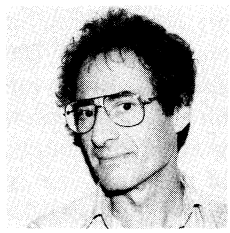
Number Theory as Gadfly

B. MAZUR, *Harvard University*

DR. MAZUR received his Ph.D. from Princeton University in 1959, and was a Junior Fellow at Harvard University from 1961–64. Since then he has been at Harvard University with frequent visits to I.H.E.S. in France. He is a member of the U.S. National Academy of Sciences and has received the Veblen Prize (in geometry) and the Cole Prize (in number theory) from the AMS.



(This is an expository article which evolved from notes written in preparation for a 40-minute talk for a general audience at the "Symposium on Number Theory," held in Washington D.C. on May 4, 1989 under the auspices of the Board on Mathematical Sciences of the National Research Council. To make the text more informative the original version has been supplemented with lots of commentary, a section (§4) has been added which may be useful to readers familiar with the classical theory of modular forms, and an appendix has been added which is meant for an even more specialized audience. I am thankful to P. Diaconis, J. Mazur, K. Ribet and J.-P. Serre, who read early drafts of this paper, and whose suggestions were very helpful to me.)

1. Introduction. When a friend saw the title to my talk he asked if what I had in mind was the well-known fact that number theory has an annoying habit: the field produces, without effort, innumerable problems which have a sweet, innocent air about them, tempting flowers; and yet... the quests for the solutions of these problems have been known to lead to the creation (from nothing) of theories which spread their light on all of mathematics[1], have been known to goad mathematicians on to achieve major unifications of their science[2], have been known to entail painful exertion in other branches of mathematics to make those branches serviceable[3]. Number theory swarms with bugs, waiting to bite the tempted flower-lovers who, once bitten, are inspired to excesses of effort!

Well, perhaps that summarizes the general aim of my talk—but, to put it more gently, I want to spend a few minutes considering one example (a conjecture, in fact) which shows how Number Theory can sometimes contrive to be a helpful, and possibly inspirational, goad to the rest of the Mathematical Sciences.

The most celebrated of all deceptively simple (and still unsolved!) problems in Number Theory is surely *Fermat's Last Theorem*[4]. Its curious history (whose statement first occurs as a marginal commentary on the equation arising from the Pythagorean theorem) is so well known, it needn't be rehearsed here. Professional mathematicians, after Fermat, have approached Fermat's Last Theorem with

[1] e.g., Kummer's theory of ideals

[2] e.g., Grothendieck's theory of schemes

[3] e.g., The theory of group representations, and in particular, the "Langlands program"

[4] For a detailed account of the recent work on this see Oesterlé's Bourbaki report [O] listed in the References for §4.

mixed feelings. Kummer, for example, called it "more of a joke than a pinnacle of science," and he came as close as anyone has come to proving it! Gauss wouldn't work on it, even when urged by his friends to do so in order to get the cash reward which had been offered for its solution.

But Fermat's Last Theorem has always been the darling of the amateur mathematicians, and as things have progressed, it seems that they are right to be enamored of it: Despite the fact that it resists solution, it has inspired a prodigious amount of first-rate mathematics. Despite the fact that its truth hasn't a *single* direct application (even within number theory!) it has, nevertheless, an interesting *oblique* contribution to make to number theory: its truth would follow from some of the most vital and central conjectures in the field. Although others are to be found, Fermat's Last Theorem presents an unusually interesting "test" for these conjectures.

My aim is to describe, in some detail, one of these grand conjectures (due to Shimura, Taniyama and Weil[5]) which, even though still unproved, plays a structural and deeply influential role in much of our thinking and our expectations in Arithmetic. Thanks to recent work of Frey, Serre, and Ribet[6], it has a large number of applications, Fermat's Last Theorem among them. As I shall not have time to make clear, but hope, at least, to make believable, the conjecture of Shimura-Taniyama-Weil is a profoundly unifying conjecture—its very statement hints that we may have to look to diverse mathematical fields for insights or tools that might lead to its resolution[7].

As we shall see, the conjecture of Shimura-Taniyama-Weil would imply a strange and important connection between the elliptic curves that arise in Arithmetic (we'll get to that shortly!) and the *Hyperbolic Plane*.

I mentioned above that the conjecture of Shimura-Taniyama-Weil has as one of its consequences, Fermat's Last Theorem. As everyone knows, Fermat's Last Theorem is an assertion about the family of (Diophantine[8]) equations

$$X^N + Y^N = Z^N$$

[5]The fact that this conjecture has also been referred to as the *Weil conjecture*, the *Taniyama-Weil conjecture*, and the *Taniyama conjecture* points to the difficulty in assigning to it a clear attribution. It was originally formulated as a problem by Taniyama in a conference in 1955 and was published in Japanese, in Sûgaku 7 (1956) p. 269. A more precise formulation corresponding to the modern form of the conjecture—involving important information concerning the conductor—was implicitly suggested by subsequent work of Weil which had the effect of bringing the problem to the attention of a large audience. The most precise version of this conjecture to date, which brings in the crucial issue of fields of definition, incorporates work of Shimura, Eichler, and others (see footnote 15 below, and the technical appendix at the end). For a moving evocation of the life and times of Taniyama (as well as an English translation of the original statement of Taniyama's problems) see the article "Yutaka Taniyama and his time, very personal recollections," Goro Shimura, *Bull. London Math. Soc.* 21 (1989) 186–196.

[6]See [S], [Fr 1, Fr 2], and [R] listed in the references at the end of §4.

[7]It does not seem unnatural to look to differential geometry for progress with this conjecture, or to partial differential equations and the study of the eigenvalue problem for elliptic operators, or to the representation theory of reductive groups.... It would be no surprise if ideas from the classical theory of one complex variable and the Mellin transform were relevant, or of Algebraic Geometry.... But perhaps one should also look in the direction of Kac-Moody algebras, loop groups, or $\mathscr{D}$-modules, perhaps to ideas that have been, or will be, imported from Physics....

[8]The adjective "Diophantine" is in honor of the Alexandrian mathematician Diophantos (perhaps A.D. 250) and signals vaguely a type of equation not dissimilar from those Diophantos considered.

for $N = 3, 4, 5, \ldots$, or with little loss of generality, for odd prime exponents N. But despite the elegance, and evident symmetries of the above family of equations, there is no denying that, after all, it is merely *one* family of Diophantine equations. What is so excellent about these particular equations?

Now, the urge to put any *single* instance in an appropriate general context before dealing with it mathematically is strong. Consider, for example, the way François Viète, the modern inventor of algebra, expressed that urge at the end of his treatise (c. 1591) by saying that algebra appropriated for itself "the proud problem of problems: which is to *leave no problem unsolved.*" More to the point, consider the most celebrated of recent Diophantine results, valid in a *truly* general context: the theorem of Faltings (conjectured originally by Mordell) which asserts that any algebraic equation in two variables, and of genus[9] greater than or equal to 2, has only a finite number of rational solutions.

