Why Ellipses Are Not Elliptic Curves

ADRIAN RICE
Randolph-Macon College
Ashland, VA 23005-5505
arice4@rmc.edu

EZRA BROWN
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0123
ezbrown@math.vt.edu

After circles, ellipses are probably the most familiar curves in all of mathematics. Like circles, they are a special subclass of the so-called conic sections, or curves obtained by slicing a cone with a plane, and their applications are many and varied. For example, thanks to Johannes Kepler and his laws of planetary motion, astronomers know ellipses as the orbits of planets and many comets about the sun. In acoustics, architects have used the reflection property of ellipses—namely, that a light ray originating at one focus is reflected off the ellipse to the other focus—to construct whispering galleries in such places as St. Paul's Cathedral in London and Statuary Hall in the U. S. Capitol. Ellipses even find their way into modern medicine, where the reflection property is the basis for lithotripsy, a medical procedure for treating kidney stones and gall stones without invasive surgery. They also crop up from time to time in bad jokes mathematicians often like to tell: "What shape is a kiss?" "A lip tickle!"

In analytic geometry we learn that an ellipse is the set of all points in the plane the sum of whose distances from two fixed points is a given positive constant. Using this definition along with the distance formula, we may derive equations for ellipses which, in general, are of the form $Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$, where $B^2 - 4AC < 0$. However, we may translate and rotate the axes as necessary to obtain the familiar equation of an ellipse centered at the origin with semimajor axis a and semiminor axis b, namely,

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

which looks like FIGURE 1.

This beautiful object is certainly a curve, and its shape is evidently elliptical, so you would think that mathematicians would call it an "elliptic curve." But they do not. The name "elliptic curve" is reserved for a very different class of curves and, as with ellipses, we can define these in more than one way.

We can start with curves in the real plane $\mathbb{R}^2$, but much of the theory of elliptic curves depends on seeing them in $\mathbb{C}^2$. All in good time: for now, we give a simplified definition, namely that an elliptic curve is the set of solutions to an equation of the form $E(x, y) = 0$, where $E(x, y)$ is a cubic polynomial in x and y. We require further that

Math. Mag. **85** (2012) 163–176. doi:10.4169/math.mag.85.3.163. © Mathematical Association of America

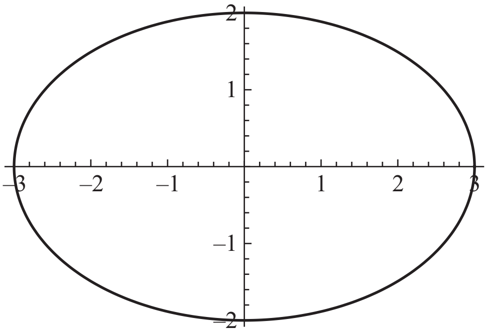

Figure 1 An ellipse with semimajor axis $a = 3$ and semiminor axis $b = 2$

$E(x, y)$ is *nonsingular*, which means that at no point do the function E and its partial derivatives E_x and E_y all vanish simultaneously. We may apply transformations, such as translations and other more complicated ones, to show that for our purposes, elliptic curves all have the form

$$y^2 = p(x),$$

where $p(x)$ is a cubic polynomial with no repeated roots. FIGURE 2 has two examples.

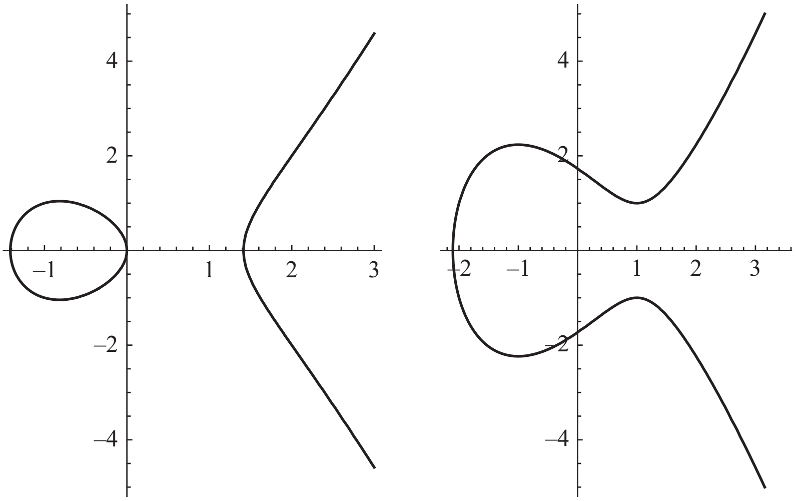

Figure 2 Two elliptic curves: $y^2 = x^3 - 4x$ (left) and $y^2 = x^3 - 3x + 3$ (right)

Elliptic curves come from algebraic geometry, and their applications show up in various systems of public key cryptography, in the factorization of large integers, in primality testing, and most famously in the proof of Fermat's Last Theorem. There are two principal facets to the study of elliptic curves, namely the discrete (which arises mainly from problems in number theory and abstract algebra) and the continuous (coming principally from the realms of calculus and complex analysis). The paper "Three Fermat trails to elliptic curves" [**5**] is a look at the history of the subject's discrete side via the congruent numbers problem, Fermat's Last Theorem, and the search for nontrivial integer solutions of $x^4 + ax^2y^2 + y^4 = z^2$, showing how each of these problems found solutions by recasting them as questions involving elliptic curves.

This paper has a very different focus and motivation. It deals with the history of the continuous side of the subject, from attempts to rectify the ellipse all the way up to the Weierstrass $\wp$-function. It thus serves as a companion piece to [**5**]. But why is such an article necessary? Well, when comparing FIGURES 1 and 2, many might be tempted to ask why it is that ellipses and elliptic curves look nothing like each other, yet have names that sound so similar. And they are quite right to wonder, because elliptic curves have almost nothing to do with ellipses at all. Why then are they called elliptic curves?

