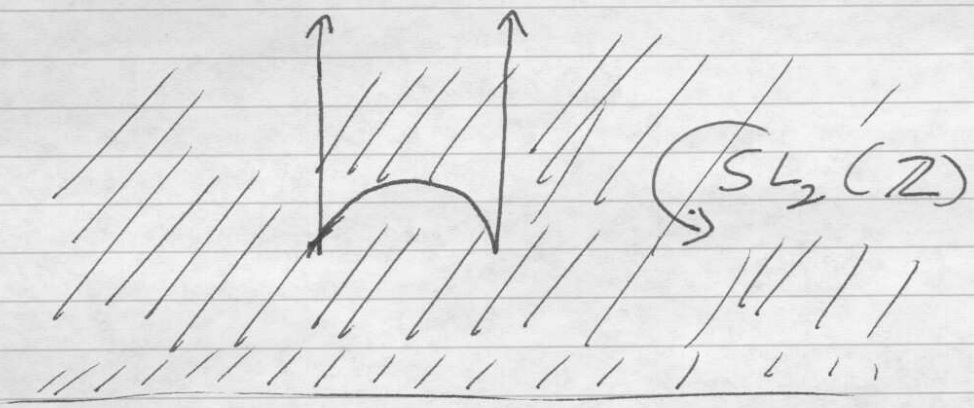
Number Theory of the Upper Half Plane

Kiran S. Kedlaya, MIT



Based on: "Number Theory as Gadfly", Barry Mazur, American Mathematical Monthly, 98 (Aug-Sep 1991), 593-610.

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.

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 **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 $\mathbb{C}$.

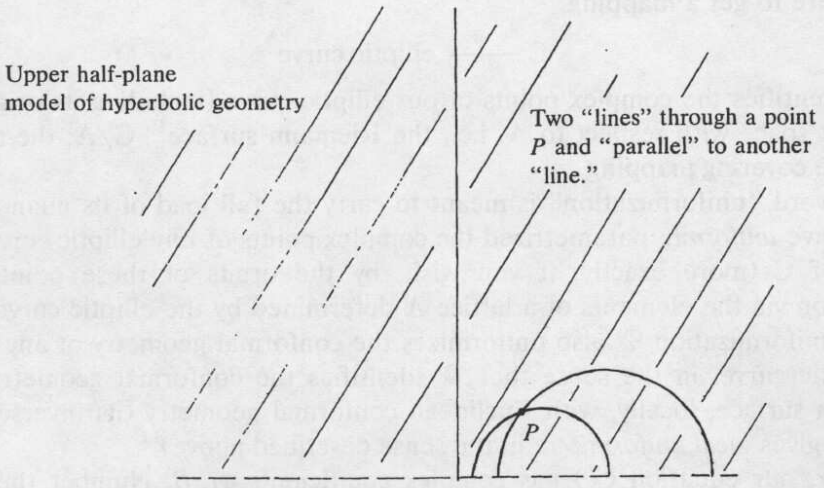


FIG. 10

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, $S : z \mapsto -1/z$.

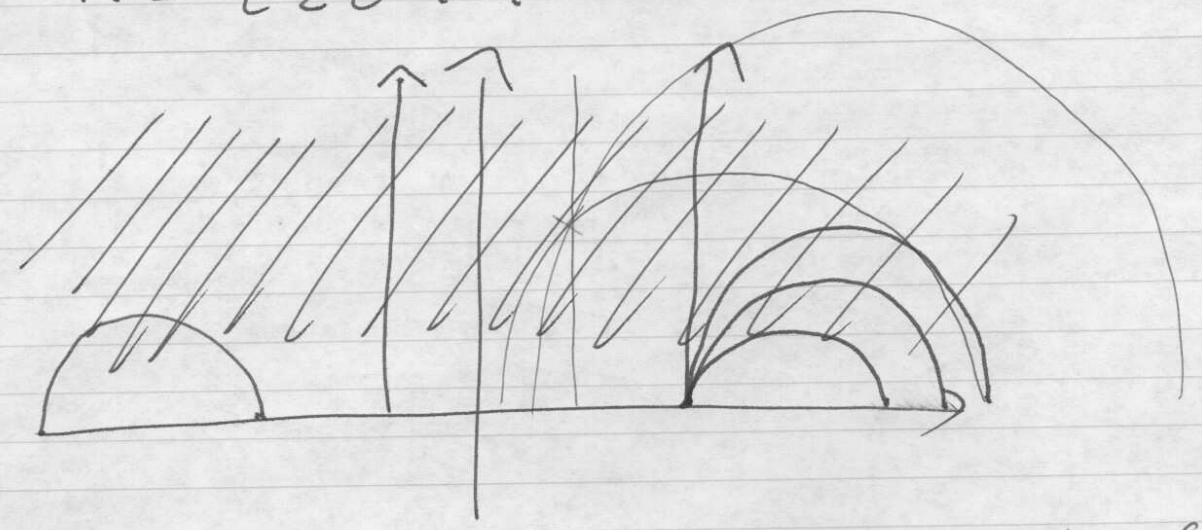
[18]Consider matrices of real numbers of determinant equal to 1, i.e.,

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

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.