But what *is* an appropriate general context in which to place the Fermat family of Diophantine equations, and what *is* the appropriate Diophantine question to ask? Despite the fact that the Fermat problem has been with us for three centuries I don't believe that we have any thoroughly comfortable answer, even to this modest question. One may always take recourse (in cases where it is not clear how to "correctly" generalize a problem) in the reliable method of *kicking the problem a bit*, to get a "nearby" one.... A relatively conservative move in this direction, in the case of Fermat's equation, is to allow general coefficients in the equation, say one coefficient for starters—for example, fix a nonzero integer A, and consider the family

$$A \cdot X^N + Y^N = Z^N,$$

and then ask: Is there an exponent N_0 such that for exponents N (or for prime exponents N) greater than N_0 there is no triple of integers (X, Y, Z), none zero, solving the above equation? In this slight perturbation of Fermat's original problem a few minutes of reflection will convince one to be circumspect in framing precise conjectures... (e.g., consider $A = 2$). Nevertheless the conjecture of Shimura-Taniyama-Weil has an impressive power of prediction concerning the

---

About Diophantos' personal history little is known, save what can be gleaned from the following problem which occurs in a collection, the *Palatine Anthology*, compiled, scholars believe, no more than a century after his death:

> Here you see the tomb containing the remains of Diophantos, it is remarkable: artfully it tells the measures of his life. The sixth part of his life God granted him for his youth. After a twelfth more his cheeks were bearded. After an additional seventh he kindled the light of marriage, and in the fifth year he accepted a son. Elas, a dear but unfortunate child, half of his father he was and this was also the span a cruel fate granted it. He consoled his grief in the remaining four years of his life. By this devise of numbers, tell us the extent of his life.

[9]The *genus* g of an algebraic curve is a nonnegative integer which was originally introduced by Riemann and defined by "topological means." It also has an "algebraic" definition, and as such is an intrinsic invariant of the field of algebraic functions on the curve. If the curve is given as the locus of zeroes of a homogeneous form of degree d in three variables in the projective plane then $g \leqslant (d-1)(d-2)/2$ with equality holding if and only if the curve has no singularities. In contrast to the genus, however, the *degree* is not given by the field of functions of the curve alone: it is defined in terms of the representation of the curve in projective space. For this reason it is more natural to look to the *genus* rather than to the *degree as* an invariant which determines "Diophantine behavior."

nature of nontrivial integer solutions for these families. For example,[10] using work of Frey, Serre, and Ribet, the Shimura-Taniyama-Weil conjecture would imply that the above equation has no such solutions for prime exponents $N > 7$ if A is any power of 3, or of 5, or of 7 (or of 11, 13, 17, 19, 23, 29, 53 or 59, for that matter[11], provided that N doesn't divide A), and it would guarantee the existence of an N_0 such that for prime exponents $N \geqslant N_0$ there are no such solutions if A is any power of any odd prime not of the form $2^n \pm 1$ (i.e., if A is neither a power of a Mersenne nor of a Fermat prime).

In summary, the conjecture of Shimura-Taniyama-Weil seems to be getting into the thorny thicket of these Diophantine issues—seems to be giving reasons why some (but not all!) of these equations cannot have solutions—seems to be beginning to put such Diophantine problems in a "context."

It also relates them to the extraordinary geometric questions to which we shall now turn.

2. Euclidean and non-Euclidean covering mappings. One of the mysteries of the Shimura-Taniyama-Weil conjecture, and its constellation of equivalent paraphrases, is that although it is undeniably a conjecture "about arithmetic," it can be phrased variously, so that: in one of its guises, one thinks of it as being also deeply "about" integral transforms in the theory of one complex variable; in another as being also "about" geometry[12].

The more striking of these two formulations is the geometric one. To explain it we need to review a few basics of geometry: *symmetries*, *orbits*, *orbit spaces*, *covering mappings*, and the interesting concept of "*uniformization*" . . . . We'll build things up slowly by first considering these notions in a relatively simple context (on the *Real Line*), and then in the two contexts (*Euclidean* and *Non-Euclidean*) necessary for the actual "geometric" statement of the conjecture.

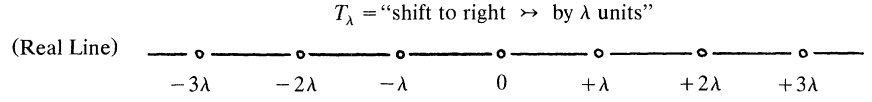


FIG. 1. The "lattice" $\Lambda = \{0, \pm\lambda, \pm 2\lambda, \ldots\}$ is the orbit of 0 under the translation T_λ.

(I). On the real line.

Let λ be a positive real number and let T_λ denote the operation on the real line $\mathbb{R}$ consisting of "shifting all points in $\mathbb{R}$ to the right" by λ (synonymously: *translation* by λ; in equations: $T_\lambda(x) = x + \lambda$). We view T_λ as a *symmetry* of $\mathbb{R}$. The iterated translates of 0 by T_λ and by its inverse $T_{-\lambda}$ give us a discrete evenly spaced "lattice" Λ in $\mathbb{R}$ consisting of all integral multiples of λ.

[10]See thme. 2 and subsequent remarks in section 4.3 of J.-P. Serre's [S] listed in the references at the end of §4.

[11]And with more work, one could surely produce more such consequences of Shimura-Taniyama-Weil.

[12]The equivalence of these two formulations is due to Weil, following upon work of Hecke.

The *orbit* of a real number x with respect to the lattice Λ, is the collection of real numbers obtained from x by translating x by iterates of T_λ and by its inverse $T_{-\lambda}$ (that is, the orbit is the collection of real numbers $x + N \cdot \lambda$ where N runs through all integers). The lattice Λ itself is the orbit containing the origin. Given a fixed real number A, any orbit for Λ has a unique representative in the interval $A \leqslant x < A + \lambda$.

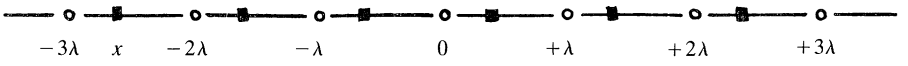


FIG. 2. The black squares mark the "orbit" of some general real number x.

The *orbit space* of $\mathbb{R}$ with respect to the symmetry T_λ (or one might also say: with respect to the lattice Λ) is simply the collection of orbits (with respect to Λ). Call the orbit space $\mathbb{R}/\Lambda$. One can think of the orbit space as obtained from $\mathbb{R}$ by "identifying" any point x in $\mathbb{R}$ with any iterated translate of x by T_λ and $T_{-\lambda}$, i.e., with all points $y \in \mathbb{R}$ such that $y = x + N\lambda$ for some integer N. The easiest way to visualize this is by wrapping the real line (viewed as a helix in the figure below) around a circle (say a unit circle in the complex plane) in such a manner that points on $\mathbb{R}$ of distance λ apart map to the sample point of the circle:

$$z \mapsto e^{2\pi i z/\lambda}. \qquad (*)$$

The circle, then, may be taken to be the orbit space. The displayed mapping $(*)$ can be called a *covering mapping* and has the property that the pre-image with respect to $(*)$ of any point on the circle consists in precisely one orbit with respect to Λ.