The answer lies in the word *almost*. There *is* a connection between ellipses and elliptic curves, but it's not at all obvious and is the result of a connected but distinctly nonlinear sequence of mathematical events. The simplest mathematical reason why ellipses are not elliptic curves is that their algebraic forms are fundamentally different: as we have seen, ellipses are quadratic, elliptic curves are cubic.

But this is not a particularly interesting answer. Nor does it explain how such different geometrical objects ended up with such similar-sounding names. To *really* answer the question properly, we need to look back at the history and development of these concepts. We will therefore take a stroll through the history of mathematics, encountering first the ellipse, moving on to elliptic integrals, then to elliptic functions, jumping back to elliptic curves, and eventually making the connection between elliptic functions and elliptic curves. We will then finally be in a position to find out why no elliptically-shaped planar curves may ever be called elliptic curves.

From ellipses to elliptic integrals

It all started, as many mathematical stories do, in ancient Greece and with one of the three classical construction problems, known as the Duplication Problem: given a cube with a certain volume V, construct a cube of volume $2V$ using only a compass and a straightedge. In modern notation, if a is the edge of the original cube, the goal is to construct a line segment of length $a\sqrt[3]{2}$. One early geometrical solution, ascribed to Hippocrates of Chios (ca. 460–380 BCE), involved determining two lengths x and y that satisfy the proportions

$$a : x = x : y = y : 2a.$$

Considering the three proportions separately and treating x and y as variables, the 4th century BCE mathematician Menaechmus showed that these proportions yield the curves $x^2 = ay$, $y^2 = 2ax$, and $xy = 2a^2$, which we recognize as equations of two parabolas and a hyperbola. For, if we multiply the first equation by x and the third equation by a, we are led to the equations $x^3 = axy = 2a^3$; hence, $x = a\sqrt[3]{2}$.

Menaechmus showed that any two of these equations imply the third, and that the lengths x and y are indeed the lengths required to produce the length $a\sqrt[3]{2}$. (We note that the Greek geometers developed several other constructions for finding $a\sqrt[3]{2}$, including Archytas of Tarentum's ingenious method involving the intersection of a cylinder, a torus with zero interior diameter, and a right circular cone. Pierre Wantzel (1814–1848) finally proved in 1837 that constructing $a\sqrt[3]{2}$ using only compass and straightedge is impossible—but that's another story.) Menaechmus went on (some say) to describe these as conic sections, discovering the ellipse in the process. Around 300 BCE Euclid wrote *Conics*, a major work, now known only through later commentaries, and containing a number of theorems on various properties of ellipses.

But it was Apollonius of Perga (ca. 262–190 BCE) who, in his eight-volume treatise *On Conics* [**10**], provided the most exhaustive study to date of the subject, as well as giving them the name ellipse, from the Greek *elleipsis*, meaning "falling short." So comprehensive was Apollonius's work that for nearly two millennia, it contained the majority of what was known on the subject. But there were gaps, and by the 17th century, mathematicians finally began to develop techniques that could fill them.

One question that Apollonius could not answer precisely was how to find the arc length of an ellipse. Geometric techniques were insufficient, as ellipses are curved shapes, but the invention of the integral calculus in the 1660s and 1670s provided a marvelous new tool for answering this question. With the introduction of this new technique, the question of finding the precise lengths of various curves, including the ellipse, became a major open problem for mathematicians. The arc length formula is one of the standard topics in courses on integral calculus: if $y = f(x)$ is continuous and has a continuous derivative on the interval $[a, b]$, then the length L_a^b of the curve is given by

$$L_a^b = \int_a^b \sqrt{1 + (f'(x))^2}\, dx.$$

But the first attempts from that era to find the arc length of an ellipse involved series, not integrals. For example, in 1669, Isaac Newton (1642–1727) expressed the arc length of an ellipse as an infinite series; other series-based expressions followed from the great Swiss genius Leonhard Euler (1707–1783) in 1733 and the Scottish mathematician Colin Maclaurin (1698–1746) in 1742. Why did they avoid integration?

To understand why, let's try using integration to find the arc length of an ellipse between, say, x_0 and x_1 and see what happens. Let a and b be positive numbers with $a > b$, and consider the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

We solve this equation for y and take the positive square root, which yields the function $y = f(x) = b\sqrt{a^2 - x^2}/a$. We calculate $\sqrt{1 + (f'(x))^2}$ and simplify the resulting messy expression by setting $k = \sqrt{a^2 - b^2}/a$; this transforms the resulting arc length integrand into $\sqrt{(a^2 - k^2x^2)/(a^2 - x^2)}$, and the arc length formula becomes

$$L_{x_0}^{x_1} = \int_{x_0}^{x_1} \sqrt{\frac{a^2 - k^2x^2}{a^2 - x^2}}\, dx.$$

Thus the total arc length, L, of the ellipse is given by

$$L = 4 \int_0^a \frac{\sqrt{a^2 - k^2x^2}}{\sqrt{a^2 - x^2}}\, dx.$$

Unfortunately, this integral cannot be evaluated directly. The same is true if we use trigonometric functions to parameterize the curve as $x = a \sin t$, $y = b \cos t$, for $0 \leq t \leq 2\pi$, when the integral becomes

$$L = 4a \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 t}\, dt.$$

This integral, in either the algebraic or the trigonometric form, is commonly known as an *elliptic integral*. To be more precise, this particular integral is called an *elliptic integral of the second kind*. The first part of this name arises because we are trying to determine the arc length of an *ellipse* via *integration*. But why "of the second kind"? And how many kinds are there?

