Elliptic Curves

John Stillwell

American Mathematical Monthly, Volume 102, Issue 9 (Nov., 1995), 831-837.

Stable URL:
http://links.jstor.org/sici?sici=0002-9890%28199511%29102%3A9%3C831%3AEC%3E2.0.CO%3B2-7

THE EVOLUTION OF...

Edited by **Abe Shenitzer**
Mathematics, York University, North York, Ontario M3J1P3, Canada

Elliptic Curves

John Stillwell

In recent years, elliptic curves have played a leading role in number theory, most famously in Wiles' program to prove Fermat's last theorem. However, since these developments are highly technical, it may be useful to look back to earlier times, when elliptic curves led a simpler life. For about 1500 years, from the time of Diophantus to Newton, elliptic curves were known only as curves defined by certain cubic equations. This put them just a step beyond the conic sections, and some of their geometric and arithmetic properties can in fact be viewed as generalisations of properties of conics. In particular, it is possible to find rational solutions of both quadratic and cubic equations by simple geometric constructions.

It was only with the development of calculus, in the 17th century, that sharp differences between conics and elliptic curves began to emerge. Conic sections can be parametrised by rational functions. For example, the circle $x^2 + y^2 = 1$ is parametrised by

$$x = \frac{1 - t^2}{t + t^2}, y = \frac{2t}{1 + t^2}$$

but the elliptic curves cannot. Their simplest parametrising functions are *elliptic functions*, which arise in calculus as the inverses of elliptic integrals, so-called because a typical example is the integral for the arc length of the ellipse. It is for this fairly accidental reason that they are called elliptic curves—an unfortunate accident since the ellipse itself is *not* an elliptic curve.

The difference between conics and elliptic curves was "felt" in the 17th century in the apparent intractability of elliptic integrals, though the parametrisation of cubic curves was not known at that time. The idea of inverting elliptic integrals to create elliptic functions had to wait until the early 19th century. The nonrationality of elliptic curves was not fully understood until the mid-19th century, when the introduction of complex coordinates revealed a *topological* difference between them and conics. This brings us within sight of the modern view of elliptic curves—a remarkable synthesis of number theory, geometry, algebra, analysis and topology. In what follows I shall attempt to describe what led up to this state of affairs.

Diophantus. Very little is known about Diophantus except that he lived sometime between 150 AD and 350 AD and was a wizard at finding rational solutions to polynomial equations in two or more variables. His *Arithmetica* (available in the

English edition of Heath [4]), contains the solutions of hundreds of equations, among them the following instructive examples.

1. A rational solution of $x^2 + y^2 = 16$, other than an obvious one such as $x = 0$, $y = 4$, is found by solving the simultaneous equations

$$x^2 + y^2 = 16,$$

$$y = 2x - 4,$$

which yield the solution $x = 16.5$, $y = 12/5$ (Heath [4], p. 145).

2. A rational solution of $x^3 - 3x^2 + 3x + 1 = y^2$, other than the obvious one $x = 0$, $y = 1$, is found by solving the simultaneous equations

$$x^3 - 3x^2 + 3x + 1 = y^2,$$

$$y = \tfrac{3}{2}x + 1,$$

which yield the solution $x = 21/4$, $y = 71/8$ (Heath [3], p. 242).

How did Diophantus choose the linear equations in these two examples? The simplest explanation is geometric, although he makes no mention of geometry.

In the first example the linear equation represents a line through the "obvious" rational point $(0, 4)$. Its slope is not important, since any line through $(0, 4)$ with rational slope t will meet the circle at a second rational point $(8t/(1 - t^2), (4t^2 - 4)/(1 + t^2))$. Conversely, all rational points on the circle are obtainable in this way, so Diophantus has essentially *parametrised* the rational points on the circle by rational functions of a rational parameter t.

