BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
S 0273-0979(07)01153-6
Article electronically published on April 9, 2007

A NORMAL FORM FOR ELLIPTIC CURVES

HAROLD M. EDWARDS

Abstract. The normal form $x^2+y^2=a^2+a^2x^2y^2$ for elliptic curves simplifies formulas in the theory of elliptic curves and functions. Its principal advantage is that it allows the *addition law,* the group law on the elliptic curve, to be stated explicitly

$$X=\frac{1}{a}\cdot\frac{xy'+x'y}{1+xyx'y'},\quad Y=\frac{1}{a}\cdot\frac{yy'-xx'}{1-xyx'y'}.$$

The j-invariant of an elliptic curve determines 24 values of a for which the curve is equivalent to $x^2+y^2=a^2+a^2x^2y^2$, namely, the roots of $(x^8+14x^4+1)^3-\frac{j}{16}(x^5-x)^4$. The symmetry in x and y implies that the two transcendental functions $x(t)$ and $y(t)$ that parameterize $x^2+y^2=a^2+a^2x^2y^2$ in a natural way are essentially the same function, just as the parameterizing functions $\sin t$ and $\cos t$ of the circle are essentially the same function. Such a parameterizing function is given explicitly by a quotient of two simple theta series depending on a parameter τ in the upper half plane.

Part I. The Addition Formula

1. Why elliptic functions?

The double periodicity of elliptic functions, the property by which they are often defined today, is not what attracted attention to them in the first place. Abel, whose own contribution to the study of elliptic functions was enormous, began his long memoir [1] about them with the observation that "the first idea of [elliptic] functions was given by the immortal Euler, when he demonstrated that the equation with variables separated

$$\frac{dx}{\sqrt{\alpha+\beta x+\gamma x^2+\delta x^3+\epsilon x^4}}+\frac{dy}{\sqrt{\alpha+\beta y+\gamma y^2+\delta y^3+\epsilon y^4}}=0 \tag{1.1}$$

can be integrated algebraically." Thus, at least in Abel's opinion, the original motive of the theory was the generalization to fourth degree polynomials $f(x)$ of the integration of differentials $\frac{dx}{\sqrt{f(x)}}$ for second degree polynomials $f(x)$, which had led to some of the most important transcendental functions studied by early calculus. Abel wanted to extend the repertory of transcendental functions available to mathematics.

Today the theory is viewed more in terms of elliptic *curves* than elliptic *functions,* and interest centers on their *group structures.* When an elliptic curve is realized as a *cubic* curve and when a point of the curve is chosen to serve as the identity of

Received by the editors December 27, 2005.
2000 *Mathematics Subject Classification.* Primary 54C40, 14E20; Secondary 46E25, 20C20.
Key words and phrases. Elliptic curves, elliptic functions, Riemann surfaces of genus one.

the group operation, the group structure can be described in terms of the sets of three points in which lines intersect the curve, a description that is now well known and widely taught. The connection between this now-familiar group structure and Euler's "algebraic integration" that inspired Abel is far from obvious, but in fact the two are aspects of the same phenomenon.

A third approach to that phenomenon was developed by Abel in the introduction to his Paris memoir [2], where he sketched a broad generalization of the group construction. Instead of intersecting a cubic curve with lines, he intersected an *arbitrary* curve with an *arbitrary* family of auxiliary curves. As the parameters in the defining equation of the auxiliary curve vary, the intersection points vary along the given curve. (See Section 10 for a fuller discussion.) Abel discovered that, under suitable conditions, N intersection points move in this way with $N-g$ degrees of freedom, where g depends only on the given curve, not on N or on the family of auxiliary curves that is used, provided the family is sufficiently general. This g is—again under suitable conditions—the genus of the given curve. When that curve is a nonsingular cubic and the auxiliary curves are lines, there are $N=3$ intersection points that move with $N-g=2$ degrees of freedom because two of the intersection points can be chosen arbitrarily. Therefore, g is 1 in this case. It was surely this interpretation of the addition on a nonsingular cubic and its generalization to elliptic curves presented in other ways, such as the curve $y^2 = 1-x^4$, that motivated Abel's more general construction.

The g constraints on the motion of N points along the given curve are expressed in the general case by g linearly independent differential equations determined by the *holomorphic differentials* on the curve (see [4], Essay 4.6). For an elliptic curve of the form $z^2 = \alpha + \beta x + \gamma x^2 + \delta x^3 + \epsilon x^4$, the holomorphic differentials are simply the constant multiples of $\frac{dx}{\sqrt{\alpha+\beta x+\gamma x^2+\delta x^3+\epsilon x^4}}$, the differential that appears in the equation (1.1) that Euler "integrated algebraically." In this way, Abel's construction connects Euler's integration to the addition operation.

This paper presents a fourth view of the above phenomenon that incorporates the three that have been mentioned: It integrates (1.1) (see Section 9 below), it expresses the addition operation in explicit algebraic form (Section 8), and it constructs the $(N-1)$-dimensional families of motions of sets of N points on an elliptic curve that are described by Abel's construction (Section 10).

Elliptic *functions* are meromorphic functions that parameterize elliptic *curves.* The classic book of Hurwitz and Courant [7] presents the theory of elliptic functions in two ways, giving both the Jacobi notation (which is similar to Abel's but more fully developed) and the Weierstrass notation (which is favored by most modern treatments). The simple explicit form of the addition law developed in Part II leads to a normalization of elliptic functions in Part III that seems preferable to both of these. In this normalization, the parameterizing functions, like the parameterizing functions $t \mapsto (\sin t, \cos t)$ of the circle, are essentially the same function; specifically, $x^2 + y^2 = a^2 + a^2x^2y^2$ has the parameterization $t \mapsto (\psi(t-\frac{1}{2}), \psi(t))$ where $\psi(t)$ is the elliptic function given by formula (15.1) for a complex number τ in the upper half plane that is found by solving a certain transcendental equation (Section 22).

2. The addition formula for $x^2 + y^2 + x^2y^2 = 1$

Euler's very first paper [5] on the theory of elliptic functions contains formulas that strongly suggest[1] an explicit "addition formula" in the special case of the elliptic curve $x^2 + y^2 + x^2y^2 = 1$. This curve, which becomes $z^2 = 1 - x^4$ when one sets $z = y(1 + x^2)$, was of great interest to Gauss; the last entry of his famous *Tagebuch* relates to it, as does his reference to "the transcendental functions which depend on the integral $\int \frac{dx}{\sqrt{1-x^4}}$" in Article 335 of the *Disquisitiones Arithmeticae.* In notes published posthumously in his *Werke* [6], Gauss stated explicitly the formulas Euler had hinted at decades earlier, putting them in the form

$$S = \frac{sc' + s'c}{1 - ss'cc'}, \qquad C = \frac{cc' - ss'}{1 + ss'cc'}. \tag{2.1}$$

Gauss's choice of the letters s and c brings out the analogy with the addition laws for sines and cosines. (The numerators *are* the addition laws for sines and cosines.) He in fact defines two transcendental functions $s(t)$ and $c(t)$ with the property that (2.1) expresses $(S, C) = (s(t + t'), c(t + t'))$ in terms of $(s, c) = (s(t), c(t))$ and $(s', c') = (s(t'), c(t'))$. The definition of $s(t)$ takes the implicit form $t = \int_0^{s(t)} \frac{dx}{\sqrt{1-x^4}}$ analogous to $t = \int_0^{\sin t} \frac{dx}{\sqrt{1-x^2}}$, while $c(t) = \sqrt{\frac{1-s(t)^2}{1+s(t)^2}}$ (with $c(0) = 1$) is analogous to $\cos t = \sqrt{1 - \sin^2 t}$ (with $\cos 0 = 1$).

These remarkable Euler-Gauss formulas apply only to the specific curve $s^2 + c^2 + s^2c^2 = 1$, but they are a special case of a formula that describes the group law of an arbitrary elliptic curve.

3. An addition formula in the general case

Elliptic curves and elliptic functions have been studied ever since Euler's time, and that study has often been intense, as it is now. In view of this long history, it seems unlikely that anything fundamentally new remains to be discovered in the most elementary parts of the theory. Nonetheless, I have not been able to find the following generalization of (2.1) in the literature.[2] If it is not new, it is certainly not as well known as it deserves to be.

Theorem 3.1. *If a is a constant for which $a^5 \neq a$, the formulas*

$$X = \frac{1}{a} \cdot \frac{xy' + yx'}{1 + xyx'y'}, \quad Y = \frac{1}{a} \cdot \frac{yy' - xx'}{1 - xyx'y'} \tag{3.1}$$

describe the addition formula for the elliptic curve $x^2 + y^2 = a^2 + a^2x^2y^2$.

As will be shown in Section 5, every elliptic curve is equivalent—in an appropriate sense—to one in this form $x^2 + y^2 = a^2 + a^2x^2y^2$, so (3.1) can be used to describe explicitly the addition law on any elliptic curve.

Formula (2.1) is the case $a = \sqrt{i}$, $x = \sqrt{i} \cdot s$ and $y = \sqrt{i} \cdot c$ of (3.1).

In Section 4, an elliptic curve is defined to be one of the form $z^2 = f(x)$ in which $f(x)$ is a polynomial of degree 3 or 4 with distinct roots. Setting $z = y(1 - a^2x^2)$ puts $x^2 + y^2 = a^2 + a^2x^2y^2$ in the form $z^2 = (a^2 - x^2)(1 - a^2x^2)$; when $a \neq 0$,

[1] See especially his Theorem 6 and its Corollary 3.

[2] Abel's form of the addition law is in his formula (10) of [1]. It serves the same purpose as (3.1) but is less simple and less symmetrical. Jacobi's form of the addition law is similar to Abel's; see §18 of [9]. The addition formula for the Weierstrass $\mathfrak{P}$-function is more complicated. See [7], II, 1, §8. Also, see [10], III, 4; and [13], I, 4.

the polynomial $a^2x^4-(a^4+1)x^2+a^2$ on the right has degree 4, so the equation describes an elliptic curve provided this polynomial has distinct roots, which is true if and only if $(a^4+1)^2-4a^4=(a^4-1)^2$ is nonzero. In short, $a^5\neq a$ if and only if $x^2+y^2=a^2+a^2x^2y^2$ is an elliptic curve.

As in Gauss's case, (3.1) can be regarded as expressing $(X,Y)=(x(t+t'), y(t+t'))$ in terms of $(x,y)=(x(t),y(t))$ and $(x',y')=(x(t'),y(t'))$, where $x(t)$ and $y(t)$ are the transcendental functions defined by $t=\int_0^{x(t)}\frac{dx}{\sqrt{(a^2-x^2)(1-a^2x^2)}}$ and $y(t)=\sqrt{\frac{a^2-x(t)^2}{1-a^2x(t)^2}}$ (with $y(0)=a$). However, the theorem is purely algebraic, and the proof does not need transcendental functions or even complex numbers, as is shown by the proof in Section 8.

Part II. The Algebraic Theory of Elliptic Curves

4. The field of rational functions on a curve

The key to dealing with algebraic curves over a ground field that is not algebraically closed is to abandon the notion of *points of* a curve and to work instead with *rational functions on* the curve. These rational functions form a field, the algebraic properties of which describe geometric properties of the curve in many cases. Two curves are *birationally equivalent* if their fields of rational functions are isomorphic.

This paper deals only with *elliptic curves,* which will be defined[3] to be curves that can be presented in the form $z^2=f(x)$, where $f(x)$ is a polynomial of degree 3 or 4 with distinct roots that has coefficients in an algebraic number field. Thus, an elliptic curve is presented by giving (1) an algebraic number field K and (2) a polynomial $f(x)$ of degree 3 or 4 with coefficients in K that is relatively prime to its derivative.

The field of rational functions on $z^2=f(x)$ is, very concretely, the field whose elements are represented by expressions of the form $r(x)+s(x)z$, where $r(x)$ and $s(x)$ are rational functions of x with coefficients in K (i.e., quotients of polynomials in x with coefficients in K in which the denominator is not zero), added in the obvious way and multiplied by multiplying in the usual way and using the relation $z^2=f(x)$ to eliminate z^2. Such a field is *an elliptic function field.*

The field of rational functions on the curve $x^2+y^2=a^2+a^2x^2y^2$ is not an "elliptic function field" in this sense, but it becomes one when $z=y(1-a^2x^2)$ is used to put it in the form $z^2=(a^2-x^2)(1-a^2x^2)$ as above. Since $y=\frac{z}{1-a^2x^2}$, there is a birational equivalence between $x^2+y^2=a^2+a^2x^2y^2$ and $z^2=(a^2-x^2)(1-a^2x^2)$ in the sense that x and y can be expressed rationally in terms of x and z and conversely. In this way, the equation $x^2+y^2=a^2+a^2x^2y^2$ of Theorem 3.1 determines an elliptic function field.

[3] It would be more satisfying to define an elliptic curve to be a curve of genus one in the sense that the genus of a curve is defined in [4], but the justification of this definition would be too great a digression. That definition is, however, equivalent the one used here, as can be seen by applying the Riemann-Roch theorem—see [4]—to conclude that a function field of genus one contains an element with just two simple poles. It follows first that the field can be defined by a relation $\chi(x,y)=0$ whose degree in y is 2 and then (complete the square) that the relation can be put in the form $z^2=f(x)$ where $f(x)$ is a polynomial with distinct roots. Finally, a curve defined by a relation of this form has genus one if and only if $\deg f$ is 3 or 4, as is shown in [4] (Example 6 of Essay 4.5).

An element of an elliptic function field is called a *constant* if it is a root of a polynomial with integer coefficients. As is easily seen, an element $r(x) + s(x)z$ is constant if and only if $r(x)$ is an element of K and $s(x)$ is zero. (In the terminology of [4], 1 and z are a normal basis of the function field.) In short, the field K, viewed in the obvious way as a subfield of the function field, is the field of constants.[4]

Extending K—adjoining constants—changes the function field determined by $z^2 = f(x)$, so it does not give a birationally equivalent "curve", even though the field of functions corresponds in some sense to the same curve. For this reason, an elliptic function field will be called *equivalent* to a field that is obtained from it by extending the underlying field of constants K, or, more generally, when there is a third field to which they are both equivalent in this sense. In short, elliptic function fields are *equivalent* if they can be made isomorphic by adjoining enough constants.

The fact that the ground field K is not algebraically closed means that geometrical constructions often require an extension of the field of constants—for example to realize the zeros and poles of a given rational function as *points.* When function fields that are equivalent in the sense just defined are regarded as describing the same curve, one can *adjoin constants as needed* without changing the "geometry".

5. The normal form

Proposition 5.1. *An elliptic function field is equivalent, in the sense of the last section, to the field of rational functions on $x^2 + y^2 = a^2 + a^2x^2y^2$ for some a.*

Proof. Let K be an algebraic number field, and let $f(x)$ be a polynomial of degree 4 with coefficients in K that has distinct roots. Let constants be adjoined to K, if necessary, to split $f(x)$ into linear factors, say $f(x) = c(x - \alpha_1)(x - \alpha_2)(x - \alpha_3) \times(x - \alpha_4)$, and let $\sqrt{c}$ be adjoined, if necessary, to put the defining equation $z^2 = f(x)$ in the form $v^2 = (x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)$ where $v = \frac{z}{\sqrt{c}}$. Thus, if the $f(x)$ in the defining relation has degree 4, one can assume without loss of generality that it is $(x - \alpha_1)(x - \alpha_2)(x - \alpha_3)(x - \alpha_4)$ for distinct elements α_1, α_2, α_3, α_4 of an algebraic number field K.

