MODULARITY OF ELLIPTIC CURVES, AS THEY RELATE TO THE PROOF OF FERMAT'S LAST THEOREM

Jenna Le
Freshman Seminar 21n
Prof. William Stein
5/17/03

[Note: This paper is intended to be understandable to anyone with a knowledge of high-school mathematics, multivariable calculus, and linear algebra.]

I. Introduction to elliptic curves

Let F(x,y) be a degree-three polynomial in two variables, x and y. A curve with the equation F(x,y)=0 is an *elliptic curve* if and only if it contains at least one point with rational coefficients, and the partial derivatives of F are never simultaneously equal to 0. Elliptic curves are exciting to study because they are the next simplest kind of curve after lines and conics. People know almost everything there is to know about lines and conics, but there are still many unsolved problems dealing with elliptic curves. Another reason why elliptic curves are interesting is because the points on any elliptic curve constitute a *group*[1].

By doing certain changes of variables, every elliptic curve with rational coefficients can be expressed in the form $y^2 = x^3 + ax^2 + bx + c$. The changes of variables that are necessary to convert an arbitrary cubic equation to this form are very tedious, so I will not elaborate on them.

More generally, every elliptic curve can be expressed in the form $y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6$. An equation of this form is called a *Weierstrass equation*. For a given elliptic curve E, the Weierstrass equation with the smallest discriminant is called the *minimal Weierstrass equation* of the elliptic curve E. This equation is sometimes used instead of the equation of the form $y^2 = x^3 + ax^2 + bx + c$.

I will now describe the group law of the points on an elliptic curve E that is expressed in the form $y^2 = x^3 + ax^2 + bx + c$. If P and Q are two points on E and if *l* is the line joining them, then P+Q is defined to be the reflection across the x-axis of the third point at which *l* intersects E. Since we can tell by looking at the equation of E that E is symmetric across the x-axis, it is clear that P+Q lies on E. This proves that the group has *closure*. The *identity element* of the group is the "point at infinity" (the point where all vertical lines meet), which we denote by the letter $\mathbf{0}$. The *inverse* of the point (x,y) is the point (x,-y).

[1] A *group* G is a set of objects endowed with one operation (e.g., addition or multiplication) and satisfying four properties: (1) G contains an *identity element*, (2) G contains the *inverse* of every object it contains, (3) the group operation of G is *associative*, and (4) G is *closed* under its group operation. For future reference, a subset of a group that is in itself a group is called a *subgroup*.

II. Reducing an elliptic curve modulo p and computing its conductor

In this paper, we are concerned with elliptic curves whose coefficients are rational numbers. By multiplying the entire equation by the coefficents' greatest common denominator, we come up with a cubic whose coefficients are integers. If we like, we can reduce all these coefficients modulo some prime number p, to see what happens. If E_p has no *singularities* (points at which both partial derivatives are zero), then p is called a "prime of good reduction" for the equation of E that we are using. If E_p has one or more singularities, then p is called a "prime of bad reduction" for the equation of E that we are using. There are two types of singularities that primes of bad reduction might cause, *nodes* (places where the curve crosses itself) and *cusps*.

It is easy to show that an elliptic curve with equation $y^2 = x^3 + ax^2 + bx + c$ has singularities if and only if $f(x) = x^3 + ax^2 + bx + c$ has multiple roots, which is the same thing as saying that the *discriminant* of f is 0. Since only finitely many primes can divide the discriminant, there are only finitely many primes p such that the discriminant is equal to 0 mod p. Therefore, every elliptic curve E has only finitely many primes of bad reduction.

If we are using a minimal Weierstrass equation to define E, we can associate to E an integer that is called the *conductor* of E. The conductor of E is an integer whose only prime divisors are E's primes of bad reduction. If E_p has a node, then the conductor of E is divisible by p but not by p^2. If E_p has a cusp and $p>3$, then the conductor of E is divisible by p^2 (but not by p^3). However, if E_p has a cusp for $p = 2$ or 3, then the problem is not so simple, and we must resort to a much more complicated algorithm (discovered by John Tate) to compute the conductor of E. Of course, we can get a computer program like PARI to perform Tate's algorithm for us, since that's what computers are for.

To use PARI to find the conductor of a curve with equation $y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6$, enter the command "ellglobalred(ellinit([a_1,a_2,a_3,a_4,a_6])." Your computer will then output a number, followed by a vector and another number. For our purposes, the vector and the second number will be irrelevant; the first number will be the conductor of the curve.

