GENERALIZED NON-ABELIAN RECIPROCITY LAWS: A CONTEXT FOR WILES'S PROOF

AVNER ASH AND ROBERT GROSS

ABSTRACT. In as elementary a way as possible, we place Wiles's proof of Fermat's Last Theorem into the context of a general description of reciprocity conjectured to obtain between algebraic varieties defined over $\mathbf{Q}$ and Hecke eigenvectors in the homology of the spaces of lattices in $\mathbf{R}^n$.

We shall find the Cube of the Rainbow,
Of that, there is no doubt.
But the Arc of a Lover's conjecture
Eludes the finding out.

Emily Dickinson

During the last few decades, the field of number theory has been increasingly permeated by the theory of automorphic forms and automorphic representations. This phenomenon often goes under the rubric of the "Langlands program," although it involves the work of many mathematicians, including R. Langlands, J.-P. Serre, and G. Shimura, to name only three of many. Recently, this program became especially prominent because it forms a background to Wiles's proof of Fermat's last theorem.

As explained by Langlands in [14], parts of this program can be viewed as a vast generalization of reciprocity laws familiar in number theory, such as quadratic reciprocity and Artin reciprocity. Some of the ultimate conjectures along these lines are spelled out in Clozel's article [4], linking up *motives* and *automorphic representations*; see also Gelbart [11].

Both of the italicized terms in the previous sentence are rather technical objects. Our goal in this article is to describe a version of these generalized nonabelian reciprocity conjectures: a version accessible to a reader who knows nothing beyond basic algebra and the definition of homology groups. Because of these self-imposed limitations, we shall be unable to state the conjectures in their full strength or with total precision. To compensate for these necessary shortcomings, we have included three basic examples that should give the flavor of the conjectures. The first involves quadratic reciprocity, the second a modular elliptic curve, and the third a (probably) automorphic algebraic surface. Moreover, the conjecture we do state is strong enough to allow its application to Fermat's last theorem. At the appropriate place, we shall discuss the term "reciprocity" and what may be thought of as being reciprocated.

By reducing the prerequisites, we hope we have increased the number of readers who can obtain some glimpses of this beautiful landscape of generalized reciprocity. We also believe the formulation of the conjectures in terms of spaces of lattices in $\mathbf{R}^n$ has a certain surprising beauty of its own. We should point out, however, that it is very unlikely that progress can be made in *proving* the conjectures on this elementary level.

We begin with a quick review of Wiles's proof of Fermat's last theorem. For a good survey of this together with references to the work of Wiles and his predecessors necessary to the proof, see [15].

Date: April 7, 2000.
1991 *Mathematics Subject Classification.* 11-02.

Let n be a prime greater than 3, and suppose that $(a, b, c) \in \mathbf{Z}^3$ is a nontrivial solution of the Fermat equation $a^n + b^n = c^n$. It was noticed that one could take these three integers and use them to define a particular elliptic curve E,

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

where we think of E as lying in the xy-plane.

Elliptic curves have been the object of intense study for much of this century, though many things about them remain unknown. One particular property of elliptic curves is that they can be *modular*. We shall give below a definition of modularity from our point of view that is equivalent to the standard one. Wiles proved that this particular elliptic curve is modular; Ribet had earlier proved that E is not modular. Therefore, it cannot exist, and so a, b, and c cannot exist either.

The work of Wiles [23], Taylor and Wiles [21], Diamond [7], and Conrad, Diamond, and Taylor [5] shows that a rather large class of elliptic curves is modular. Just recently, Christophe Breuil, Brian Conrad, Fred Diamond, and Richard Taylor have announced a proof of the modularity conjecture [6], which asserts that every elliptic curve defined over $\mathbf{Q}$ is modular.

This conjecture is part of a vast program which sets up a conjectural correspondence between two large sets. On one side of the correspondence, roughly speaking, is the collection of systems of simultaneous polynomial equations with rational coefficients. A member of this collection is called a "variety defined over $\mathbf{Q}$," and is relatively simple to comprehend. (More precisely, this side of the correspondence contains motives, which we shall discuss below.) For example, elliptic curves defined over $\mathbf{Q}$ are varieties, because they can be described in general by an equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6,$$

where the a_i are integers, and the discriminant Δ of the elliptic curve, which is a polynomial function of the numbers a_i, is non-zero. Points on the curve over the ring R are simply solutions $(x, y) \in R^2$ to the given equation.

The formula for the discriminant is quite complex, and may be found in [19, p. 46]; if we restrict to the simpler family of curves given by the equation

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

then $\Delta = -16(4A^3 + 27B^2)$. Except for a constant factor, this is the same as the discriminant of the cubic polynomial $x^3 + Ax + B$. For future reference, we mention a more subtle number N called the *conductor*, which consists of a product of the primes dividing Δ raised to certain powers.

The other side of the correspondence is harder to grasp; it consists of certain automorphic representations of reductive groups defined over $\mathbf{Q}$. (Linking these two collections is something even harder to get an elementary handle on: representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, which we shall discuss further in the Appendix.) To simplify our exposition, in this article we shall consider only a certain subset of automorphic representations: those which are "geometric" for $\mathrm{GL}(n)$ with constant coefficients. This particular subset is still rich enough to allow us to define the concept of modularity for elliptic curves. This narrowing of scope allows us to replace automorphic representations by simpler equivalent objects, called $\mathbf{H}$-eigenelements, to be explained below. We denote by α an element of this collection.