The terminology and classification were introduced by the French mathematician Adrien-Marie Legendre (1752–1833). Legendre was fascinated by various interesting types of integrals which could not be computed by regular means—today we would call these "non-elementary integrals." These are integrals of functions $f(x)$ for which f does not have an antiderivative expressible in terms of elementary functions—polynomial, rational, algebraic, trigonometric, logarithmic or exponential. Starting in the 1750s, Euler had derived a great many results about these kinds of integrals, but it was Legendre who turned the subject into a systematic theory.

For 40 years from 1786, Legendre worked with many kinds of nonelementary integrals. He finally realized that the integrals arising from the above arc-length calculations could be expressed as one of three fundamental types, which we now define as *elliptic integrals of the first, second, and third kind,* respectively:

$$F(\phi) = \int_0^{\phi} \frac{dt}{\sqrt{1 - k^2 \sin^2 t}},$$

$$E(\phi) = \int_0^{\phi} \sqrt{1 - k^2 \sin^2 t}\, dt, \quad \text{and}$$

$$\Pi(\phi) = \int_0^{\phi} \frac{dt}{\left(1 + n \sin^2 t\right) \sqrt{1 - k^2 \sin^2 t}}.$$

Here, k, or the *modulus*, is a value in [0, 1]. Strictly speaking, the modulus k is a real constant such that k^2 is not equal to 0 or 1, but most texts on elliptic integrals restrict k such that $0 < k < 1$. The upper limit of the elliptic integrals, the amplitude ϕ, can be any real number, although it makes sense to focus on values in $[0, \pi/2]$. As for the word *amplitude*, this expression arose from Legendre's usage in referring to the physical applications which drove much of his work on elliptic integrals. Finally, in the elliptic integral of the third kind, n is taken to be a real constant, usually assumed to be nonzero because the case $n = 0$ reduces to the elliptic integral of the first kind.

Between 1825 and 1828, Legendre published a three-volume treatise [**16**] on these elliptic integrals (which, confusingly for us, he called elliptic *functions*), containing much of his four decades of work on the subject. How ironic, then, that just as Legendre was finishing his life's work, two young mathematicians were just beginning theirs with ideas that would render many of Legendre's techniques obsolete. Those two mathematicians were Niels Henrik Abel (1802–1829) and Carl Gustav Jacobi (1804–1851).

From elliptic integrals to elliptic functions

Both Abel and Jacobi wrote Legendre's elliptic integrals using the substitution $x = \sin t$, to give

$$F(u) = \int_0^{u} \frac{dx}{\sqrt{\left(1 - x^2\right)\left(1 - k^2 x^2\right)}},$$

$$E(u) = \int_0^{u} \frac{1 - k^2 x^2}{\sqrt{(1 - x^2)(1 - k^2 x^2)}}\, dx, \quad \text{and}$$

$$\Pi(u) = \int_0^{u} \frac{dx}{\left(1 + n x^2\right)\sqrt{\left(1 - x^2\right)\left(1 - k^2 x^2\right)}} \quad (\text{where } |u| \le 1),$$

as the elliptic integrals of the first, second, and third kind, respectively. Indeed, to this day, elliptic integrals are still defined as those whose integrands are rational functions involving square roots of cubic or quartic polynomials. But it was Abel who realized that these integrals, although interesting and important, were not the most significant thing to be studying. (Gauss had realized this 30 years before, but did not publish his findings.) Consider the well-known integral

$$u = f(x) = \int_0^x \frac{dt}{\sqrt{1-t^2}},$$

which, as we learn in calculus, is the inverse sine function. Abel argued that the function $f(x)$ defined by this integral was not as convenient to use as its inverse, $x = f^{-1}(u) = \sin u$. Likewise, he said, we should turn our attention from elliptic integrals to their inverses, which we now call *elliptic functions*.

Jacobi took this idea and ran with it [**12**]. He noticed that if $k = 0$ in the first kind of elliptic integral, we would simply get the inverse sine function. So for nonzero k, he defined the inverse of the first elliptic integral to be what he called the "sine amplitude" or sn u. Now, just as in regular trigonometry, where everything else can be built on the sine function, Jacobi went on to define further elliptic functions, such as the "cosine amplitude" function $\text{cn}\, u = \sqrt{1 - \text{sn}^2 u}$, and the "delta amplitude" function $\text{dn}\, u = \sqrt{1 - k^2 \text{sn}^2 u}$. He soon found that his new elliptic functions had many similar properties to the familiar trigonometric functions. For example, the regular sine function is periodic with period 2π, so that for any integer n, $\sin(x + 2\pi n) = \sin x$. But Jacobi's sine amplitude function was *doubly periodic;* in other words, there were two distinct numbers α and β (both complex, with $\alpha/\beta \notin \mathbb{R}$) such that

$$\text{sn}(u + m\alpha) = \text{sn}(u + n\beta) = \text{sn}\ u.$$

In 1835, Jacobi proved that no single-valued function that is either analytic or *meromorphic* (that is, analytic except possibly at locations called *poles*, where a denominator vanishes to finite order) could ever have more than two independent periods. In fact, the only such functions to have two such periods were the elliptic functions. By 1847, a young German prodigy by the name of Ferdinand Gotthold Eisenstein (1823–1852) had taken the innovative step of starting with the periods to define the elliptic functions via infinite series (see [**7**], [**21**]). From there, he proved a startling result that made a connection between elliptic functions and a particular kind of cubic curve, to whose history we now turn.

The pre-history of elliptic curves

Having traced the study of the ellipse—particularly its arc length—to what we now call elliptic functions, let's back up and trace another story which originated with the Greeks and helps us understand elliptic curves.

