i.e., the positive integer solutions to $a^2 + b^2 = c^2$. In fact, if a, b, c are any integers satisfying $a^2 + b^2 = c^2$ and $c \neq 0$, then $(a/c, b/c)$ will be a rational point on C.

Conversely, if (x, y) is a rational point on C. then by choosing a common denominator for x and y one can write $x = a/c$ and $y = b/c$ for some integers a, b, c with $c \neq 0$, and the relation $x^2 + y^2 = 1$ implies $a^2 + b^2 = c^2$. If moreover x and y are nonzero, then a, b, c will all be nonzero, and $(|a|, |b|, |c|)$ will be a Pythagorean triple.

It would be nice to have a description of *all* the rational points on C, because then we would have a description of all the Pythagorean triples. Our goal now is to find such a description using geometry!

Consider the following construction. Start with the rational point $P = (-1, 0)$ on C. Fix a rational number t. Draw the line L_t with slope t passing through P. This line will intersect the circle at a second point Q_t (which depends on the number t).

By "pure thought" (no calculation), one can see that Q_t must have rational coordinates, because its x-coordinate will arise as the solution to a quadratic equation *which already has one rational root*, namely the x-coordinate of P, and then the y-coordinate of Q_t also will be rational (either by the same argument with y-coordinates, or by using the equation of L_t).

For the incredulous, here is a full calculation of Q_t. The equation of L_t (in point-slope form) is $y = t(x+1)$. To intersect this with C, which is $x^2 + y^2 = 1$, substitute $y = t(x+1)$

to obtain

$$\begin{aligned} x^2 + t^2(x+1)^2 &= 1 \\ (x^2-1) + t^2(x+1)^2 &= 0 \\ (x+1)\left[(x-1) + t^2(x+1)\right] &= 0. \end{aligned}$$

It was not just luck that the quadratic polynomial in x factored: the point is that it *had* to have $x+1$ as a factor, because we already knew that there was a point with x-coordinate -1 in the intersection $L_t \cap C$, namely $P = (-1,0)$. Anyway, discarding the root $x = -1$ and solving for the other possible x-coordinate, we see that Q_t has x-coordinate $(1-t^2)/(1+t^2)$, and y-coordinate

$$y = t(x+1) = t\left[\frac{1-t^2}{1+t^2} + 1\right] = \frac{2t}{t^2+1}$$

so

$$Q_t = \left(\frac{1-t^2}{1+t^2}, \frac{2t}{t^2+1}\right).$$

Since t is rational, Q_t has rational coordinates.

We now claim that every rational point on the circle C other than $P = (-1,0)$ arises as Q_t for exactly one rational number t. In other words, we obtain a parameterization of all the rational points on C (except P). Recall that $\mathbb{Q}$ denotes the set of all rational numbers.

Theorem 1. *The map*

$$\begin{aligned} \mathbb{Q} &\to \{\textit{rational points on } x^2+y^2 = 1 \textit{ other than } (-1,0)\} \\ t &\mapsto Q_t = \left(\frac{1-t^2}{1+t^2}, \frac{2t}{t^2+1}\right) \end{aligned}$$

is a bijection (one-to-one correspondence).

Proof. There is a natural candidate for the inverse map, namely, the map

$$\begin{aligned} \{\text{rational points on } x^2+y^2 = 1 \text{ other than } (-1,0)\} &\to \mathbb{Q} \\ (r,s) &\mapsto \frac{s}{r+1} \end{aligned}$$

sending a rational point $Q = (r,s)$ on $x^2+y^2 = 1$ other than $(-1,0)$ to the slope of the line $\overleftrightarrow{PQ}$.

To show that the two maps are indeed inverse bijections, it suffices to show that the composition of the two maps in either order is the identity map.

Given $t \in \mathbb{Q}$, if we construct Q_t, and then take the slope of the line $\overleftrightarrow{PQ_t}$, we get t back, by definition of Q_t.

On the other hand, if we start with a rational point $Q \neq P$ on C, compute the slope t of the line $L = \overleftrightarrow{PQ}$, and then construct Q_t, then $Q_t = Q$ because Q is the intersection point other than P of C with the line L through P with slope t. This completes the proof. □

If m and n are positive integers with $m > n$, and we take $t = n/m$, then we obtain the point

$$Q_t = \left(\frac{m^2-n^2}{m^2+n^2}, \frac{2mn}{m^2+n^2}\right),$$

so $(m^2-n^2, 2mn, m^2+n^2)$ is a Pythagorean triple.