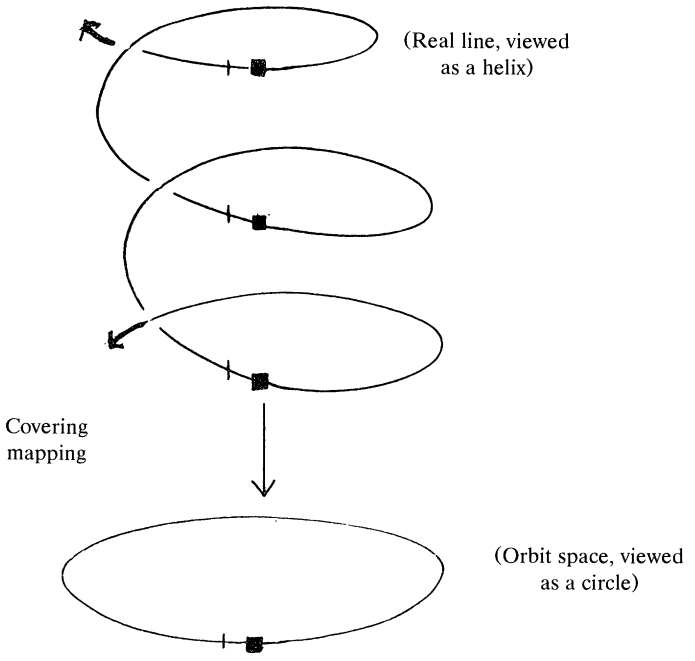
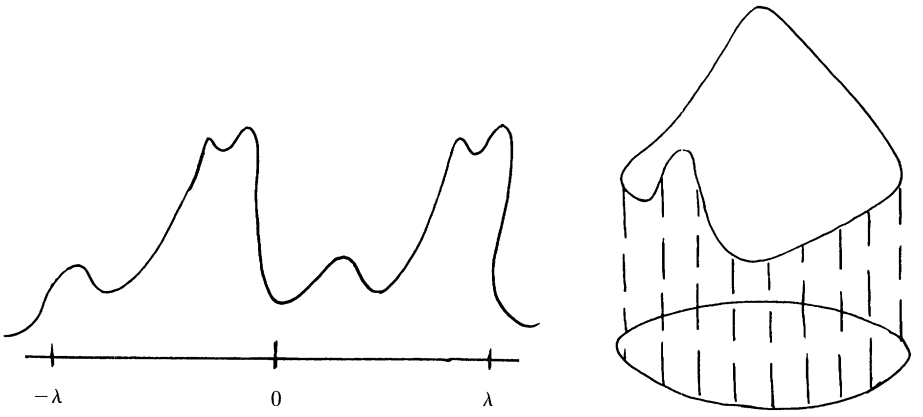


FIG. 3

If a function f on $\mathbb{R}$ is *invariant under the translation* T_λ—which means that $f(T_\lambda x) = f(x)$ for all x, i.e., $f(x + \lambda) = f(x)$—we say that f is "*periodic*" with period λ.



Graph of a periodic function on the line. Now viewed as function on the orbit space.

FIG. 4

The circle as orbit space is a proper realm on which to consider periodic functions with period λ. Any such function f may be viewed in a natural way as defined on the orbit space, and conversely any function on the orbit space may be viewed as coming from a periodic function on $\mathbb{R}$ with period λ.

(II). Double periodicity on the (Euclidean) complex plane—the setting for the classical theory of elliptic functions.

Now let us pass from the real line $\mathbb{R}$ to the complex plane $\mathbb{C}$. Instead of considering only one translation, as we did with $\mathbb{R}$, it is natural in this (two-dimensional) context to consider as "symmetries" two translations T_{λ_1} and T_{λ_2} acting on the complex plane

$$T_{\lambda_1} : x \mapsto x + \lambda_1 \qquad T_{\lambda_2} : x \mapsto x + \lambda_2,$$

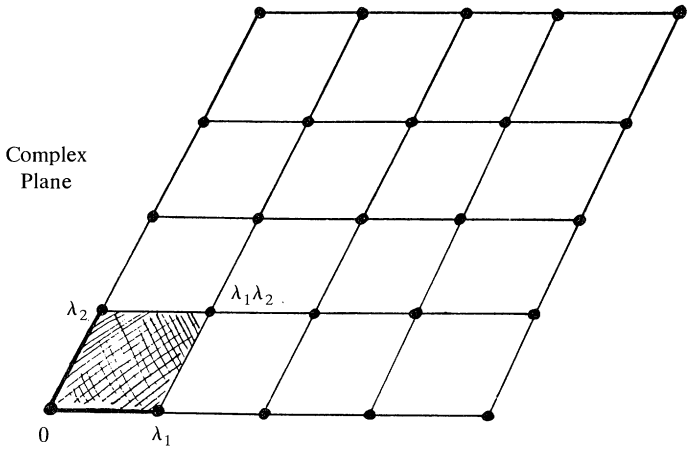


FIG. 5

and the various symmetries obtainable by iterations of them ($x \mapsto N_1 \cdot \lambda_1 + N_2 \cdot \lambda_2$, for integers N_1 and N_2). Here λ_1, λ_2 are two complex numbers such that 0, λ_1, and λ_2 are not collinear (and therefore "generate" a lattice Λ as in FIGURE 5).

An orbit with respect to Λ (that is, the set of points in the complex plane which can be obtained by the iterated application of the translations $T_{\lambda_1}, T_{\lambda_2}$ and their inverses to a single point in the complex plane) is simply a "displaced lattice" as in FIGURE 6 below. Any such orbit has exactly one representative in the (half-closed) parallelogram as in the next figure, whose vertices are 0, λ_1, λ_2, and $\lambda_1 + \lambda_2$.

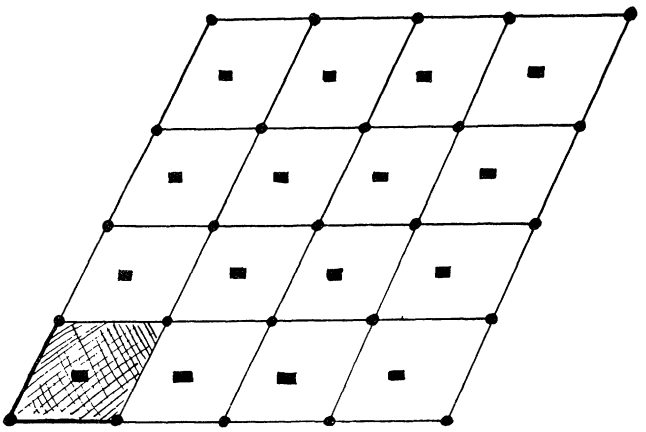


FIG. 6. The ■'s mark an orbit with respect to the lattice.

As in the example of $\mathbb{R}$ given previously, by the orbit space with respect to the two symmetries T_{λ_1} and T_{λ_2} (or, we may say, with respect to the lattice Λ) we mean *the collection of orbits* with respect to Λ. We may visualize this orbit space as the parallelogram with vertices $0, \lambda_1, \lambda_2$, and $\lambda_1 + \lambda_2$ "wrapped up" as shown in FIGURE 7 below. Topologically it is a "torus," i.e., the surface of a doughnut. Call