In our introduction, you learned that an elliptic curve is a curve of the form $y^2 = p(x)$, where $p(x)$ is a cubic polynomial with no repeated roots. Although such cubic curves were not studied in detail until the late 1600s, two different problems from the apparently unrelated area of number theory, both going back many centuries, mark the origin of questions involving these curves. (In what follows, we will call these cubic curves *elliptic curves*, although they did not receive this name until the early twentieth century.)

The first problem comes from Diophantus of Alexandria's *Arithmetica* [**11**], written some time during the third or fourth century CE. Problem 24 of Book IV reads as follows: “To divide a given number into two numbers such that their product is a cube minus its side.” If we call Diophantus’ given number a, the task is to find X and Y such that

$$Y(a-Y)=X^3-X.$$

Diophantus solved the problem for $a=6$ by substituting $X=kY-1$ and choosing the value $k=3$; this causes the resulting polynomial in Y to have only a cubic and quadratic term. Ignoring the double root $Y=0$, he obtained $Y=26/27$ and $X=17/9$. Therefore, the two numbers called for in the problem are $Y=26/27$ and $a-Y=136/27$ (since $26/27+136/27=6$), and the product of those two numbers is $(17/9)^3-(17/9)$. (For additional details, see [**4**, pp. 34–35].)

We note that Diophantus’ curve $Y(a-Y)=X^3-X$ is actually an elliptic curve in disguise, for the linear substitution $y=Y-a/2, x=-X$ leads to $y^2=x^3-x+(a/2)^2$. Now, it is important to stress that Diophantus had no concept of analytic geometry or modern algebraic notation, and certainly no idea about elliptic curves. Nevertheless, his work marked the beginning of a chain of inquiry that was to have wide-reaching and deep consequences many centuries later.

The second problem related to elliptic curves dates from certain Arabic manuscripts of roughly the eighth century, and Leonardo of Pisa, better known as Fibonacci (ca. 1175–1250), made it famous in Europe. He encountered the problem in question at the court of the Holy Roman Emperor Frederick II—namely, to find a rational number r such that both r^2-5 and r^2+5 are rational squares. Fibonacci found such a number, namely $r=41/6$: sure enough, $r^2-5=(31/6)^2, r^2=(41/6)^2$ and $r^2+5=(49/6)^2$ are indeed all squares. In his 1225 book *Liber quadratorum* (The Book of Squares) [**19**], Fibonacci called the positive integer n a *congruent number* if u^2-n, u^2 and u^2+n are all nonzero squares for some rational number u.

The connection with elliptic curves lies in the fact that if n is a congruent number, then the *product* of the three nonzero rational squares u^2-n, u^2 and u^2+n is also a nonzero rational square, say, v^2. In modern terminology, this implies that (u^2, v) is a point on the curve $E_n : y^2=x^3-n^2x=x(x-n)(x+n)$ with rational coordinates that are not both zero—a so-called nonzero *rational point*. Now for every positive integer n, $x(x-n)(x+n)$ is a cubic polynomial with distinct roots, which implies that E_n is an elliptic curve. Thus, if n is a congruent number, then the elliptic curve E_n contains a nonzero rational point. (For more information about congruent numbers, see the companion paper [**5**]; in addition, Koblitz uses congruent numbers as a unifying theme throughout his excellent book [**15**] on elliptic curves.)

Both of these ancient problems resurfaced in the early seventeenth century, when the French mathematician Claude-Gaspar Bachet de Meziriac (1581–1638) made a Latin translation of Diophantus’s *Arithmetica* and published it in 1621 [**3**]. This translation contained an appendix, which included Fibonacci’s congruent numbers problem, as well as some original results about Diophantine equations. One of the latter was the following theorem, which we give in modern notation. Fix an integer c and consider the equation $y^2=x^3+c$. If (x, y) is a solution to this equation with x and y both rational numbers, i.e., a *rational solution*, then

$$\left(\frac{x^4-8cx}{4y^2}, \frac{-x^6-20cx^3+8c^2}{8y^3}\right)$$

is also a rational solution.

The impact that Bachet’s translation of the *Arithmetica* had on the history of mathematics is a direct consequence of the fact that Pierre de Fermat (1601–1665) acquired a copy of it around 1630. Fermat’s chief mathematical contributions lie in his work in number theory: his introduction of the ideas of divisibility and primality gave the subject its definitive flavor, its most tantalizing question, and the direction it has taken for the better part of the last four centuries. Fermat's copy of Bachet's translation, which was reproduced and published by his son Samuel in 1670 [**9**], contained copious notes, annotations, and conjectures, including the famous Fermat Conjecture that if the integer n is greater than 2, then the equation $x^n + y^n = z^n$ has no integer solutions with $xyz \neq 0$—ultimately proved by Andrew Wiles and Richard Taylor in 1994.

Among Fermat's collected works we find several references to problems involving what we would now call elliptic curves, in particular, his conjecture that the only integers satisfying the equation $y^2 = x^3 - 2$ are $(x, y) = (3, \pm 5)$, and that the only integers satisfying $y^2 = x^3 - 4$ are $(x, y) = (2, \pm 2)$ and $(x, y) = (5, \pm 11)$.

Fermat's remarkable work might have gone unnoticed, except that during the 1730s, Leonhard Euler obtained a copy of Fermat's collected works. He was so struck by this body of mathematics that he proceeded to verify nearly all of Fermat's conjectures, including Fermat's statement about integer points on the curves $y^2 = x^3 - 2$ and $y^2 = x^3 - 4$. Euler expanded the scope of number theory far beyond Fermat's work, and his influence gave number theory its status as a legitimate field of mathematical inquiry [**8**]. Euler also did quite a bit of work on the congruent numbers problem and, as noted earlier, derived many results about elliptic integrals. The latter included formulas for adding these integrals, which provided a starting point for Legendre's work on the same subject in the 1780s.