The first side of the correspondence will now consist of a more narrowly defined collection of varieties that will still include elliptic curves. We shall denote such a variety by V. Our subprogram of the general reciprocity program is the following correspondence:

$$\begin{array}{ccc} \left(\begin{array}{c}\text{certain varieties (including elliptic}\\ \text{curves) defined over } \mathbf{Q}\end{array}\right) & \leftrightsquigarrow & (\ \mathbf{H}\text{-eigenelements}\ ) \\ V & \longleftrightarrow & \alpha \end{array}$$

We need to explain more precisely the terms and nature of this correspondence. First, the left-hand side: by multiplying through by the common denominator of all fractions used to define the variety V, we may assume that in fact the equations used to define V have coefficients in $\mathbf{Z}$ rather than in $\mathbf{Q}$. If R is any ring, we can then use $V(R)$ to mean the solutions to the equations which are in R. In other words, if V is defined by polynomials $\{g_i(x_1, \ldots, x_m) = 0, i = 1, \ldots, n\}$, then

$$V(R) = \{(a_1, \ldots, a_m) \in R^m \mid g_i(a_1, \ldots, a_m) = 0 \text{ for } i = 1, \ldots, n\}.$$

In particular, we can take R to be the field $\mathbf{F}_{p^m}$, the field of p^m elements, where p is any prime and m is any positive integer. Because the field is finite, we get a set of integers $\#V(\mathbf{F}_{p^m})$ for primes p and positive integers m. These integers are encoded in the Hasse-Weil zeta function of V as described below.

Unfortunately, making precise which equivalence classes of varieties we consider on the left-hand side of this correspondence is quite complex. In fact, we actually need "pieces" of varieties, or, more precisely, "pieces" of their cohomology, called *motives*. When the cohomology of a variety V breaks up into a number of motives, all of which except one are easily understood, then V as a whole can stand in for its nontrivial motivic piece in the formulas that give a description of the correspondence $V \leftrightarrow \alpha$. This is what happens in the three examples discussed in this paper. Therefore, the first-time reader might wish to skim this *ad hoc* introduction to motives. Advanced readers may consult [12] for more information about motives.

A motive is a piece of the cohomology of a variety defined over $\mathbf{Q}$. Just like the complex cohomology of a complex variety, motives have Hodge types, to which we shall need to refer for the sake of accurately stating the conjectures below. One projective variety V can "contain" many motives, and one motive can appear in various different varieties. A motive M gives rise to an L-function $L(M, t)$ with an Euler product

$$L(M, t) = L_\infty(M, t) \prod_p L_p(M, t),$$

where p runs over all prime numbers. If the variety V yields the set of motives $\{M_i\}$, then

$$\prod_i L(M_i, t)^{\varepsilon_i} = Z(V, t) = Z_\infty(V, t) \prod_p Z_p(V, t)$$

where $Z(V, t)$ is the Hasse-Weil zeta-function of V, and ε_i is -1 or 1 depending on whether M_i appears in an odd- or even-degree cohomology group of V. The local factor $Z_p(V, t)$ is defined by

$$Z_p(V, t) = \exp\left(\sum_{m=1}^{\infty} \frac{\#V(\mathbf{F}_{p^m})}{m} t^m\right).$$

In fact, we even have the local factorization

$$Z_p(V, t) = \prod_i L_p(M_i, t)^{\varepsilon_i}$$

for each prime p.

An example could not but help to clarify these concepts. Consider the variety $V = \mathbf{P}^2$, defined over $\mathbf{Q}$. We can think of $\mathbf{P}^2$ as the disjoint union $\mathbf{A}^0 \amalg \mathbf{A}^1 \amalg \mathbf{A}^2$. Therefore, $\#\mathbf{P}^2(\mathbf{F}_{p^m}) = 1 + p^m + p^{2m}$.

We have

$$\begin{aligned}\log Z_p(V,t) &= \sum_{m=1}^{\infty} \frac{1+p^m+p^{2m}}{m} t^m \\ &= \sum_{m=1}^{\infty} \left[\frac{t^m}{m} + \frac{(pt)^m}{m} + \frac{(p^2t)^m}{m} \right] \\ &= -\left(\log(1-t) + \log(1-pt) + \log(1-p^2t) \right),\end{aligned}$$

so

$$Z_p(V,t) = \frac{1}{(1-t)(1-pt)(1-p^2t)}.$$

It is traditional to substitute $t = p^{-s}$, and view s as a complex variable. We then have

$$Z_p(V,p^{-s}) = \left((1-p^{-s})(1-p^{1-s})(1-p^{2-s}) \right)^{-1},$$

and therefore

$$\prod_p Z_p(V,p^{-s}) = \zeta(s)\zeta(s-1)\zeta(s-2).$$

In this case, the three motives corresponding to V are exactly $H^i(V)$ for $i = 0, 2, 4$, and $L(M_i, p^{-s}) = \zeta(s - i/2)$.