The linear equation in the second example has an even stronger geometric smell. It is the *tangent* to $x^3 - 3x^2 + 3x + 1 = y^2$ at the "obvious" rational point $(0, 1)$. Here there is no option about the slope because a line has to meet a cubic curve in *two* rational points for its third intersection to be rational. When only one rational point is known, this forces us to use the tangent, which is the line through two "coincident" points.

It is possible, of course, that Diophantus discovered these facts purely algebraically, and did not notice their geometric interpretation. However, that would be a truly amazing departure from the Greek mathematical culture of his time. Even in the more algebraic culture of the 17th century, Fermat and Newton immediately recognised Diophantus' work as geometry, with Newton [6] explicitly interpreting Diophantus' solutions as chord and tangent constructions. Later discoveries added more weight to the geometric interpretation, as we shall see below.

Fermat and Newton. Fermat was the first mathematician to make significant progress in number theory beyond Diophantus. Among his many discoveries were methods for proving *non*existence of integer or rational solutions for certain equations. For example, he proved that there are no positive rationals a, b, c such that

$$a^4 \pm b^4 = c^2$$

This implies in particular that no positive integer fourth powers sum to a fourth power (the $n = 4$ case of Fermat's last theorem), but it is also a statement about an elliptic curve. It says that there are no nontrivial rational points on the curve

$$y^2 = 1 - x^4,$$

since a rational point $(p/r, q/r)$ with $p, q \neq 0$ and

$$\frac{p^2}{r^2} = 1 - \frac{q^4}{r^4}$$

gives nonzero integers $a = r, b = q, c = pr$ with $a^4 - b^4 = c^2$.

Now I know I said that elliptic curves are cubics, but they are cubic *in a suitable coordinate system*. Any quartic curve of the form

$$y^2 = (x - \alpha)(x - \xi)(x - \gamma)(x - \delta)$$

can be rewritten

$$\left(\frac{y}{x - \alpha^2}\right)^2 = \left(1 - \frac{\beta - \alpha}{x - \alpha}\right)\left(1 - \frac{\gamma - \alpha}{x - \alpha}\right)\left(1 - \frac{\delta - \alpha}{x - \alpha}\right)$$

and hence it is cubic in the coordinates

$$X = \frac{1}{x - \alpha}, Y = \frac{y}{x - \alpha^2}.$$

In particular, $y^2 = 1 - x^4$ is a cubic $Y^2 = 4X^3 - 6X^2 + 4X - 1$ in the coordinates $X = 1/(1 - x), Y = y/(1 - x)^2$. Notice that this is an appropriate coordinate change from the point of view of number theory, because it makes the rational points (x, y) on one curve correspond to the rational points (X, Y) on the other. Such a coordinate change is called *birational*.

Newton made the surprising discovery that all cubic equations in x and y can be reduced to the form

$$Y^2 = X^3 + aX + b$$

by a birational coordinate transformation. In fact, the transformations he used were simply projections. He called this "genesis of curves by shadows". His result can be viewed as an analogue of the well known theorem that second degree curves are conic sections and hence, in nondegenerate cases, projections of the circle. The degenerate cubic curves are those for which the right hand side $X^2 + aX + b$ has a repeated factor. The corresponding repeated root $X = \alpha$ is either a double point (Fig. 1) or cusp (Fig. 2) of the curve, and by drawing a line of slope t through this point we obtain the coordinates of the general point on the curve as rational functions of t.

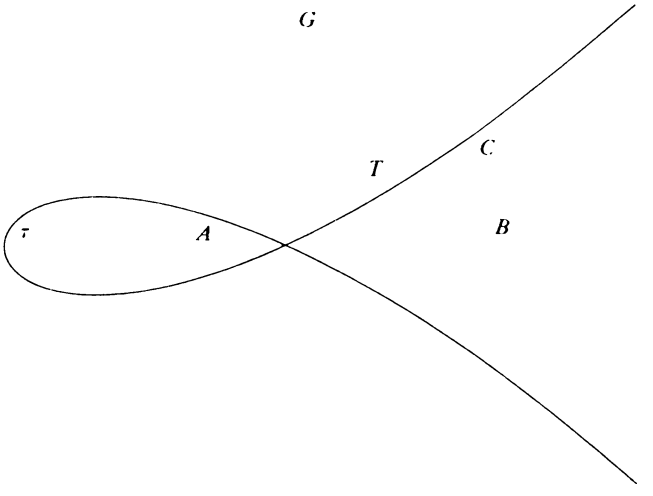

Figure 1. Cubic with double point.