III. The Shimura-Taniyama Conjecture

Now, one thing that you immediately wonder after you have reduced E modulo p is: What is the relationship between p and N_p, the number of points mod p on E_p? In trying to determine what this relationship is, mathematicians looked at the sequence of numbers $\{a_p(E)\}_p$, where $a_p(E) = p - N_p + 1$, and tried to find a pattern in the sequence. No pattern was immediately obvious, and mathematicians were stymied by this problem for a very long time. Then, in 1955, the brilliant Japanese mathematician Yutaka Taniyama conjectured that, for every elliptic curve E, there exists a *modular form* such that the sequence $\{a_p(E)\}_p$ is related to the coefficients in the q-expansion of that modular form. At first, nobody believed

Taniyama because modular forms and elliptic curves belong to two completely different branches of mathematics (complex analysis and number theory, respectively) and seem to have absolutely nothing to do with each other. However, an overwhelming amount of examples supporting Taniyama's claim quickly piled up, and people began to acknowledge that he might be right. Goro Shimura helped Taniyama refine his conjecture, and it came to be known as the *Shimura-Taniyama Conjecture.*

Incidentally, Andre Weil helped publicize the conjecture by mentioning the idea in a paper he published in 1967. (Towards the end of the paper, he naively asked the reader to prove the conjecture as an exercise.) For this reason, some people call it the *Shimura-Taniyama-Weil Conjecture*.

To understand just how brilliant Taniyama must have been to espy this deep connection between modular forms and elliptic curves, it is necessary to know a bit about modular forms.

IV. Modular forms and their q-expansions

An *unrestricted modular form* is an analytic function whose domain consists of the complex numbers whose imaginary parts are positive. An unrestricted modular form's range is the set of all complex numbers. f(z) is an *unrestricted modular form of weight w* if and only if $f((az+b)/(cz+d)) = (cz+d)^w f(z)$ for all 4-tuples (a,b,c,d) such that a, b, c, and d are the entries of a 2x2 integer matrix whose determinant is 1. If we plug a=b=d=1 and c=0 into the above equation, we discover that f(z + 1) = f(z). This means that f is a periodic function (its period is 1). Therefore, like all periodic functions, f can be expressed as a Fourier series. A *Fourier series* is a way of expressing a periodic function as an infinite series of terms that have to do with sines and cosines.

Let z = J+Ki. Now let's expand f(z) into a Fourier series in the variable J. We get:

$$f(z) = \sum_{n=-\infty}^{\infty} c_n(K) e^{2\pi i n J},$$

where $c_n(K)$ = the integral from –½ to ½ of $f(J+Ki)e^{-2\pi i n J}dJ$.

As Knapp shows in his book *Elliptic Curves*, $e^{2\pi i n J} = e^{2\pi i n(z-Ki)} = e^{2\pi n(iz+K)} = e^{2\pi n K + 2\pi n i z} = e^{2\pi n K} e^{2\pi n i z}$. Thus, f(z) can be rewritten like this:

$$\infty$$

$$f(z) = \sum_{n=-\infty} c_n(K) e^{2\pi nK} e^{2\pi niz}.$$

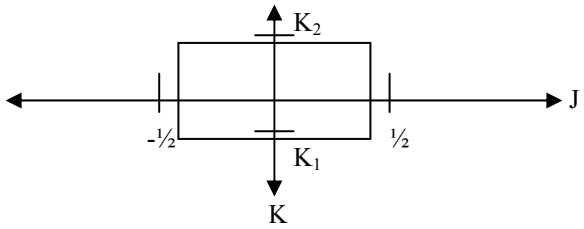
All we need to do now is show that $c_n(K)e^{2\pi nK}$ is a constant in terms of K. After that, we will be able to give this constant a nice name like $a_n(f)$, and then we can express f(z) in the following form, which is called the *q-expansion* of f:

$$f(z) = \sum_{n=-\infty}^{\infty} a_n(f) e^{2\pi niz}.$$

An unrestricted modular form for which $a_n(f) = 0$ for all negative numbers n is called a *modular form*.

Below, I have paraphrased Knapp's proof of why $c_n(K)e^{2\pi nK}$ is a constant in terms of K, as it appears in his book *Elliptic Curves* (pp. 224-225). It is rather technical, so, if you don't like calculus, feel free to skip over it and go directly to Part V.

To show that $c_n(K)e^{2\pi nK}$ is a constant in terms of K, Knapp first rewrites it as the integral from $-½$ to ½ of $f(J+Ki)e^{-2\pi in(J+Ki)}dJ$. He then tells us to visualize a rectangle whose horizontal sides stretch from J= -½ to ½ and whose vertical sides stretch from $K = K_1$ to K_2, where K_1 and K_2 are arbitrary real numbers.



$c_n(K)e^{2\pi nK}$ is clearly the bottom part of the line integral of $f(J+Ki)e^{-2\pi in(J+Ki)}$ around this rectangle. Note that one of the vertical sides of the rectangle is the integral from K_2 to K_1 of $f(-½+Ki)e^{-2\pi in(-½+Ki)}dK$, while the other vertical side is the integral from K_1 to K_2 of $f(½+Ki)e^{-2\pi in(½+Ki)}dK$. Since f(z) = f(z+1), f(-½+Ki) = f(½+Ki). Also, $e^{-2\pi in(½+Ki)}$ =

$= e^{-2\pi in(-½+Ki)}e^{-2\pi in} = e^{-2\pi in(-½+Ki)} * 1 = e^{-2\pi in(-½+Ki)}$. Therefore, the two vertical sides cancel each other out. Since the line integral around the whole rectangle is 0 (by a theorem of Cauchy), the bottom part of the line integral must equal the top part. In other words,

The integral from –½ to ½ of $f(J+K_1i)e^{-2\pi in(J+K_1i)}dJ$ =

= the integral from –½ to ½ of $f(J+K_2i)e^{-2\pi in(J+K_2i)}dJ$.