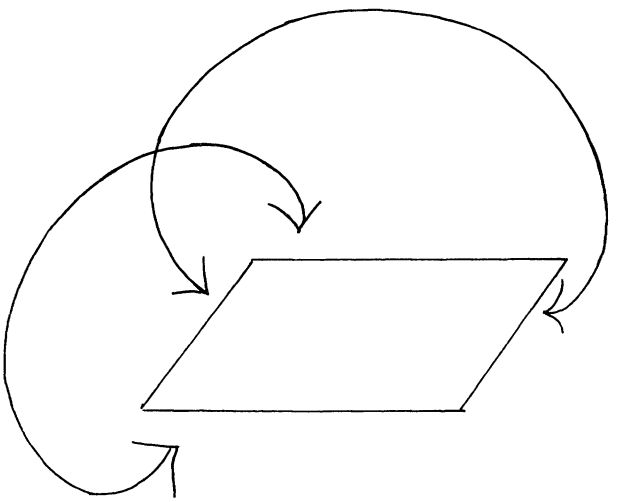


FIG. 7

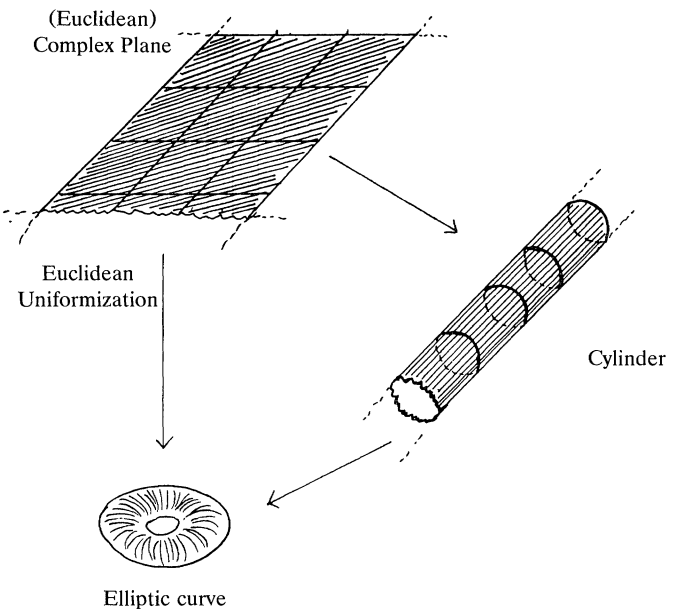


FIG. 8. The "covering mapping" which brings the complex plane $\mathbb{C}$ to the orbit space $\mathbb{C}/\Lambda$ may also be visualized as a two-stage process, where in the first stage the plane is wrapped around a cylinder, and in the second stage the cylinder is wrapped around a torus.

the orbit space $\mathbb{C}/\Lambda$. The mapping which sends each point in the complex plane to the orbit which contains it is our *covering mapping* $\mathbb{C} \to \mathbb{C}/\Lambda$:

We wish to think of the orbit space $\mathbb{C}/\Lambda$ as inheriting a "conformal geometry" (and an orientation) from the standard Euclidean geometry of the complex plane $\mathbb{C}$ via this natural mapping. A *conformal geometry* on a smooth surface is a "geome-

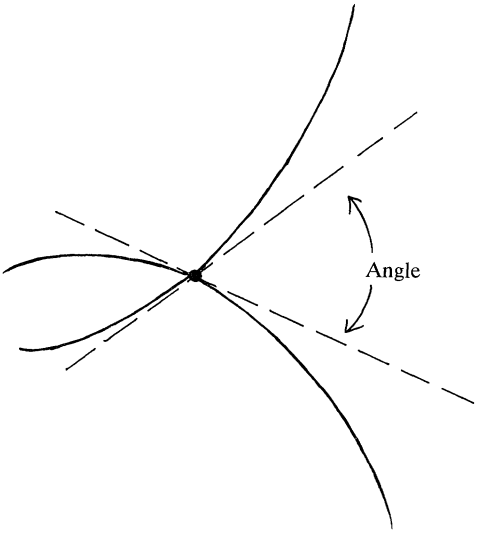


FIG. 9. In conformal geometry, there is no invariant notion of "length" of an arc, but there is a notion of angle.

try" where there is *no* intrinsic notion of "length of arc," but if we are given two, real, smooth pieces of arc traced on the surface which intersect at a point (as in FIGURE 9) it makes sense to talk about their *angle of intersection*.

To endow a smooth surface with an orientation and a (locally Euclidean) conformal geometry is the same as to consider it as a *Riemann surface*.[13] {Here is a sketch of the modern formulation of this notion, originating essentially with Herman Weyl: to give a smooth surface the structure of *Riemann surface* has to specify at each point of the surface a preferred class of complex-valued functions defined locally about the point, called *analytic functions*, where this specification is done in the following manner: Among the analytic functions about each point there is at least one, call it z, which identifies some neighborhood of the point with an open set U of the complex plane so that the specified class of *analytic functions* on that neighborhood is identified with the class of standard analytic functions on U. Moreover, the condition of being an 'analytic function' is a *local condition* in the sense that a function defined on an open subset of the surface is 'analytic' if and only if it is 'analytic' on each subset of any open cover of that open subset. Given, then, a Riemann surface we may say what it means for a complex-valued function on it to be 'analytic,' or 'meromorphic' (a ratio of 'analytic functions').

A function which plays the role required of the function denoted z above, which is defined locally about a point of the surface and, moreover, which vanishes at that point is sometimes called a *local uniformizer* at the point in question. The local isomorphism between an open subset of a Riemann surface and $U \subset \mathbb{C}$ given by a 'local uniformizer' z identifies the "conformal geometry" on that open subset with the Euclidean conformal geometry on U}.

With Weierstrass let us ask for (*doubly*) *periodic*[14] functions on the complex plane: functions which are periodic with respect to both T_{λ_1} and T_{λ_2}. These doubly periodic functions are precisely the meromorphic functions on the quotient space $\mathbb{C}/\Lambda$, meromorphic in the sense of its Riemann surface structure. As Weierstrass showed, the problem of constructing all such periodic functions can be elegantly settled by the construction of one special periodic function, the so-called $\wp$-function attached to $\Lambda, \wp_\Lambda(z)$ (call it $\wp$ for short)[15] which satisfies a nonlinear differential equation of the form

$$(\wp')^2 = 4 \cdot \wp^3 + A \cdot \wp + B,$$

where A and B are complex numbers, and $\wp'(z) = d\wp(z)/dz$. This turns out to be quite an adventitious mathematical construction! An unexpected number of problems are solved in a single stroke: First, the stated problem finds its solution in this construction, *for any doubly periodic* (*with respect to* Λ) *meromorphic function can be obtained as a rational function in* $\wp_\Lambda$ *and* $\wp'_\Lambda$. But it is also the case that any equation

$$Y^2 = 4 \cdot X^3 + A \cdot X + B \qquad (A, B \in \mathbb{C}) \qquad (*)$$

[13]See Felix Klein's classic expository account: On Riemann's Theory of Algebraic Functions and Their Integrals (A Supplement to the Usual Treatises). Dover.

[14](meromorphic)

[15]given by: $\wp_\Lambda(z) = 1/z^2$ plus the infinite sum of the terms $\{1/(z-\lambda)^2 - 1/\lambda^2\}$ where λ ranges through the nonzero elements of the lattice Λ.