In general, for any smooth projective variety V defined by equations with integral coefficients, and for almost all primes p, $Z_p(V,t)$ is a rational function with

$$Z_p(V,t) = \frac{P_{1,p}(t)P_{3,p}(t)\cdots}{P_{0,p}(t)P_{2,p}(t)\cdots}$$

with the factorization

$$P_{i,p}(t) = \prod_r (1 - \beta_{r,i}^{(p)} t).$$

Let $\overline{V} = V(\overline{\mathbf{F}_p})$. Then the Frobenius element ϕ_p acts on the étale cohomology of $\overline{V}$, with $\beta_{r,i}^{(p)}$ the reciprocal eigenvalues of ϕ_p. We have $|\beta_{r,i}^{(p)}| = p^{i/2}$. Finally,

$$\#V(\mathbf{F}_{p^m}) = \sum_{r,i} (-1)^i (\beta_{r,i}^{(p)})^m.$$

Now, a motive M corresponds first to a choice of a single index i and a positive integer n, and then for each prime p a choice of a subset S_p of $\{\beta_{r,i}^{(p)}\}$ of cardinality n such that S_p is the set of reciprocal eigenvalues of ϕ_p acting on a "geometrically defined piece" of the étale cohomology of V. For example, the "piece" might be all of $H^i(V)$. See the discussion of [10] later in this paper for an example of a motive that is not all of $H^i(V)$.

We can now define the L-function of a motive M by setting

$$L_p(M,s) = \prod_{\beta \in S_p} (1 - \beta p^{-s})^{-1}, \qquad \text{for almost all } p.$$

Note the customary change of variable from t to s. We omit the definition for "bad" primes and for L_∞, but we mention that $L_\infty(M,s)$ depends on the Hodge type of M. Then

$$L(M,s) = L_\infty(M,s) \prod_p L_p(M,s).$$

Now we can say: the left-hand side of the correspondence consists of motives of dimension n with Hodge type $(n-1,0) \oplus (n-2,1) \oplus \cdots \oplus (0,n-1)$.

On the right-hand side of the correspondence, $\mathbf{H}$ will denote a ring of operators acting on certain homology groups, and this action has eigenelements. The eigenvalues of the operators in $\mathbf{H}$ acting on α will also be a set of numbers, and the reciprocity conjecture specifies a relationship between these eigenvalues and the numbers $\#V(\mathbf{F}_{p^m})$. These relationships are precisely given by an equality of L-functions; but first we have to define the α and their L-functions.

What are these mysterious $\mathbf{H}$-eigenelements? We begin with a review of a common definition:

Definition. A *lattice* $\Lambda \subset \mathbf{R}^n$ is a free abelian group generated by a basis of $\mathbf{R}^n$.

Our next definition is less common, but also simple to comprehend:

Definition. A *level N structure* on an n-dimensional lattice Λ is a group isomorphism $\phi : \Lambda/N\Lambda \to (\mathbf{Z}/N\mathbf{Z})^n$.

We could consider the collection of all lattices in $\mathbf{R}^n$ with level N structure, but that turns out to be too large a class. Instead, we consider two such lattices to be equivalent if one can be obtained from the other by proper Euclidean motion (that is, an orthogonal transformation of determinant 1), and positive homothety (change of scale), where we always view $\mathbf{R}^n$ as endowed with its usual Euclidean structure.

Definition. $L_n(N) = \{(\text{lattice } \Lambda \subset \mathbf{R}^n, \text{level } N \text{ structure } \phi)\} / \langle \text{proper Euclidean motion, positive homotheties} \rangle$.

Before we consider a few examples, we put a topology on our set $L_n(N)$. Given $(\Lambda, \phi) \in L_n(N)$, fix a basis for Λ, and then consider nearby points to be those lattices gotten by small perturbations of the basis in $\mathbf{R}^n$, keeping the level N structure ϕ the same. These spaces of lattices come into the game because they are locally symmetric spaces on which automorphic forms naturally live.

The simplest example is $L_1(1)$. Because our lattices are equivalent up to homothety, we can take a basis of our lattice Λ to be 1. A level 1 structure would be an isomorphism from $\Lambda/\Lambda \to \mathbf{Z}/\mathbf{Z}$, and since each group contains only the identity element, there can be only one such isomorphism. Therefore, $L_1(1)$ consists of a single point.

A more enlightening example is given by $L_1(N)$. Again, there is only one lattice Λ up to homothety, so we can take $\Lambda = \mathbf{Z}$. Our level N structure is an isomorphism $\phi : \mathbf{Z}/N\mathbf{Z} \to \mathbf{Z}/N\mathbf{Z}$. Such a map is defined by $\phi(1)$, and since ϕ must be invertible, $\phi(1)$ must be invertible, and hence in $(\mathbf{Z}/N\mathbf{Z})^*$. Therefore, $L_1(N) \cong (\mathbf{Z}/N\mathbf{Z})^*$.

In order to see how $L_n(N)$ can contain more interesting geometric information, we consider one more example: $L_2(1)$. We start with a two-dimensional lattice Λ. We take a vector of minimal length in Λ, and by means of rotations and stretching, we can take that vector to be $(1,0)$. We are free to take any linearly independent vector in Λ as the second basis vector. Take any such vector (x,y), where we may suppose that $x \neq 0$ and $y > 0$. By adding multiples of $(1,0)$ to this vector, we may suppose that $-\frac{1}{2} \leq x \leq \frac{1}{2}$. Since this vector must have length at least as large as $(1,0)$, we also know that $x^2 + y^2 \geq 1$.

Furthermore, there are two identifications we can make. Clearly, $(-\frac{1}{2}, y)$ is equivalent to $(\frac{1}{2}, y)$. Less obviously, there is another identification. If the second chosen basis vector (x,y) has length 1, we can rotate this vector to $(1,0)$, and then the vector formerly at $(1,0)$ will rotate to $(x,-y)$. After we multiply by -1 to get a positive second coordinate, we see that (x,y) is equivalent to $(-x,y)$ when $x^2 + y^2 = 1$.