There is a simple condition under which two such elliptic function fields, say the one defined by $z^2 = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)(x - \alpha_4)$ and the one defined by $v^2 = (u - \beta_1)(u - \beta_2)(u - \beta_3)(u - \beta_4)$, are equivalent, namely, the condition that there be a fractional linear transformation $x \mapsto \frac{Ax+B}{Cx+D}$ that carries $\alpha_i \mapsto \beta_i$ for $i = 1$, 2, 3, and 4. (Here, of course, A, B, C, and D are constants for which $AD \neq BC$.) This sufficient condition can be derived in the following way. By straightforward computation, $u - \beta_i = \frac{(AD-BC)(x-\alpha_i)}{(Cx+D)(C\alpha_i+D)}$ for each i when $u = \frac{Ax+B}{Cx+D}$. The product of these four formulas shows that $(u - \beta_1)(u - \beta_2)(u - \beta_3)(u - \beta_4)$ is a constant times $\frac{(x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)}{(Cx+D)^4}$. Thus $z^2 = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)(x - \alpha_4)$ implies that $(u-\beta_1)(u-\beta_2)(u-\beta_3)(u-\beta_4)$ is a constant times the square of $\frac{z}{(Cx+D)^2}$. When the square root of the constant is adjoined, if necessary, one finds a birational change of variables between (z, x) and (v, u) under which $z^2 = (x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)$ corresponds to $v^2 = (u - \beta_1)(u - \beta_2)(u - \beta_3)(u - \beta_4)$, as was to be shown.

In particular, the fractional linear transformation

$$x \mapsto \frac{(\alpha_4 - \alpha_2)(x - \alpha_3)}{(\alpha_2 + \alpha_4)(x + \alpha_3) - 2\alpha_3 x - 2\alpha_2\alpha_4}$$

[4] Chevalley emphasizes the importance of the field of constants on page 1 of [3].

carries $\alpha_2 \mapsto -1$, $\alpha_3 \mapsto 0$, and $\alpha_4 \mapsto 1$, so it shows that $z^2 = (x-\alpha_1)(x-\alpha_2) \times(x-\alpha_3)(x-\alpha_4)$ and $v^2 = (u-\phi)\cdot(u+1)\cdot u\cdot(u-1)$ define equivalent function fields when

$$\begin{aligned}\phi &= \frac{(\alpha_4-\alpha_2)(\alpha_1-\alpha_3)}{(\alpha_2+\alpha_4)(\alpha_1+\alpha_3)-2\alpha_3\alpha_1-2\alpha_2\alpha_4} \\ &= \frac{\alpha_1\alpha_4+\alpha_2\alpha_3-\alpha_1\alpha_2-\alpha_3\alpha_4}{\alpha_1\alpha_2+\alpha_2\alpha_3+\alpha_3\alpha_4+\alpha_4\alpha_1-2\alpha_1\alpha_3-2\alpha_2\alpha_4}.\end{aligned} \tag{5.1}$$

The defining relation of the elliptic function field determined by $x^2+y^2 = a^2 + a^2x^2y^2$ can be written $(\frac{z}{a})^2 = (x-a)(x-\frac{1}{a})(x+a)(x+\frac{1}{a})$, from which it follows that the field is equivalent to the one defined by $v^2 = (u-\phi)\cdot(u+1)\cdot u\cdot(u-1)$ when $\phi = \frac{-1-1-1-1}{1-1+1-1+2a^2+2a^{-2}} = -\frac{2}{a^2+a^{-2}}$.

Thus, a given relation $z^2 = (x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)$ can be transformed first to $v^2 = (u-\phi)\cdot(u+1)\cdot u\cdot(u-1)$ and then to $x^2+y^2 = a^2+a^2x^2y^2$ when a is a constant for which $\phi = -\frac{2}{a^2+a^{-2}}$—i.e., when a is a solution of $a^4 + \frac{2}{\phi}\cdot a^2 + 1 = 0$—which completes the construction when $f(x)$ has degree 4.

When $f(x)$ has degree 3 and distinct roots, it can be replaced by $f(x+c)$, if necessary, to make its constant term nonzero. Then division by x^4 puts the equation $z^2 = f(x)$ in the form $(\frac{z}{x^2})^2 = f_1(\frac{1}{x})$ where f_1 is a polynomial of degree 4 with distinct roots and the construction can proceed as before. □

6. The J-invariant

Since every elliptic function field is equivalent to one of the form $x^2+y^2 = a^2+a^2x^2y^2$, the problem of determining whether two elliptic function fields are equivalent reduces to the problem of determining whether two in this normal form are equivalent. A sufficient condition is:

Proposition 6.1. *The elliptic function field determined by $x^2+y^2 = a^2+a^2x^2y^2$ is equivalent to the one determined by $x^2+y^2 = b^2+b^2x^2y^2$ whenever b has one of the 24 values*

$$i^\epsilon a, \quad \frac{i^\epsilon}{a}, \quad i^\epsilon\cdot\frac{a-1}{a+1}, \quad i^\epsilon\cdot\frac{a+1}{a-1}, \quad i^\epsilon\cdot\frac{a-i}{a+i}, \quad i^\epsilon\cdot\frac{a+i}{a-i} \tag{6.1}$$

where i is a square root of -1 that is to be adjoined, if necessary, to the field of constants, and where ϵ is 0, 1, 2, or 3.

Proof. The values listed in (6.1) are the orbit of a under the group of fractional linear transformations of the Riemann sphere generated by the two transformations $a \mapsto ia$ and $a \mapsto \frac{a-1}{a+1}$. That this group is isomorphic to the group of the cube becomes clear when one observes that the six points 0, ±1, $\pm i$, ∞ of the Riemann sphere (these are the values that a is not permitted to have) are permuted by the group in the same way that the faces of a cube are permuted by the motions of the cube. Specifically, $a \mapsto ia$ permutes $1 \mapsto i \mapsto -1 \mapsto -i \mapsto 1$ cyclically while leaving 0 and ∞ fixed, and $a \mapsto \frac{a-1}{a+1}$ permutes $1 \mapsto 0 \mapsto -1 \mapsto \infty \mapsto 1$ cyclically while leaving i and $-i$ fixed, which is the way that the group of a cube permutes the faces when the pairs ±1, $\pm i$, and the pair $(0,\infty)$ label the three pairs of opposite faces. Therefore, not only does the group contain just 24 elements, but the orbit of any a under the action of the group contains 24 *distinct* elements—those listed in (6.1)—except that the values of a that correspond to the vertices of the cube constitute an orbit that contains only 8 distinct values of a, and the values of a

that correspond to the midpoints of the edges of the cube constitute an orbit that contains only 12 distinct values.

(The forbidden values of a correspond to the centers of the faces and constitute an orbit that contains only 6 distinct values. The 12 point orbit is the orbit of $\sqrt{i}$ because this value is invariant under the group element $a \mapsto \frac{i}{a}$ which interchanges 1 and i, -1 and $-i$, and 0 and ∞ and therefore leaves invariant the midpoints of two edges: the one between 1 and i and the one between -1 and $-i$. The 8 point orbit is the orbit of $(1+i)\cdot\frac{\sqrt{3}-1}{2}$ because this number is invariant under $a \mapsto \frac{1+ia}{1-ia}$ which cyclically permutes 0, 1, and i, as well as ∞, -1 and $-i$, and which therefore leaves invariant two vertices, the ones common to the two sets of three faces that are permuted cyclically.)

The proposition will therefore be proved if the function field of $x^2+y^2 = a^2 + a^2x^2y^2$ is shown to be equivalent to the two function fields obtained by replacing a with ia and with $\frac{a-1}{a+1}$. For this, it will suffice to show that if $b = ia$ or $b = \frac{a-1}{a+1}$, then there is a fractional linear transformation that carries the set $(a, -a, \frac{1}{a}, -\frac{1}{a})$ to the set $(b, -b, \frac{1}{b}, -\frac{1}{b})$, which is true because $x \mapsto ix$ is such a fractional linear transformation in the first case, and $x \mapsto \frac{x-1}{x+1}$ is in the second. □

The polynomial $\mathcal{K}(x) = \prod_{\iota=1}^{24}(x - a_\iota)$, where a_ι ranges over the 24 values listed in (6.1), is a polynomial of degree 24 in x with coefficients that are rational functions of a. One can find by computation that $\mathcal{K}(x) = (x^8+14x^4+1)^3 - C(a)(x^5-x)^4$ where $C(a)$ is the rational function of a determined by the condition that $\mathcal{K}(a) = 0$, which is to say that

$$C(a) = \frac{(a^8+14a^4+1)^3}{a^4(a^4-1)^4}. \tag{6.2}$$

This result is verified by the observation that if $C(a)$ is *defined* by (6.2), then $(x^8+14x^4+1)^3 - C(a)(x^5-x)^4$ is zero when $x = a$ or when a is changed to ia (obvious) or $\frac{a-1}{a+1}$ (multiply numerator and denominator of $C(\frac{a-1}{a+1})$ by $(a+1)^{24}$ and simplify to find $C(\frac{a-1}{a+1}) = C(a)$), so $\mathcal{K}(x)$ is indeed given by this formula.

The usual notation for C is $\frac{j}{16}$. The classic book [7] of Hurwitz and Courant instead uses $J = \frac{j}{1728} = \frac{C}{108}$, which is a natural normalization, because it makes J equal to 1 on the 12 point orbit and 0 on the 8 point orbit.

(When $J = 1$, $\mathcal{K}(x) = (x^8+14x^4+1)^3 - 108(x^5-x)^4 = (x^4+1)^2(x^8-34x^4+1)^2$ has twelve double roots. When $J = 0$, $\mathcal{K}(x) = (x^8+14x^4+1)^3$ has 8 triple roots. For all other finite values of J, $\mathcal{K}(x)$ has 24 distinct roots.)

The J that corresponds to an elliptic curve $z^2 = x^4 + ex^3 + fx^2 + gx + h$ is $\frac{C(a)}{108} = \frac{(a^4+14+a^{-4})^3}{108(a^2-a^{-2})^4} = \frac{((a^2+a^{-2})^2+12)^3}{108((a^2+a^{-2})^2-2)^2}$, which can be found by setting $x^4+ex^3+fx^2+gx+h = (x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)$ and noting that $\phi = -\frac{2}{a^2+a^{-2}}$ is expressed in terms of the α's in Section 4. Since J is by its nature symmetric in the α's, it can be expressed in terms of e, f, g, and h.

The explicit formula is[5]

$$J = \frac{4(2f^3 - 9efg + 27g^2 + 27e^2h - 72fh)^2}{108 \cdot \Delta} + 1$$

[5] If α_1, α_2, α_3 are all nonzero, then $z^2 = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)$ can be rewritten as $(\frac{z}{x^2})^2 = \frac{\alpha_1\alpha_2\alpha_3}{x} \cdot (\frac{1}{\alpha_1} - \frac{1}{x})(\frac{1}{\alpha_2} - \frac{1}{x})(\frac{1}{\alpha_3} - \frac{1}{x})$ which is $v^2 = (u - \frac{1}{\alpha_1})(u - \frac{1}{\alpha_2})(u - \frac{1}{\alpha_3}) \cdot u$ where $u = \frac{1}{x}$ and $v = \frac{iz}{x^2\sqrt{\alpha_1\alpha_2\alpha_3}}$. In this case, four of the roots of $\mathcal{K}(x)$ are the roots of $x^4 - \frac{2}{\phi}x^2 + 1$ where $\phi = \frac{\frac{1}{\alpha_1}\cdot 0 + \frac{1}{\alpha_2}\frac{1}{\alpha_3} - \frac{1}{\alpha_1}\frac{1}{\alpha_2} - \frac{1}{\alpha_3}\cdot 0}{\frac{1}{\alpha_1}\frac{1}{\alpha_2} + \frac{1}{\alpha_2}\frac{1}{\alpha_3} + \frac{1}{\alpha_3}\cdot 0 + 0\cdot\frac{1}{\alpha_1} - 2\frac{1}{\alpha_1}\frac{1}{\alpha_3} - 2\frac{1}{\alpha_2}\cdot 0} = \frac{\alpha_1 - \alpha_3}{\alpha_3 + \alpha_1 - 2\alpha_2}$. The six permutations of the α's all give values of ϕ for which the same is true, which accounts for all 24 roots of $\mathcal{K}(x)$ and shows that $\mathcal{K}(x) = \prod_{i=1}^{6}(x^4 - \frac{2}{\phi_i}x^2 + 1)$ where the ϕ_i are the six rational functions of α_1, α_2, α_3 given by such permutations. Since interchange of α_1 and α_3 changes the sign of ϕ, one has in fact $\mathcal{K}(x) = \prod_{i=1}^{3}(x^4 - \frac{2}{\phi_i}x^2 + 1) \cdot \prod_{i=1}^{3}(x^4 + \frac{2}{\phi_i}x^2 + 1)$ where ϕ_1, ϕ_2, ϕ_3 are the three versions of ϕ obtained by *cyclic* permutations of α_1, α_2, α_3. The first factor of this product is $x^{12} - \sigma_1 x^{10} + (3 + \sigma_2)x^8 + (-\sigma_3 - 2\sigma_1)x^6 + (3 + \sigma_2)x^4 - \sigma_1 x^2 + 1$, where the σ_i are the elementary symmetric polynomials in $\frac{2}{\phi_1}$, $\frac{2}{\phi_2}$, and $\frac{2}{\phi_3}$.

By straightforward computation, $\sigma_i = \frac{P_i}{(\alpha_1 - \alpha_3)(\alpha_2 - \alpha_1)(\alpha_3 - \alpha_2)}$ where $P_1 = 4(\alpha_1^3 + \alpha_2^3 + \alpha_3^3) + 24\alpha_1\alpha_2\alpha_3 - 6\cdot(\text{all } \alpha_i\alpha_j^2)$, $P_2 = 36(\alpha_1\alpha_2^2 + \alpha_2\alpha_3^2 + \alpha_3\alpha_1^2 - \alpha_1^2\alpha_2 - \alpha_2^2\alpha_3 - \alpha_3^2\alpha_1)$, and $P_3 = -16(\alpha_1^3 + \alpha_2^3 + \alpha_3^3) - 96\alpha_1\alpha_2\alpha_3 + 24\cdot(\text{all } \alpha_i\alpha_j^2)$ where $(\text{all } \alpha_i\alpha_j^2) = \alpha_1\alpha_2^2 + \alpha_1\alpha_3^2 + \alpha_2\alpha_1^2 + \alpha_2\alpha_3^2 + \alpha_3\alpha_1^2 + \alpha_3\alpha_2^2$. Therefore, $\sigma_1 = 2 \cdot \frac{2S_1^3 - 9S_1S_2 + 27S_3}{(\alpha_1 - \alpha_3)(\alpha_3 - \alpha_2)(\alpha_2 - \alpha_1)}$, $\sigma_2 = -36$, and $\sigma_3 = -4\sigma_1$, where S_1, S_2, and S_3 are the elementary symmetric polynomials in α_1, α_2, and α_3, and the product of the first three factors of $\mathcal{K}(x)$ is $x^{12} - \sigma_1 x^{10} + (-33)x^8 + (2\sigma_1)x^6 + (-33)x^4 - \sigma_1 x^2 + 1 = (x^{12} - 33x^8 - 33x^4 + 1) - \sigma_1(x^{10} - 2x^6 + x^2) = (x^4 + 1)(x^8 - 34x^4 + 1) - \sigma_1 x^2(x^4 - 1)^2$, while the product of the last three factors is the same with the sign of σ_1 reversed. Therefore, $\mathcal{K}(x)$ for the curve $z^2 = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)$ is given by the explicit formula $(x^4 + 1)^2(x^8 - 34x^4 + 1)^2 - \sigma_1^2 x^4(x^4 - 1)^4$, provided α_1, α_2, α_3 are all nonzero. The identity $(x^8 + 14x^4 + 1)^3 - 108(x^5 - x)^4 = (x^4 + 1)^2(x^8 - 34x^4 + 1)^2$ then implies $\mathcal{K}(x) = (x^8 + 14x^4 + 1)^3 - (\sigma_1^2 + 108)(x^5 - x)^4$, which is to say that $C = \sigma_1^2 + 108$. The denominator of σ_1^2 is the discriminant of the cubic, and the numerator is the square of $2(-2E^3 + 9EF - 27G)$ when the cubic is $x^3 + Ex^2 + Fx + G = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)$. Thus, $J = \frac{4(2E^3 - 9EF + 27G)^2}{108\cdot\Delta} + 1$ for the curve $z^2 = x^3 + Ex^2 + Fx^2 + G$, provided $G \neq 0$.