5. Elliptic curves and Bézout's theorem

In the previous section we parameterized the rational points on the circle $x^2 + y^2 = 1$ by choosing one rational point P, and then looking at the intersection of the circle with lines through P having rational slope.

Rational points on elliptic curves cannot be parameterized in the same way. What goes wrong if we try to repeat the procedure that worked for the circle? To fix ideas, let us see what happens for the elliptic curve E with equation

$$y^2 = x(x+5)(x-5).$$

(The polynomial $x(x+5)(x-5)$ has distinct roots, so this *is* an elliptic curve.) The point $S = (-4, 6)$ is on the curve E. What happens if we intersect E with the line L of slope 1 through S?

The equation of L is $y - 6 = 1(x+4)$, i.e., $y = x + 10$. Substituting this into the equation of E yields

$$\begin{aligned}(x+10)^2 &= x(x+5)(x-5)\\ 0 &= x^3 - x^2 - 45x - 100\\ 0 &= (x+4)(x^2 - 5x - 25).\end{aligned}$$

The linear factor $x + 4$ was expected; it reflects the fact that the point $S = (-4, 6)$ is one of the intersection points. But this time the leftover factor is quadratic, not linear, so there is no reason to expect the other solutions to be rational. In fact, here they are not, because the discriminant of $x^2 - 5x - 25$ is $(-5)^2 - 4(1)(-25) = 125$, which is not the square of a rational number. Hence we do not obtain rational points on E in this way.

Geometrically what has happened is that L intersects E in three points, one of which is S, and the best that can be said of the other two is that their coordinates will involve a single square root. It is not an accident that $L \cap E$ consisted of three points here, whereas the intersection of L with a circle in the previous section had two points. These are special cases of the following general result:

Bézout's Theorem (almost): It is almost true that the intersection of a curve $f(x, y) = 0$ of degree m with a curve $g(x, y) = 0$ of degree n consists of exactly mn points.

To make the theorem true, some care must be taken. For instance, the intersection of $xy = 0$ and $(x^2 + y^2)y = 0$ has infinitely many points, not $2 \cdot 3 = 6$ as predicted, because both curves contain the curve $y = 0$. Therefore one should insist that the two curves do not have a curve in common. Algebraically, this is equivalent to imposing the condition that $f(x, y)$ and $g(x, y)$ have no common factor.

With this assumption, it is now true that the curves intersect in *at most* mn points. But to get exactly mn points, three more modifications to the problem are required. As stated, the theorem gives the wrong answer for the number of real points in the intersection of $x - 2 = 0$ with $x^2 + y^2 = 1$. To get the correct number of intersection points $(1 \cdot 2 = 2)$, one should allow the intersection points $(2, \sqrt{-3})$ and $(2, -\sqrt{-3})$ with complex coordinates. Another problem is illustrated by the example in which one intersects $x - 1 = 0$ with $x^2 + y^2 = 1$. We again expect 2 intersection points, but there is only one, the point $(1, 0)$ where the line is tangent to the circle. The fix this time is to count intersection points with multiplicity: points where two curves meet tangentially count extra! The third problem is that certain curves such as $y-1=0$ and $y-2=0$ do not meet as many times as they are supposed to. One finds the "missing intersection points" by adjoining a "line of points at infinity" to the plane, to form the *projective plane* $\mathbb{P}^2$. The lines $y-1=0$ and $y-2=0$ meet at one of the points on this line at infinity. (We will not discuss this in detail here.)

The precise version of Bézout's Theorem reads as follows:

Bézout's Theorem: Let X and Y be curves of degrees m and n in the projective plane over the complex numbers. If X and Y have no curves in common, then the number of intersection points in $X \cap Y$ with complex coordinates, counted with multiplicities, equals mn exactly.

6. The addition law on elliptic curves

Let E be an elliptic curve defined over $\mathbb{Q}$. We have seen that a line L with rational slope passing through one rational point on E need not intersect E in rational points only. But if L passes through *two* rational points on E, then the third intersection point must be rational. This is because a cubic polynomial with two rational roots must have all its roots rational. One caveat is required, however: in order to be guaranteed to have three intersection points, one must interpret intersections in the sense of Bézout's Theorem; in other words, one really should work in the projective plane $\mathbb{P}^2$ over the complex numbers, and count intersection points with multiplicities. It turns out that among all the points at infinity in the projective plane, only one is on the elliptic curve; i.e., the line at infinity intersects E only in one point (with multiplicity 3, though!)