The upper half plane

$$\mathbb{H} = \{ z \in \mathbb{C} \mid \mathrm{Im}(z) > 0 \}$$



gives a model, within Euclidean geometry, of the hyperbolic plane. (The "lines" are vertical lines and semicircles centered on the real axis.)

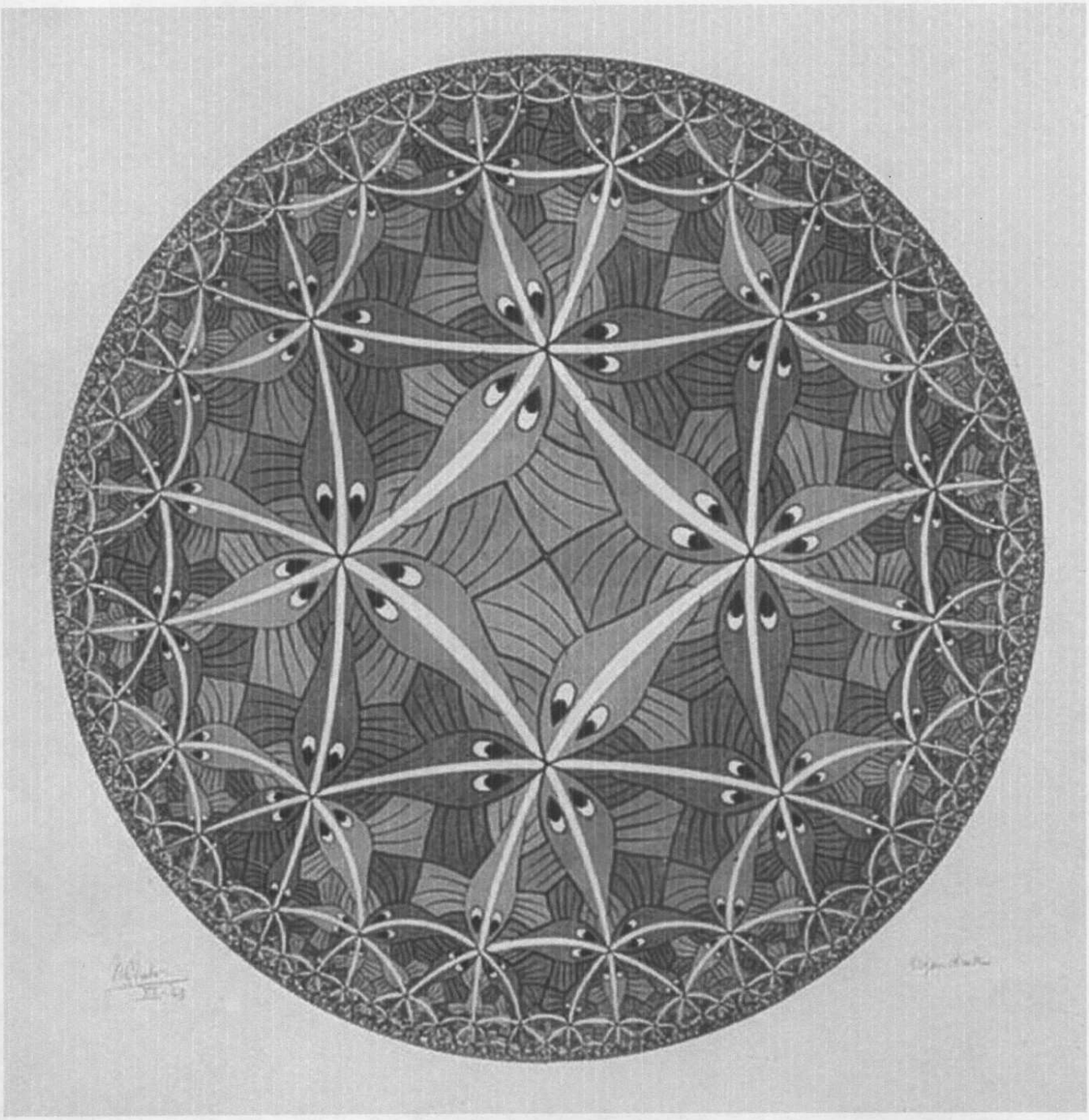
It is acted upon by the group

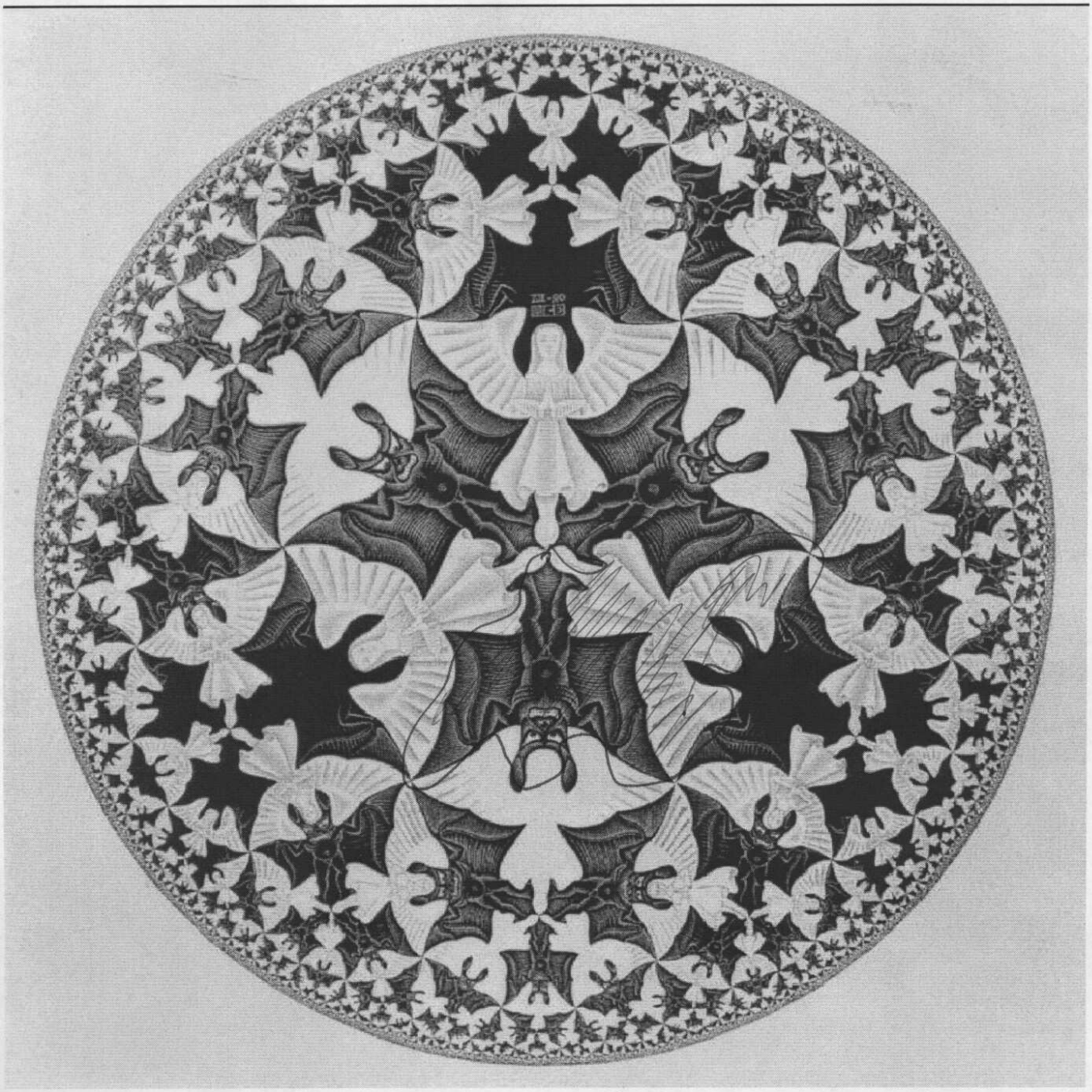
$$SL_2(\mathbb{R}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \begin{array}{l} a,b,c,d \in \mathbb{R}, \\ ad - bc = 1 \end{array} \right\}$$

by linear fractional transformations:

$$z \longmapsto \frac{az+b}{cz+d}.$$

These maps are analytic (locally described by power series) and hence conformal (angle preserving).