In other words,

$c_n(K_1)e^{2\pi nK_1} = c_n(K_2)e^{2\pi nK_2}$.

Since K_1 and K_2 are arbitrary real numbers, this means that $c_n(K)e^{2\pi nK}$ has the same value for all real numbers K. We are done going through Knapp's proof of why $c_n(K)e^{2\pi nK}$ is a constant in terms of K. Now we are ready to restate the Shimura-Taniyama Conjecture using the vocabulary of modular forms and their q-expansions.

V. The Shimura-Taniyama Conjecture again

Before I can restate the Shimura-Taniyama Conjecture, I should tell you that the 2x2 integer matrices that we use to define modular forms can be restricted to subgroups of $SL_2(\mathbf{Z})$ called *Hecke subgroups*. ($SL_2(\mathbf{Z})$ is the group of 2x2 integer matrices with determinant 1.) The Hecke subgroup $\Gamma_0(N)$ consists of all 2x2 integer matrices whose lower-left-hand entries are divisible by N. The modular forms that are defined by the Hecke subgroup $\Gamma_0(N)$ are said to have *level* N.

Recall that, in Part II, we defined $a_p(E)$ to be $p - N_p + 1$. Also, recall that we defined $a_n(f)$ to be the n^{th} coefficient of the q-expansion of the modular form f. The *Shimura-Taniyama Conjecture* states that every elliptic curve E has a weight-two modular form f associated with it such that $a_p(E) = a_p(f)$, for all "primes of good reduction" p. Furthermore, the conductor of E equals the level of f. Amazing but true! This tough conjecture was not fully proven until 1999, when Christophe Breuil, Brian Conrad, Fred Diamond, and Richard Taylor stepped forward with the long-awaited and long-sought proof.

The shorthand version of the Shimura-Taniyama Conjecture is, "All elliptic curves are modular."

> *Definition.* An elliptic curve is *modular* if and only if it has a weight-two modular form f associated with it such that $a_p(E) = a_p(f)$ for all primes p of good reduction. Furthermore, the conductor of E equals the level of f.

Deep and wonderful as the Shimura-Taniyama Conjecture is in its own right, it is most famous for the important role it played in Andrew Wiles's 1994 proof of Fermat's Last Theorem.

VI. Fermat's Last Theorem and the Frey curve

As we all know, *Fermat's Last Theorem* asserts that if n is an integer greater than or equal to 3, then there exists no triple of integers (a,b,c) that satisfies both the equation $a^n + b^n = c^n$ and the inequality $abc \neq 0$. (*Note:* we are allowed to require that a, b, and c be relatively prime. If a, b, and c shared a prime factor p, then we could get rid of p by dividing the whole equation $a^n + b^n = c^n$ by p^n.) Although Fermat claimed to have come up with a proof of this theorem, no proof was found among his papers after he died. For centuries, mathematicians all over the world tried to prove Fermat's Last Theorem themselves, but none succeeded.

Then, in 1985, Gerhard Frey, a German mathematician, suggested a method of proof that involves elliptic curves. Frey pointed out that if there existed a triple of integers (a,b,c) such that $a^n + b^n = c^n$ for $n \geq 3$ and $abc \neq 0$, then the curve defined by the equation $y^2 = x(x - a^n)(x + b^n)$ would be an elliptic curve. This assertion is proven in the following paragraph.

Since $abc \neq 0$, we know that $a \neq 0$, $b \neq 0$, and $c \neq 0$. This implies that $a^n \neq 0$, $b^n \neq 0$, and $c^n \neq 0$. Since $c^n = a^n + b^n$, this is the same as saying that $a^n \neq 0$, $b^n \neq 0$, and $a^n + b^n \neq 0$. Another way of saying this is: $a^n \neq 0$, $b^n \neq 0$, and $a^n \neq -b^n$. Since 0, a^n, and $-b^n$ are the three roots of the equation $y^2 = x(x - a^n)(x + b^n)$, our three inequalities clearly imply that the three roots of this equation are distinct. As we said earlier, a curve with three distinct roots (i.e., a curve with no multiple roots) has no singularities. Therefore, the curve $y^2 = x(x - a^n)(x + b^n)$, called the *Frey curve*, is an elliptic curve.

Notice that the Frey curve is purely hypothetical: *it does not really exist*. However, it *would* exist if Fermat's Last Theorem were not true. Therefore, proving that the Frey curve does not exist is equivalent to proving Fermat's Last Theorem.

VII. The contributions of Serre, Ribet, and Wiles

Let's now return to the topic of modularity for a moment. It is obviously impossible for an elliptic curve to be both modular and nonmodular: an elliptic curve must be one or the other, but not both. (The Shimura-Taniyama conjecture states that all elliptic curves are modular, but when Frey invented the Frey curve in 1985, the Shimura-Taniyama conjecture had not been proven yet.) If someone were to prove that the Frey curve is both modular and nonmodular, that would immediately imply that the Frey curve does not exist. And if the Frey curve does not exist, then Fermat's Last Theorem must be true.