[such that the cubic polynomial $4 \cdot X^3 + A \cdot X + B$ has no multiple roots] arises as the polynomial relating $\wp_\Lambda$ and $\wp'_\Lambda$ where $\wp_\Lambda$ is the Weierstrass $\wp$-function attached to *some* lattice Λ. The algebraic curve in the (X, Y)-plane defined by such an equation (or rather, the completion of this curve in the projective plane) is called an *elliptic curve*. There is a single point at ∞ on this curve; this point will be called the *origin* of the curve. It is an easy exercise in algebra to see that *any* smooth algebraic curve of degree 3 in the projective plane can be given, after any choice of an origin and an appropriate rational change of variables, a defining equation of the above form (called a *Weierstrass equation*). This means that if we are interested in the *algebraic* problem of finding all complex solutions of smooth homogeneous forms in three variables of degree 3 (or equivalently, of finding solutions of a "Weierstrass equation") we are led to the *analytic* construction of a specific lattice Λ in $\mathbb{C}$, of its associated $\wp$-function $\wp_\Lambda(z)$, and then the non-origin points on E are given by $(X, Y) = (\wp_\Lambda(z), \wp'_\Lambda(z))$ where z ranges through the points of $\mathbb{C}$ which are not in Λ. Let us call the mapping

$$\mathbb{C} - \Lambda \xrightarrow{\mathscr{U}} \textit{elliptic curve} - \textit{origin}$$
$$z \mapsto (X, Y) = (\wp(z), \wp'(z))$$

a (*Euclidean*) *uniformization* of our elliptic curve. One can, if one wishes, complete the picture to get a mapping

$$\mathbb{C} \xrightarrow{\mathscr{U}} \text{elliptic curve}$$

which identifies the complex points of our elliptic curve (including its origin) with the orbit space with respect to Λ, i.e., the Riemann surface[16] $\mathbb{C}/\Lambda$, the mapping being the covering mapping.

The word "uniformization" is meant to carry the full load of its nuances here.

We have *uniformly* parametrized the complex points of *any* elliptic curve by the points of $\mathbb{C}$ (more exactly, if you wish, by the orbits of these points under translation via the elements of a lattice Λ determined by the elliptic curve).

The uniformization $\mathscr{U}$ also uniformizes the conformal geometry of any particular elliptic curve, in the sense that $\mathscr{U}$ identifies the conformal geometry of the Riemann surface, locally, with Euclidean conformal geometry (its inverse, locally defined, gives *local uniformizers* in the sense described above)[17].

So far, our equation $(*)$ has complex coefficients A, B. Number theory will properly enter our picture when we consider elliptic curves $Y^2 = 4 \cdot X^3 + A \cdot X + B$ where the coefficients A, B are *algebraic numbers*, or more specifically, *rational numbers*. To focus on the latter case which will be our eventual particular interest, let us refer to an elliptic curve whose coefficients A, B lie in $\mathbb{Q}$ as simply

[16] It has mystified generations of students that (algebraic) *curves* can be viewed as (Riemann) *surfaces*. The clash in terminology indicates that, as its name implies, Algebraic Geometry is at the meeting-ground between algebra and geometry: one thinks, for example, of $\mathbb{C}$ as the affine line (i.e., a *curve*) if one is thinking algebraically and as the complex plane (i.e., a *surface*) if one is thinking geometrically.

[17] The uniformization $\mathscr{U}$ has the further "uniformity" property, usually incorporated in the technical definition of a covering mapping, that any small enough disc on the elliptic curve has the property that its pre-image under $\mathscr{U}$ consists in a disjoint union of components, each of which maps isomorphically onto the given disc.

an *arithmetic elliptic curve*. But before we deal with arithmetic elliptic curves we have some hyperbolic geometry to do.

(III). Periodicity on the (non-Euclidean) hyperbolic plane—the setting for the classical theory of modular functions.

Let us turn now to *hyperbolic geometry*, the (independent) discovery of Bolyai, Gauss, and Lobachevsky.

Hyperbolic geometry is a homogeneous geometry satisfying all the Euclidean axioms except for the fifth postulate, and possessing *many* lines through a given point, parallel to a given line; it now has a number of equivalent concretizations. The model particularly useful to us is the *upper half-plane model*.

Here the points of the geometry are the points $z = x + iy$ in the upper half of the complex plane $\mathbf{H}$, i.e., x can be any real number and y any *positive* real. The lines are either vertical straight lines $\{a + iy\}$ for a fixed real number a and all positive reals y, or else they are semi-circles abutting on the real axis. The upper half-plane model inherits a Riemann surface structure, and hence also a *conformal geometry* by virtue of its being an open subset in $\mathbb{C}$.

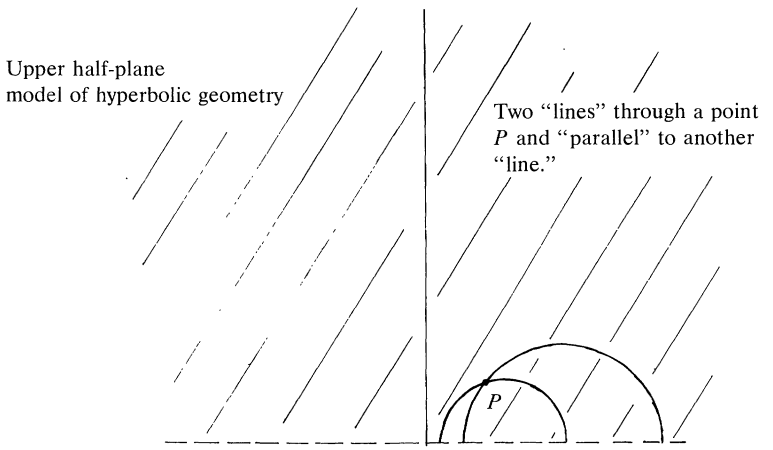


FIG. 10

The *translations* $T_b : z \mapsto z + b$ for any real number b are symmetries of hyperbolic geometry, but there are many more symmetries (in fact two other continuous parameters of them[18]), perhaps the most important single one being *inversion* with respect to the unit circle, $S : z \mapsto -1/z$.

[18]Consider matrices of real numbers of determinant equal to 1, i.e.,

$$\begin{pmatrix} a, b \\ c, d \end{pmatrix}$$

with $ad - bc = 1$. Then $z \mapsto az + b/cz + d$ is an orientation-preserving transformation of the upper half plane which is a symmetry of its hyperbolic geometry, and any orientation-preserving symmetry is given by such a matrix.

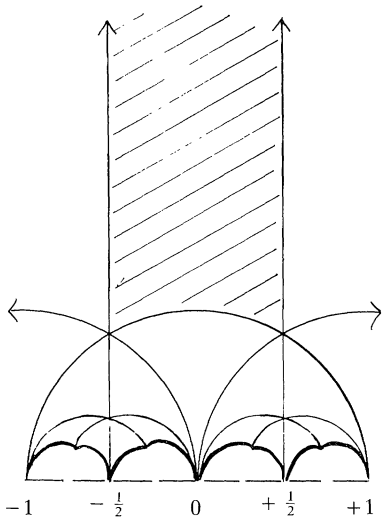


FIG. 11

FIGURE 11 is meant to illustrate the tiling of the hyperbolic plane that is gotten by systematically applying composites of iterates of the *unit translation*, $T_1 : z \mapsto z + 1$, and of the *inversion* S (and of their inverses), to the "basic tile," which is the shaded region in the figure. Let Γ be the group of symmetries of the hyperbolic plane gotten from such compositions of T_1 and S. It is a fact that Γ consists in *all* transformations of the form $z \mapsto az + b/cz + d$ where the coefficients a, b, c, d are all integers and $ad - bc = 1$. There are a number of striking differences between Γ acting on the hyperbolic plane and a lattice Λ, generated by translations T_{λ_1} and T_{λ_2}, acting on the complex plane. First, the two translations of the complex plane T_{λ_1} and T_{λ_2} commute with one another, which is not the case for *translation* and *inversion* of the hyperbolic plane, i.e., Γ is a more interesting, noncommutative, group. And second, there is a natural way of *identifying* Λ with

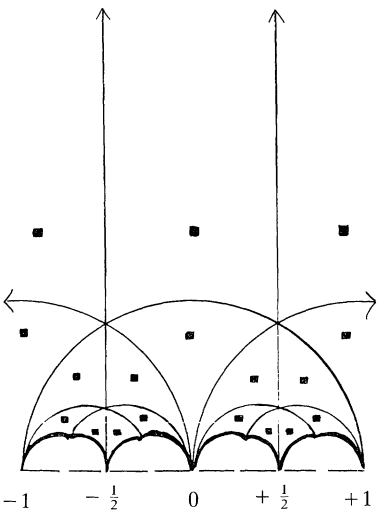


FIG. 12

an orbit in the complex plane (the orbit containing 0). There is no such identification for Γ[19]. In FIGURE 12, the "dots" (one in each tile) mark a single orbit with respect to Γ.