Changing $x^3 + Ex^2 + Fx + G$ to $(x + c)^3 + E(x + c)^2 + F(x + c) + G$—that is, changing E to $3c + E$, F to $3c^2 + 2cE + F$, and G to $c^3 + c^2E + cF + G$ does not change the discriminant of the cubic or J, so it must not change $4(2E^3 - 9EF + 27G)^2$, at least not when $G \neq 0$. But this means that the polynomial $2(3c + E)^3 - 9(3c + E)(3c^2 + 2cE + F) + 27(c^3 + c^2E + cF + G)$ is independent of c. Therefore, the formula for J is valid even when $G = 0$.

Consider now a curve of the form $z^2 = x^4 + ex^3 + fx^2 + gx + h$. Assume first that 0 is a root of the polynomial on the right, which is to say that $h = 0$. Let β_1, β_2, β_3 be the other three roots of this polynomial (which is assumed to have distinct roots). Since the given curve is equivalent to $(\frac{z}{x^2})^2 = (1 - \frac{\beta_1}{x})(1 - \frac{\beta_2}{x})(1 - \frac{\beta_3}{x}) = -\beta_1\beta_2\beta_3(\frac{1}{x} - \frac{1}{\beta_1})(\frac{1}{x} - \frac{1}{\beta_2})(\frac{1}{x} - \frac{1}{\beta_3})$, the value of J for it is $\frac{4(2E^3 - 9EF + 27G)^2}{108\cdot\Delta} + 1$ where $-E$, F and $-G$ are the elementary symmetric functions in $\frac{1}{\beta_1}$, $\frac{1}{\beta_2}$, and $\frac{1}{\beta_3}$ and Δ is the square of $(\frac{1}{\beta_1} - \frac{1}{\beta_2})(\frac{1}{\beta_1} - \frac{1}{\beta_3})(\frac{1}{\beta_2} - \frac{1}{\beta_3})$. Multiplication of numerator and denominator by $(\beta_1\beta_2\beta_3)^6$ gives the square of $\beta_1\beta_2\beta_3(\beta_2 - \beta_1)(\beta_3 - \beta_1)(\beta_3 - \beta_2)$, which is the discriminant of $x^4 + ex^3 + fx^2 + gx$, in the denominator; in the numerator, it gives $4(-2(\beta_1\beta_2 + \beta_1\beta_3 + \beta_2\beta_3)^3 + 9(\beta_1\beta_2 + \beta_1\beta_3 + \beta_2\beta_3)(\beta_1 + \beta_2 + \beta_3)(\beta_1\beta_2\beta_3) - 27\beta_1^2\beta_2^2\beta_3^2)^2 = 4(2f^3 - 9efg + 27g^2)^2$. The given formula for J follows in the case $h = 0$.

The value of J for the curve $z^2 = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)(x - \alpha_4)$ is its value for $z^2 = (x + \alpha_4 - \alpha_1)(x + \alpha_4 - \alpha_2)(x + \alpha_4 - \alpha_3)x$, which is given by the above formula with $\beta_1 = \alpha_1 - \alpha_4$, $\beta_2 = \alpha_2 - \alpha_4$, and $\beta_3 = \alpha_3 - \alpha_4$. Since the discriminants of these two quartic polynomials are the same, the formula for J takes the form $\frac{4P^2}{108\cdot\Delta} + 1$ where P is a polynomial in e, f, g, and h. Because P is symmetric of degree 6 in the α's and is $2f^3 - 9efg + 27g^2$ when $h = 0$, it is $2f^3 - 9efg + 27g^2 + re^2h + sfh$ for some integers r and s. Because the coefficient of c in $2f^3 - 9efg + 27g^2 + re^2h + sfh$ when e is changed to $4c + e$, f is changed to $6c^2 + 3ce + f$, g is changed to $4c^3 + 3c^2e + 2cf + g$, and h is changed to $c^4 + c^3e + c^2f + cg + h$ is $72fg - 27e^2g + r(8eh + e^2g) + s(3eh + fg)$, and because these changes must leave the polynomial unchanged, the values $r = 27$ and $s = -72$ follow.

where Δ is the discriminant of $x^4 + ex^3 + fx^2 + gx + h$, or, for an elliptic curve of the form $z^2 = x^3 + Ex^2 + Fx + G$, the formula is

$$J = \frac{4(2E^3 - 9EF + 27G)^2}{108 \cdot \Delta} + 1$$

where Δ is the discriminant of $x^3 + Ex^2 + Fx + G$. For a curve in the commonly used canonical form $Q^2 = 4P^3 - g_2P - g_3$, this formula becomes $J = \frac{g_2^3}{g_2^3 - 27g_3^2}$.

The proposition implies that elliptic curves with the same J-invariant are equivalent. Theorem 11.1 below states that this sufficient condition is also necessary, except that the J-invariants of different elliptic function fields need not be in the same algebraic number field, so they may not be directly comparable unless (as is the case in most of the examples that are studied) they are rational.

7. Rational points and places

A *rational point* of the field of rational functions on $x^2+y^2 = a^2+a^2x^2y^2$ is a pair of constants[6] (x_1, y_1) of the field for which $x_1^2 + y_1^2 = a^2 + a^2x_1^2y_1^2$. As constants are adjoined to the function field, more rational points may be created, but the J-invariant and the equivalence class of the field are unchanged.

Not just x and y, but *all* elements of the function field have "values" at a rational point—although the "value" might be ∞—an observation which leads to the more intrinsic description of the rational point as a *place* on the curve in the following way.

Assume for the moment that x_1 is a constant for which the equation $x_1^2 + y^2 = a^2 + a^2x_1^2y^2$ has two distinct roots, which is to say $x_1 \neq \pm a$, $x_1 \neq \pm\frac{1}{a}$. For such values of x_1, the Newton polygon algorithm[7] generates two solutions y of $x^2 + y^2 = a^2+a^2x^2y^2$ in powers of $x-x_1$, one for each possible value $\pm y_1$ of the constant term of y (which may need to be adjoined). In other words, it produces for each such point (x_1, y_1) on the curve (provided $x_1 \neq \pm a, \pm\frac{1}{a}$ and provided a root y_1 of $x_1^2 + y^2 = a^2 + a^2x_1^2y^2$ is adjoined if necessary) a formal power series $y = y_1 + c_1t + c_2t^2 + \cdots$, which, when paired with the terminating power series $x = x_1 + t$, gives a pair of formal power series (x, y) in t for which $x^2 + y^2 = a^2 + a^2x^2y^2$. The coefficients x_1, y_1, c_1, c_2, $\ldots$ of these series are constants of the function field (after y_1 is adjoined, if necessary) because, in the terminology of [4], the truncated solution $y = y_1$ of $x_1^2 + y^2 = a^2 + a^2x_1^2y^2$ is unambiguous, so application of the algorithm generates a formal power series solution $y = y_1 + c_1(x-x_1) + c_2(x-x_1)^2 + \cdots$ without requiring the adjunction of constants. But since all elements of the function field are rational functions of x and y, the two expansions $x = x_1 + t$ and $y = y_1 + c_1t + c_2t^2 + \cdots$ imply expansions in powers of t of all elements of the function field provided a finite number of terms in which the power of t is negative are allowed.[8] Thus, (x_1, y_1)

[6] More correctly, such a point is "rational over the field of constants" K of the function field. This is the natural meaning of "rational" when K is a part of the definition of the field as in Section 4.

[7] See [4], where the construction assumes y is *integral* over x. A simple modification suffices to make it applicable to $x^2 + y^2 = a^2 + a^2x^2y^2$.

[8] The needed expansions are elements in the field of quotients of the integral domain $K[[t]]$ of all formal power series in t with coefficients in K, a field which is often denoted $K((t))$. Because the reciprocal of a power series $ct^k(1 + b_1t + b_2t^2 + \cdots)$ is $c^{-1}t^{-k}(1 - (b_1t + \cdots) + (b_1t + \cdots)^2 - (b_1t + \cdots)^3 + \cdots)$, each nonzero element of $K((t))$ can be written in one and only one way in the form $t^k(c_0 + c_1t + c_2t^2 + \cdots)$ where c_0, c_1, $\ldots$ are in K, $c_0 \neq 0$, and k is an integer, which is the *order* of the element.

determines a formal expansion for each element of the field and in particular assigns an *order* to each nonzero field element at the point (x_1, y_1), namely, the exponent of t in the first nonzero term of the expansion found in this way. Naturally, the field element is said to "have a pole" at (x_1, y_1) if this order is negative, to "be finite" at (x_1, y_1) if the order is nonnegative, and to "be zero" at (x_1, y_1) if the order is positive. If the element is finite at (x_1, y_1), its "value" at (x_1, y_1) is the constant term of its expansion, or, what is the same, the "value" is the constant which differs from it by an element that is zero at the place.

When $y_1 \neq \pm a, \pm\frac{1}{a}$, all field elements can be expanded in powers of $s = y - y_1$ in the analogous way and orders at (x_1, y_1) can be defined accordingly. These orders *coincide* with the orders determined by expansions in the parameter $t = x - x_1$ in all cases in which both are defined, because the expansion of $s = y - y_1$ in powers of t has the form $s = c_1t + c_2t^2 + \cdots$, so $t = b_1(c_1t + c_2t^2 + \cdots) + b_2(c_1t + c_2t^2 + \cdots)^2 + \cdots$, which implies $b_1c_1 = 1$, $b_1c_2 + b_2c_1^2 = 0$, $\ldots$, and in particular implies $c_1 \neq 0$, so that substitution of $s = c_1t + \cdots$ in an expansion $d_ks^k + d_{k+1}s^{k+1} + \cdots$ (where k may be negative) gives an expansion that begins $d_kc_1^kt^k + \cdots$. Therefore, the order of the element at (x_1, y_1) is k, whether t or s is used to determine it. More generally, the Newton polygon algorithm can be used to expand all field elements in powers of *any* local parameter at (x_1, y_1)—any element of the function field whose order at (x_1, y_1) is 1—and the assignment of orders at (x_1, y_1) to elements of the field is independent of the choice of the parameter.

As was just shown, the order of any element of the function field at (x_1, y_1) can be found using the parameter $y - y_1$ whenever $y_1 \neq \pm a, \pm\frac{1}{a}$. In particular, this method assigns orders to all field elements at the rational points $(a, 0)$ and $(-a, 0)$ for which the parameter $x - x_1$ cannot be used. (At these points, the Newton polygon algorithm gives an expansion of y in powers of $x^{1/2}$, but these expansions will not be needed.)

Literally speaking, there are no rational points where $x = \pm\frac{1}{a}$, because $(\frac{1}{a})^2 + y^2 = a^2 + y^2$ has no solution y (by assumption, $a^4 \neq 1$). Similarly, there are no rational points where $y = \pm\frac{1}{a}$. However, it is natural to regard the curve $x^2 + y^2 = a^2 + a^2x^2y^2$ as having four[9] "points at infinity", namely, $(\pm\frac{1}{a}, \infty)$ and $(\infty, \pm\frac{1}{a})$, because division of the defining equation by x^2y^2 gives another equation of the same form in which x is replaced by $\frac{1}{x}$ and y is replaced by $\frac{1}{y}$, so that $(x, y) = (\pm\frac{1}{a}, \infty)$ and $(\infty, \pm\frac{1}{a})$ can be regarded as the rational points where $(\frac{1}{x}, \frac{1}{y}) = (\pm a, 0)$ and $(0, \pm a)$, respectively. The *places* corresponding to these four points at infinity are determined accordingly.

In summary, each rational point of $x^2 + y^2 = a^2 + a^2x^2y^2$ (the four points at infinity included) gives rise to a place—a way of assigning orders to all elements of the function field—that describes the point in a more coordinate-free way.

[9] If one "projectivizes" the curve $x^2 + y^2 = a^2 + a^2x^2y^2$ by writing its equation in the form $x^2t^2 + y^2t^2 = a^2t^4 + a^2x^2y^2$, it has only *two* points at infinity, namely, $(t, x, y) = (0, 1, 0)$ and $(0, 0, 1)$. This discrepancy is noted in the discussion of the last entry of Gauss's *Tagebuch* in [8], where Gauss's choice to count four points at infinity on $x^2 + y^2 + x^2y^2 = 1$ is contrasted with the count of only two given by projectivization. I believe that Gauss's formula (2.1) and the symmetry $(x, y) \to (\frac{i}{x}, \frac{i}{y})$ of the curve make his count more convincing.

8. Algebraic proof of the addition formula

Let X and Y be defined as in (3.1), except let x' and y' be replaced by x_1 and y_1 to avoid the awkward notations $(x')^2$ and $(y')^2$. When the letter P is used to abbreviate xx_1yy_1, multiplication of the desired equation $X^2+Y^2=a^2+a^2X^2Y^2$ by $a^2(1-P^2)^2$ puts it in the form $(xy_1+yx_1)^2(1-P)^2+(yy_1-xx_1)^2(1+P)^2 = a^4(1-P^2)^2+(xy_1+yx_1)^2(yy_1-xx_1)^2$. It is to be shown that this equation is a consequence of the assumptions that $x^2+y^2=a^2+a^2x^2y^2$ and $x_1^2+y_1^2=a^2+a^2x_1^2y_1^2$. In other words, it is to be shown that the polynomial Δ defined by

$$\begin{aligned}(xy_1+yx_1)^2(1-P)^2&+(yy_1-xx_1)^2(1+P)^2\\ &=(xy_1+yx_1)^2(yy_1-xx_1)^2+a^4(1-P^2)^2+\Delta\end{aligned}\tag{8.1}$$

is a sum of multiples of $R=x^2+y^2-a^2-a^2x^2y^2$ and $R_1=x_1^2+y_1^2-a^2-a^2x_1^2y_1^2$.

The left-hand side of this equation can be rewritten $(x^2y_1^2+2P+y^2x_1^2)(1-2P+P^2)+(y^2y_1^2-2P+x^2x_1^2)(1+2P+P^2)$. Combine the parts that contain $1+P^2$ in the second factors on the one hand and the parts that contain $2P$ in the second factors on the other to find $(x^2y_1^2+2P+y^2x_1^2+y^2y_1^2-2P+x^2x_1^2)(1+P^2)+(-x^2y_1^2-2P-y^2x_1^2+y^2y_1^2-2P+x^2x_1^2)(2P)$. The first term is $(x^2+y^2)(x_1^2+y_1^2)(1+P^2)$ and the second is $((x^2-y^2)(x_1^2-y_1^2)-4P)(2P)=2P(x^2-y^2)(x_1^2-y_1^2)-8P^2$.