In the meantime, during the 1670s, Newton used the recently developed tools of analytic geometry to try to classify cubic curves, in particular those of the form $y^2 = ax^3 + bx^2 + cx + d$ [**17**]. In doing so, he explained the mysteries behind both Diophantus' *Arithmetica* problem and Bachet's theorem about rational solutions to $y^2 = x^3 + c$. He pointed out that both Diophantus and Bachet were essentially intersecting a line with a cubic curve, and that in general, such an intersection consists of three points. However, if the line is tangent to the curve, then two of those three points are the same. FIGURE 3 tells the story.

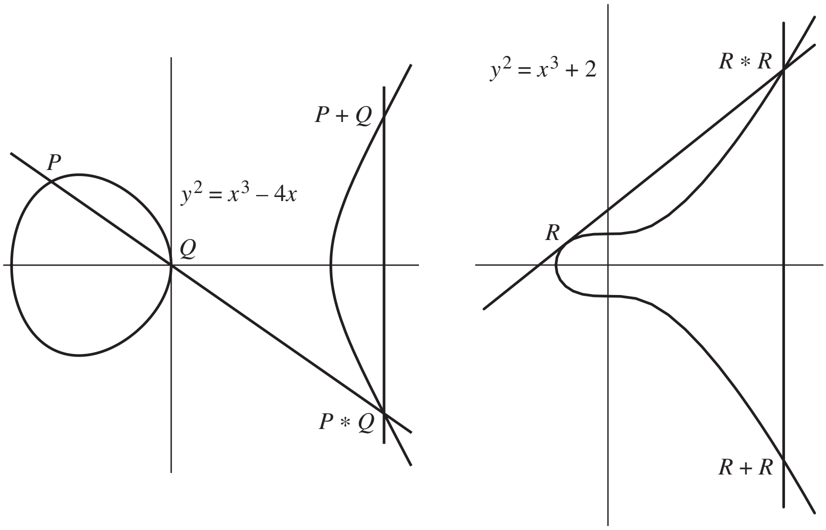

Figure 3 Chord and tangent addition

For the curve on the left, the line through P and Q intersects the curve in the third point $P * Q$; for the curve on the right, the tangent to the curve at R intersects the curve in the "third point" $R * R$. Using this star $*$ operation we can define an addition of the points on our curve. Namely, we define the sum $P + Q$ and $R + R$ to be the reflections of $P * Q$ and $R * R$, respectively, in the x-axis. This so-called chord and tangent addition gives the elliptic curve a group structure, which future researchers would find extremely useful—but that's another story, for which see [**5**].

Newton's insight ultimately led to general formulas for the addition of points on curves of the form $y^2 = ax^3 + bx^2 + cx + d$ (see [**14**, p. 10], for details), but it would first require the discovery of an amazing connection between such curves and elliptic functions. And it is precisely that connection which brings us back to the groundbreaking work of Gotthold Eisenstein in 1847.

From elliptic functions to elliptic curves

In order to appreciate Eisenstein's work, let's begin with an infinite series. Now it is probably not obvious, but it is true that

$$\sum_{m=-\infty}^{\infty} (z + m\pi)^{-2} = (\sin z)^{-2}.$$

(To begin to informally convince yourself of this, note that replacing z by $z + 2\pi$ on both sides of the equation leaves the relationship unchanged. For a formal proof, see [**1**, p. 11].) And since all trigonometry is ultimately based on the sine function, the whole subject could in theory be founded just as well on the above infinite series. Using this as his inspiration, Eisenstein constructed a new function out of a *doubly* infinite series, namely

$$\sum_{m,n=-\infty}^{\infty} (z + m\omega_1 + n\omega_2)^{-2}$$

where $\omega_1, \omega_2 \in \mathbb{C}$, and $\omega_1/\omega_2 \notin \mathbb{R}$. A bit of algebra reveals that this convergent series has two distinct periods, ω_1 and ω_2.

As previously noted, Jacobi had proved that the only single-valued meromorphic (i.e., analytic everywhere except for poles) functions with two linearly independent periods are the elliptic functions. Indeed, the modern definition of an elliptic function is a single-valued, meromorphic function f, defined on $\mathbb{C}$, for which there exist two distinct complex numbers ω_1 and ω_2 such that ω_1/ω_2 is not a real number and $f(z + \omega_1) = f(z + \omega_2) = f(z)$. And since Eisenstein's series-based function is defined over $\mathbb{C}$, is single-valued, meromorphic and doubly-periodic, it therefore has to be an elliptic function.

It was then that Eisenstein came up with a massively important result (see [**21**, pp. 22–24]). He proved that all elliptic functions of the form

$$y(z) = \sum_{m,n=-\infty}^{\infty} (z + m\omega_1 + n\omega_2)^{-2} - \sum_{\substack{m,n=-\infty \\ (m,n)\neq(0,0)}}^{\infty} (m\omega_1 + n\omega_2)^{-2}$$

must satisfy differential equations of the form

$$\left[y'(z)\right]^2 = p(y(z)),$$

where p is a cubic polynomial (depending on ω_1 and ω_2) with no repeated roots.

Does the phrase “a cubic polynomial with no repeated roots” ring a bell? If it does not, go back and re-read our introductory section: we’ll wait for you.

* * *

That’s right—we defined an elliptic curve to be a curve of the form $y^2 = p(x)$, where $p(x)$ is a cubic polynomial with no repeated roots. We thus see that Eisenstein’s work connects elliptic functions with elliptic curves.