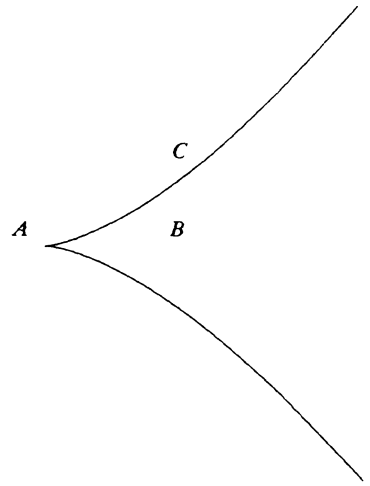

Figure 2. Cubic with cusp.

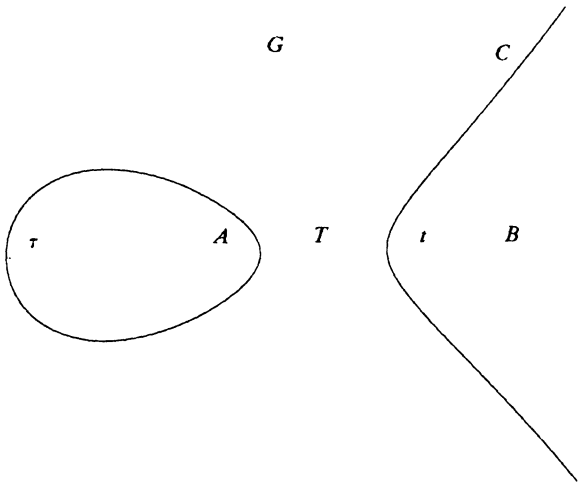

Figure 3. Nonsingular cubic.

The curves for which $X^3 + aX + b$ has no repeated factor cannot be parametrised by rational functions, and are what we now call elliptic curves (Fig. 3).

Elliptic integrals. Early in the development of integral calculus, mathematicians encountered the problem of "rationalising" square roots of polynomials. For example, to find the area or arc length of a circle one finds an integral involving $\sqrt{1 - x^2}$. This can be rationalised by the "Diophantine" substitution $x = (1 - t^2)/1 + t^2$, and fact Jakob Bernoulli [1], in a similar situation, actually attributed the substitution to Diophantus. He used it to obtain the expression

$$\frac{\pi}{4} = \int_0^1 \frac{dt}{1 + t^2},$$

whence he obtained the famous series

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

by expanding $1/(1 + t^2)$ in a geometric series and integrating term by term.

Integrals involving square roots of cubic or quartic polynomials proved more intractable. They were called *elliptic integrals* because one of them expresses the arc length of the ellipse. Cubics and quartics were lumped together because of birational equivalences between them, as noted above for $y^2 = 1 - x^4$ and $Y^2 = 4X^3 - 6X^2 + 4X - 1$. Such integrals arise from a great number of natural geometric and mechanical problems, so a lot of effort was expended on them, but without success.

Perhaps the first to see why rationalisation might be impossible was Jakob Bernoulli [2], who noted that a rationalisation of $\sqrt{1 - x^4}$, at least by a rational function $x = f(t)$ with *rational* coefficients, would violate Fermat's theorem on the nonexistence of positive integer solutions of $a^4 \pm b^4 = c^2$. In fact, it can be shown that $\sqrt{1 - x^4}$ cannot be rationalised by *any* rational function $x = f(t)$, by repeating Fermat's argument with polynomials in place of integers, so Jakob Bernoulli was on the right track. However, this type of argument was not used until the 19th century, so the nature of elliptic integrals remained unclear until then (when ideas not only from number theory, but also from analysis and topology, were directed at the problem).