On the right-hand side, the first term $(xy_1+yx_1)^2(yy_1-xx_1)^2$ can be written $(x^2y_1^2+y^2x_1^2+2P)(y^2y_1^2+x^2x_1^2-2P)=(x^2y_1^2+y^2x_1^2)(y^2y_1^2+x^2x_1^2)+2P(y^2y_1^2+x^2x_1^2-x^2y_1^2-y^2x_1^2)-4P^2=x^2y^2y_1^4+x^4x_1^2y_1^2+y^4x_1^2y_1^2+x^2y^2x_1^4+2P(x^2-y^2)(x_1^2-y_1^2)-4P^2$. Therefore, subtraction of $2P(x^2-y^2)(x_1^2-y_1^2)-8P^2$ from both sides of (8.1) results in

$$\begin{aligned}&(x^2+y^2)(x_1^2+y_1^2)(1+P^2)\\ &\quad=x^2y^2y_1^4+x^4x_1^2y_1^2+y^4x_1^2y_1^2+x^2y^2x_1^4+4P^2+a^4(1-P^2)^2+\Delta\\ &\quad=x^2y^2(y_1^4+2x_1^2y_1^2+x_1^4)+x_1^2y_1^2(y^4+2x^2y^2+x^4)+a^4(1-P^2)^2+\Delta\\ &\quad=x^2y^2(y_1^2+x_1^2)^2+x_1^2y_1^2(y^2+x^2)^2+a^4(1-P^2)^2+\Delta.\end{aligned}\tag{8.2}$$

When $(1-P^2)^2$ is rewritten as

$$\begin{aligned}&(1-P^2)^2=(1+P^2)^2-4P^2\\ &\quad=(1+P^2)(1+x^2y^2+x_1^2y_1^2+P^2)-(1+P^2)(x^2y^2+x_1^2y_1^2)-4P^2\\ &\quad=(1+P^2)(1+x^2y^2)(1+x_1^2y_1^2)\\ &\qquad\quad-x^2y^2-x_1^2y_1^2-2P^2-2P^2-x^2y^2P^2-x_1^2y_1^2P^2\\ &\quad=(1+P^2)(1+x^2y^2)(1+x_1^2y_1^2)\\ &\qquad\quad-x^2y^2(1+2x_1^2y_1^2+x_1^4y_1^4)-x_1^2y_1^2(1+2x^2y^2+x^4y^4)\\ &\quad=(1+P^2)(1+x^2y^2)(1+x_1^2y_1^2)-x^2y^2(1+x_1^2y_1^2)^2-x_1^2y_1^2(1+x^2y^2)^2\end{aligned}\tag{8.3}$$

equation (8.2) shows Δ is

$$\begin{aligned}&\big((x^2+y^2)(x_1^2+y_1^2)-(a^2+a^2x^2y^2)(a^2+a^2x_1^2y_1^2)\big)(1+P^2)\\ &\quad+x^2y^2\big((a^2+a^2x_1^2y_1^2)^2-(x_1^2+y_1^2)^2\big)+x_1^2y_1^2\big((a^2+a^2x^2y^2)^2-(x^2+y^2)^2\big),\end{aligned}\tag{8.4}$$

thereby showing that $x^2+y^2 = a^2+a^2x^2y^2$ and $x_1^2+y_1^2 = a^2+a^2x_1^2y_1^2$ imply $\Delta = 0$.

The needed fact can be expressed entirely in terms of polynomial algebra in the following way. Since $RR_1 = (x^2+y^2)(x_1^2+y_1^2) - (x^2+y^2)(a^2+a^2x_1^2y_1^2) - (x_1^2+y_1^2)(a^2+a^2x^2y^2)+(a^2+a^2x^2y^2)(a^2+a^2x_1^2y_1^2)$, the difference $(x^2+y^2)(x_1^2+y_1^2)-(a^2+a^2x^2y^2)(a^2+a^2x_1^2y_1^2)$ can be written $RR_1+(x^2+y^2)(a^2+a^2x_1^2y_1^2)+(x_1^2+y_1^2)(a^2+a^2x^2y^2) - 2(a^2+a^2x^2y^2)(a^2+a^2x_1^2y_1^2) = RR_1+R(a^2+a^2x_1^2y_1^2)+R_1(a^2+a^2x^2y^2)$, giving the explicit equation

$$X^2+Y^2 = a^2+a^2X^2Y^2 + \frac{\Delta}{a^2(1-x^2x_1^2y^2y_1^2)^2} \tag{8.5}$$

where Δ is the polynomial in x, x_1, y, and y_1 given by

$$\begin{aligned}(RR_1{+}R(a^2+a^2x_1^2y_1^2)+R_1(a^2+a^2x^2y^2))(1+P^2)\\ - x^2y^2R_1(a^2+a^2x_1^2y_1^2+x_1^2+y_1^2) - x_1^2y_1^2R(a^2+a^2x^2y^2+x^2+y^2)\end{aligned} \tag{8.6}$$

in which $R = x^2+y^2-a^2x^2y^2$ and $R_1 = x_1^2+y_1^2-a^2-a^2x_1^2y_1^2$. Equation (8.5) proves:

Theorem 8.1. *Let K be an algebraic number field and let a be an element of K for which $a^5 \neq a$. To the field $K(x, x_1)$ of rational functions in x and x_1 with coefficients in K, adjoin square roots z and z_1 of $(a^2-x^2)(1-a^2x^2)$ and $(a^2-x_1^2)(1-a^2x_1^2)$, respectively. The formulas $y = \frac{z}{1-a^2x^2}$ and $y_1 = \frac{z_1}{1-a^2x_1^2}$ define elements of this extension field which generate the extension over $K(x, x_1)$ and which satisfy $x^2+y^2 = a^2+a^2x^2y^2$ and $x_1^2+y_1^2 = a^2+a^2x_1^2y_1^2$. The formulas* (3.1) *(with x_1 and y_1 in place of x' and y') determine elements X and Y of this extension field that satisfy $X^2+Y^2 = a^2+a^2X^2Y^2$.*

The function field constructed in the statement of this theorem is the field of rational functions on an algebraic surface—namely, the product of two copies of the algebraic curve $x^2+y^2 = a^2+a^2x^2y^2$.

If, instead of x_1 being an indeterminate and y_1 being an algebraic element adjoined to the field $K(x_1)$, both x_1 and y_1 are elements of K for which $x_1^2+y_1^2 = a^2+a^2x_1^2y_1^2$, formula (3.1) determines a pair of elements X and Y in the field of rational functions on $x^2+y^2 = a^2+a^2x^2y^2$ with coefficients in K that satisfy $X^2+Y^2 = a^2+a^2X^2Y^2$. Therefore, it determines an *automorphism* of this field of rational functions whose restriction to K is the identity.

Since places are intrinsic to the curve, an automorphism of the curve induces a permutation of the rational points on the curve. Therefore, the automorphism of the curve determined by a rational point (x_1, y_1) as in the preceding paragraph determines a permutation of the rational points on the curve. Specifically, this permutation of the rational points is the one that carries (x_2, y_2) to the rational point whose (x, y)–coordinates are $(\frac{1}{a}\cdot\frac{x_1y_2+y_1x_2}{1+x_1x_2y_1y_2}, \frac{1}{a}\cdot\frac{y_1y_2-x_1x_2}{1-x_1x_2y_1y_2})$, provided none of the four coordinates is ∞, because these coordinates are the constant terms of the series expansions obtained by setting $x = x_2+\cdots$ and $y = y_2+\cdots$ (where, in both cases, all omitted terms contain t) in the rational functions $\frac{1}{a}\cdot\frac{xy_1+yx_1}{1+xx_1yy_1}$ and $\frac{1}{a}\cdot\frac{yy_1-xx_1}{1-xx_1y_1}$. (This formula can also be interpreted in cases in which a coordinate is ∞. For example, if $x_1 = \infty$, then $y_1 = \pm\frac{1}{a}$; if $y_1 = \frac{1}{a}$, then the addition formula should be interpreted to mean $(\frac{1}{a}\cdot\frac{y_2}{x_2\cdot\frac{1}{a}\cdot y_2}, \frac{1}{a}\cdot\frac{-x_2}{-x_2\cdot\frac{1}{a}\cdot y_2}) = (\frac{1}{x_2}, \frac{1}{y_2})$ because one ignores terms in the numerator and denominator that do not contain $x_1 = \infty$.)

In this way, one can define a *binary operation* on the rational points on the curve $x^2 + y^2 = a^2 + a^2x^2y^2$. Clearly it is commutative, the rational point $(0, a)$ is an identity, and any rational point (x_1, y_1) has the inverse $(-x_1, y_1)$ (even the rational points $(\infty, \pm\frac{1}{a})$ and $(\pm\frac{1}{a}, \infty)$). To complete the proof that the rational points are given a group structure by the addition formula, it remains only to show that the binary operation is associative. The associativity of the operation is clearly indicated by the fact that the operation expresses the automorphism of the curve that carries the place $(0, a)$ to the place (x_1, y_1), because the composition of automorphisms is associative; however, the proof must also make use of the fact that these automorphisms preserve the differential $\frac{dx}{y(1-a^2x^2)}$, because $(x, y) \mapsto (-x, y)$ is an automorphism of the curve that leaves $(x, y) = (0, a)$ fixed, so automorphisms are not determined by their effects on one point without this further condition. Associativity is proved in the next section.

That the group law (3.1) on the rational points of an elliptic curve $x^2 + y^2 = a^2 + a^2x^2y^2$ coincides with the group law on this curve as defined in the usual way by the chord-and-tangent construction will be proved in Section 10.

9. Euler's integration

Theorem 9.1. *The rational functions X and Y defined by* (3.1) *satisfy the differential equation*

$$\frac{dX}{Y(1-a^2X^2)} = \frac{dx}{y(1-a^2x^2)} + \frac{dx_1}{y_1(1-a^2x_1^2)}.$$

If one treats x_1 and y_1 as constants, the map $(x, y) \to (X, Y)$ is a morphism of function fields, and the problem is to show that the pullback of $\frac{dX}{Y(1-a^2X^2)}$ is $\frac{dx}{y(1-a^2x^2)}$. The same equation with constant (x, y) and variable (x_1, y_1) then will follow by symmetry, and the full equation of the theorem is simply the sum of the two equations obtained in this way.

The fact that an elliptic curve has genus one implies that there is just a one-dimensional space of holomorphic differentials. A holomorphic differential is one that has no poles, so the pullback of a holomorphic differential is holomorphic, a condition that determines the pullback of $\frac{dX}{Y(1-a^2X^2)}$ up to a constant multiple. When x_1 and y_1 are fixed, one has $\frac{dX}{dx} = \frac{(y_1+x_1\frac{dy}{dx})(1+xx_1yy_1)-x_1y_1(y+x\frac{dy}{dx})(xy_1+yx_1)}{a(1+xx_1yy_1)^2}$ which gives $\frac{dX}{dx} = \frac{y_1(1-a^2x_1^2)}{a}$ when $(x, y) = (0, a)$. (Differentiation of $x^2 + y^2 = a^2 + a^2x^2y^2$ gives $\frac{dy}{dx} = -\frac{x(1-a^2y^2)}{y(1-a^2x^2)}$, which shows that $\frac{dy}{dx} = 0$ at this point.) Thus, the pullback of $\frac{dX}{Y(1-a^2X^2)}$ agrees with $\frac{dx}{y(1-a^2x^2)}$ at that point, so they must be equal.

Proof. The argument just given explains the idea that underlies the theorem. The proof, on the other hand, is probably more convincing if it is treated as an exercise in differential calculus. The calculation will make use of the formula

(9.1)

$$\frac{(1-x_1^2x^2)(1-y_1^2x^2)}{1-P^2} = \frac{1-(x_1^2+y_1^2)x^2+x_1^2y_1^2x^4}{1-P^2} = \frac{1-(a^2+a^2x_1^2y_1^2)x^2+x_1^2y_1^2x^4}{1-P^2}$$

$$= \frac{1-a^2x^2+x_1^2y_1^2x^2(x^2-a^2)}{1-P^2} = \frac{1-a^2x^2+x_1^2y_1^2x^2(-y^2+a^2x^2y^2)}{1-P^2} = 1-a^2x^2$$

and the formula $\frac{(1-y_1^2y^2)(1-x_1^2y^2)}{1-P^2} = 1-a^2y^2$ that follows from this one by symmetry. The above formula $\frac{dX}{dx} = \frac{(y_1+x_1\frac{dy}{dx})(1+P)-x_1y_1(y+x\frac{dy}{dx})(xy_1+yx_1)}{a(1+P)^2}$ can be transformed to

(9.2)

$$\begin{aligned}
&\frac{y_1+x_1\frac{dy}{dx}-x_1x^2y_1^2\frac{dy}{dx}-y_1x_1^2y^2}{a(1+P)^2} = \frac{(y_1+x_1\frac{dy}{dx}-x_1x^2y_1^2\frac{dy}{dx}-y_1x_1^2y^2)y(1-a^2y^2)}{a(1+P)^2y(1-a^2y^2)}\\
&= \frac{(y_1-y_1x_1^2y^2)y(1-a^2x^2)+(x_1-x_1x^2y_1^2)\frac{dy}{dx}y(1-a^2x^2)}{a(1+P)^2y(1-a^2y^2)}\\
&= \frac{y_1y(1-a^2x^2)(1-x_1^2y^2)-x_1x(1-x^2y_1^2)(1-a^2y^2)}{a(1+P)^2y(1-a^2x^2)}\\
&= \frac{y_1y(1-x_1^2x^2)(1-y_1^2x^2)(1-x_1^2y^2)-x_1x(1-x^2y_1^2)(1-y_1^2y^2)(1-x_1^2y^2)}{(1-P^2)a(1+P)^2y(1-a^2x^2)}\\
&= \frac{(1-y_1^2x^2)(1-x_1^2y^2)(y_1y(1-x_1^2x^2)-x_1x(1-y_1^2y^2))}{(1-P)a(1+P)^3y(1-a^2x^2)}\\
&= \frac{(1-y_1^2x^2-x_1^2y^2+P^2)(y_1y-x_1xP-x_1x+y_1yP)}{(1-P)a(1+P)^3y(1-a^2x^2)}\\
&= \frac{((1+P)^2-(y_1x+x_1y)^2)(y_1y-x_1x)(1+P)}{(1-P)a(1+P)^3y(1-a^2x^2)} = \frac{(1-a^2X^2)Y}{y(1-a^2x^2)}.
\end{aligned}$$

Thus $\frac{dX}{Y(1-a^2X^2)} = \frac{dx}{y(1-a^2x^2)}$, as was to be shown. □

The theorem implies that X *is an integral of* $\frac{dx_1}{y_1(1-a^2x_1^2)} + \frac{dx_2}{y_2(1-a^2x_2^2)} = 0$ *in the sense Abel intended* in the statement quoted in Section 1—that is, the curves $X =$ const. are curves along which the differential is zero. Because $\frac{dY}{X(1-a^2Y^2)} = \frac{-dX}{Y(1-a^2X^2)}$, Y is also an integral, so (X, Y) can be interpreted as a function from the direct product of the curve with itself to the curve—the group operation—whose level curves integrate $\frac{dx_1}{y_1(1-a^2x_1^2)} + \frac{dx_2}{y_2(1-a^2x_2^2)} = 0$. When $y(1-a^2x^2)$ is rewritten as $\sqrt{a^2x^4-(a^4+1)x^2+a^2}$, this differential equation is in Abel's form (1.1). In fact, since Section 5 shows that every elliptic curve is equivalent to one in this form, *every* differential equation in Abel's form is integrated in this way, provided only that the fourth degree polynomial under the radical sign has distinct roots (and that its coefficients are algebraic numbers).