After making these identifications, our space of lattices is topologically equivalent to a sphere with one point removed. In fact, taking the strip

$$\left\{(x,y) \in \mathbf{R}^2 \,\middle|\, -\tfrac{1}{2} \leq x \leq \tfrac{1}{2}, x^2 + y^2 \geq 1, y > 0\right\},$$

and identifying the left and right vertical edges, gives a cylinder. Then glueing (x, y) to $(-x, y)$ for those points (x, y) with $x^2 + y^2 = 1$ collapses the circle at the bottom of the cylinder to an arc. Topologically, we now have an open cup, which is homeomorphic to a sphere minus a point. (A more detailed discussion can be found in many places in the literature; see, for example, [16, Ch. VII.1], for a different explanation and a picture.) As before, the level 1 structure does not add any further detail, since ϕ will be a map from the trivial group to the trivial group.

Although the general picture is obviously much more complex, certain features of this situation will remain true. For all $n \geq 1$, $N \geq 1$, $L_n(N)$ will be a "nice" topological space with $\#(\mathbf{Z}/N\mathbf{Z})^*$ components, classified by $\det \phi$. The situation for $n = 2$ is especially favorable: $L_2(N)$ is the disjoint union of Riemann surfaces, and can be defined by equations with $\mathbf{Q}$-coefficients. There are formulas for the genus of $L_2(N)$; see [18, Ch. 1]. If $n \geq 3$, then $L_n(N)$ is not an algebro-geometric space, but only a V-manifold (a manifold if $N \geq 3$).

We next consider the homology of these spaces: $H_d(L_n(N), \mathbf{C})$, which is the group of d-dimensional cycles modulo d-dimensional boundaries. This is always a finite-dimensional vector space.

There is extra structure, given by the action of the *Hecke algebra* $\mathbf{H}$, defined as follows. Let p be any prime not dividing N, and let k be any integer between 0 and n, inclusive. We can then define

$$\begin{aligned} T_p^{(k)} : L_n(N) &\longrightarrow L_n(N) \\ (\Lambda, \phi) &\longmapsto \{(\Lambda', \psi) \mid \Lambda' \subset \Lambda, \Lambda/\Lambda' \cong (\mathbf{Z}/p\mathbf{Z})^k, \psi = \phi|_{\Lambda'}\} \end{aligned}$$

There are several observations to be made about these operators $T_p^{(k)}$.

1. Because $p \nmid N$, ψ is again a level N structure.
2. $T_p^{(k)}$ is not a function, but rather a one-to-many map. In fact, the image of (Λ, ϕ) is finite, and the cardinality is given by the number of k-planes in $(\mathbf{Z}/p\mathbf{Z})^n$.
3. The action on the homology groups is in fact a function. In fact, $T_p^{(k)}$ maps a cycle to the union of all of its images, but that union is again a cycle. Therefore, we do have a well-defined function $T_p^{(k)} : H_d(L_n(N), \mathbf{C}) \to H_d(L_n(N), \mathbf{C})$.
4. These functions $T_p^{(k)}$ all commute. Therefore, there are simultaneous eigenclasses α. If we write $T_p^{(k)}(\alpha) = a_p^{(k)}\alpha$, then one can prove that these numbers $a_p^{(k)}$ are algebraic integers. For a fixed α, they in fact generate a finite-degree extension of $\mathbf{Q}$.
5. If $N|N'$, and $\alpha \in H_d(L_n(N), \mathbf{C})$ is an $\mathbf{H}$-eigenelement, there always exists $\alpha' \in H_d(L_n(N'), \mathbf{C})$, an $\mathbf{H}$-eigenelement with eigenvalues equal to those of α for those primes p not dividing N'.
6. In principle, these eigenclasses α and numbers $a_p^{(k)}$ are computable. For $n = 1$ and any N, the computation is easy. For $n = 2$ and relatively small N, the computation is not impossible; there is much help coming from the theory of modular forms and elliptic curves. For $n = 3$ and relatively small N, a computer can provide the answers [1]. For $n = 4$ and very small N, there is some research in progress by the first author, Paul Gunnells, and Mark McConnell.

In theory, whenever one can get a correspondence between a variety V (really a motive M) and such an α, there is information to be gained. Such correspondences are "generalized reciprocity laws," so called because quadratic reciprocity can be interpreted as such a correspondence, as we shall see shortly.

The term "reciprocity" seems to go back to Legendre, as quoted on p. 328 of [22]. Originally, the term referred to reciprocity between two primes p and q: whether or not p was a square modulo q being determined according to a simple rule depending on whether or not q was a square modulo p. Eventually, this was interpreted as a property of ϕ_p, the Frobenius element at p, restricted to the Galois group of $\mathbf{Q}(\sqrt{\pm q})$ over $\mathbf{Q}$ and *vice versa*. Later, the term "reciprocity" was extended to

a variety of rules that told how ϕ_p acted in various situations; see [24] for a good introduction. An early example of a non-abelian reciprocity law was given by Shimura [17]. Its modern use is explained by Langlands [14, pp. 408–409] as including assertions "that an L-function defined by diophantine data, that is, by an algebraic variety over a number field, is equal to an L-function defined by analytic data, that is, by an automorphic form." See also Tate's article [20] in the same volume, and the Appendix below, for some examples of what kind of concrete information a reciprocity law can provide.