Elliptic functions. In the 1820s, Abel and Jacobi finally saw what to do with elliptic integrals—*Invert* them. Instead of studying the integral

$$u = g^{-1}(x) = \int_0^x \frac{dt}{\sqrt{t^3 + at + b}},$$

say, study its inverse function $x = g(u)$. The gain in simplicity is comparable to studying the function $x = \sin u$ instead of the integral $\sin^{-1} x = \int_0^x (dt/\sqrt{1 - t^2})$. In particular, instead of a multi-valued integral $g^{-1}(x)$, one has a *periodic function* $x = g(u)$.

The difference between $\sin u$ and $g(u)$ is that the periodicity of $g(u)$ cannot be properly seen until complex values of the variables are admitted, at which stage it emerges that $g(u)$ has *two* periods. That is, there are nonzero $\omega_1, \omega_2 \in \mathbf{C}$, with $\omega_1/\omega_2 \notin \mathbf{R}$, such that

$$g(u) = g(u + \omega_1) = g(u + \omega_2).$$

The two periods can be brought to light in various ways. One method, originating with Eisentein [1847] and commonly used today, is to write down a function that obviously has periods ω_1 and ω_2, namely

$$g(u) = \sum_{m,n \in \mathbf{Z}} \frac{1}{(u + m\omega_1 + n\omega_2)^2},$$

and derive its properties by manipulation of infinite series. Eventually one finds that $g^{-1}(x)$ is an integral of the type we started with.

A more insightful approach though harder to make rigorous, is to study the behaviour of the integrand $1/\sqrt{t^3 + at + b}$ as t varies over the complex plane. Following Riemann [7], and viewing the 2-valued "function" $1/\sqrt{t^3 + at + b}$ as a 2-sheeted surface over $\mathbf{C}$, one finds that there are two independent closed paths of

integration, over which the integrals are ω_1 and ω_2. This accounts for the periods ω_1 and ω_2 of the inverse function $g(u)$.

Since $g(u) = x$, it follows by basic calculus that

$$g'(u) = \frac{dx}{du} = \frac{1}{du/dx} = \frac{1}{1/\sqrt{x^3 + ax + b}} = \sqrt{x^3 + ax + b} = y,$$

so $x = g(u)$, $y = g'(u)$ gives a parametrisation of the curve $y^2 = x^3 + ax + b$. With a little more work it can be shown that $u \mapsto (g(u), g'(u))$ is in fact a continuous one-to-one correspondence between $\mathbf{C}/\langle \omega_1, \omega_2 \rangle$ and the curve. $\mathbf{C}/\langle \omega_1, \omega_2 \rangle$ is the quotient of $\mathbf{C}$ by the subgroup generated by ω_1 and ω_2 and is topologically a *torus*, hence so is the curve $y^2 = x^3 + ax + b$. This is the deeper reason why elliptic curves are not rationally parametrisable—a curve parametrised by rational functions $x = p(u)$, $y = q(u)$ is the topological image of the completed plane $\mathbf{C} \cup \{\infty\}$ of u values, and $\mathbf{C} \cup \{\infty\}$ is topologically a *sphere*.

Another consequence of the parametrisation $x = g(u)$, $y = g'(u)$ is that the curve $y^2 = x^3 + ax + b$ is an abelian group. The "sum" of points with parameter values u_1, u_2 is simply the point with parameter value $u_1 + u_2$. Under this definition of sum, the curve is isomorphic to the group $\mathbf{C}/\langle \omega_1, \omega_2 \rangle$. Amazingly, there is an equivalent definition of the sum that Diophantus would have understood (and which helps to explain why elliptic functions are useful in number theory): the sum of the points P_1 and P_2 is simply the reflection, in the x-axis, of the third point on the curve collinear with P_1 and P_2 (Fig. 4). For an explanation