The associative law for the composition (3.1) is a consequence of Theorem 9.1 in the following way. Let (X_1, Y_1) be the composition of (x_1, y_1) and (x_2, y_2) according to this formula, and let $(\mathcal{X}_1, \mathcal{Y}_1)$ be the composition of (X_1, Y_1) and (x_3, y_3). The other way of composing (x_1, y_1), (x_2, y_2), and (x_3, y_3) is to let (X_2, Y_2) be the composition of (x_2, y_2) and (x_3, y_3) and $(\mathcal{X}_2, \mathcal{Y}_2)$ be the composition of (x_1, y_1) and (X_2, Y_2). It is to be shown that $\mathcal{X}_1 = \mathcal{X}_2$ and $\mathcal{Y}_1 = \mathcal{Y}_2$.

The four "functions" $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{Y}_1$, and $\mathcal{Y}_2$ are naturally regarded as rational functions on the 3-dimensional algebraic surface that is the direct product of the curve $x^2+y^2 = a^2+a^2x^2y^2$ with itself three times, which is the field $K(x_1, x_2, x_3)$ of rational functions in x_1, x_2, and x_3 with coefficients in K to which are adjoined the square roots z_i of $a^2x_i^4-(a^4+1)x^2+a^2$ for $i = 1$, 2, and 3, a field extension of degree 8 in which every element has a unique representation as a polynomial of degree less than 2 in z_1, z_2, and z_3 whose coefficients are in $K(x_1, x_2, x_3)$. Two applications of Theorem 9.1 imply

$$\begin{aligned} \frac{d\mathcal{X}_1}{\mathcal{Y}_1(1-a^2\mathcal{X}_1^2)} &= \frac{dX_1}{Y_1(1-a^2X_1^2)} + \frac{dx_3}{y_3(1-a^2x_3^2)} \\ &= \frac{dx_1}{y_1(1-a^2x_1^2)} + \frac{dx_2}{y_2(1-a^2x_2^2)} + \frac{dx_3}{y_3(1-a^2x_3^2)}. \end{aligned} \tag{9.3}$$

In the same way

$$\frac{d\mathcal{X}_2}{\mathcal{Y}_2(1-a^2\mathcal{X}_2^2)} = \frac{dx_1}{y_1(1-a^2x_1^2)} + \frac{dx_2}{y_2(1-a^2x_2^2)} + \frac{dx_3}{y_3(1-a^2x_3^2)}.$$

Therefore, the pairs $(\mathcal{X}_1, \mathcal{Y}_1)$ and $(\mathcal{X}_2, \mathcal{Y}_2)$ integrate the same differential, both satisfy $\mathcal{X}^2 + \mathcal{Y}^2 = a^2 + a^2\mathcal{X}^2\mathcal{Y}^2$, and both are $(0, a)$ when $(x_1, y_1) = (x_2, y_2) = (x_3, y_3) = (0, a)$.

Thus, when all "functions"—elements of the field of rational functions on the 3-dimensional surface—are expressed as polynomials of degree less than 2 in z_1, z_2, and z_3 whose coefficients are in $K(x_1, x_2, x_3)$, the pairs $(\mathcal{X}_1, \mathcal{Y}_1)$ and $(\mathcal{X}_2, \mathcal{Y}_2)$ have the same values when $x_1 = x_2 = x_3 = 0$ and have the same partial derivatives with respect to x_1, x_2, and x_3 at all points. Therefore, they are identical, as was to be shown.

10. Algebraic variations

The construction of Abel mentioned in Section 1 that initiated the study of curves of higher genus can be described in a very heuristic way as follows.[10] Let a set of N points on a planar curve $\chi(x, y) = 0$ be given, where χ is a polynomial with integer coefficients. To construct an "algebraic variation" of the given points along $\chi = 0$, choose an auxiliary curve $\theta(x, y) = 0$ that contains a large number of variable coefficients. Choose values for the variable coefficients of θ that make $\theta(x, y) = 0$ at all of the N given points, so that $\chi = 0$ and $\theta = 0$ intersect in the given N points. Because θ has many coefficients, it is of high degree, so $\theta = 0$ will intersect $\chi = 0$ in many points other than the required N. Let these other intersection points be called the extraneous intersection points. When the parameters in θ are allowed to vary in such a way that $\theta = 0$ continues to intersect $\chi = 0$ in the extraneous intersection points, the motion of the original N points of intersection along $\chi = 0$ is an *algebraic variation* of them.

This very general (heuristic) construction generalizes the addition operation on a nonsingular cubic in the following way. Let $\chi(x, y) = 0$ be a nonsingular cubic (for example, $y^2 - x^3 - x = 0$) and let P and Q be two given points on $\chi(x, y) = 0$. Choose the coefficients a, b, and c of $\theta(x, y) = ax + by + c$ to make the line $\theta(x, y) = 0$ pass through P and Q. Since $\chi(x, y)$ is cubic, the line $\theta(x, y) = 0$ intersects $\chi(x, y) = 0$ in a third, extraneous, point; call it R. The algebraic variations of the pair (P, Q) are the pairs of points $(\mathcal{P}, \mathcal{Q})$ in which lines that pass through R intersect $\chi(x, y) = 0$.

In terms of the addition operation on $\chi(x, y) = 0$, the algebraic variations $(\mathcal{P}, \mathcal{Q})$ of (P, Q) are described by the formula

$$\mathcal{P} + \mathcal{Q} = P + Q \tag{10.1}$$

because the "sum" of $\mathcal{P}$ and $\mathcal{Q}$ depends only on R.

[10] See the Introduction to [2]. My book [4] describes the construction in Essay 4.1. See also the Historical Sketch that is an appendix to [12].

Abel's great realization was that the number of constraints satisfied by the algebraic variations of N points along a curve *depends only on the curve.* For the circle $x^2 + y^2 - 1 = 0$ there are *no* constraints—a set of N points on the circle can be varied in all possible ways along the circle—but for a nonsingular cubic there is a single constraint which in the case $N = 2$ is given by (10.1). For arbitrary N it was shown in Essay 4.6 of [4] that for an elliptic curve $z^2 = f(x)$ the single constraint is described by the differential equation

$$\frac{dx_1}{z_1} + \frac{dx_2}{z_2} + \cdots + \frac{dx_N}{z_N} = 0, \tag{10.2}$$

where (x_1, z_1), (x_2, z_2), ..., (x_N, z_N) are the N given points on $z^2 = f(x)$ and dx_i is the variation in the x-coordinate of the ith point. More generally, for a curve of genus g the algebraic variations are the variations that satisfy the g linearly independent differential equations determined by the holomorphic differentials.

The operation of addition on an elliptic curve "integrates" (10.2) for all N in the same way that it integrates this equation when $N = 2$. Construct an equivalence between the given elliptic curve $z^2 = f(x)$ and an elliptic curve in the normal form $x^2 + y^2 = a^2 + a^2x^2y^2$. Let (x_1, y_1), (x_2, y_2), ..., (x_N, y_N) be the N given points in this new coordinate system; the composition of all N points using the addition formula (3.1) gives two explicit rational functions X and Y of $x_1, y_1, x_2, y_2, \ldots, x_N, y_N$ with the property that the level surfaces of X or Y or any nonconstant rational function of X and Y, which are hypersurfaces of codimension 1 in the direct product of the curve with itself N times, are the N-tuples of points that can be reached from one another by algebraic variations.

In short, the condition $X = \text{const.}$ gives an *explicit* algebraic description of the algebraic variations of the N given points on the curve.

11. Equivalent curves have the same J-invariant

Theorem 11.1. *Let K_a and K_b be algebraic number fields containing elements a and b, respectively, for which $a^5 \neq a$ and $b^5 \neq b$, and for which the elliptic function fields determined by $x^2 + y^2 = a^2 + a^2x^2y^2$ and $u^2 + v^2 = b^2 + b^2u^2v^2$ over K_a and K_b, respectively, are equivalent in the sense of Section 4. Then the J-invariants of these two function fields are conjugate algebraic numbers in the sense that they are roots of the same irreducible polynomial with integer coefficients.*

Proof. The assumption of equivalence implies a third algebraic number field K and embeddings of K_a and K_b in K for which the function fields over K determined by the equations $x^2 + y^2 = a_1^2 + a_1^2x^2y^2$ and $u^2 + v^2 = b_1^2 + b_1^2u^2v^2$ are isomorphic, where a_1 and b_1 are the images in K of a and b under the respective embeddings. An isomorphism of function fields implies an isomorphism of their fields of constants, so, at the cost of replacing b_1 with one of its conjugates under the Galois group of K over the rationals, one can assume that a and b are in K and that $x^2 + y^2 = a^2 + a^2x^2y^2$ and $u^2 + v^2 = b^2 + b^2u^2v^2$ define function fields over K that are isomorphic under an isomorphism that is the identity on K. It will be shown that with these stronger assumptions the two J-invariants are *equal.*

Such an isomorphism of two elliptic function fields over K that is the identity on K implies a one-to-one correspondence between the rational points on the two curves, because values x_1 and y_1 of x and y determine values (possibly ∞) for all elements of the function field and in particular determine values u_1 and v_1

of u and v, and conversely. *One can assume without loss of generality that the given isomorphism carries the rational point* $(x, y) = (0, a)$ *to the rational point* $(u, v) = (0, b)$, because if $(x, y) = (x_1, y_1)$ is the rational point that corresponds to $(u, v) = (0, b)$ under the given automorphism, then the addition formula gives an automorphism of the field that carries $(x, y) = (x_1, y_1)$ to $(x, y) = (0, a)$ and therefore gives new xy-coordinates in which the rational points $(x, y) = (0, a)$ and $(u, v) = (0, b)$ correspond under the isomorphism.

With this stronger assumption, *the rational point* $(u, v) = (0, -b)$ *corresponds under the isomorphism to one of the three rational points* $(x, y) = (0, -a)$, $(\infty, \frac{1}{a})$, *or* $(\infty, -\frac{1}{a})$. The reason is that the four points $(x, y) = (0, \pm a)$, $(\infty, \pm\frac{1}{a})$ are the preimage of the identity $(x, y) = (0, a)$ under the *doubling* map, and this map is intrinsic to the curve once the identity $(x, y) = (0, a)$ is specified. That is, the pair of field elements $(X, Y) = (\frac{2xy}{a(1+x^2y^2)}, \frac{y^2-x^2}{a(1-x^2y^2)})$ given by the addition formula when the two input pairs are the same generate a subfield—the elements that can be expressed rationally in terms of X and Y—that is isomorphic to the whole field but has index 4 in it. This subfield is intrinsic in the sense that it must correspond under the isomorphism to the subfield of the function field of $u^2 + v^2 = b^2 + b^2u^2v^2$ generated by the functions $(U, V) = (\frac{2uv}{b(1+u^2v^2)}, \frac{v^2-u^2}{b(1-u^2v^2)})$ because an expression of a rational function of x and y as a rational function of X and Y corresponds under the isomorphism to an expression of a rational function of u and v as a rational function of U and V. This subfield has index 4 because when x^2 is written as $\frac{a^2-y^2}{1-a^2y^2}$ and denominators are cleared, the equation $aY(1 - x^2y^2) = y^2 - x^2$ becomes an equation of degree 4 in y with coefficients in the subfield; each root y gives a unique value of x by virtue of $aX(1 + x^2y^2) = 2xy$. Each rational point on $x^2 + y^2 = a^2 + a^2x^2y^2$ implies values in K of X and Y, and this mapping of rational points is 4-to-1 because the values of X and Y at $(x, y) = (x_1, y_1)$ are the same as they are at $(-x_1, -y_1)$ and $(\pm\frac{1}{x_1}, \pm\frac{1}{y_1})$. In particular, the four points whose doubles are $(X, Y) = (0, a)$ are as stated above and the desired conclusion follows.

One can assume without loss of generality that the given isomorphism between the xy-curve and the uv-curve carries $(u, v) = (0, -b)$ *to* $(x, y) = (0, -a)$ as well as $(u, v) = (0, b)$ to $(x, y) = (0, a)$. As was just shown, if it does not carry $(u, v) = (0, -b)$ to $(x, y) = (0, -a)$, it must carry it to one of $(x, y) = (\infty, \pm\frac{1}{a})$. Therefore, it will suffice to show that there is an isomorphism of the xy-curve with another curve of the same form $X^2 + Y^2 = A^2 + A^2X^2Y^2$ that carries either of these points to $(X, Y) = (0, -A)$ while carrying $(x, y) = (0, a)$ to $(X, Y) = (0, A)$.

The fractional linear transformation $\lambda(z) \mapsto \frac{1+iz}{1-iz}$ has order 3. (As was noted in the proof of Proposition 6.1, it corresponds to the permutation $0 \mapsto 1 \mapsto i \mapsto 0$ and $\infty \mapsto -1 \mapsto -i \mapsto \infty$ of the "faces of the cube".) When A is defined to be $\lambda(a)$ (an element of K) and Y is defined to be $\lambda(y)$ (an element of the function field of $x^2 + y^2 = a^2 + a^2x^2y^2$ over K), then

$$\frac{A^2 - Y^2}{1 - A^2Y^2} = \frac{(y - a)(ay + 1)}{(y + a)(1 - ay)},$$

as one can find by direct computation. This element of the function field of $x^2+y^2 = a^2 + a^2x^2y^2$ is a *square*, namely,

$$\frac{(y - a)(ay + 1)}{(y + a)(1 - ay)} \cdot \frac{(ay + 1)(y + a)}{(ay + 1)(y + a)} = \frac{(ay + 1)^2(y^2 - a^2)}{(y + a)^2(1 - a^2y^2)} = \Big(\frac{ay + 1}{y + a} \cdot ix\Big)^2.$$

Thus, setting $X = ix\cdot\frac{ay+1}{y+a}$ describes a presentation of the function field of $x^2+y^2 = a^2+a^2x^2y^2$ over K as the function field of $X^2+Y^2 = A^2+A^2X^2Y^2$ over K. At the point $(x, y) = (0, a)$, Y has the value $\lambda(a) = A$, so this point corresponds to $(X, Y) = (0, A)$. At the point $(x, y) = (\infty, -\frac{1}{a})$, $Y = \lambda(y)$ has the value $\lambda(-\frac{1}{a}) = \frac{1-\frac{i}{a}}{1+\frac{i}{a}} = \frac{ia+1}{ia-1} = -\lambda(a) = -A$, so $(x, y) = (\infty, -\frac{1}{a})$ corresponds to $(X, Y) = (0, -A)$. At $(x, y) = (\infty, \frac{1}{a})$, Y has the value $\lambda(\frac{1}{a}) = -\frac{1}{A}$, so setting $\mathcal{A} = \lambda(A)$, $\mathcal{Y} = \lambda(Y)$, and $\mathcal{X} = iX\cdot\frac{AY+1}{Y+A}$ gives a presentation of the field as a field of the same form in which $(x, y) = (\infty, \frac{1}{a})$ corresponds to $(\mathcal{X}, \mathcal{Y}) = (0, -\mathcal{A})$ and $(x, y) = (0, a)$ corresponds to $(\mathcal{X}, \mathcal{Y}) = (0, \mathcal{A})$.