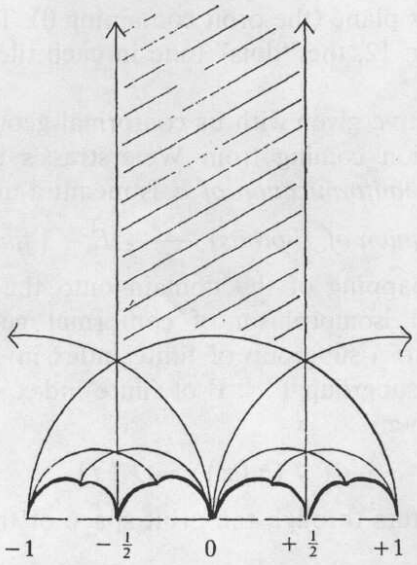


FIG. 11

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

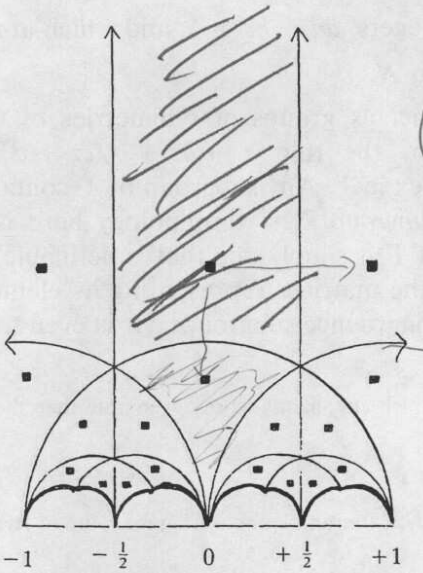
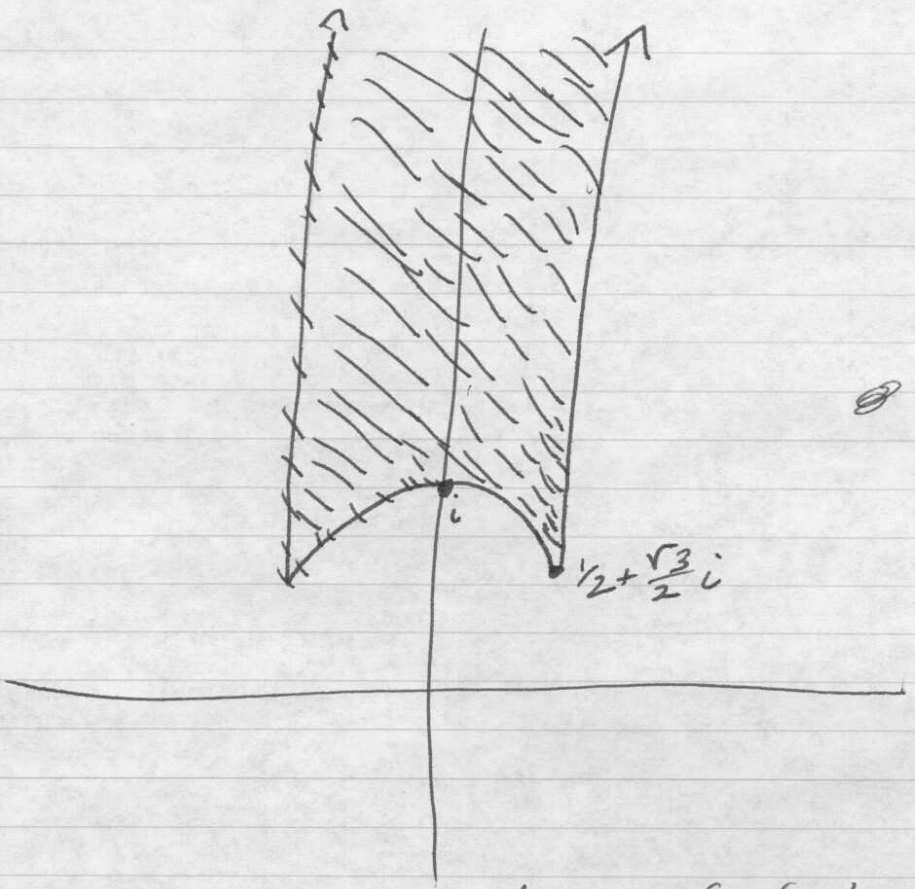


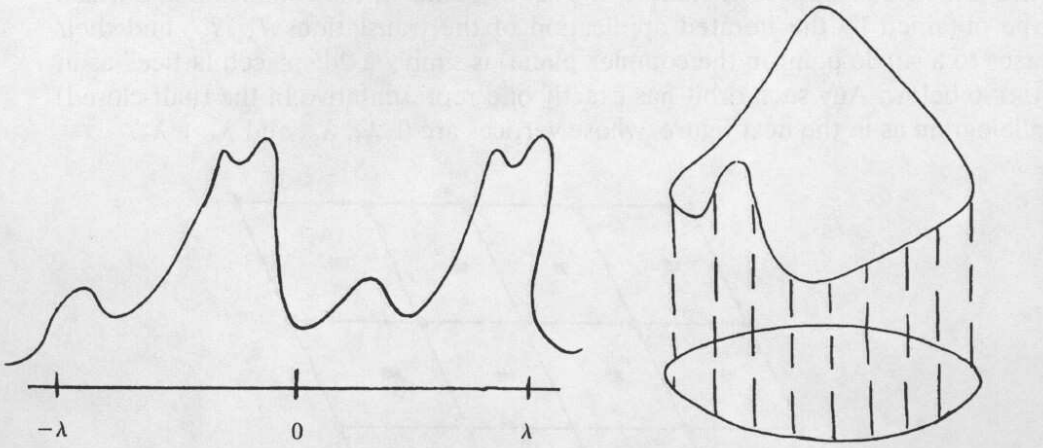
FIG. 12



The above region (including half of the boundary) is a fundamental domain for $SL_2(\mathbb{Z})$ - each point of $\mathcal{H}$ can be moved to exactly one point in the region.