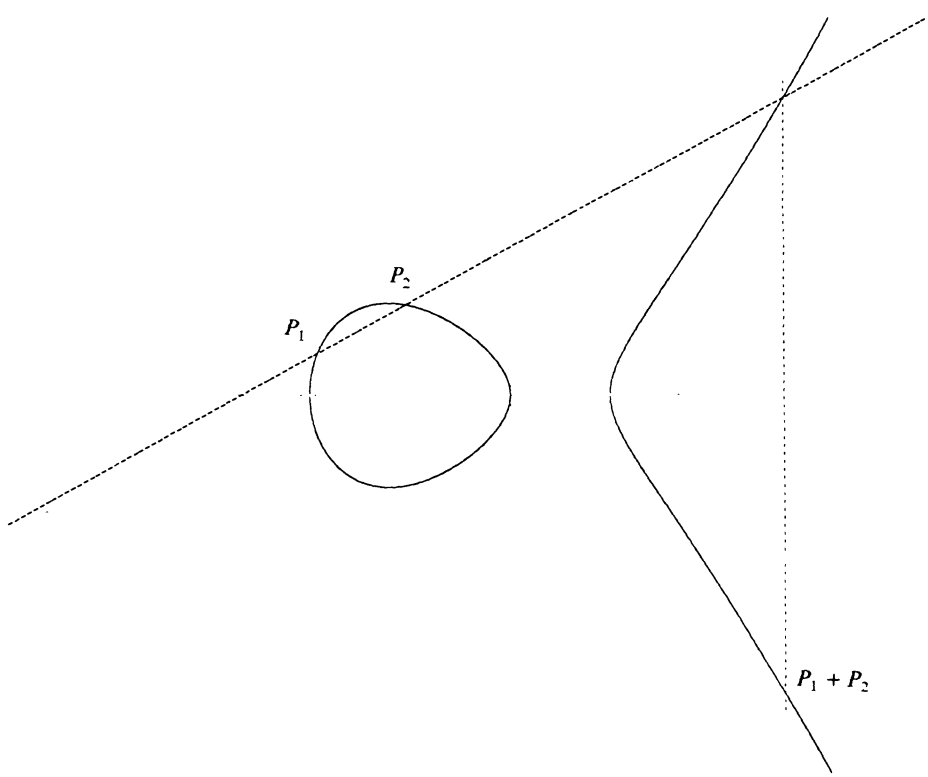

Figure 4. Addition of points on an elliptic curve (from Koblitz [5]).

of this face we must refer the reader to a recent book on elliptic curves, such as Koblitz [5]. In the same book you will find many beautiful modern results on elliptic curves, motivated by ancient problems in number theory and geometry.

REFERENCES

1. Bernoulli, Jakob (1696) Positionum de seriebus infinitis pars tertia. *Werke*, **4**, 85–106.
2. Bernoulli, Jakob (1704) Positionum de seriebus infinitis ... pars quinta. Werke, **4**, 127–147.
3. Eisenstein, G. (1847) Beitrâge zur Theorie der elliptischen Functionen. *J. reine angew. Math.* **35**, 137–274.
4. Heath, T. L. (1910) *Diophantus of Alexandria*, Cambridge University Press.
5. Koblitz, N. (1985) *Introduction to Elliptic Curves and Modular Forms*, Springer-Verlag, New York.
6. Newton, I. (late 1670s) De resolutione quaestionum circa numeros. *Math. Papers* **4**, 110–115.
7. Riemann, G. B. H. (1851) Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse. *Werke*, 2nd ed., 3–48.

Department of Mathematics
Monash University
Clayton 3168
AUSTRALIA
stillwell@monash.edu.au.

Without the concepts, methods and results found and developed by previous generations right down to Greek antiquity one cannot understand either the aims or the achievements of mathematics in the last fifty years.

—H. Weyl (in 1950)