We can now offer a precise conjecture. An $\mathbf{H}$-eigenvector corresponds to a motive as follows. We can define an L-function corresponding to the $\mathbf{H}$-eigenelement α by defining

$$L_p(\alpha, s) = 1 - a_p^{(1)} p^{-s} + a_p^{(2)} p^{1-s} - a_p^{(3)} p^{3-s} + \cdots + (-1)^n a_p^{(n)} p^{\frac{1}{2}n(n-1)-s}.$$

for almost all p (that is, for those finite primes p not dividing the level N), with other factors for primes dividing N and for ∞, and then define

$$L(\alpha, s) = L_\infty(\alpha, s) \prod_p L_p(\alpha, s).$$

Conjecturally, α corresponds to a motive M in the sense that $L(\alpha, s) = L(M, s)$. More precisely:

Conjecture. (i) *Given an absolutely irreducible motive M of dimension n with Hodge type $(n-1,0)\oplus(n-2,1)\oplus\cdots\oplus(0,n-1)$ with conductor N, there is a cuspidal $\mathbf{H}$-eigenclass $\alpha \in H_*(L_n(N), \mathbf{C})$ such that $L(\alpha, s) = L(M, s)$.*

(ii) *Given a cuspidal $\mathbf{H}$-eigenclass $\alpha \in H_*(L_n(N), \mathbf{C})$, then there is a motive M of dimension n and conductor N and Hodge type $(n-1,0)\oplus(n-2,1)\oplus\cdots\oplus(0,n-1)$, such that $L(\alpha, s) = L(M, s)$.*

We make some observations.

1. Part (i) is Question 4.16 in [4], and part (ii) is Conjecture 4.5 in [4]. We have restricted each of Clozel's formulations to the case of a trivial coefficient module for the homology of $L_n(N)$.
2. Other Hodge types occur if we allow the cohomology of $L_n(N)$ with non-trivial coefficient modules.
3. The conductor N of a motive M is a well-defined attribute of M, independent of the Conjecture.
4. "Cuspidal" is a technical term whose definition is beyond the scope of this paper. Roughly speaking, an $\mathbf{H}$-eigenclass $\alpha \in H_*(L_n(N), \mathbf{C})$ is cuspidal if it cannot be "induced" from any space of lattices in $\mathbf{R}^m$ for $m < n$. If $n = 2$ or 3, then the concept simplifies: α is non-cuspidal if for any compact subset K of $L_n(N)$, α is homologous to a cycle supported outside of K. If α is cuspidal, then it is known that $L(\alpha, s)$ has an analytic continuation to the entire s-plane, and therefore the Conjecture implies that $L(M, s)$ also has an analytic continuation.
5. The equality of L-functions is equivalent to the equality of the local factors: $L_p(\alpha, s) = L_p(M, s)$ for all p, and $L_\infty(\alpha, s) = L_\infty(M, s)$.
6. In part (ii), M need not be absolutely irreducible; for example, $\alpha \in H_1(L_2(N), \mathbf{C})$ could be associated to an elliptic curve with complex multiplication.

Let us return to our example of $L_1(N)$, and show how we can use the conjecture to deduce quadratic reciprocity. The group $H_0(L_1(N), \mathbf{C})$ is isomorphic to $H_0((\mathbf{Z}/N\mathbf{Z})^*, \mathbf{C})$, and this homology group consists of cycles $\zeta_f = \sum f(x)x$ for functions $f : (\mathbf{Z}/N\mathbf{Z})^* \to \mathbf{C}$. Let us work out how the operator $T_p^{(1)}$ behaves.

Because we are free to scale our lattice by a scalar, we can take Λ to be $\mathbf{Z}$, with distinguished generator $g = 1$ (although, for clarity, we shall continue to write g). The level N structure ϕ is determined by $\phi(g) = a \in (\mathbf{Z}/N\mathbf{Z})^*$.

We know that $T_p^{(1)}(\Lambda, \phi) = (p\Lambda, \phi|_{p\Lambda}) = (\Lambda', \psi)$. Since Λ' has distinguished generator pg, we have $\psi(pg) = p\psi(g) = pa$, so

$$T_p^{(1)}\zeta_f = \sum f(x)px = \sum f(p^{-1}x)x,$$

where p^{-1} is the inverse of p modulo N. Therefore $(T_p^{(1)}f)(x) = f(p^{-1}x)$.

Suppose that α is a simultaneous eigenelement for $T_p^{(1)}$ for all p not dividing N. We may scale α so that $\alpha(1) = 1$. We know that $(T_p^{(1)}\alpha)(x) = \alpha(p^{-1}x)$, and we also have $(T_p^{(1)}\alpha)(x) = a_p\alpha(x)$ (we need not write $a_p^{(1)}$, since this example is one-dimensional). Therefore, $\alpha(p^{-1}x) = a_p\alpha(x)$. Set $x = 1$, and we have $\alpha(p^{-1}) = a_p$. Now set $x = p$, and we have $\alpha(p) = a_p^{-1}$. We can now use induction to conclude that $\alpha(p^k) = a_p^{-k}$. This formula holds for the image of any prime p in $(\mathbf{Z}/N\mathbf{Z})^*$, which implies that $\alpha(xy) = \alpha(x)\alpha(y)$. In other words, the eigenelement α is a character of $(\mathbf{Z}/N\mathbf{Z})^*$, and the eigenvalue $a_p = \alpha(p^{-1})$. Notice that the eigenfunction α also determines the eigenvalue a_p.