If E is an elliptic curve given with its conformal geometry, it has, of course, its *Euclidean* uniformization coming from Weierstrass's theory as described in II above. By a *hyperbolic uniformization of* E is meant a mapping

$$H - \{\textit{finite union of } \Gamma\textit{-orbits}\} \xrightarrow{\mathscr{V}} E - \{\textit{finite set of points}\},$$

which is a covering mapping of the domain onto the range[20], which preserves orientation, is a local isomorphism of conformal geometries[21], and which is periodic with respect to a subgroup of finite index in Γ. This last condition just means that there is a subgroup $\Gamma' \subset \Gamma$ of finite index such that if $z \mapsto \gamma(z)$ is a transformation in Γ', then

$$\mathscr{V}(\gamma(z)) = \mathscr{V}(z).$$

In other terms, $\mathscr{V}$ factors through the orbit space of the upper-half plane under the action of Γ'.[22]

For any such Γ', there are only a finite number of distinct elliptic curves admitting a hyperbolic uniformization periodic with respect to Γ'. Moreover, it is a consequence of a theorem of Bely that an elliptic curve admits a hyperbolic uniformization *if and only if* it has a Weierstrass equation with coefficients A, B which are algebraic numbers, i.e., $A, B \in \overline{\mathbb{Q}}$.

3. The conjecture of Shimura-Taniyama-Weil. A *hyperbolic uniformization* of E was defined to be a covering mapping periodic with respect to *any* subgroup of finite index in Γ. There is, however, a specific class of subgroups of finite index in Γ which plays a *dominant role* in arithmetic. To explain why this class should play the role it does (and there are interesting geometric reasons for this) would lead us far afield. But the definition of the class is simple enough: For a positive integer N, let $\Gamma(N)$ denote the group of matrices

$$\begin{pmatrix} a, b \\ c, d \end{pmatrix},$$

where a, b, c, d are integers, $ad - bc = 1$, and which are congruent to the identity matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ modulo N.

The groups $\Gamma(N)$ act as groups of symmetries of the hyperbolic plane (the above matrix acting by the rule $z \mapsto az + b/cz + d$), these groups all being subgroups of finite index in Γ. Any subgroup of Γ containing one of the $\Gamma(N)$'s is called a *congruence subgroup*. The terminology here is quite straightforward: a *congruence subgroup* of Γ is simply one that is definable by congruence conditions on the coefficients of the matrices representing its elements. Not all subgroups of finite index in Γ are congruence subgroups. (It is even true that, at least according

[19]There are two orbits which are slightly more degenerate than the rest.

[20]As in footnote 17.

[21]Equivalently: is an analytic (unramified) covering mapping of Riemann surfaces.

[22]The uniformization $\mathscr{V}$ extends to an analytic mapping of the upper half-plane onto the complement of a finite set of points in E.

to one natural way of counting them, relatively few subgroups of finite index are congruence subgroups.) But this gives rise to the key.

DEFINITION. *Let E be an elliptic curve.* A hyperbolic uniformization (*of E*) of arithmetic type *is a hyperbolic uniformization of the elliptic curve E which is periodic with respect to a congruence subgroup* $\Gamma' \subset \Gamma$.

Although (by Weierstrass) *any* elliptic curve admits a Euclidean uniformization (and, in fact with respect to a lattice $\Lambda \subset \mathbb{C}$ unique up to complex scalar change), and (by Bely) an elliptic curve admits a hyperbolic uniformization if and only if it can be defined by a Weierstrass equation with coefficients *A, B which are algebraic numbers*, the Shimura-Taniyama-Weil conjecture asserts, further, that

Any arithmetic *elliptic curve* (*i.e., any elliptic curve whose defining equation can be taken with coefficients in* $\mathbb{Q}$) *admits a* hyperbolic uniformization of arithmetic type.[23,24]

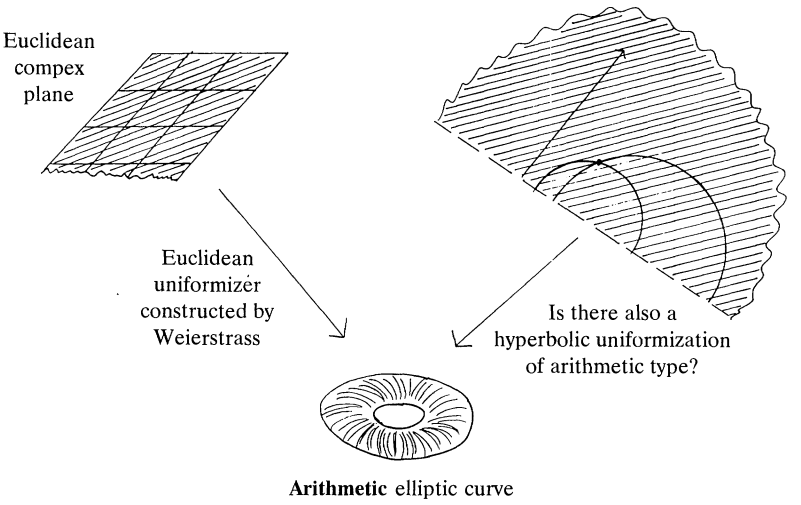


FIG. 13

[23]The formulation we have just given of the conjecture would make it seem "unfalsifiable." But in fact, there are more precise versions of the conjecture which predict, given an arithmetic elliptic curve E, exactly which $\Gamma(N)$ would be involved in a hyperbolic uniformization of arithmetic type for E—these precise versions are known (by the work of Hecke, Eichler, Shimura, Weil, Deligne, Carayol, Faltings, and others) to be equivalent to the one given here. A technical point relevant to this equivalence is treated briefly in an appendix to this expository article. There are also stronger conjectures by Langlands (concerning automorphic representations of reductive groups) and by Serre (concerning 2-dimensional representations of Galois groups over $\mathbb{Q}$) which imply the conjecture of Shimura-Taniyama-Weil.

[24]As Serre remarked, it might be illuminating to formulate a precise conjectural characterization of the class of elliptic curves (necessarily definable over $\overline{\mathbb{Q}}$) which admit hyperbolic uniformizations of arithmetic type. The conjecture of Shimura-Taniyama-Weil asserts, of course, that any elliptic curve definable over $\mathbb{Q}$ admits such a uniformization. Among the elliptic curves definable over quadratic number fields, for example, a *necessary* condition for them to have such a uniformization is that they be $\mathbb{C}$-isogenous to their conjugate (cf. Goro Shimura, Class fields over real quadratic fields and Hecke operators, 95 (1972) 130–190, where the case of real quadratic fields is analyzed and examples are given).