A decade and a half later, in 1863, the famous analyst Karl Weierstrass (1815–1897) used this to define perhaps the most famous elliptic function of all, the Weierstrass $\wp$-function [**20**]:

$$\wp(z) = \wp(z;\omega_1,\omega_2) = z^{-2} + \sum_{\substack{m,n=-\infty \\ (m,n)\neq(0,0)}}^{\infty} \left[(z - m\omega_1 - n\omega_2)^{-2} - (m\omega_1 + n\omega_2)^{-2}\right].$$

As one would expect of an elliptic function, the $\wp$-function is doubly periodic with periods ω_1 and ω_2. But so are its derivative $\wp'(z)$, and its second derivative $\wp''(z)$, and so on. In fact all of its derivatives are elliptic functions with periods ω_1 and ω_2. Furthermore—and this is the amazing bit—*every* single elliptic function with periods ω_1 and ω_2 can be written as a rational function of $\wp(z)$ and $\wp'(z)$. (For a proof of this standard result, see [**2**, p. 189].) In other words, just as the sine function is the basis for all other trigonometric functions, so is the Weierstrass $\wp$-function the basis of all other elliptic functions.

By means of a clever argument using series, Weierstrass was able to show that the differential equation that his function satisfied was indeed a cubic, just as Eisenstein had proved, namely

$$\left[\wp'(z)\right]^2 = 4\wp^3(z) - g_2\wp(z) - g_3,$$

where g_2 and g_3 are special constants depending only on ω_1 and ω_2. It is therefore not hard to see that the point $(\wp(z), \wp'(z))$ lies on the cubic curve

$$y^2 = 4x^3 - g_2x - g_3.$$

Now in calculus, you learn about parametric equations and how they can describe a curve. By a *parameterization* of a curve C, we mean a continuous bijection from a set of numbers to the set of all points on C. For example, letting $x = a\sin t$ and $y = b\cos t$ with $t \in [0, 2\pi)$ gives a familiar parameterization of the standard ellipse $x^2/a^2 + y^2/b^2 = 1$. In the same way, we see that setting $x = \wp(z)$ and $y = \wp'(z)$ gives a parameterization of the cubic curve $y^2 = 4x^3 - g_2x - g_3$.

(You may wonder about the domain of the set of complex numbers z needed for the parameterization of the cubic. We’ll get to that later. We also note that in order to rigorously prove parameterization, one must show the existence of a continuous, bijective map. Those interested in the details should consult [**15**, pp. 22–26].)

Now, cubic curves of the form $y^2 = ax^3 + bx^2 + cx + d$ had been well known for years. Indeed, as we mentioned earlier, Isaac Newton carried out a major study of them in the 1670s. But it wasn’t until 1834 that Jacobi pointed out a possible connection between cubic curves and elliptic functions [**13**], followed by Eisenstein’s proof of such a relationship in 1847. Then in 1864, a German mathematician by the name of Alfred Clebsch (1833–1872) introduced the idea above of using elliptic functions to parameterize cubic curves [**6**], and Weierstrass linked a clever addition formula for elliptic functions to the addition of points on these cubic curves. Finally, in a landmark paper of 1901 [**18**], Henri Poincaré (1854–1912) tied all these ideas together, effectively marking the birth of a new area of study. And because they require elliptic functions for their parameterization, these curves became known as elliptic curves.

Why ellipses are not elliptic curves

We have thus seen the historical path that led from the ellipse, first of all to elliptic integrals (one of which expresses the arc length of an ellipse), then to elliptic functions (obtained by inverting an elliptic integral), and finally to elliptic curves (which require elliptic functions for their parameterization). All of this leads to the question we posed at the beginning: why are ellipses not elliptic curves? The answer lies firstly in extending the domain of both curves from the reals to the complex numbers, and secondly in the matter of their respective parameterizations. We have already mentioned that ellipses may be parameterized by trigonometric functions, and this holds as much for ellipses in $\mathbb{C}^2$ as it does for those in $\mathbb{R}^2$. But in $\mathbb{C}^2$ such curves are best described, not as curves at all, but as *surfaces*. In particular, curves parameterizable by singly-periodic complex-valued functions are topologically equivalent to spheres. Here is one way to see this.

As functions in $\mathbb{R}^2$, the single periodicity of the sine and its derivative, cosine, means that their domain is $\mathbb{R}$ mod $2\pi\mathbb{Z}$. Now on the real line, $2\pi\mathbb{Z}$ is a one-dimensional lattice, so geometrically, $\mathbb{R}$ mod $2\pi\mathbb{Z}$ is a circle. If we change the domain to $\mathbb{C}$ and map to $\mathbb{C}^* \times \mathbb{C}^*$, where $\mathbb{C}^* = \mathbb{C} \cup \{\infty\}$, here's what happens. Since the ellipse is parameterized by $\sin z$ and $\cos z$, its (complex) domain is divided into infinitely long vertical strips with real width 2π, as in the following figure, where on the left we see the plane with the vertical lines $\mathrm{Re}(z) = \pi$ and $\mathrm{Re}(z) = -\pi$ drawn. The periodicity of sine and cosine means that every distinct point in the complex plane corresponds to a distinct point in this strip; thus the ellipse, since it is parameterized by these two functions, is completely described by how they map the points in this strip. Since we are mapping to a compact set, we can identify all points in the strip such that $\mathrm{Im}(z) = \pm\infty$ with a single point, ∞. If we further identify the two lines $\mathrm{Re}(z) = \pi$ and $\mathrm{Re}(z) = -\pi$, this transforms the strip into a surface with ∞ represented by a point at the top, the origin by a point at the bottom, and a meridian line, corresponding to the identified vertical lines, joining the two points along the surface. As we see on the right, this resulting geometric figure is topologically equivalent to a sphere.