Let us move now to the other side of the correspondence. Take any non-zero integer W, and let V be the variety defined by the equation $x^2 - W = 0$. We compute the Hasse-Weil L-function $Z_p(V, t)$ for primes $p \nmid W$: If the quadratic residue symbol $\left(\frac{W}{p}\right) = 1$, then $\#V(\mathbf{F}_{p^m}) = 2$ for all positive integers m. Therefore,

$$\begin{aligned} Z_p(V,t) &= \exp\left(\sum_{m=1}^{\infty} \frac{2}{m} t^m\right) \\ &= \exp(2 \cdot (-\log(1-t))) \\ &= \frac{1}{(1-t)^2} \\ &= \frac{1}{(1-t)\left(1 - \left(\frac{W}{p}\right)t\right)}. \end{aligned}$$

If $\left(\frac{W}{p}\right) = -1$, then $\#V(\mathbf{F}_{p^{2m}}) = 2$, and $\#V(\mathbf{F}_{p^{2m+1}}) = 0$. Therefore,

$$\begin{aligned} Z_p(V,t) &= \exp\left(\sum_{m=1}^{\infty} \frac{2}{2m} t^{2m}\right) \\ &= \exp\left(\sum_{m=1}^{\infty} \frac{(t^2)^m}{m}\right) \\ &= \exp(-\log(1-t^2)) \\ &= \frac{1}{1-t^2} \\ &= \frac{1}{(1-t)\left(1 - \left(\frac{W}{p}\right)t\right)}. \end{aligned}$$

We now have

$$\prod_p Z_p(V, p^{-s}) = \zeta(s)L(\chi, s),$$

where $L(\chi, s)$ is the Dirichlet L-series defined by the function $\chi(p) = \left(\frac{W}{p}\right)$.

The "interesting" part of the cohomology of V corresponds to the L-series $L(\chi, s)$, and this is the finite part of the L-series of the motive M that we study. Precisely, for each p of good reduction, we must make a choice S_p of a subset of $\left\{1, \left(\frac{W}{p}\right)\right\}$, and we select $S_p = \left\{\left(\frac{W}{p}\right)\right\}$. The conductor N of the variety M is a divisor of $4W$ (in fact, it might be considerably less than $4|W|$, since we have not as yet asked that W be square-free; in addition, if $W \equiv 1 \pmod 4$, then the factor of 4 is unnecessary). The conjecture tells us that there is an eigenelement $\alpha \in H_0((\mathbf{Z}/4W\mathbf{Z})^*, \mathbf{C})$ so that $\alpha(p) = \left(\frac{W}{p}\right)$. (This is where the fact that the eigenelement α also determines the eigenvalue a_p comes into play, along with the helpful observation that we can ignore the exponent of -1 in our formulas for $\alpha(p)$ because $-1^{-1} = -1$. We have also used the fifth observation following the definition of $T_p^{(k)}$, with $N' = 4|W|$.)

Suppose to start that $W = -1$. Since α can be thought of as a character on $(\mathbf{Z}/4\mathbf{Z})^*$, we can conclude that $\left(\frac{-1}{p}\right)$ is defined by the residue class of $p \pmod 4$. Since $\left(\frac{-1}{3}\right) = -1$, and $\left(\frac{-1}{5}\right) = 1$, we have deduced the usual formula for $\left(\frac{-1}{p}\right)$.

Next, take $W = 2$, and we now have that α is defined $\pmod 8$. Computation of $\left(\frac{2}{p}\right)$ for $p = 3$, 5, 7, and 17 give the usual formula for $\left(\frac{2}{p}\right)$.

Next, suppose that $p \equiv q \pmod 4$, with $p > q$, and let $W = (p-q)/4$. Then $p \equiv q \pmod{4W}$, which means that $\alpha(p) = \alpha(q)$, or $\left(\frac{W}{p}\right) = \left(\frac{W}{q}\right)$. We have

$$\left(\frac{p}{q}\right) = \left(\frac{4W+q}{q}\right) = \left(\frac{4W}{q}\right) = \left(\frac{W}{q}\right) = \left(\frac{W}{p}\right) = \left(\frac{4W}{p}\right) = \left(\frac{p-q}{p}\right) = \left(\frac{-q}{p}\right) = \left(\frac{-1}{p}\right)\left(\frac{q}{p}\right),$$

which implies most of the usual formula for quadratic reciprocity.

To derive the remaining case, observe that the congruence

$$x^2 - W \equiv 0 \pmod{4W-1}$$

always has the solution $x \equiv 2W$. This tells us that if W is positive and p is any prime dividing $4W-1$, then $\alpha(p) = 1$, so $\alpha(4W-1) = 1$.

Now, suppose that $p+q \equiv 0 \pmod 4$, and let $W = (p+q)/4$. Our preceding observation implies that $\alpha(p) = \alpha(q)$, which in turn implies that $\left(\frac{W}{p}\right) = \left(\frac{W}{q}\right)$, and, reasoning as before, we can conclude that $\left(\frac{p}{q}\right) = \left(\frac{q}{p}\right)$.

We next take a more complex example, to explain (finally!) what it means for an elliptic curve to be modular. For instance, we shall consider the curve

$$E \qquad : \qquad y^2 + y = x^3 - x^2$$

with discriminant $\Delta = -11$ and conductor $N = 11$. For any elliptic curve V given by a non-singular cubic equation in the plane, we write $\hat{V}$ for the complete curve, that is, $V \cup \{\infty\}$. We count the number of solutions of this equation modulo p for various primes $p \neq 11$, and add 1 for the "point at ∞," to get $\#\hat{V}(\mathbf{F}_p)$.