It is the confluence of *two* uniformizations, the Euclidean one, and the (conjectural) hyperbolic one of arithmetic type, that puts an exceedingly rich geometric structure on an arithmetic elliptic curve, and that carries deep implications for arithmetic questions.

4. Why does the Shimura-Taniyama-Weil conjecture imply Fermat's Last Theorem? This is not the place to go into the reams of evidence in favor of the conjecture, or of the overarching Langlands Program of which this conjecture is a minute part. But those readers who have borne patiently with me up to this point deserve some hints, however brief, of the connection between the conjecture which has been the focus of our talk and Fermat's Last Theorem.

Suppose, then, that we are given a *counterexample* to Fermat's Last Theorem, i.e., we are given a particular solution to a Fermat equation, $a^p + b^p = c^p$, where a, b, c are nonzero integers and the exponent p is an odd prime number. We might safely suppose that $p > 150{,}000$ since Fermat's last theorem has been established for odd exponents smaller than that bound. The subsequent argument, however, works for all primes $p \geqslant 5$, which is all that we assume at this point. We can always arrange the equation, by permuting (a, b, c) with appropriate changes of sign, if necessary, to get b to be even and a to be of the form $4k - 1$. Hellegouarch [**H**] and, more recently, Frey [**Fr** 1, 2] consider the arithmetic elliptic curve given by the cubic equation

$$y^2 = x(x - a^p)(x + b^p) \qquad (**)$$

which at first sight may not seem remarkable, but... (as is suggested in [**Fr** 2], formulated and set up in [**S**], and concluded in [**R**])

THEOREM. *The arithmetic elliptic curve* $(**)$ *does not admit a hyperbolic uniformization of arithmetic type.*

COROLLARY. *The conjecture of Shimura-Taniyama-Weil implies that the arithmetic elliptic curve* $(**)$ *does not exist, i.e., the conjecture of Shimura-Taniyama-Weil implies Fermat's Last Theorem.*

For a sketch of the proof of the above theorem, the reader might consult the Bourbaki report of Oesterlé [**O**]. The proof makes essential use of the classical and also the more modern arithmetic theory of modular forms. One supposes that $(**)$ does admit a hyperbolic uniformization of arithmetic type. Then, if ω is a regular differential 1-form on the elliptic curve, pulling ω back to the upper half-plane via the hyperbolic uniformization one gets a differential 1-form on the upper half-plane which after suitable normalization can be written as $f(z)\,dz$ where $f(z)$ is a cuspidal modular form f of weight 2 (the form f is modular for one of Hecke's groups and is an eigenform of the Hecke operators) with integral Fourier coefficients:

$$f(z) = 1 \cdot e^{2\pi iz} + \alpha_2 \cdot e^{4\pi iz} + \alpha_3 \cdot e^{6\pi iz} + \cdots \qquad (\alpha_j \in \mathbb{Z}).$$

Using an approach suggested by a conjecture of Serre, and using the specific form of the discriminant and the conductor of $(**)$[25], Ribet has shown that the

[25]The discriminant of the cubic equation $(**)$ is a *perfect pth power* times a power of 2 (it is precisely $(a^p b^p c^p/2^8)^2$) and the conductor of $(**)$ is an even square-free number.

Fourier coefficients of $f, \alpha_1 = 1, \alpha_2, \alpha_3, \cdots$ are congruent modulo p to the Fourier coefficients of a modular form φ of weight two on Hecke's group $\Gamma_0(2)$. Ribet's argument is startling in its originality and makes use, among many other things, of the very mysterious "Drinfeld switch" which occurs in the description of the "bad fibers" of Shimura curves.

But by our good fortune there is only one modular form of weight two on $\Gamma_0(2)$, the Eisenstein series. That is, we know exactly what φ is, and therefore we know the Fourier coefficients of f, α_j modulo p.

Now, the theorem of Eichler-Shimura links the Fourier coefficients α_j modulo p to questions of $\mathbb{Q}$-rational p-torsion on the arithmetic elliptic curve $(**)$. From our explicit knowledge of α_j mod p, we deduce that there exists a rational point of order p on $(**)$[26].

But any arithmetic elliptic curve, which like $(**)$ has all four points of order 2 rational over $\mathbb{Q}$, cannot have any rational p-torsion for $p \geqslant 5$ (see [**M**]). Contradiction!

I hope, in these few minutes to have given some sense of the eclectic spirit of this conjecture, of how broadly it reaches out towards realms of mathematics that one might, at first, believe to be remote from Arithmetic, and yet how it gets to the heart of Arithmetic matters.

References for §4.

[**Fr** 1]. G. Frey, Rationale Punkte auf Fermatkurven and getwisten Modulkurven, *J. Crelle*, 331 (1982) 185–191.

[**Fr** 2]. _____, Links between stable elliptic curves and certain Diophantine equations, Ann. Univ. Saraviensis, Ser. Math. 1.

[**H**]. Y. Hellegouarch, Courbes elliptiques et équations de Fermat, Thèse, Bésancon, 1972.

[**M**]. B. Mazur, Modular curves and the Eisenstein ideal, *Publ. Math. I.H.E.S.*, 47 (1977) 33–186.

[**O**]. J. Oesterlé, Nouvelles approaches du "Théorème" de Fermat, Séminaire Bourbaki 87/88 *n* 694, *Astérisque*, 161–162 (1988) 165–186.

[**R**]. K. Ribet, On modular representations of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ arising from modular forms, *Inv. Math.*, 100 (1990) 431–476.

[**S**]. J.-P. Serre, Sur les représentations modulaires de degré 2 de $\mathrm{Gal}(\overline{Q}/Q)$, *Duke Math J.*, 54 (1987) 179–230.

Technical appendix: questions of fields of definition. These questions have not been dwelt upon, in the expository account above, because one of the points of my exposition was to emphasize the purely "geometric" nature of the Shimura-Taniyama-Weil conjecture. Nevertheless, the more standard expression of this conjecture (and, in fact, the form of the conjecture that has direct arithmetic consequences) is that given any elliptic curve E over $\mathbb{Q}$, there is a nonconstant mapping *defined over* $\mathbb{Q}$ from a modular curve $X_0(N)$ onto E. Here $X_0(N)$ is viewed as an algebraic curve over $\mathbb{Q}$ (given its "canonical model" structure over $\mathbb{Q}$, as initiated in [**Sh** 1]).

The point of this appendix is to make clear that we have not "cheated" in the statement we gave in the body of our expository account. That is, we will show that the "Shimura-Taniyama-Weil conjecture" as given in our expository text is equivalent to the statement in the preceding paragraph.

[26]Or at least on an arithmetic elliptic curve isogenous to $(**)$ via an isogeny of degree p rational over $\mathbb{Q}$.

Specifically, we shall sketch the proof that if an elliptic curve defined over $\mathbb{Q}$ admits a nonconstant mapping from $X(N)$ *defined over* $\mathbb{C}$, for some N, then it admits a nonconstant mapping from $X_0(N)$ *defined over* $\mathbb{Q}$ as well (but possibly for a *different* value of N). Now this fact is surely a "folk-theorem" either known to, or else easily derivable by, any of the experts. We record some brief indications of its proof here simply because it doesn't seem to be anywhere in the literature (but see [**Ri**] for very closely allied results).