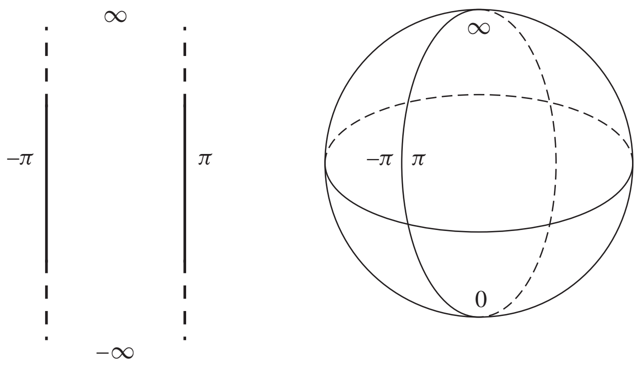

Figure 4 The complex co-domain of an ellipse is a sphere

However, just as the elliptic integral representing the arc length of an ellipse cannot be evaluated using regular calculus techniques, elliptic curves cannot be parameterized by elementary functions. The simplest functions that will successfully parameterize elliptic curves are elliptic functions, and it is this parameterization that is the key to understanding why ellipses are not elliptic curves.

Suppose we have an elliptic curve

$$y^2 = ax^3 + bx^2 + cx + d.$$

Then the ordered pairs (x, y) that work in this equation can be written as $(f(z), f'(z))$, where $f(z)$ is an elliptic function with periods ω_1 and ω_2. Since $f(z)$ is elliptic, it is a rational function of $\wp(z)$ and $\wp'(z)$, as is its derivative $f'(z)$, which is also an elliptic function with the same periods. Not only that, but it can be shown that the correspondence $(x, y) \leftrightarrow (f(z), f'(z))$ between all points (x, y) on the elliptic curve and complex numbers $z \in \mathbb{C}/\Lambda$, where

$$\Lambda = \{m\omega_1 + n\omega_2 : m, n \in \mathbb{Z}\},$$

is one-to-one and onto; see, for example, [**15**, pp. 22–26]. In other words, the functions $x = f(z)$ and $y = f'(z)$ parameterize the elliptic curve, and the fact that f and f' have periods ω_1 and ω_2 means that there is a one-to-one correspondence between all points on the curve and the equivalence classes

$$z + \Lambda = \{z + m\omega_1 + n\omega_2 : m, n \in \mathbb{Z}\}.$$

The double-periodicity of f and f' thus partitions the complex plane into a lattice that looks something like FIGURE 5.

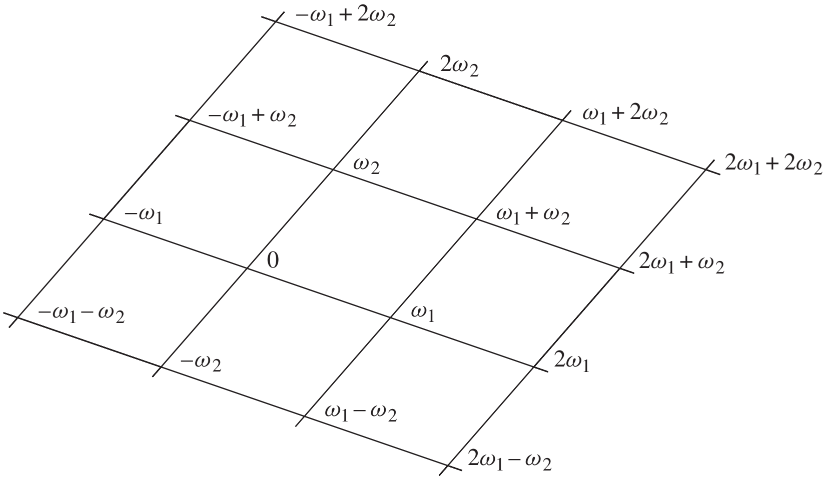

Figure 5 The lattice Λ generated by ω_1 and ω_2

Elliptic curves, which require elliptic functions for their parameterization, are thus isomorphic to the space $\mathbb{C}/\Lambda$ of these equivalence classes. On the other hand, as we have seen, ellipses, since they can be parameterized by elementary trigonometric functions, are isomorphic to the sphere. But $\mathbb{C}/\Lambda$ is not topologically equivalent to a sphere at all. Instead, it is a torus! To see this, look at FIGURE 6. The top left shows one of the infinitely many parallelograms which make up the lattice in FIGURE 5. (Since the parallelograms are all congruent to each other, it doesn't really matter which one we choose.) The top right shows this parallelogram rolled up into a cylinder, so that edge C meets edge A. Imagine that this resulting cylinder is made of material flexible enough