In fact, this is exactly how Fermat's Last Theorem was proven. In 1986, with some help from Jean-Pierre Serre, Kenneth Ribet proved that the Frey curve is nonmodular. Then, in 1994, to great acclaim, the Englishman Andrew Wiles (with help from his student Richard Taylor) proved that the Frey

curve is modular. Together, Ribet's and Wiles's proofs show that the Frey curve does not exist. Thus, the combined efforts of Taniyama, Shimura, Frey, Serre, Ribet, and Wiles, Taylor, and many others resulted in a proof of Fermat's Last Theorem over three centuries after Fermat's death.

Actually, Wiles did more than prove that the Frey curve is modular; he proved that *all* semistable elliptic curves are modular. By definition, an elliptic curve is *semistable* if and only if its conductor has no perfect square divisors. If we look again at the definition of a conductor, we can see that an elliptic curve E is semistable if and only if E_p does not have a cusp for any prime p>3. Since having a cusp is the same thing as having a triple root, an alternative definition of a semistable elliptic curve is "an elliptic curve E for which E_p does not have a triple root for any prime p>3."

It is easy to show that the Frey curve is semistable. First, let E be the Frey curve. Next, assume E is not semistable. Then E_p has a triple root for some prime p>3. Since 0, a^n, and b^n are the roots of the Frey curve, E_p has a triple root if and only if $0 \equiv a^n \equiv -b^n \pmod p$. If $0 \equiv a^n \equiv -b^n \pmod p$, then p divides both a^n, $-b^n$, and c^n (because $c^n = a^n + b^n$). If p divides a^n, $-b^n$, and c^n, then p divides a, b, and c. This contradicts our original assumption that a, b, c are relatively prime. Hence, the Frey curve is semistable.

A great many elliptic curves are semistable. So, by proving that all semistable elliptic curves are modular, Wiles made significant progress in humankind's search for a proof of the Shimura-Taniyama Conjecture.

*VIII. The Galois group Gal(L/**Q**) and how it acts on E[m]*

Before we could ever hope to understand Ribet's proof of the Frey curve's nonmodularity, we first need to know a bit about Galois groups. In this section, I will present the definitions of the Galois group Gal(L/**Q**) and the group E[m]. Afterwards, I will prove that if T is a function in Gal(L/**Q**), then T maps E[m] to E[m]. Additionally, I will prove that T preserves the group law of the elliptic curve E.

As you know, the letter **Q** represents the field of rational numbers. If r is a root of a polynomial whose coefficients are in **Q**, then r is called an *algebraic number*. The field of algebraic numbers is called the *algebraic closure of **Q*** and will be represented in this paper by the letter L. Observe that L contains **Q**.

Let a and b be any two elements of L, and let c be any element of **Q**. Consider a bijective map T: $L \rightarrow L$ that satisfies the following properties: $T(a)+T(b)=T(a+b)$, $T(ab)=T(a)T(b)$, and $T(c)=c$. It it is easy to show that the set of all functions T that satisfy these properties is a *group*, which we call the *Galois group* Gal(L/**Q**)[2]. Notice that the properties imply $(T(a))^{-1} = T(a^{-1})$.

Now let's go back to looking at an elliptic curve E whose equation is of the form $y^2 = x^3 + ax^2 + bx + c$. Fix a positive integer m; then consider all points P on E for which $mP = \mathcal{O}$. A point P that

[2] In general, if E is a field containing the field F, the *Galois group* Gal(E/F) is the group of all bijective maps U: $E \rightarrow E$ such that $\forall$ a,b $\in$ E and $\forall$ c $\in$ F, the following properties hold: U(a)+U(b)=U(a+b), U(ab)=U(a)U(b), and U(c)=c.

satisfies this description is called an *m-division point*. The set of all m-division points is denoted E[m]; it is a subgroup of E. Notice that the coordinates of points in E[m] need not be real: often, they are complex.

Since the coordinates of every point in E[m] are algebraic numbers[3], each function T in Gal(L/**Q**) maps the points in E[m] to other points with coefficients in L. It is clear that the image points lie on the curve E because

$$\begin{aligned}(T(x))^3 + a(T(x))^2 + bT(x) + c &= T(x^3) + aT(x^2) + bT(x) + c \\ &= T(x^3) + T(a)T(x^2) + T(b)T(x) + T(c) \\ &= T(x^3 + ax^2 + bx + c) \\ &= T(y^2) \\ &= (T(y))^2.\end{aligned}$$

The same sort of reasoning can be used to show that, in general, given a polynomial g with rational coefficients, $g(x,y) = 0$ implies $g(T(x),T(y)) = 0$. In other words, if the point $P=(x,y)$ lies on the curve defined by the equation $g(x,y) = 0$, then so does the point $T(P)=(T(x),T(y))$.