For any elliptic curve V defined over $\mathbf{Q}$, set $\#\hat{V}(\mathbf{F}_p) = 1 + p - a_p$. It is plausible, though not obvious, that one might want to write the number of solutions in this way: for roughly half of the p possible values of x, one expects the left-hand side to have a solution, but in those cases, one can expect there to be 2 solutions, since $y^2 + y$ is a quadratic polynomial. Therefore, one expects roughly p solutions to the congruence, plus an additional solution for the point at infinity, and then

we can consider a_p to measure how far the actual number of solutions deviates from the expected number.

We can also motivate the expression $1 - a_p + p$ by thinking in terms of motives. The elliptic curve V affords the motives $H^0(\hat{V})$, $H^1(\hat{V})$, and $H^2(\hat{V})$. The motives in dimensions 0 and 2 are essentially trivial, and contribute 1 and p, respectively (via the same analysis that gives the formula $\#\mathbf{P}^1(\mathbf{F}_{p^m}) = 1 + p^m$). The number a_p corresponds to the non-trivial motive in dimension 1.

The statement that an elliptic curve V of conductor N is "modular" is equivalent to the fact that there is an eigenclass $\alpha \in H_1(L_2(N), \mathbf{C})$ such that

$$T_p^{(1)}(\alpha) = a_p \alpha, \qquad p \nmid N.$$

If you know the usual definition of "modular elliptic curve," you can see this as follows: The set $H_1(L_2(N), \mathbf{C})$ is closely connected to the "space of modular forms of weight 2 and level N." In fact, by a theorem of Eichler and Shimura [18, Ch. 7], any cuspidal $\mathbf{H}$-eigenelement $\alpha \in H_1(L_2(N), \mathbf{C})$ has the same $\mathbf{H}$-eigenvalues as some newform of weight 2 and level N.

One can check that the elliptic curve E given above is modular by finding the corresponding α, or equivalently the weight 2 modular form of level 11; see for instance [17, 13]. If we set $q = e^{2\pi i \tau}$, and

$$\eta(\tau) = e^{\pi i \tau/12} \prod_{n=1}^{\infty} (1 - q^n),$$

then the modular form corresponding to E is

$$\begin{aligned} f(\tau) &= \eta(\tau)^2 \eta(11\tau)^2 \\ &= \sum a_n q^n \\ &= q - 2q^2 - q^3 + 2q^4 + q^5 + 2q^6 - 2q^7 - 2q^9 - 2q^{10} \\ &\quad + q^{11} - 2q^{12} + 4q^{13} + 4q^{14} - q^{15} - 4q^{16} - 2q^{17} + 4q^{18} + 2q^{20} + \cdots \end{aligned}$$

It is amusing to check a few of the a_p's by hand. For instance, if $p = 2$, we compute the solutions $\hat{E}(\mathbf{F}_2) = \{(0,0), (0,1), (1,0), (1,1), \infty\}$. Therefore, $\#\hat{E}(\mathbf{F}_2) = 5 = 1 + 2 - (-2)$, so a_2 should equal -2. It does: the coefficient of q^2 in $f(\tau)$ is -2.

The conjectures stated above imply the well-known "modularity conjecture": *every elliptic curve defined over* $\mathbf{Q}$ *is modular.*

Notice that in our first example, using $L_1(N)$, we didn't include a "point at infinity" on V because V was already complete, since it consisted simply of 2 geometric points. In general, including points at infinity compactifies the geometric space and simplifies the formulas.

For our final example, we move to a higher dimension. $H_3(L_3(N), \mathbf{C})$ has been computed for some small values of N in [1, 2, 10, 9], and certain eigenclasses have been experimentally "related" to Galois representations. Van Geeman and Top have related a few of these classes to varieties in [10]. Here is an example.

Consider $\hat{V}$ to be the variety given by

$$t^2 = xy(x^2 - 1)(y^2 - 1)(x^2 - y^2 + 2xy)$$

along with points at ∞. Then there is an $\mathbf{H}$-eigenclass in $H_3(L_3(128), \mathbf{C})$ with

$$T_p^{(k)} \alpha = a_p^{(k)} \alpha, \qquad k = 0, 1, 2, 3$$

so that $\#\hat{V}(\mathbf{F}_{p^m})$, for $m \geq 1$, corresponds in a precise though rather complicated way to the pair $(a_p^{(1)}, a_p^{(2)})$ for all primes $p \leq 67$. (Note that $a_p^{(0)} = a_p^{(3)} = 1$ for all p.)

In particular, define

$$N_{p^m}(V) := \#\{(x,y,z) \in \mathbf{F}_{p^m}^3 \mid t^2 = xy(x^2-1)(y^2-1)(x^2-y^2+2xy)\}$$
$$N_{p^m}(E) := \#\{(v,w) \in \mathbf{F}_{p^m}^2 \mid w^2 = v(v^2+2v-1)\}$$

The second quantity corresponds to points at ∞, and may be considered "understood," as E is a modular elliptic curve. Then

$$\sum_{i=1}^{6} \alpha_i^m = N_{p^m}(V) + 2N_{p^m}(E) - p^{2m} - 2p^m\left(1 + \left(\frac{2}{p}\right)^m\right)$$

where the numbers α_i are determined by the equation

$$\begin{aligned}\prod_{i=1}^{6}(1-\alpha_i X) &= X^6 - c_1X^5 + c_2X^4 - c_3X^3 + p^2c_2X^2 - p^4c_1X + p^6 \\ &= \left(X^3 - \chi(p)b_pX^2 + p\overline{b_p}X - \chi(p)p^3\right)\left(X^3 - \chi(p)\overline{b_p}X^2 + pb_pX - \chi(p)p^3\right)\end{aligned}$$

where $b_p = a_p^{(1)}$, $\overline{b_p} = a_p^{(2)}$ and $\chi(p) = \left(\frac{-2}{p}\right)$. The motive in question here is a 6-dimensional piece of $H^2(\hat{V})$, which itself is 34-dimensional.