The quotient $SL_2(\mathbb{Z})\backslash\mathcal{H}$ (the set of orbits of $\mathcal{H}$ under $SL_2(\mathbb{Z})$ - glue boundaries in the above picture) is topologically a sphere minus one point (at infinity).

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

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

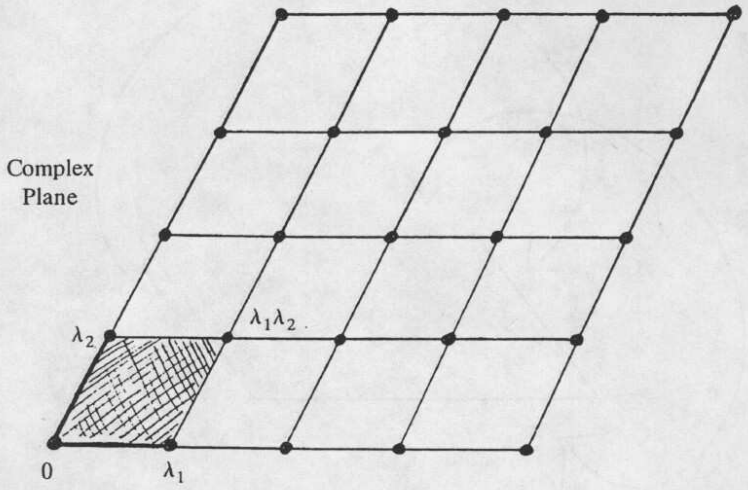


FIG. 5

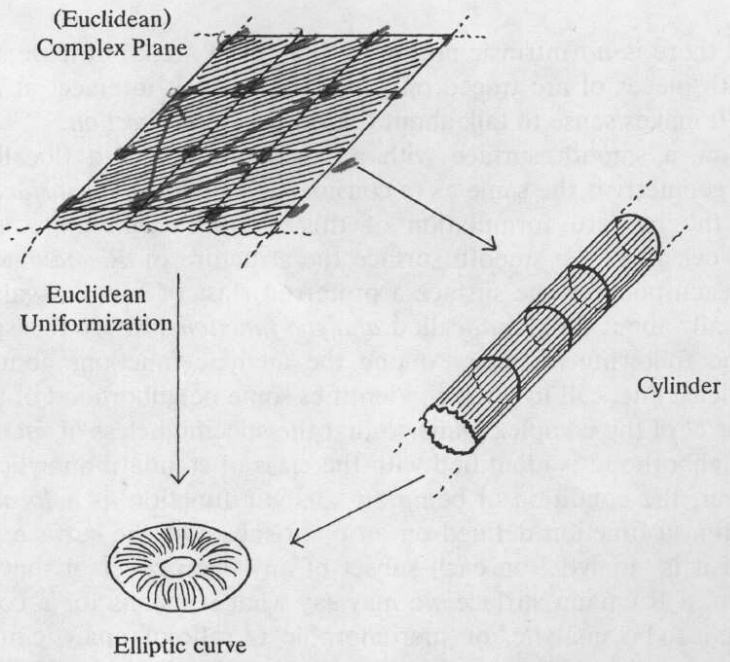


FIG. 8. The "covering mapping" which brings the complex plane $\mathbb{C}$ to the orbit space $\mathbb{C}/\Lambda$ may also be visualized as a two-stage process, where in the first stage the plane is wrapped around a cylinder, and in the second stage the cylinder is wrapped around a torus.