Let K be a field (characteristic 0 is all we shall need) and consider the category $\mathscr{A}_K$ whose objects are abelian varieties over K, and whose morphisms are $\mathbb{Q}$-vector spaces obtained as follows: If A, B are abelian varieties over K, and $\mathrm{Hom}_K(A, B)$ is the abelian group of K-homomorphisms of abelian varieties, then the $\mathbb{Q}$-vector space of morphisms from A to B in $\mathscr{A}_K$ is $\mathrm{Hom}_K(A, B) \otimes \mathbb{Q}$. Thus, two abelian varieties are isomorphic in $\mathscr{A}_K$ if and only if they are K-isogenous. Let A be an abelian variety over K, and let M be a finite-dimensional $\mathbb{Q}$-vector space with continuous $\mathrm{Gal}(\overline{K}/K)$-action (i.e., the action factors through a finite quotient), where $\overline{K}$ is an algebraic closure of K. Then $A \otimes M$ is the unique object of $\mathscr{A}_K$ representing the contravariant functor $B \mapsto \{\mathrm{Hom}_{\overline{K}}(B, A) \otimes M\}^{\mathrm{Gal}(\overline{K}/K)}$ where the superscript $\mathrm{Gal}(\overline{K}/K)$ refers to the elements invariant under the (diagonal) action of $\mathrm{Gal}(\overline{K}/K)$ on $\mathrm{Hom}_{\overline{K}}(B, A) \otimes M$.

Let N be a positive integer, and $X_1(N)_{/\mathbb{Q}}$ the canonical model over $\mathbb{Q}$ of the modular curve attached to the classical group $\Gamma_1(N)$, as in [**Sh** 3]. Let $J_1(N)_{/\mathbb{Q}}$ be the jacobian of $X_1(N)_{/\mathbb{Q}}$, which we view as object of the category $\mathscr{A}_K$. Let E be an elliptic curve over $\mathbb{Q}$, and let $L/\mathbb{Q}$ be a field extension.

Definition. *The elliptic curve E is called* modular ***over*** *L if there is an integer N and a nonconstant mapping $X(N)_{/L} \to E_{/L}$ (rational over L).*

Proposition. *The elliptic curve E is modular over L if and only if E is modular over $\mathbb{Q}$ and more specifically, if and only if there is a nonconstant mapping $X_0(N)_{/\mathbb{Q}} \to E_{/\mathbb{Q}}$ (rational over $\mathbb{Q}$).*

Sketch of proof. We may assume that E does not have complex multiplication over $\mathbb{C}$, for if it did, then it would be modular over $\mathbb{Q}$ [**Sh** 4].

If E is modular over a given field, then it is also modular over any extension field, and if E is modular over $\mathbb{C}$ then it is modular over $\overline{\mathbb{Q}}$, so we need to prove that if E is modular over $\overline{\mathbb{Q}}$, then it is also modular over $\mathbb{Q}$. Recall that there is a nonconstant mapping $X_1(N^2) \to X(N)$ defined over $\overline{\mathbb{Q}}$, so (after a possible change of level N, and reduction from field of definition $\overline{\mathbb{Q}}$ to a suitable finite Galois extension $L/\mathbb{Q}$) we may assume that $E_{/\mathbb{Q}}$ is an elliptic curve without complex multiplications, admitting an L-rational nontrivial mapping $X_1(N) \to E$. Let $G = \mathrm{Gal}(L/\mathbb{Q})$.

By consideration of the Weil restriction from L to $\mathbb{Q}$ of this homomorphism of abelian varieties, and using the fact that $J_1(N)_{/\mathbb{Q}}$ is isogenous (over $\mathbb{Q}$) to a product of abelian varieties A_f for newforms f (Prop. 2.3 of [**Ri**]; for more about the varieties A_f see [**Sh 2, 3**]) we see that there is an irreducible G-module M and a newform f with Fourier coefficients in $\mathbb{C}$ such that A_f is isomorphic in the category $\mathscr{A}_{\mathbb{Q}}$ to $E \otimes M$ (i.e., these abelian varieties are isogenous over $\mathbb{Q}$). Let $F \subset \mathbb{C}$ be the coefficient field of the newform f (compare [**Ri**]). Then $[F : \mathbb{Q}] = \dim A_f = \dim_{\mathbb{Q}}(M)$. We also have that $F = \mathrm{End}_{\mathscr{A}_{\mathbb{Q}}}(A_f)$, by (Cor. 4.2 of [**Ri**]).

Using the $\mathscr{A}_{\mathbb{Q}}$-isomorphism between A_f and $E \otimes M$ and the equality in the previous sentence, we get an imbedding of F into $\mathrm{End}_{\mathscr{A}_L}((E_{/L}) \otimes M) = \mathrm{End}_{\mathbb{Q}}(M)$, where $\mathrm{End}_{\mathbb{Q}}(M)$ means the ring of endomorphisms of the $\mathbb{Q}$-vector space M. This makes M an F-vector space of dimension 1. Since the action of F is rational over $\mathbb{Q}$, it follows that it commutes with the action of G on $(E_{/L}) \otimes M$. It therefore follows that the action of G on the F-vector space M is via a character $\chi : G \to F^* \subset \mathbb{C}^*$. Let $\bar{\chi} : G \to \mathbb{C}^*$ denote the conjugate character, and $f \otimes \bar{\chi}$ the twisted modular form. Using the Eichler-Shimura relations one can show that for almost all prime numbers p, the p-th Fourier coefficient of $f \otimes \bar{\chi}$ is equal to $a_p = 1 + p - N_p$, where N_p is the number of rational points of the (good) reduction of the elliptic curve E to the prime field of characteristic p. It follows that there is a newform φ whose field of Fourier coefficients is $\mathbb{Q}$ and such that for almost all prime numbers p, the p-th Fourier coefficient of φ is a_p. Consequently, the abelian variety A_φ uniformized by the newform φ is an elliptic curve over $\mathbb{Q}$ whose associated l-adic Galois representation (for any l) is isomorphic to that of $E_{/\mathbb{Q}}$. At this point we may deduce (and it suffices, in our situation, to appeal to Theorem 6.1 of [**Ri**] rather than to the general Isogeny Theorem proved by Faltings [**Fa**]) that $E_{/\mathbb{Q}}$ is isogenous over $\mathbb{Q}$ to A_φ.

Remark. It is known that if E is modular over $\mathbb{Q}$, then there is a surjective mapping defined over $\mathbb{Q}$ $X_0(N) \to E$, where N is the conductor of E. See [**St**] for a beautiful discussion of refined questions concerning uniformization of elliptic curves by modular curves.

REFERENCES

[**Fa**]. G. Faltings, Endlichkeitssätze für abelsche Varietäten über Zahlkörpern, *Inv. Math.*, 73 (1983) 349–366.

[**Ri**]. K. Ribet, Twists of modular forms and endomorphisms of Abelian varieties, *Math. Annalen*, 253 (1980) 43–62.

[**Sh** 1]. G. Shimura, Correspondances modulaires et les fonctions ζ de courbes algébriques, *J. Math. Soc. Japan*, 10 (1958) 1–28.

[**Sh** 2]. _____, On the factors of the Jacobian variety of a modular function field, *J. Math. Soc. Japan*, 25 (1973) 523–544.

[**Sh** 3]. _____, Introduction to the Arithmetic Theory of Automorphic Functions, Princeton University Press, 1971.

[**Sh** 4]. _____, On elliptic curves with complex multiplication as factors of the jacobians of modular function fields, *Nagoya Math. J.*, 43 (1971) 199–208.

[**St**]. G. Stevens, Stickelberger elements and modular parametrizations of elliptic curves, *Inv. Math.*, 98 (1989) 75–106.