Finally, when the given isomorphism of $u^2+v^2 = b^2+b^2u^2v^2$ with $x^2+y^2 = a^2+a^2x^2y^2$ carries $(u, v) = (0, b)$ to $(x, y) = (0, a)$ and $(u, v) = (0, -b)$ to $(x, y) = (0, -a)$, it carries points whose doubles are $(u, v) = (0, -b)$ to points whose doubles are $(x, y) = (0, -a)$. But these points are $(u, v) = (\pm b, 0)$, $(\pm\frac{1}{b}, \infty)$ and $(x, y) = (\pm a, 0)$, $(\pm\frac{1}{a}, \infty)$, respectively, as is easily checked. Therefore, b must have one of the values $\pm a$, $\pm\frac{1}{a}$, all of which are in the list (6.1), so the J-invariants are equal, as was to be shown.

12. The holomorphic parameter at the origin

The transcendental function $x(t)$ defined by $t = \int_0^{x(t)} \frac{dx}{\sqrt{(a^2-x^2)(1-a^2x^2)}}$ will be used in Part III to parameterize the curve $x^2+y^2 = a^2+a^2x^2y^2$ using complex analytic functions. It can be treated *algebraically* by developing it as a formal power series $x(t) = b_1t+b_2t^2+b_3t^3+\cdots$ with coefficients in K that can be determined in the following way.

At the place $(x, y) = (0, a)$, all elements of the function field over K defined by $x^2+y^2 = a^2+a^2x^2y^2$ can be expanded in powers of x (possibly with a finite number of negative powers). They can also be expanded in powers of t for any formal power series t in x of order 1 at $(0, a)$. Let t be defined implicitly by an equation $x = b_1t+b_2t^2+\cdots$ in which the coefficients $b_1 \neq 0$, b_2, b_3, $\ldots$ are to be determined. Once values are given to the b_i, substitution of the expansion of x in powers of t in the expansion of y in powers of x will give an expansion y, and therefore expansions of all elements of the function field, in powers of t.

When the integral that defines $x(t)$ is differentiated with respect to t, one finds $1 = x'(t)\frac{1}{\sqrt{(a^2-x^2)(1-a^2x^2)}} = \frac{x'(t)}{y(1-a^2x^2)}$ or, more simply, $x'(t) = y(1-a^2x^2)$. The equations

$$y^2 = \frac{a^2-x^2}{1-a^2x^2} = (a^2-x^2)(1+a^2x^2+a^4x^4+a^6x^6+\cdots) \tag{12.1}$$

and

$$\frac{dx}{dt} = y(1-a^2x^2) \tag{12.2}$$

can be used to find the desired expansion $x = b_1t+b_2t^2+b_3t^3+\cdots$ in the following way.

Heuristic considerations suggest that the expansion of x in powers of t will contain only *odd* powers of t and the expansion of y only *even* powers, but for purposes of constructing an expansion of x in powers of t with the desired property

$\frac{dx}{dt} = y(1 - a^2x^2)$, one can simply stipulate that the expansions have the form

$$x = b_1t + b_3t^3 + b_5t^5 + b_7t^7 + \cdots,$$
$$y = a + c_2t^2 + c_4t^4 + c_6t^6 + \cdots \tag{12.3}$$

and note that (12.1) and (12.2) suffice to determine

$$c_0, b_1, c_2, b_3, c_4, b_5, c_6, b_7, c_8, \ldots$$

in succession, given $c_0 = a$. Specificially, when these coefficients are known up to, but not including, b_{2n-1}, then y is known mod t^{2n} (it contains no term in t^{2n-1} and the previous terms are known) and x^2 is also known mod t^{2n} (the first unknown term in the expansion of x^2 is $2b_1b_{2n-1}t^{2n}$), so the left side of (12.2) is known mod t^{2n}, which determines the coefficient $(2n-1)b_{2n-1}$ of t^{2n-2} in the right side, thereby determining b_{2n-1}. Then x^2 is known mod t^{2n+2} (it has no term in t^{2n+1} and the previous terms are known), so the right side of (12.1) is known mod t^{2n+2}, which is sufficient to determine the coefficient $2ac_{2n}$ of t^{2n} on the right, thereby determining c_{2n}.

The coefficients can be described more fully by saying that they have the form $b_{2n+1} = \frac{a\beta_{2n+1}}{(2n+1)!}$ and $c_{2n} = \frac{a\gamma_{2n}}{(2n)!}$ where β_{2n+1} and γ_{2n} are *polynomials in* a^4 *with integer coefficients.* With the understanding that $\gamma_0 = 1$, they are given by the formulas

$$\beta_{2n+1} = \gamma_{2n} - a^4\Big(\sum \frac{(2n)!}{i!j!k!}\gamma_i\beta_j\beta_k\Big)$$

where the sum is over all triples of nonnegative integers (i, j, k) for which $i+j+k = 2n$, i is even, and j and k are odd, and

$$\gamma_{2n} = -\beta_{2n-1} + a^4\Big(\sum \frac{(2n-1)!}{i!j!k!}\beta_i\gamma_j\gamma_k\Big)$$

where the sum is over all triples (i, j, k) in which $i + j + k = 2n - 1$, i is odd, and j and k are even. The first few terms of the series are

$$x = at - \frac{a(a^4+1)}{6}t^3 + \frac{a(a^8+14a^4+1)}{120}t^5 + \cdots$$

and

$$y = a + \frac{a(a^4-1)}{2}t^2 + \frac{a(5a^8-6a^4+1)}{24}t^4 + \cdots .$$

□

Part III. The Theory of Elliptic Functions

13. Double periodicity

One of Abel's main objectives was[11] to extend the study of transcendental functions [1] beyond the trigonometric and logarithmic functions. His treatise on elliptic functions is devoted to the study of the function $\Phi(t)$ defined implicitly for t near

[11] See the introduction of [1] and the title of his Paris memoir [2].

zero by[12]

$$t = \int_0^{\Phi(t)} \frac{dx}{\sqrt{(1-c^2x^2)(1+e^2x^2)}} \tag{13.1}$$

where c^2 and e^2 are positive[13] real constants. The integrand is real and positive for $0 \le x \le \frac{1}{c}$, so $\Phi(t)$ increases as t goes from 0 to $\int_0^{\frac{1}{c}} \frac{dx}{(1-c^2x^2)(1+e^2x^2)}$, a number to which Abel gives the name $\frac{\omega}{2}$. He clearly has in mind the analogy with the formula $\int_0^1 \frac{dx}{\sqrt{1-x^2}} = \frac{\pi}{2}$ which, when it is used to define $\frac{\pi}{2}$, determines the period 2π of the trigonometric functions. In a similar way, Abel's ω has the property that 2ω is a period of Φ, although he does not prove this property of ω right away. Instead, he observes that changing x to ix in the defining formula interchanges e and c and leads to a second, purely imaginary, period of Φ, which he calls $2i\tilde{\omega}$. In short, he *begins* with the double periodicity of Φ.

It must be remembered that the theory of functions of a complex variable was then in its infancy. Abel dealt primarily in *formulas*—he put the defining relation in the form (13.1) because it "made the formulas simpler," and he followed his derivation of his form of the addition formula with the statement[14] that "one can deduce a crowd [*une foule*] of others" (formulas). After the first sections of the paper, the requirement that c^2 and e^2 be positive reals seems to be forgotten. I believe that once he had established the formulas in this case, he would comfortably assume them to be true for all complex numbers c and e for which $(1-c^2x^2)(1+e^2x^2)$ has distinct roots; he notes in his introduction that Legendre assumed e^2 was negative, not positive, and he would surely not have intended to exclude all of the cases covered by Legendre. From this point of view, it is natural that he would first establish a real period and a purely imaginary period in the case in which $c^2 > 0$, $e^2 > 0$. Then he could use his version of the addition formula to represent the value of Φ throughout the complex plane and deduce formulas that describe all of its periods, which is exactly what he did.

14. Parameterizing elliptic curves

The defining equation (13.1) of Abel's function $\Phi(t)$ can also be written, as Abel himself wrote it[15] in his introduction, in the form

$$\Phi'(t) = \sqrt{(1-c^2\Phi(t)^2)(1+e^2\Phi(t)^2)}.$$

In other words, the functions $x(t) = \Phi(t)$ and $z(t) = \Phi'(t)$ parameterize the elliptic curve $z^2 = (1-c^2x^2)(1+e^2x^2)$, which suggests that Abel's $\Phi(t)$ can be obtained as a byproduct of a general solution of the problem: *Parameterize elliptic curves.*

Part I shows that any given elliptic curve is equivalent to the 24 elliptic curves of the form $x^2 + y^2 = a^2 + a^2x^2y^2$ in which a ranges over the 24 roots of $\mathcal{K}(x) = (x^8 + 14x^4 + 1)^3 - 108J(x^5 - x)^4$, J being the J-invariant of the given curve. In

[12] Abel used the letter α instead of t for the independent variable, and the letter ϕ instead of Φ for the function. The change from ϕ to Φ is made here to avoid confusion with the ϕ introduced in Section 16.

[13] He first gives the denominator of the integrand the form $\sqrt{(1-x^2)(1-c^2x^2)}$ that Legendre used, but then says, "M. Legendre takes c^2 to be positive, but I have noticed that the formulas become simpler when one takes c^2 to be negative and equal to $-e^2$." He then changes 1 to c^2 for the sake of symmetry.

[14] [1], §3.

[15] Except that he wrote α for t and ϕ for Φ.

Part I, however, a was an algebraic number and the curves were algebraic curves. Abel's $\Phi(t)$ belongs in an altogether different realm, the realm of functions of a complex variable, but the statement and proof of the result adapt immediately to Abel's case: If $f(x)$ is a polynomial of degree 3 or 4 with complex coefficients and with distinct roots, the curve $z^2 = f(x)$ is equivalent[16] to each of the 24 curves of the form $x^2 + y^2 = a^2 + a^2x^2y^2$ in which a is a complex root of the same $\mathcal{K}(x)$, J being the complex number given by the formulas of Section 6. Here the exact meaning of "curve" can remain unspecified because the meaning of "equivalent" is clear: Two curves $z^2 = f(x)$ and $v^2 = g(u)$ are equivalent if there is a fractional linear transformation of the Riemann sphere that carries the four roots of $f(x)$ to the four roots of $g(u)$. (A polynomial of degree 3 should be regarded as a polynomial of degree 4 with one root at ∞.)

Since an equivalence between elliptic curves implies explicit birational formulas[17] relating their coordinates, a parameterization of an elliptic curve implies a parameterization of any curve equivalent to it in the sense just defined, and the problem of parameterizing elliptic curves becomes the problem of parameterizing elliptic curves of the special form $x^2 + y^2 = a^2 + a^2x^2y^2$, where a is a complex number (for which $a^5 \neq a$), and, even more narrowly, of parameterizing just one of each set of 24 equivalent curves of this form.

In a certain sense, this parameterization problem is solved by the expansions of x and y in powers of t in Section 12. If a is a given complex number ($a^5 \neq a$) and if these power series converge for $|t| < \delta$ (as they do for any given a when δ is sufficiently small), then the functions $x(t)$ and $y(t)$ of a complex variable t that they define satisfy $x^2 + y^2 = a^2 + a^2x^2y^2$ on the disk $|t| < \delta$ and therefore satisfy this equation in whatever region of the t-plane their definitions can be extended to. Seen from this point of view, the problem is simply the *analytic continuation* of the functions $x(t)$ and $y(t)$ defined by these series to the entire t-plane.

Like Riemann's analytic continuation of $\zeta(s)$ to the entire complex plane in [11], these analytic continuations will be accomplished by finding "an expression of the function that is always valid."

15. A doubly periodic function

In addition to making possible the simple statement (3.1) of the addition formula on an elliptic curve, the normal form $x^2 + y^2 = a^2 + a^2x^2y^2$ has the advantage of dealing with x and y symmetrically, so that the parameterizing functions $x(t)$ and $y(t)$ of the last section are in essence the *same function* just as $\sin t$ and $\cos t$ are essentially the same function. The symmetries of this function determine it and show that it must be, in essence, the function with the "always valid" expression (15.1) for a suitably chosen value of τ. The formula is a quotient of θ-functions, but there is no need to invoke the theory of θ-functions because the needed properties can be deduced directly.

[16] Now that the ground field is the complex numbers, one can even say that the fields are birationally equivalent.

[17] See Section 5.

Theorem 15.1. *Given a complex number τ in the upper half plane $\Im\tau > 0$, the formula*

$$\psi(t) = \frac{\sum_{n \text{ odd}} e^{i\pi(\frac{n^2}{2}\cdot\tau + nt)}}{\sum_{n \text{ even}} e^{i\pi(\frac{n^2}{2}\cdot\tau + nt)}} \tag{15.1}$$

(the variable of summation n ranges over all integers, positive and negative, with odd integers in the numerator and even integers in the denominator) defines a meromorphic function $\psi(t)$ of a complex variable t with the following properties:

(1) $\psi(t+1) = -\psi(t)$.
(2) $\psi(t+\tau) = \frac{1}{\psi(t)}$.
(3) *The periods 2 and 2τ of $\psi(t)$ implied by (1) and (2) are a basis of the periods in the sense that if γ is a complex number for which $\psi(t+\gamma) = \psi(t)$ for all t, then there are integers m and n for which $\gamma = 2m + 2n\tau$.*
(4) *The only zeros of $\psi(t)$ in the period parallelogram $\{r + s\tau : 0 \le r < 2, 0 \le s < 2\}$ are at $\frac{1}{2}$ and $\frac{3}{2}$. Therefore, the only poles in this parallelogram are at $\frac{1}{2} + \tau$ and $\frac{3}{2} + \tau$.*
(5) $\psi(\frac{\tau}{2}) = 1$.
(6) $\psi(\frac{\tau}{2} - \frac{1}{2}) = i$.
(7) *Properties 1-5 determine $\psi(t)$.*

Proof of convergence. The summand in the sums in the numerator and denominator is $w^{n^2/4}z^{n/2}$ where $w = e^{2\pi i\tau}$ and $z = e^{2\pi it}$, so its modulus is e^k where $k = \frac{n^2}{4}\Re(2\pi i\tau) + \frac{n}{2}\Re(2\pi it) = -n(\frac{n\pi}{2}\cdot\Im\tau + \pi\cdot\Im t)$, which is less than e^{-n} whenever n is large enough that $\frac{n\pi}{2}\cdot\Im\tau + \pi\cdot\Im t > 1$. Therefore, both sums converge to entire functions. (In fact, they converge extremely rapidly when $|t|$ is small and $\Im\tau$ is at all large.) The quotient of these entire functions is meromorphic. □

Lemma 15.2. *The denominator of (15.1) is a constant times the infinite product*

$$\prod_{n=1}^{\infty}(1 + w^{2n-1}z^{-1})(1 + w^{2n-1}z),$$

where $w = e^{2\pi i\tau}$ and $z = e^{2\pi it}$.