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-

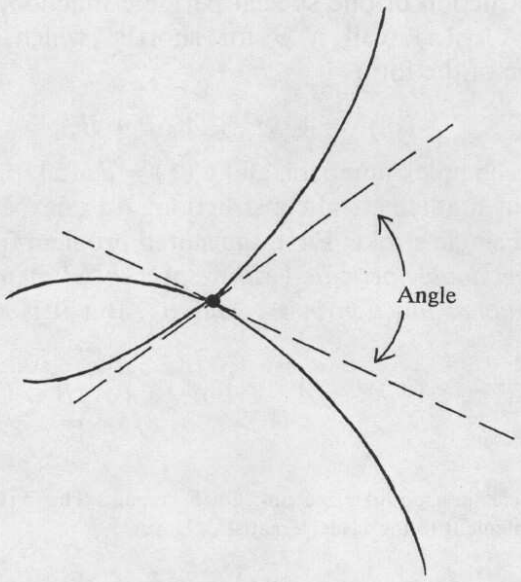
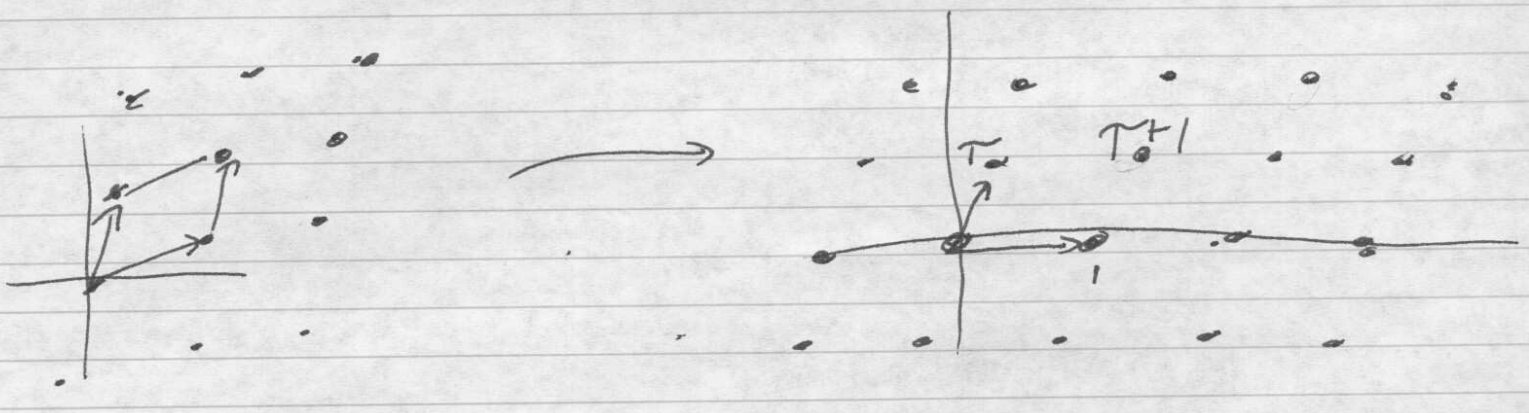


FIG. 9. In conformal geometry, there is no invariant notion of "length" of an arc, but there is a notion of angle.

The quotient $SL_2(\mathbb{Z})\backslash\mathcal{H}$ classifies <u>lattices in $\mathbb{C}$ up to homothety</u>.

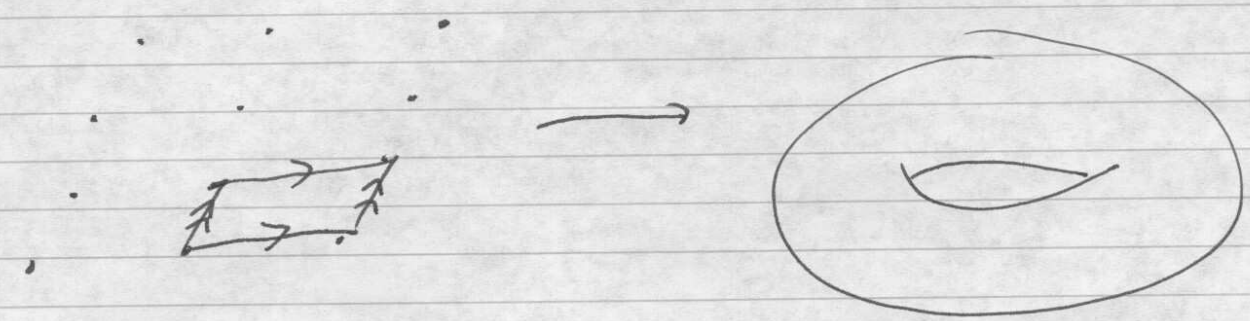
$z \longrightarrow \alpha z \quad (\alpha \in \mathbb{C}^*)$



each lattice can be rescaled so as to be generated by 1 and τ for some $\tau \in \mathcal{H}$; if $\tau' \in \mathcal{H}$ also, then the two lattices are homothetic if and only if

τ' ∈ image of τ under $SL_2(\mathbb{Z})$.

In other words, $SL_2(\mathbb{Z})\backslash\mathcal{H}$ is the <u>moduli space</u> of lattices up to homothety, or of <u>complex tori</u>.



Interlude: doubly periodic functions

For Λ a lattice in $\mathbb{C}$, Weierstrass constructed a complex analytic function $\wp : \mathbb{C} - \Lambda \to \mathbb{C}$

which is <u>doubly periodic</u>:

$$\wp(z + \lambda) = \wp(z) \quad \text{for all } \lambda \in \Lambda$$

with poles at Λ

such that any doubly periodic analytic function from $\mathbb{C}-\Lambda$ to $\mathbb{C}$ can be written as a polynomial in $\wp$ and $\wp'$.
Moreover, $\wp$ satisfies a nonlinear differential equation:

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

for some $A, B \in \mathbb{C}$.