It is possible to use the geometric description of the group law that was given in Part I to find explicit formulas for the coordinates of P+Q for all points P and Q. According to these explicit formulas, each of the coordinates of mP can be expressed as the quotient of two polynomials with rational coefficients. For example, the y-coordinate of mP can be expressed as $f(x,y)/g(x,y)$, where f and g are polynomials with rational coefficients. Clearly, $mP = \mathcal{O}$ implies that $g(x,y) = 0$. But $g(x,y) = 0$ implies that $g(T(x),T(y)) = 0$, which in turn implies that $mT(P) = \mathcal{O}$. Hence, if P is in E[m], then so is T(P). In other words, not only do the image points lie on the curve E (as we showed in the previous paragraph), but in fact they are elements of the subgroup E[m].

We have just shown that each function T in Gal(L/**Q**) maps E[m] to E[m]. Now we will show that each function T in Gal(L/**Q**) preserves the group law of E[m]. The explicit formulas tell us that if $P = (x_1,y_1)$, $Q = (x_2,y_2)$, and $P+Q = (x_3,y_3)$, then $x_3 = j(x_1,y_1,x_2,y_2) / k(x_1,x_2)$, where j and k are polynomials with rational coefficients. So, the x-coordinate of $T(P)+T(Q)$ is $j(T(x_1),T(y_1),T(x_2),T(y_2)) / k(T(x_1),T(x_2))$. However,

$$j(T(x_1),T(y_1),T(x_2),T(y_2)) / k(T(x_1),T(x_2)) = T(j(x_1,y_1,x_2,y_2)) / T(k(x_1,x_2))$$

[3] This is true because, according to the explicit formulas, the y-coordinate of mP can be expressed as $f(x,y)/g(y)$, where f and g are polynomials with rational coefficients and g depends *only* on the y-coordinate of P. $mP = \mathcal{O}$ implies that $g(y) = 0$, which implies that the y-coordinate of a point in E[m] is an algebraic number. And if the y-coordinate is an algebraic number, then so is the x-coordinate.

$$= T(j(x_1,y_1,x_2,y_2)) * (T(k(x_1,x_2)))^{-1}$$
$$= T(j(x_1,y_1,x_2,y_2)) * T(k(x_1,x_2)^{-1})$$
$$= T(j(x_1,y_1,x_2,y_2) * k(x_1,x_2)^{-1})$$
$$= T(j(x_1,y_1,x_2,y_2) / k(x_1,x_2))$$
$$= T(x_3).$$

So, the x-coordinate of T(P)+T(Q) is $T(x_3)$.

Similarly, it can be shown that the y-coordinate of T(P)+T(Q) is $T(y_3)$. Thus, $T(P)+T(Q) = (T(x_3),T(y_3)) = T(P+Q)$, which means that T preserves the group law of E. We have hereby finished proving everything we set out to prove in this section.

*IX. A matrix representation of Gal(L/**Q**)*

Every elliptic curve E is associated with a unique lattice L in the complex plane. (The method by which one finds the lattice L is too complicated to be included in this paper.) The lattice L is generated by two complex numbers σ and τ: i.e., L is the set of all complex numbers of the form $a\sigma+b\tau$, where a and b are integers.

A function called the *Weierstrass ℘ function* defines a one-to-one correspondence between the set $S = \{a\sigma+b\tau$, where a and b are real numbers between 0 and 1$\}$ and the points with complex coefficients on E. The set S can be pictured as a parallelogram in the complex plane, whose vertices are located at the points 0, σ, τ, and σ+τ. In their book *Rational Points on Elliptic Curves* (pp. 43-45), Silverman and Tate show how this parallelogram can be used to prove that E[m] is the direct product of two cyclic groups of order m. I do not feel like paraphrasing this proof, so if you want to see it, pick up a copy of Silverman's and Tate's book.

What do we mean when we say that E[m] is the direct product of two cyclic groups? We mean that the entire group E[m] can be generated by just two of its elements. Two elements of E[m], ω_1 and ω_2, are *generators* of E[m] if and only if, for all m-division points P, there exist a and b in (**Z**/m**Z**) such that $P = a\omega_1 + b\omega_2$.

Let T be an element of Gal(L/**Q**). In Part VIII, we showed that if P is in E[m], so is T(P). Hence, for all P in E[m], there exist a and b in (**Z**/m**Z**) such that $T(P) = a\omega_1 + b\omega_2$. In particular, there exist a_T, b_T, c_T, and d_T in (**Z**/m**Z**) such that $T(\omega_1) = a_T\omega_1 + b_T\omega_2$ and $T(\omega_2) = c_T\omega_1 + d_T\omega_2$. These two equations can be expressed in matrix form, like this:

$$\begin{pmatrix} & \end{pmatrix} \quad \begin{pmatrix} & \end{pmatrix} \begin{pmatrix} \\ \end{pmatrix}$$

$$\begin{matrix} T(\omega_1) \\ \\ T(\omega_2) \end{matrix} = \begin{matrix} a_T & b_T \\ \\ c_T & d_T \end{matrix} \begin{matrix} \omega_1 \\ \\ \omega_2 \end{matrix} .$$