Proof of the lemma. This infinite product converges because its nth term has the form $(1 + rw^{2n})(1 + sw^{2n})$ where r and s are fixed and the modulus of w is less than 1. It defines a complex analytic function of z for all z other than 0 and ∞. Let $\sum_{n=-\infty}^{\infty} C_n z^n$ be the Laurent expansion of this function. Changing t to $t + 2\tau$ changes z to $z\cdot e^{2\pi i\cdot 2\tau} = zw^2$ and therefore changes the factors $1 + w^{2n-1}z^{-1}$ to $1 + w^{2n-3}z^{-1}$ and changes the factors $1 + w^{2n-1}z$ to $1 + w^{2n+1}z$, so the new product contains only one factor $1 + w^{-1}z^{-1}$ that was not in the original product and fails to contain just one factor $1 + wz$ of the original product, which implies that

$$\sum_{n=-\infty}^{\infty} C_n\cdot(zw^2)^n = \frac{1 + w^{-1}z^{-1}}{1 + wz}\sum_{n=-\infty}^{\infty} C_n z^n.$$

The factor in front of the sum on the right is $\frac{1}{wz}$, so multiplication by wz gives

$$\sum_{n=-\infty}^{\infty} C_n\cdot w^{2n+1}\cdot z^{n+1} = \sum_{n=-\infty}^{\infty} C_n\cdot z^n.$$

Equating the coefficients in these two Laurent expansions gives $C_{n+1} = C_n \cdot w^{2n+1}$. When D_n is defined to be $C_n w^{-n^2}$ one finds $D_{n+1} = C_{n+1} w^{-n^2-2n-1} = C_n w^{-n^2} = D_n$. Thus, D_n is independent of n—call it D—and the infinite product in the lemma is $\sum_{n=\infty}^{\infty} C_n z^n = D \cdot \sum_{n=-\infty}^{\infty} w^{n^2} z^n = D \cdot \sum_{m \text{ even}} w^{m^2/4} z^{m/2}$, as was to be shown. □

Proof of Properties 1–7. The poles of $\psi(t)$ are at the zeros of the infinite product in the lemma, which occur when $z^{\pm 1} = -w^{2n-1}$; these are the points at which $e^{\pm 2\pi i t} = e^{\pi i + 2\pi i(2n-1)\tau}$, which is to say that $\pm t = \frac{1}{2} + (2n-1)\tau + m$ is the condition for t to be a pole of $\psi(t)$, where m is an integer and n is a positive integer. In conclusion, the poles of $\psi(t)$ are at the points $t = \frac{1}{2} + m + (2n-1)\tau$ for integer m and n, as Property 4 states.

Property 1 follows from the observation that changing t to $t+1$ changes the summand $e^{i\pi(\frac{n^2}{2}\cdot\tau + nt)}$ in (15.1) to $(-1)^n e^{i\pi(\frac{n^2}{2}\cdot\tau+nt)}$, thereby multiplying summands in the numerator by -1 and leaving summands in the denominator unchanged.

If the numerator and denominator of $\psi(t+\tau) = \frac{\sum_{n \text{ odd}} e^{i\pi(\frac{n^2}{2}\cdot\tau+nt+n\tau)}}{\sum_{n \text{ even}} e^{i\pi(\frac{n^2}{2}\cdot\tau+nt+n\tau)}}$ are multiplied by $e^{i\pi\frac{\tau}{2}+t}$, the result is $\frac{\sum_{n \text{ odd}} e^{i\pi(\frac{(n+1)^2}{2}\cdot\tau+(n+1)t)}}{\sum_{n \text{ even}} e^{i\pi(\frac{(n+1)^2}{2}\cdot\tau+(n+1)t)}}$, from which Property 2 follows.

If p is a period of $\psi(t)$, then adding p to a pole of $\psi(t)$ must give a pole of $\psi(t)$, so p must have the form $m + 2n\tau$ for integers m and n. Since Property 1 shows that 1 is not a period of $\psi(t)$, m must be even, and Property 3 follows.

Property 5 can be proved by setting $n = -m-1$ in the numerator of $\psi(\frac{\tau}{2}) = \frac{\sum_{n \text{ odd}} e^{i\pi(\frac{n^2}{2}\cdot\tau+\frac{n\tau}{2})}}{\sum_{n \text{ even}} e^{i\pi(\frac{n^2}{2}\cdot\tau+\frac{n\tau}{2})}} = \frac{\sum_{n \text{ odd}} e^{\frac{i\pi\tau}{2}(n^2+n)}}{\sum_{n \text{ even}} e^{\frac{i\pi\tau}{2}(n^2+n)}}$ to find $\frac{\sum_{m \text{ even}} e^{\frac{i\pi\tau}{2}((-1-m)^2+(-1-m))}}{\sum_{n \text{ even}} e^{\frac{i\pi\tau}{2}(n^2+n)}} = 1$.

Property 6 follows when the same trick is applied to

(15.2)
$$\begin{aligned}\psi(\frac{\tau}{2}-\frac{1}{2}) &= \frac{\sum_{n \text{ odd}} e^{\frac{i\pi\tau}{2}\pi(n^2+n)}\cdot(-i)^n}{\sum_{n \text{ even}} e^{\frac{i\pi\tau}{2}(n^2+n)}\cdot(-i)^n} = \frac{\sum_{m \text{ even}} e^{\frac{i\pi\tau}{2}\pi(m^2+m)}\cdot(-i)^{-1-m}}{\sum_{n \text{ even}} e^{\frac{i\pi\tau}{2}(n^2+n)}\cdot(-i)^n} \\ &= i \cdot \frac{\sum_{m \text{ even}} e^{\frac{i\pi\tau}{2}\pi(m^2+m)}\cdot i^m}{\sum_{n \text{ even}} e^{\frac{i\pi\tau}{2}(n^2+n)}\cdot i^{-n}} = i,\end{aligned}$$

because $i^n = i^{-n}$ when n is even.

If two doubly periodic meromorphic functions have the same poles and zeros, their quotient has no poles or zeros; therefore, their quotient is an analytic function that is defined and nonzero in the entire complex plane. Such a function is represented by a power series with an infinite radius of convergence. In particular, its modulus is bounded on the disk of radius R for any R. For R sufficiently large, the disk includes a period parallelogram, so the modulus of the quotient is bounded on the period parallelogram and therefore bounded on the entire complex plane, which implies the quotient is a constant. In short, the zeros and poles of a doubly periodic function determine the function up to a nonzero constant multiple. Thus, a meromorphic function with Properties 1–4 is $c\psi(t)$ where c is its value at $\frac{\tau}{2}$. □

16. A transcendental equation that solves the parameterization problem

Let a complex number τ in the upper half plane be given; let $\psi(t)$ be the meromorphic function of t, depending on τ, defined in the last section; and let $\phi(t) = \psi(t-\frac{1}{2})$. The functions $\phi(t)^2+\psi(t)^2$ and $1+\phi(t)^2\psi(t)^2$ are doubly periodic with periods 2 and 2τ, and they have the same zeros and poles, namely, double poles where $\phi(t)$ or $\psi(t)$ has a pole (these are the points $t = \frac{m}{2} + (2n+1)\tau$ where m and n are integers, of which there are 4 in each period parallelogram) and zeros at the points where $\psi(t)$ is a power of i (these are the points $t = \frac{m}{2}+\frac{(2n+1)\tau}{2}$ where m and n are integers, of which there are 8 in each period parallelogram). Therefore, their quotient is a constant. At $t = 0$ the first is $\psi(0)^2$ and the second is 1, so the functions $\phi(t)$ and $\psi(t)$ identically satisfy the equation $\phi(t)^2 + \psi(t)^2 = \psi(0)^2 + \psi(0)^2\phi(t)^2\psi(t)^2$.

In other words, *the function $t \mapsto (\phi(t), \psi(t))$ for a given τ maps the complex t-plane in a doubly periodic way onto the Riemann surface $x^2 + y^2 = a^2 + a^2x^2y^2$ for $a = \psi(0)$.*

In this way, the problem of parameterizing $x^2+y^2 = a^2+a^2x^2y^2$—and therefore the problem of parameterizing any elliptic curve when it is regarded as a Riemann surface—reduces to: Given a complex number a with $a^5 \neq a$, find a complex number τ for which

$$a = \frac{\sum_{n \text{ odd}} e^{\frac{i\pi n^2}{2}\cdot\tau}}{\sum_{n \text{ even}} e^{\frac{i\pi n^2}{2}\cdot\tau}}. \tag{16.1}$$

17. A functional equation for ψ

Let $\psi(t)$ be written as $\psi(t, \tau)$ to show its dependence on τ as well as t. The functional equation

$$\psi(\frac{t}{\tau}, -\frac{1}{\tau}) = \frac{1-\psi(t,\tau)}{1+\psi(t,\tau)} \tag{17.1}$$

holds for this function (for all complex t and all τ in the complex upper half plane). The functions on either side of this equation are characterized by the properties that they (1) are doubly periodic and $(2, 2\tau)$ is a basis of the periods, (2) the zeros in the period parallelogram are at $\frac{\tau}{2}$ and $\frac{3\tau}{2}$, (3) the poles in the period parallelogram are at $\frac{\tau}{2}+1$ and $\frac{3\tau}{2}+1$, and (4) the value at $\frac{1}{2}$ is 1.

In the case of $\psi(\frac{t}{\tau}, -\frac{1}{\tau})$, adding 1 to t adds $\frac{1}{\tau}$ to the first argument of the function, which subtracts $-\frac{1}{\tau}$, which takes the function to its reciprocal. In particular, as a function of t it has the period 2. Adding τ to t adds 1 to the first argument of $\psi(\frac{t}{\tau}, -\frac{1}{\tau})$, which changes its sign. In particular, as a function of t it has the period 2τ. Its zeros are the points where $\frac{t}{\tau} = \frac{1}{2}+m+2n(-\frac{1}{\tau})$ for integer m and n, and its poles are the points where $\frac{t}{\tau} = \frac{1}{2}+m+(2n-1)(-\frac{1}{\tau})$, which shows that the function has properties (1)–(3). Finally, its value at $\frac{1}{2}$ is $\psi(\frac{1}{2\tau}, -\frac{1}{\tau}) = \psi(-\frac{1}{2\tau}, -\frac{1}{\tau})^{-1} = 1$.

In the case of $\frac{1-\psi(t,\tau)}{1+\psi(t,\tau)}$, the periods are clearly the same as the periods of $\psi(t, \tau)$. Its value at $t = \frac{1}{2}$ is 1 because $\psi(\frac{1}{2}, \tau) = 0$. The zeros occur where $\psi(t, \tau) = 1$. The two places in the period parallelogram of $\psi(t, \tau)$ where the value 1 occurs are $t = \frac{\tau}{2}$ and $t = \frac{3\tau}{2}$, so (2) holds. Similarly, (3) holds because the places where $\psi(t, \tau) = -1$ are obtained from the places where $\psi(t, \tau) = 1$ by adding 1 to t.

18. The action of the modular group

The *modular group* can be regarded as the group of transformations of the upper half plane generated by $\tau \mapsto \tau + 1$ and $\tau \mapsto -\frac{1}{\tau}$. An element S of the modular group acts on the function $\psi(t, \tau)$, carrying it to $\psi(t, S\tau)$.

Theorem 18.1. *The orbit of $a = \psi(0, \tau)$ under this action of the modular group consists of the complex numbers listed in* (6.1).

Proof. Adding 1 to τ multiplies the summand $e^{i\pi n^2\tau/2}$ in the numerator and denominator of $\psi(0, \tau)$ by $e^{i\pi n^2/2} = i^{n^2}$, which is i when n is odd and 1 when n is even. In short, $\psi(0, \tau + 1) = i\psi(0, \tau)$. On the other hand, the functional equation (17.1) implies that $\psi(0, -\frac{1}{\tau}) = \frac{1-\psi(0,\tau)}{1+\psi(0,\tau)}$. Therefore, the modular group carries $a = \psi(0, \tau)$ to all of its images under compositions of the two fractional linear transformations $a \mapsto ia$ and $a \mapsto \frac{1-a}{1+a}$, which are the 24 values listed in (6.1). □

The kernel of the above homorphism from the modular group to the group of 24 fractional linear transformations of a is a normal subgroup of index 24 in the modular group. Because it contains $\tau \mapsto \tau + 4$, it is what is called the congruence subgroup of level 4, containing those elements of the modular group whose matrix representations[18] are congruent to $\pm I \mod 4$.

19. A fundamental domain for a

A fundamental domain of the action of the group of the cube on the surface of the cube is given by an isosceles triangle formed on a face of the cube by joining the center of the face to the ends of one of its edges. In terms of the correspondence between the surface of a cube and the points of the Riemann sphere as in Section 6, the face of the cube corresponding to 0 is the curvilinear "square" one of whose sides is a segment of the circle through i with center -1, namely, the segment of that circle that lies between the ray from the origin that bisects the fourth quadrant and the ray from the origin that bisects the first quadrant, and whose remaining sides are obtained by rotations of 90° around the origin. Thus, a fundamental domain for the action of the group of the cube on the orbits (6.1) is the region $\mathcal{D}$ bounded by the line segments from 0 to $(1 \pm i)\frac{\sqrt{3}-1}{2}$ and the circular arc joining the two points $(1 \pm i)\frac{\sqrt{3}-1}{2}$ that passes through $\sqrt{2} - 1$. (The region shaded in Figure 1.)

Every orbit (6.1) then contains exactly one point of $\mathcal{D}$, except that points on the boundary of $\mathcal{D}$ are in the same orbits as their complex conjugates. (The points 0 and $\sqrt{2} - 1$ on the intersection of the boundary of $\mathcal{D}$ and the real axis are the only points of their orbits that lie in $\mathcal{D}$.)

To parameterize all elliptic curves, it will suffice to solve the transcendental equation (16.1) for all complex numbers $a \neq 0$ in $\mathcal{D}$.

20. Solution of the transcendental equation

When τ is chosen in such a way that $a = \psi(0)$, the functions $\phi'(t)$ and $\psi(t)(1 - a^2\psi(t)^2)$ have the same zeros and poles, so their quotient is a constant, say

$$\phi'(t) = \mu \cdot \psi(t)(1 - a^2\phi(t)^2).$$

[18] The modular group is the quotient group $SL(2, \mathbf{Z})/\{\pm I\}$ in a natural way.

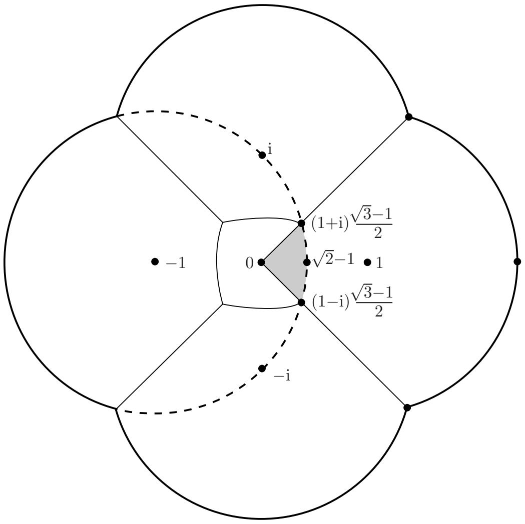

FIGURE 1.