Fine print: $\wp$ is given by the infinite sum

$$\frac{1}{z^2} + \sum_{\lambda \in \Lambda - \{0\}} \left[\frac{1}{(z-\lambda)^2} - \frac{1}{\lambda^2} \right].$$

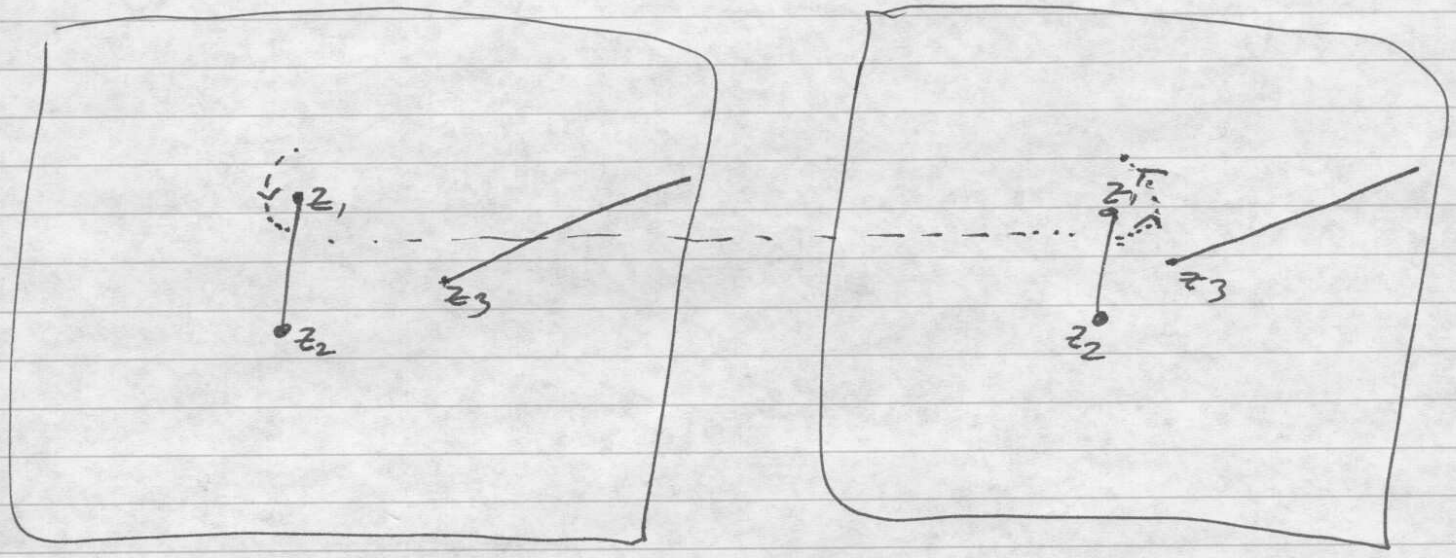
$$\wp' = -2 \sum_{\lambda \in \Lambda} \frac{1}{(z-\lambda)^3}$$

The map $z \longmapsto (\wp(z), \wp'(z))$ identifies $\mathbb{C}/\Lambda$ (the quotient) with the complex solutions (x, y) of the equation

$$y^2 = 4x^3 + Ax + B$$

plus a point at infinity; these form an elliptic curve.

In particular, the points of the elliptic curve form a complex torus. Another way to see this: glue two slitted copies of $\mathbb{C}$ so that $\sqrt{4x^3+Ax+B}$ becomes a well-defined function on the result.



z_1, z_2, z_3 = roots of $4z^3 + Az + B = 0$

Aside: the points of $\mathbb{C}/\Lambda$ have an addition law, so via $\wp$, we get one on the elliptic curve also.

That law has a geometric interpretation: the points

$(x_1, y_1), (x_2, y_2), (x_3, y_3)$ on the elliptic curve $y^2 = 4x^3 + Ax + B$

add up to zero if and only if they are collinear, ie; if

$$\det \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix} = 0.$$

This can be used to "generate" points on the elliptic curve; this was already detected by Diophantos!

In short, the quotient $SL_2(\mathbb{Z})\backslash\mathcal{H}$ is the

moduli space of elliptic curves

The quotients $\Gamma\backslash\mathcal{H}$, for Γ a congruence subgroup of $SL_2(\mathbb{Z})$, have related interpretations. e.g. if

$$\Gamma_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \begin{array}{l} a,b,c,d \in \mathbb{Z} \\ ad-bc=1 \\ a \equiv 1 \pmod N,\ c \equiv 0 \pmod N \end{array} \right\}$$

then $\Gamma\backslash\mathcal{H}$ parametrizes elliptic curves plus a choice of a point on the curve whose N-th multiple is the point at infinity (= the zero element for the addition law)

To get to number theory, follow a general principle:

moduli spaces (of geometric objects) have interesting geometry <u>in their own right!</u>

The study of the rational points on elliptic curves

$$y^2 = 4x^3 + Ax + B$$

with $A, B \in \mathbb{Q}$ is as old as recorded history. For instance, rational points on

$$y^2 = 4x^3 - 4n^2 x$$

give rise to Pythagorean right triangles (with all integer side lengths) and area $n \square$

More recently, it was linked to Fermat's Last Theorem... more on that later.