For an example, let us go back to the elliptic curve E of the previous section with equation $y^2 = x(x+5)(x-5)$. Let us find the third intersection point U of E with the line L through $S=(-4,6)$ and $T=(0,0)$. The equation of L is $y=(-3/2)x$, so the x-coordinates of the points in $E \cap L$ are solutions to

$$\begin{aligned}
[(-3/2)x]^2 &= x(x+5)(x-5) \\
0 &= x^3 - (9/4)x^2 - 25x \\
0 &= x(x+4)(x-25/4).
\end{aligned}$$

As usual, the factors x and $x+4$ had to be there, because S and T are in $E \cap L$. Now we know that the x-coordinate of U is $25/4$, and using the equation of L we find that $U=(25/4,-75/8)$.

Using this operation of taking two rational points and producing a third, we can develop a way to "add" two points. One says that the set of rational points on E can be given the structure of an abelian group. This means that there is an operation $+$ that takes two rational points P,Q on E and produces a new rational point $P+Q$ on E, such that the following axioms are satisfied:

- $(P+Q)+R = P+(Q+R)$ for all rational points P,Q,R on E.
- There exists a rational point O on E such that $P+O=P$ and $O+P=P$ for all rational points P on E.
- For any rational point P on E there exists a point Q (also called $-P$) such that $P+Q=O$ and $Q+P=O$.
- $P+Q=Q+P$ for all rational points P and Q on E.

The specific addition rule on the elliptic curve is characterized by the following rules:

(1) The point O mentioned in the abelian group axioms is the unique point on E at infinity mentioned earlier.

(2) If a line L intersects E in three rational points P, Q, R (listed with multiplicity), then $P + Q + R = O$.

Note: a line passes through O if and only if it is vertical or is the "line at infinity."

As an example, let us compute $S + T$, where $S = (-4, 6)$ and $T = (0, 0)$. We already found that the line $y = (-3/2)x$ intersects E in the points S, T, and $U = (25/4, -75/8)$. Therefore, by Rule 2, $S + T + U = 0$. Thus $S + T = -U$. The vertical line $x = 25/4$ intersects E in the three points U, $V = (25/4, 75/8)$ and O, so $U + V + O = O$. Hence $V = -U$, so $S + T = V = (25/4, 75/8)$.

In general, the recipe for adding two points A and B on an elliptic curve E is as follows: draw the line L through A and B. (If $A = B$, draw the line tangent to E at A, in order to get a line that intersects E at A with multiplicity at least 2.) Find a third point C such that $L \cap E$ consists of A, B, and C (listed with multiplicity if necessary). If $C = O$, then $A + B$ equals O; if $C \neq O$, $A + B$ equals the reflection of C in the x-axis.

7. Generating all the rational points

It turns out that for some elliptic curves over $\mathbb{Q}$, such as $y^2 = x^3 - x$, there are only finitely many rational points, while for others, such as the example $y^2 = x(x+5)(x-5)$ above, there are infinitely many.

But in any case there is a deep theorem, proved by Mordell, that says that the group of rational points on an elliptic curve E is "finitely generated." This means there is a finite list of rational points $\mathcal{S}$ on E such that all rational points on E can be generated from the points in $\mathcal{S}$ by iteratively applying $+$ to pairs of points.

On the other hand, it is not known whether there exists an *algorithm* that takes the equation of an elliptic curve and outputs a finite list $\mathcal{S}$ of generating points as above. Researchers in number theory have spent about 70 years trying to prove the existence of such an algorithm, but the problem is still unsolved!

8. Beyond elliptic curves

A special case of an even deeper theorem of Faltings shows that if X is a nonsingular[3] curve defined over $\mathbb{Q}$ of degree *greater than 3*, then there are only finitely many rational points on X. Faltings was awarded the Fields Medal (the mathematical equivalent of the Nobel Prize) for proving this theorem.

From the algorithmic point of view, however, things are still very mysterious: it is not known whether there is a method for actually *listing* the rational points on a given nonsingular curve of degree greater than 3.

For example, the French mathematician Jean-Pierre Serre challenged the mathematical community many years ago to prove that the eight obvious rational points on $x^4 + y^4 = 17$ are the only ones; it took until 2001 for this to be proved.

[3] We have not defined this term, but loosely speaking it means that X has no "corners" or points where the curve "crosses itself."

9. Further reading

If you would like to read more about elliptic curves, you could try the book *Rational points on elliptic curves* by Silverman and Tate, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1992. It is written at a fairly high level, though: it is really intended for junior and senior undergraduate mathematics majors.

These notes were originally developed for the Berkeley Math Circle.

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA
E-mail address: poonen@math.mit.edu
URL: http://math.mit.edu/~poonen/