Let P and Q be two arbitrary points in E[m]. We know there exist e, f, g, and h in (**Z**/m**Z**) such that

$$\begin{pmatrix} P \\ \\ Q \end{pmatrix} = \begin{pmatrix} e & f \\ \\ g & h \end{pmatrix} \begin{pmatrix} \omega_1 \\ \\ \omega_2 \end{pmatrix} .$$

Therefore,

$$\begin{pmatrix} T(P) \\ \\ T(Q) \end{pmatrix} = \begin{pmatrix} e & f \\ \\ g & h \end{pmatrix} \begin{pmatrix} T(\omega_1) \\ \\ T(\omega_2) \end{pmatrix} = \begin{pmatrix} e & f \\ \\ g & h \end{pmatrix} \begin{pmatrix} a_T & b_T \\ \\ c_T & d_T \end{pmatrix} \begin{pmatrix} \omega_1 \\ \\ \omega_2 \end{pmatrix} .$$

We say that $\begin{pmatrix} a_T & b_T \\ \\ c_T & d_T \end{pmatrix}$ is the *matrix representation* of the transformation T.

Notice that, since T is a linear transformation by definition, its matrix representation is determined by the way it acts on the "basis" of (**Z**/m**Z**) x (**Z**/m**Z**): i.e., the way it acts on ω_1 and ω_2.

We have just shown that every element T of Gal(L/**Q**) can be represented by a 2x2 matrix with entries in (**Z**/m**Z**). Let's call that matrix $\rho_{E,m}(T)$. Now we will show that if S and T are two elements of Gal(L/**Q**), then $\rho_{E,m}(S \circ T)$ is equal to the matrix product of $\rho_{E,m}(S)$ and $\rho_{E,m}(T)$.

$$\begin{pmatrix} \\ \\ \end{pmatrix} \begin{pmatrix} \\ \\ \end{pmatrix} \qquad \begin{pmatrix} \\ \\ \end{pmatrix}$$

$$a_{S\circ T} \quad b_{S\circ T} \quad \omega_1 \qquad\qquad S(T(\omega_1))$$

$$\qquad\qquad\qquad\quad = \qquad\qquad$$

$$c_{S\circ T} \quad d_{S\circ T} \quad \omega_2 \qquad\qquad S(T(\omega_2))$$

$$= \begin{pmatrix} a_S & b_S \\ c_S & d_S \end{pmatrix} \begin{pmatrix} T(\omega_1) \\ T(\omega_2) \end{pmatrix}$$

$$= \begin{pmatrix} a_S & b_S \\ c_S & d_S \end{pmatrix} \begin{pmatrix} a_T & b_T \\ c_T & d_T \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}. \square$$

One thing remains to be pointed out before we can go on to the next section. It is a well-known fact of linear algebra that the trace of a matrix is invariant under a change of basis. Hence, the trace of the matrix $\rho_{E,m}(T)$ is, in an important way, characteristic of T.

X. The Frobenius automorphism σ_p

Let H be the set of all functions S in Gal(L/**Q**) such that ρ(S) is the 2x2 identity matrix I_2. If T is an element of Gal(L/**Q**) and S is an element of H, then $T\circ S\circ T^{-1}$ is an element of H, because

$$\begin{aligned} \rho(T\circ S\circ T^{-1}) &= \rho(T)\rho(S)\rho(T^{-1}) \\ &= \rho(T)\, I_2\, \rho(T^{-1}) \\ &= \rho(T)\rho(T^{-1}) \\ &= \rho(T)T^{-1}) \\ &= \rho(\text{identity}) \\ &= I_2. \end{aligned}$$

Furthermore, it is easy to show that if $S \neq R$, then $T)S)T^{-1} \neq T)R)T^{-1}$. This proves that H is a *normal subgroup* of Gal(L/**Q**). By the Fundamental Theorem of Galois Theory, this implies that there exists a field K such that K contains **Q**, K is contained in L, and the Galois group Gal(K/**Q**) is isomorphic to Gal(L/**Q**)/H. For every "prime of good reduction" p such that p does not divide m, there exists a very

useful element of Gal(K/**Q**), called the *Frobenius automorphism* σ_p. In the next few paragraphs, I will describe this element of Gal(K/**Q**).

Let p be a prime of good reduction that does not divide m. Consider the ring of all *algebraic integers* in K (i.e., the set of all numbers in K that are roots of polynomials whose coefficients are integers and whose leading coefficient is 1). It is often useful to look at *prime ideals*[4] of this ring that contain p. Let's choose one of these prime ideals and call it q. Take a moment to convince yourself that Gal(K/**Q**) maps ideals to ideals and that, moreover, it maps prime ideals to prime ideals. The subgroup of Gal(K/**Q**) consisting of functions that map q to itself is called the *decomposition subgroup* D_q.

Keeping the definition of D_q in mind, let's now consider the field R/q; let's call this field $\mathbf{F}_q$. The elements of $\mathbf{F}_q$ are *equivalence classes* of algebraic integers. Two algebraic integers are in the same equivalence class if and only if the difference between them is an element of q.