The number theory comes in when we restrict to

$$SL_2(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \begin{array}{l} a, b, c, d \in \mathbb{Z} \\ ad - bc = 1 \end{array} \right\}$$

(the <u>modular group</u>) or to subgroups of $SL_2(\mathbb{Z})$ defined by congruence conditions, e.g.

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \begin{array}{l} a, b, c, d \in \mathbb{Z} \\ ad - bc = 1 \\ c \equiv 0 \pmod{N} \end{array} \right\}.$$

We will be interested in the quotients of $\mathcal{H}$ by such subgroups — these are the <u>modular curves</u>.

(Why "curves" and not "surfaces"? As complex-analytic manifolds, they are <u>one</u>-dimensional.)

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]

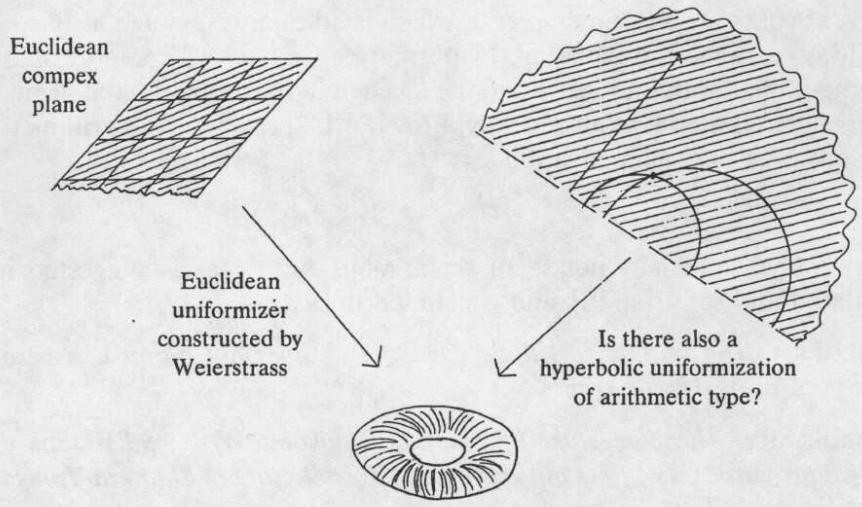


FIG. 13

[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).

The modularity of elliptic curves [conjectured by Shimura, Taniyama, Weil]

[Breuil-Conrad-Diamond-Taylor, after Wiles]

For any elliptic curve $E: y^2 = 4x^3 + Ax + B$ over $\mathbb{Q}$, there is a <u>covering map</u> $X_0(N) \to E$ for a certain integer N (the <u>conductor</u> of E).

(This can also sometimes happen when A, B are <u>algebraic</u> numbers – roots of rational polynomials – but the situation is not fully understood.)

Modularity continued

There also exists a (cuspidal) modular form of level N, i.e. a function $f: \mathcal{H} \to \mathbb{C}$ (complex analytic) such that

$$f(z) = f\left(\frac{az+b}{cz+d}\right)(cz+d)^{-2} \quad \text{for } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$$

and $\lim_{z \to \infty} f\left(\frac{az+b}{cz+d}\right) \to 0$ for all $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})$.

which can be written as $\sum_{n=1}^{\infty} a_n q^n \qquad (q = e^{2\pi i z})$

such that the number of solutions of the congruence

$$y^2 \equiv 4x^3 + Ax + B \pmod{p} \qquad x, y \in \{0, \ldots, p-1\}$$

equals $p - a_p$ for all but finitely many p.

$$y^2 + y = x^3 - x^2 - 10x - 20$$

$\Gamma_0(11)$ $\qquad q \prod_{n=1}^{\infty} (1-q^n)^2 (1-q^{11n})^2$

$$q - 2q^2 - q^3 + 2q^4 + q^5 + 2q^6 - 2q^7 - \cdots$$

Fermat's Last Theorem: for $n \geq 3$ the equation

$$a^n + b^n = c^n$$

has no solutions in nonzero integers a, b, c.

This reduces to the case $n = p$ prime (the case $n = 4$ having been resolved by Fermat). Also, we may take b even and $a \equiv -1 \pmod 4$.

Theorem [Hellegouarch, Frey, Serre, Ribet]:

if (a, b, c) were a counterexample to FLT as above, the elliptic curve

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

could not satisfy the modularity theorem!!

Hence modularity (or even the special case due to Wiles, for "semistable" elliptic curves) implies FLT!