Specifically, the functions on the two sides of this equation have periods generated by $(2, 2\tau)$, have double poles where $\phi(t)$ has poles (these are the points in the period parallelogram where $t = \tau$, $\tau + 1$) and are zero at the places where $\phi(t)$ has double values (these are the points where $\phi(t) = \pm a$, $\pm\frac{1}{a}$, the points in the period parallelogram where $t = \frac{1}{2}$, $\frac{3}{2}$, $\frac{1}{2} + \tau$, and $\frac{3}{2} + \tau$). In the function on the right, the poles of $\psi(t)$ where $t = \frac{1}{2} + \tau$ and $\frac{3}{2} + \tau$ are cancelled by the double zeros of $1 - a^2\phi(t)^2$ at these points, leaving simple zeros at these points in addition to the simple zeros of $\psi(t)$ at the points $t = \frac{1}{2}$, $\frac{3}{2}$.

Therefore, the pullback of the differential $\frac{dx}{y(1-a^2x^2)}$ under the parameterization $t \mapsto (\phi(t), \psi(t))$ is $\mu \cdot dt$, and the integral $\int_{(0,a)}^{(a,0)} \frac{dx}{y(1-a^2x^2)}$ is $\int_0^{\frac{1}{2}} \mu \cdot dt = \frac{\mu}{2}$. In other words, $\mu = 2\int_{(0,a)}^{(a,0)} \frac{dx}{y(1-a^2x^2)}$, where the integral is along the path on the Riemann surface from $(0, a)$ to $(a, 0)$ that is parameterized by ϕ and ψ as t moves along a path in the t-plane from $t = 0$ to $t = \frac{1}{2}$. (The t-plane is simply connected and $\frac{dx}{y(1-a^2x^2)}$ is holomorphic, so the integral is independent of the path.) *When a is in the fundamental domain $\mathcal{D}$* this definite integral can be written as $\int_0^a \frac{dx}{y(1-a^2x^2)}$, where the path of integration is the line segment in the x-plane from 0 to a and where y is the function of x defined on the disk $|x| < |a|$ by the formula $y = \sqrt{\frac{a^2-x^2}{1-a^2x^2}}$ and the condition that $y = a$ when $x = 0$. (The poles of $\frac{a^2-x^2}{1-a^2x^2}$ at $x = \pm\frac{1}{a}$ lie outside the disk when a is in $\mathcal{D}$, and the zeros at $x = \pm a$ are on the boundary of the disk. The integrand has a pole at $x = a$, but the definite integral, though improper, is convergent.) In summary, then, μ is given by the formula

$$\mu = 2\int_0^a \frac{dx}{y(1-a^2x^2)} = 2\int_0^a \frac{dx}{\sqrt{(a^2-x^2)(1-a^2x^2)}}.$$

Similarly, the integral of $\frac{dx}{y(1-a^2x^2)}$ along the Riemann surface from $(0, a)$ to the 4-section point[19] $(i, 1)$, when the path is parameterized by a path in the t-plane from 0 to $\frac{\tau}{2}$, is $\int_0^{\frac{\tau}{2}} \mu\, dt = \frac{\mu\tau}{2}$. Again, when a is in $\mathcal{D}$, this integral is unambiguously indicated by the integral $\int_0^i \frac{dx}{y(1-a^2x^2)}$, because $y = \sqrt{\frac{a^2-x^2}{1-a^2x^2}}$ is determined all along the line segment from 0 to i in the x-plane by the conditions that its value at $x = 0$ be a and that it depend continuously on x. (In short, the line segment avoids the zeros and poles of $\frac{a^2-x^2}{1-a^2x^2}$.)

Therefore, provided a is in $\mathcal{D}$, the required value of τ can be expressed in terms of a as

$$\tau = \frac{\frac{\mu\tau}{2}}{\frac{\mu}{2}} = \frac{\int_0^i \frac{dx}{y(1-a^2x^2)}}{\int_0^a \frac{dx}{y(1-a^2x^2)}} \tag{20.1}$$

where the definite integrals are taken along line segments in the x-plane, the integrand is a at $x = 0$, and the integrand is determined along the paths of integration by continuity.

Because it expresses τ in terms of a, this formula solves the problem.

21. A bisection method

The definite integrals that determine τ in formula (20.1) are of the form

$$\int_{(0,a)}^{(\phi(t),\psi(t))} \frac{dx}{y(1-a^2x^2)}$$

where the path of integration is determined by the parameterization $t \mapsto (\phi(t), \psi(t))$. Although the integral determines a complex number, the direct computation of that number is impractical. It can be made practical by using the fact that $t \mapsto (\phi(t), \psi(t))$ is a group homomorphism to establish the following bisection method.

The homomorphism property of $t \mapsto (\phi(t), \psi(t))$ is the addition formula (3.1), which implies

$$\phi(2t) = \frac{1}{a} \cdot \frac{2\phi(t)\psi(t)}{1 + \phi(t)^2\psi(t)^2}, \qquad \psi(2t) = \frac{1}{a} \cdot \frac{\psi(t)^2 - \phi(t)^2}{1 - \phi(t)^2\psi(t)^2},$$

so $\phi(t)$ and $\psi(t)$ can be found by solving algebraic equations when $\phi(2t)$ and $\psi(2t)$ are known. Specifically, let $(\phi(2t), \psi(2t))$ be written (X, Y) and $(\phi(t), \psi(t))$ be written (x, y); then $a(1 - x^2y^2)Y = y^2 - x^2$, and multiplication of this equation by $1 - a^2x^2$ and use of $y^2(1 - a^2x^2) = a^2 - x^2$ gives $aY(1 - a^2x^2) - aYx^2(a^2 - x^2) = a^2 - x^2 - x^2(1 - a^2x^2)$, which is to say $(aY - a^2)x^4 + 2(1 - a^3Y)x^2 + (aY - a^2) = 0$, from which it is clear that if x is one solution, then the other three are $-x$ and $\pm\frac{1}{x}$. Moreover, once x is known, $y^2 = \frac{a^2-x^2}{1-a^2x^2}$ is known and the equation $X = \frac{1}{a} \cdot \frac{2xy}{1+x^2y^2}$ determines y.

In this way, knowledge of $(\phi(2t), \psi(2t))$ reduces to *four* the possible values of $(\phi(t), \psi(t))$. Algebraically, the four are indistinguishable, but when the path from $(0, a)$ to $(\phi(2t), \psi(2t))$ stays near $(0, a)$, the correct value of $(\phi(t), \psi(t))$ can be found on the basis of topological considerations.

[19] In Section 11, the bisection points were shown to be $(0, -a), (\infty, \pm\frac{1}{a})$. The 4-section points are the 12 points whose doubles are bisection points, namely, $(\pm a, 0), (\pm\frac{1}{a}, \infty), (\pm i, \pm 1)$, and $(\pm 1, \pm i)$.

Let $B = \frac{a^3Y-1}{aY-a^2}$, where $Y = \psi(2t)$, so that the problem is to choose $x = \phi(t)$ from among the four roots of $x^4 - 2Bx^2 + 1$. When t is not too large, Y is near a and $|B|$ is large, so $x^2 = B \pm B\sqrt{1 - B^{-2}}$ can be expanded as a convergent series $B \pm B(1 - \frac{1}{2} \cdot B^{-2} - \cdots)$, which gives two values of x^2, one near $\frac{1}{2}B^{-1}$ and the other near $2B$. When t is not too large, x^2 will be small and therefore not near $2B$, so x^2 must be given by

$$x^2 = \frac{1}{2} \cdot B^{-1} + \frac{1}{2} \cdot \frac{1}{4} \cdot B^{-3} + \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{3}{6} \cdot B^{-5} + \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{3}{6} \cdot \frac{5}{8} \cdot B^{-7} + \cdots .$$

One can choose between the square roots of this known x^2 on the basis of the fact that when t is not too large, x must be near $\frac{X}{2}$, after which y is determined as above.

Because

$$\int_{(0,a)}^{(\phi(2t),\psi(2t))} \frac{dx}{y(1-a^2x^2)} = \int_0^{2t} \mu\, dt = 2\int_0^t \mu\, dt = 2\int_{(0,a)}^{(\phi(t),\psi(t))} \frac{dx}{y(1-a^2x^2)},$$

this determination of $(\phi(t), \psi(t))$ "bisects" the integral on the left in the sense that it reduces the evaluation of the definite integral to the evaluation of a definite integral over "half" the path of integration.

22. Numerical solution for τ given a

When the definite integrals in the numerator and denominator of (20.1) are both bisected by the method of Section 21, their ratio will still be τ, but the definite integrals will be over shorter paths and therefore easier to evaluate. After n such bisections, the formula becomes

$$\tau = \frac{\int_0^{x'_n} \frac{dx}{y(1-a^2x^2)}}{\int_0^{x_n} \frac{dx}{y(1-a^2x^2)}} \tag{22.1}$$

where $x_0 = a$, x_1, x_2, ... is the sequence of x-coordinates of the points (x_k, y_k) obtained by starting with the direct path from $(0, a)$ to $(a, 0)$ and repeatedly "bisecting," and where $x'_0 = i$, x'_1, x'_2, ... is the analogous sequence of x-coordinates of points (x'_k, y'_k) obtained by starting with the direct path from $(0, a)$ to $(i, 1)$ and "bisecting" it. For large n, the integrals are easy to evaluate because the paths of integration are extremely short.

For example, when a is $.4 + .1i$, which lies in $\mathcal{D}$, implementation of the above method on a programmable calculator gives the sequences (x_k, y_k) and (x'_k, y'_k) shown in Table 1. The values have been rounded to 4 places but were computed with greater accuracy. Note that the y's approach a and each x is roughly half its predecessor.

To a first approximation, the integrand is constant over these short paths of integration, and the constant is the same $\frac{1}{a}$ in the numerator and denominator, so to a first approximation[20] $\tau \approx \frac{x'_4}{x_4} \approx .1596 + 1.007i$. Direct evaluation of $\psi(0)$ for

[20] The integral $\frac{\pi}{2} = \int_0^1 \frac{dx}{\sqrt{1-x^2}}$ can be computed in an analogous way. For small values of x, the doubling formulas $X = 2xy$, $Y = y^2 - x^2$ for the sine and cosine imply $2\int_0^x \frac{dx}{\sqrt{1-x^2}} = \int_0^X \frac{dx}{\sqrt{1-x^2}}$ where $\sqrt{1-X^2} = Y = y^2 - x^2$ and $\sqrt{1-x^2} = y$. Thus, $\sqrt{1-X^2} = 1 - 2x^2$, so X determines x via $x = \sqrt{\frac{1-\sqrt{1-X^2}}{2}}$. Application of this method to $\frac{\pi}{2} = \int_0^1 \frac{dx}{\sqrt{1-x^2}}$ five times,

Table 1.

k	x_k	y_k		x'_k	y'_k
0	$.4+.1i$	0		i	1
1	$.2832+.0717i$	$.2832+.0717i$		$-.0246+.3658i$	$.5355+.0864i$
2	$.1534+.0392i$	$.3697+.0928i$		$-.0149+.1689i$	$.4328+.0972i$
3	$.0782+.0200i$	$.3924+.0982i$		$-.0077+.0828i$	$.4081+.0993i$
4	$.0393+.0101i$	$.3981+.0995i$		$-.0039+.0412i$	$.4020+.0998i$

this τ—which is easy because of the rapid convergence of the series in the numerator and denominator—verifies that it is near $a = .4 + .1i$.

Closer approximations to τ can be found by further bisection and by expanding the integrals in infinite series in their upper limits of integration, but because $\psi(0)$ is easy to compute with great accuracy, a very rough approximation to τ is all that is needed to find arbitrarily close approximations to a solution of $\psi(0) = a$—in other words, to determine τ as a complex number—using simple interpolation. For example, if one starts with so rough a guess as $\tau = .1 + i$ from above and tries $\tau = .9i$ and $\tau = .1 + i$, one finds $\psi(0) = .483$ and $.410 + .064i$, respectively. Since the latter is closer to $.4 + .1i$, let τ_0 be the former and τ_1 the latter, and use the approximation $\frac{\tau_{n+1}-\tau_n}{\psi_{n+1}-\psi_n} \approx \frac{\tau_n-\tau_{n-1}}{\psi_n-\psi_{n-1}}$ together with the desired value $\psi_{n+1} = a$ to find the iteration formula

$$\tau_{n+1} = \tau_n + (a - \psi_n) \cdot \frac{\tau_n - \tau_{n-1}}{\psi_n - \psi_{n-1}},$$

which for the above τ_0 and τ_1 gives $\tau_3 \approx .15789 + 1.00403i$, for which $\psi(0)$ is very near $.4 + .1i$, and further iterations converge to even better approximations.

About the author

Harold M. Edwards is Professor Emeritus at New York University. He has received the Whiteman and Steele Prizes of the AMS and is the author of seven books: *Advanced Calculus, Riemann's Zeta Function, Fermat's Last Theorem, Galois Theory, Divisor Theory, Linear Algebra,* and *Essays in Constructive Mathematics.*

References

1. N. H. Abel, *Recherches sur les fonctions elliptiques*, Crelle, vols. 2, 3, Berlin, 1827, 1828; Oeuvres, I, pp. 263–388.
2. N. H. Abel, *Mémoire sur une propriété générale d'une classe très-étendue de fonctions transcendantes*, Mémoires présenteés par divers savants à l'Académie des sciences, Paris, 1841. Also *Oeuvres Complètes*, vol. 1, 145–211.
3. C. Chevalley, *Introduction to the Theory of Algebraic Functions of One Variable,* Mathematical Surveys, No. 6, AMS, New York, 1951. MR0042164 (13:64a)
4. H. M. Edwards, *Essays in Constructive Mathematics*, Springer, New York, 2004. MR2104015 (2005h:00010)

for example, gives $\frac{\pi}{64} = \int_0^x \frac{dx}{\sqrt{1-x^2}}$ where $x = .049067674328\ldots$; the infinite series expansion $\int \frac{dx}{\sqrt{1-x^2}} = x + \frac{1}{2} \cdot \frac{x^3}{3} + \frac{1\cdot 3}{2\cdot 4} \cdot \frac{x^5}{5} + \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6} \cdot \frac{x^7}{7} + \cdots$ of the integral then gives π to 10 decimal places.

5. L. Euler, *Observationes de Comparatione Arcuum Curvarum Irrectificabilium*, Novi Comm. Acad. Sci. Petropolitanae, vol. 6, pp. 58–84, 1761, Opera, ser. 1, vol. 20, pp. 80–107, Eneström listing 252.
6. C. F. Gauss, *Werke,* Vol. 3, p. 404.
7. A. Hurwitz and R. Courant, *Vorlesungen über allgemeine Funktionentheorie und elliptische Funktionen*, Springer, Berlin, 1922, 1925, 1964. MR0173749 (30:3959)
8. K. Ireland and M. Rosen, A Classical Introduction to Modern Number Theory, Springer, New York, 1990. MR1070716 (92e:11001)
9. C. G. J. Jacobi, *Fundamenta Nova Theoriae Functionum Ellipticarum*, Regiomonti (Königsberg), 1829 (Math. Werke, vol. 1, pp. 49–241).
10. A. W. Knapp, *Elliptic Curves*, Princeton Univ. Press, Mathematical Notes 40, 1992. MR1193029 (93j:11032)
11. B. Riemann, *Über die Anzahl der Primzahlen unter einer gegebenen Grösse*, Monatsber. der Berliner Akad., Nov. 1859; *Werke*, 145–153.
12. I. R. Shafarevich, *Basic Algebraic Geometry*, Springer, Berlin, 1974. MR0366917 (51:3163)
13. J. H. Silverman and J. Tate, *Rational Points on Elliptic Curves*, Springer, New York, 1992. MR1171452 (93g:11003)

Department of Mathematics, New York University, 251 Mercer Street, New York, New York 10012