It is useful to compare the field $\mathbf{F}_q$ with the field $\mathbf{F}_p$. The elements of $\mathbf{F}_p$ are equivalence classes of integers; two integers are in the same equivalence class if and only if the difference between them is a multiple of p. These descriptions make it clear that $\mathbf{F}_q$ contains $\mathbf{F}_p$. Additionally, there is a Galois group Gal($\mathbf{F}_q/\mathbf{F}_p$), and it is easy to show that the function $f_p(x) = x^p$ is an element of this Galois group. (Hint: Expand $(x+y)^p$ using the Binomial Theorem.)

To each element of D_q, we can associate an element of Gal($\mathbf{F}_q/\mathbf{F}_p$). Namely, if the function g(x) is an element of D_q, then we can associate to it the function h(x) = [g(x)], which, when its domain is restricted to algebraic integers in K, is an element of Gal($\mathbf{F}_q/\mathbf{F}_p$). (The brackets denote the equivalence class that g(x) is in.) It is true (though it is too complicated to prove here) that this map from D_q and Gal($\mathbf{F}_q/\mathbf{F}_p$) is an isomorphism. Thus, there is exactly one element of D_q that is mapped to $f_p(x)$. This element of D_q is called the *Frobenius automorphism* σ_p.

Since D_q is a subgroup of Gal(K/**Q**) and Gal(K/**Q**) is isomorphic to Gal(L/**Q**)/H, there exists an element of Gal(L/**Q**) whose restriction to K is σ_p. Hence, just like every other element of Gal(L/**Q**), σ_p has a matrix representation. The trace of the matrix $\rho_{E,m}(\sigma_p)$ is independent of our choice of prime ideal q lying over p. (Why is this true? Well, if q and r are two prime ideals whose Frobenius elements are $\sigma_{p,q}$ and $\sigma_{p,r}$ respectively, then there is a theorem stating that there exists a matrix A such that $(A)(\sigma_{p,q})(A^{-1}) = \sigma_{p,r}$. From this fact, it can be proved via linear algebra that $\sigma_{p,q}$ and $\sigma_{p,r}$ have the same trace.)

Interestingly, the trace of $\rho_{E,m}(\sigma_p)$ is congruent to $a_p(E)$ modulo m. Recall from Part III that $a_p(E) = p - N_p + 1$, where N_p is the number of points mod p on E_p. Recall, also, that the number $a_p(E)$ appears prominently in the Taniyama-Shimura Conjecture. I bet you didn't expect it to turn up in this seemingly irrelevant discussion about Galois groups and Frobenius automorphisms!

---

[4] An *ideal* I of a ring R is a subset of R that is a group satisfying the following property: if a ∈ I and b ∈ R, then ab ∈ I and ba ∈ I. A *prime ideal* J is an ideal such that if c,d ∈ R and cd ∈ J, then c ∈ J or d ∈ J.

In the next few paragraphs, I will present a proof of the congruence $tr(\rho_{E,m}(\sigma_p)) = a_p(E) \pmod m$.

Let V be a set whose only elements are the coordinates of the points on the curve E_p. The *algebraic closure* of V is the set of all numbers that are roots of polynomials with coefficients in V. The *algebraic closure* of E_p is the set of all points (x,y) such that x or y is an element of the algebraic closure of V. In this paper, we will use the letter A to represent the algebraic closure of E_p.

Consider the function $Frob_p(x,y) = (x^p,y^p)$. (This is essentially the Frobenius automorphism we were discussing earlier.) It is easy to show that the function $Frob_p$: A → A is an element of the Galois group Gal(L/**Q**). Hence, it is the type of function that is called an *automorphism*. Because every automorphism that acts on the algebraic closure of a field fixes the base field, $Frob_p(E_p) = E_p$. In other words, E_p is the kernel of $Frob_p - I$, where I is the identity transformation. It follows that N_p is the number of elements in the kernel of $Frob_p - I$. Since the number of elements in the kernel of a transformation is the determinant of the matrix of that transformation, $N_p = \det(\rho_{E,m}(Frob_p - I))$.

Using linear algebra, it is child's play to prove that, for all 2x2 matrices A, det(A-I) = det(A) – tr(A) + 1. The reader may wish to prove this as an exercise.

Although it is too complicated to prove here, it is a fact that $\det(\rho_{E,m}(Frob_p)) = p \pmod m$. Hence,

$$\begin{aligned} N_p &= \det(\rho_{E,m}(Frob_p - I)) \\ &= \det(\rho_{E,m}(Frob_p)) - tr(\rho_{E,m}(Frob_p)) + 1 \\ &= p - tr(\rho_{E,m}(Frob_p)) + 1 \pmod m. \end{aligned}$$

Rearranging terms, we get: $tr(\rho_{E,m}(Frob_p)) = p - N_p + 1 \pmod m$. Our proof is now complete.

XI. Ribet's proof of Serre's Level-Lowering Conjecture

Let us now return to the problem of proving that the Frey curve is nonmodular. In 1985, Jean-Pierre Serre made a famous conjecture that goes something like this:

Consider a modular form f, whose level is N. According to a theorem of Shimura, we can associate to such a modular form an *abelian variety*. An abelian variety is a *group* that is also the solution set of a system of polynomial equations. Since every elliptic curve is a group as well as the solution set of a polynomial equation F(x,y) = 0, elliptic curves are said to be one-dimensional abelian varieties. But there exist higher-dimensional abelian varieties as well.