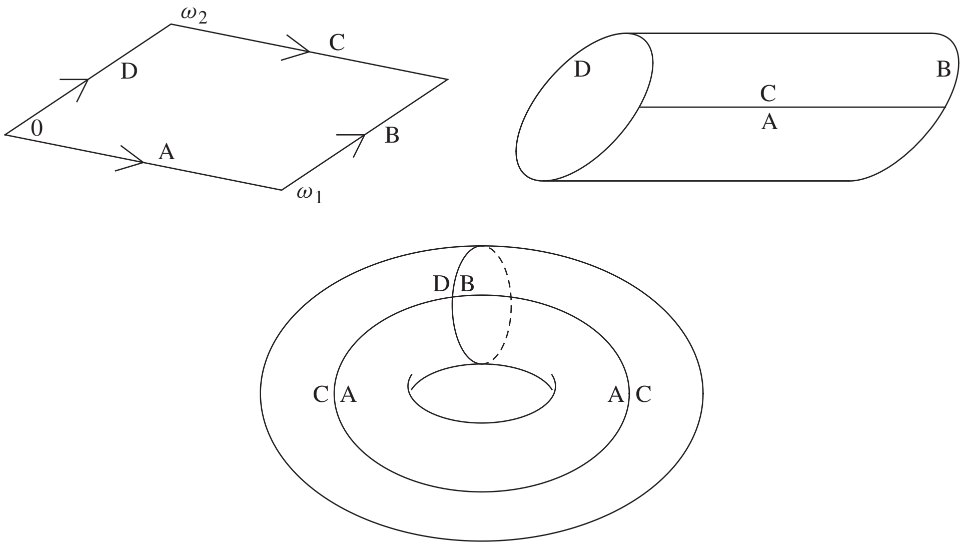

Figure 6 $\mathbb{C}/\Lambda$ is a torus

to be bent round so that edge B and edge D can be joined together, as in the bottom figure. It is clear that the resulting surface formed by this rolled-up parallelogram is a torus. Thus, since every point on an elliptic curve can be mapped to one of these "periodic parallelograms," which in turn can be transformed into tori, every elliptic curve is topologically equivalent to a torus.

Now, a torus is a surface that is completely incapable of being (legitimately) transformed into a sphere, meaning that no curve isomorphic to a sphere could possibly belong to a set of objects isomorphic to $\mathbb{C}/\Lambda$. This gives a visually very obvious—and mathematically, very profound—reason why elliptic curves are not parameterizable by any elementary functions.

It also tells us why ellipses are not (and never could be) elliptic curves!

REFERENCES

1. G. E. Andrews, R. Askey, and R. Roy, *Special Functions*, Cambridge University Press, 2000.
2. J. V. Armitage and W. F. Eberlein, *Elliptic Functions*, Cambridge University Press, 2006.
3. C.-G. Bachet, *Diophanti Alexandrini Arithmeticorum*, Sebastiani Cramoisy, Paris, 1621.
4. I. G. Bashmakova, *Diophantus and Diophantine Equations*, Mathematical Association of America, 1997.
5. E. Brown, Three Fermat trails to elliptic curves, *College Math. J.* **31** (May, 2000) 162–172. http://dx.doi.org/10.2307/2687483
6. A. Clebsch, Über einen Satz von Steiner und einige Punkte der Theorie der Curven dritter Ordnung, *J. für die reine und angewandte Mathematik* **63** (1864) 94–121. http://dx.doi.org/10.1515/crll.1864.63.94
7. F. G. Eisenstein, Beiträge zur Theorie der elliptischen Funktionen, *J. für die reine und angewandte Mathematik* **35** (1847) 137–274. http://dx.doi.org/10.1515/crll.1847.35.137
8. L. Euler, *Leonhardi Euleri Opera Omnia*, Ser. 1, vol. 2–5, B. G. Teubner, Leipzig and Berlin, 1911–13.
9. P. Fermat, *Oeuvres*, vol. 1. Gauthier-Villars, Paris, 1891–1896.
10. M. N. Fried and S. Unguru, *Apollonius of Perga's Conica: Text, Context, Subtext*, Brill, Leiden, 2001.
11. T. L. Heath, *Diophantus of Alexandria*, Cambridge University Press, 1910.
12. C. G. J. Jacobi, *Fundamenta nova functionarum ellipticarum*, Borntraeger, 1829.
13. ———, De usu theoriae integralium ellipticorum et integralium abelianorum in analysi diophantea, *Werke* **2** (1834) 53–55.
14. A. W. Knapp, *Elliptic Curves*, Princeton University Press, 1992.
15. N. Koblitz, *Introduction to Elliptic Curves and Modular Forms,* Springer-Verlag, 1984.
16. A.-M. Legendre, *Traité des fonctions elliptiques*, 3 vols., Huzard-Courcier, Paris, 1825–28.

17. I. Newton, De resolutione quaestionum circa numeros, *Mathematical Papers* **4** (1670s) 110–115.
18. H. Poincaré, Sur les propriétés arithmétiques des courbes algébriques, *Journal de mathématiques pures et appliquées* (series 5) **7** (1901) 161–233.
19. L. E. Sigler (transl.), *Fibonacci's Liber Quadratorum*, Academic Press, 1987.
20. K. Weierstrass, Vorlesungen über die Theorie der elliptischen Funktionen, *Mathematische Werke*, vol. 5, 1863.
21. A. Weil, *Elliptic Functions According to Eisenstein and Kronecker*, Springer-Verlag, 1976.

Summary Elliptic curves are a fascinating area of algebraic geometry with important connections to number theory, topology, and complex analysis. As their current ubiquity in mathematics suggests, elliptic curves have a long and fascinating history stretching back many centuries. This paper presents a survey of key points in their development, via elliptic integrals and functions, closing with an explanation of why no elliptically-shaped planar curved line may ever be called an elliptic curve.

ADRIAN RICE is Professor of Mathematics at Randolph-Macon College in Ashland, Virginia. His research specialty is the history of mathematics, focusing on nineteenth- and early twentieth-century mathematics in particular. His most recent book, *Mathematics in Victorian Britain*, co-edited with Raymond Flood and Robin Wilson, was published by Oxford University Press in 2011. In his spare time, he enjoys music, travel, and spending time with his wife and three-year-old son.

EZRA (BUD) BROWN grew up in New Orleans and has degrees from Rice and LSU. Since 1969, he has been at Virginia Tech in Blacksburg, Virginia, where he is currently Alumni Distinguished Professor of Mathematics. He does research in number theory and combinatorics, and his book, *Biscuits of Number Theory*, co-edited with Art Benjamin, was published by the MAA in 2009. He plays piano jazz, has been in six operas, goes kayaking with his wife, and occasionally bakes biscuits for his students.