Using methods somewhat similar to the ones we discussed in Sections VIII and IX, we can use the abelian variety of f to come up with a matrix representation of Gal(L/**Q**). To help understand this better, let's only consider the special case where the abelian variety of f is an elliptic curve E. In this special case,

as you know, we can come up with a matrix representation $\rho_{E,m}$ of Gal(L/**Q**). We can choose m to be whatever we want, so let's make it an odd prime number. Then we can come up with *another* matrix representation of Gal(L/**Q**), which is called the *semisimplication* of $\rho_{E,m}$. The semisimplification of $\rho_{E,m}$ depends only on f and m, so we can name it $\rho_{f,m}$.

Let's now suppose that $\rho_{f,m}$ lacks a property called *ramification* at a prime $p \neq m$ that divides N. Also, suppose p^2 does not divide N. Serre's conjecture states that there exists a modular form g, whose level is N/p, and a prime number h such that $\rho_{g,h}$ is isomorphic to $\rho_{f,m}$. This conjecture is sometimes called *Serre's Level-Lowering Conjecture*.

Using the property $tr(\rho_{f,m}(\sigma_p)) = a_p(f) \pmod m$ that we proved earlier, Ribet proved Serre's Level-Lowering Conjecture in 1986. Keeping Ribet's results in mind, let's now talk again about the Frey curve $y^2 = x(x - a^m)(x + b^m)$. Recall that m is the exponent in Fermat's Last Theorem. (We can require that m be an odd prime. Suppose m were composite. Then, either m would be a power of 2, or m would equal qr, where r is an odd prime and $q \neq 1$. If m=qr, then $a^m + b^m = c^m$ implies $(a^q)^r + (b^q)^r = (c^q)^r$, which means that we would be able to use r (an odd prime) instead of m. If m were a power of 2, then it would equal 4s for some integer s. Then $a^m + b^m = c^m$ implies $(a^s)^4 + (b^s)^4 = (c^s)^4$. But it is easy to prove that there is no counterexample to Fermat's Last Theorem in which 4 is the exponent. So m cannot be a power of 2.)

Let's denote the Frey curve by the letter E. Then let's consider the matrix representation $\rho_{E,m}$ of Gal(L/**Q**). It turns out that this representation is unramified at all odd primes $p \neq m$.

Now, suppose the Frey curve is modular. In other words, suppose that the Frey curve can be associated with a modular form f, whose level is some positive integer N. According to a theorem of Goro Shimura and Martin Eichler, if the abelian variety associated to a modular form f is an elliptic curve E, then the level of f is equal to the conductor of E. So, N is both the level of f *and* the conductor of the Frey curve. And since the Frey curve is semistable, we know that p^2 does not divide N, for all primes p that divide N. This means that we can apply Serre's Level-Lowering Conjecture.

According to the conjecture, we can find a modular form g whose level is lower than the level of f. By repeating this level-lowering process over and over, we can divide out all the odd prime factors of N, and we will eventually find a modular form whose level is 2. But modular forms with level 2 do not exist. This is a contradiction; hence, the Frey curve cannot possibly be modular.

Ribet's proof of the Frey's curve nonmodularity paved the way for Wiles's proof of Fermat's Last Theorem.

BIBLIOGRAPHY

Daney, Charles. *Galois Representations and Elliptic Curves*. 12 Apr. 1996. 20 Apr. 2003 <http://www.mbay.net/~cgd/flt/flt07.htm>.

---. *Proof of Fermat's Last Theorem, The*. 13 Mar. 1996. 17 May 2003 < http://www.mbay.net/~cgd/flt/flt08.htm>.

Darmon, Henri. "A Proof of the Full Shimura-Taniyama-Weil Conjecture is Announced." *Notices of the AMS* December 1999: 1397-1401.

Eric Weisstein's World of Mathematics. 17 Apr. 2003. 20 Apr. 2003 <www.mathworld.com>.

Hall, Marshall, Jr. *The Theory of Groups*. New York: Macmillan, 1959.

Hellegouarch, Yves. *Invitation to the Mathematics of Fermat-Wiles*. San Diego: Academic Press, 2002.

Knapp, Anthony W. *Elliptic Curves*. Princeton, N.J.: Princeton University Press, 1992.

Lang, Serge. "Some History of the Shimura-Taniyama Conjecture." *Notices of the AMS* Nov. 1995: 1301-1307.

Ribet, Kenneth A. "Galois Representations and Modular Forms." *Bulletin of the AMS* Oct. 1995: 375-402.

Ribet, Kenneth A. and Brian Hayes. "Fermat's Last Theorem and Modern Arithmetic." *American Scientist* March-Apr. 1994: 144-156.

Silverman, Joseph H. and John Tate. *Rational Points on Elliptic Curves*. New York: Springer-Verlag, 1992.