There is as yet no proof of this correspondence for all p, though there is no reason other than lack of computer time that the computations cannot be continued for larger primes.

The status of the generalized reciprocity conjecture, restricted as we have treated it, that is, for homology with trivial coefficients, is summarized neatly by the number n. For $n = 1$, it is essentially equivalent to class field theory for $\mathbf{Q}$ or the theory of cyclotomic fields. For $n = 2$, the results of Breuil, Conrad, Diamond, Taylor, and Wiles cited in the introduction, along with results of Eichler and Shimura [8, 18], give a good piece of the picture. For $n \geq 3$, the conjecture is mostly unproven.

Appendix

In this appendix, we introduce Galois representations, which are really at the heart of the conjectures described above. We also give one more example. In this one, the motive conjectured to exist is not yet known.

A motive affords a compatible series of ℓ-adic representations of $\mathrm{G}_{\mathbf{Q}}$, the Galois group of $\overline{\mathbf{Q}}$ over $\mathbf{Q}$. In an ℓ-adic representation ρ, for each finite unramified prime p, the characteristic polynomial F_p of a Frobenius element ϕ_p at p is well-defined and gives information about how $\rho(\phi_p)$ acts in the representation. Thus a formula expressing F_p in some other terms can be thought of as a generalized reciprocity law. Moreover, as we have seen above, F_p is closely related to the $\#V(\mathbf{F}_{p^m})$ for the variety V from which the given motive comes.

For a last concrete example, consider the Hecke eigenclass α on $L_3(61)$ from [1]. If we set $\omega = (1+\sqrt{-3})/2$, then the first few $a_p^{(1)}$'s were computed to be

p	2	3	5	7	11
$a_p^{(1)}$	$1-2\omega$	$-5+4\omega$	$-2+4\omega$	-6ω	$-2+2\omega$

For each p, $a_p^{(2)}$ is the complex conjugate of $a_p^{(1)}$. We then conjecture the existence of a continuous 3-dimensional λ-adic representation ρ of $\mathrm{G}_{\mathbf{Q}}$ (where ℓ is a rational prime, and λ a prime in some number field above ℓ) unramified outside 61ℓ such that

$$F_p = X^3 - a_p^{(1)}X^2 + pa_p^{(2)}X - p^3. \qquad (*)$$

In [3], we reduced all the coefficients modulo $\sqrt{-3}$ and looked for a $\overline{\rho} : \mathrm{G}_{\mathbf{Q}} \to \mathrm{GL}(3, \mathbf{F}_3)$. This would be the weakest possible check of the conjecture for $\ell = 3$. Knowing the $\overline{F_p}$'s allowed us to infer how $\overline{\rho}(\phi_p)$ must act, and hence (eventually) how p must split in the fixed field of $\ker \overline{\rho}$. For instance, if $p = 11$, then $a_p^{(1)} \equiv a_p^{(2)} \equiv -1 \pmod 3$, and $p \equiv -1 \pmod 3$, so $F_p \equiv X^3 + X^2 + X + 1$. Thus, $\overline{\rho}(\phi_p)$ must be a matrix in $\mathrm{GL}(3, \mathbf{F}_3)$ with that characteristic polynomial. Juggling such information, we determined the only possible $\overline{\rho}$, and showed that it matched the given data. We have no idea how to show that $\overline{\rho}$ is really attached to α (that is, that $(*)$ holds for all $p \nmid 3 \cdot 61$), nor how to find ρ.

To summarize this example: we have here reciprocity between a certain series of representations of $\mathrm{G}_{\mathbf{Q}}$ and the Hecke information contained in α. It is relatively easy to compute α and then to conjecture these rather deep properties of $\mathrm{G}_{\mathbf{Q}}$.

In the case of the proof of Fermat's last theorem, a putative solution of the Fermat equation leads to a certain representation of $\mathrm{G}_{\mathbf{Q}}$. It is this representation that is shown not to exist by proving that it should be controlled through reciprocity by a certain Hecke eigenclass in $L_2(2)$, which is readily checked to be non-existent.

Acknowledgements. We would like to thank Fred Diamond, Nicholas Katz, Barry Mazur, David Rohrlich, Joseph Silverman, Glenn Stevens, and the referee for comments on earlier drafts of this paper.

References

[1] A. Ash, D. Grayson, and P. Green, *Computations of cuspidal cohomology of congruence subgroups of* $SL(3, \mathbf{Z})$, J. Number Th. **19** (1984), pp. 412–436.
[2] A. Ash and M. McConnell, *Double cuspidal cohomology for principal congruence subgroups of* $GL(3, \mathbf{Z})$, Math. Comp. **59** (1992), pp. 673–688.
[3] A. Ash, R. Pinch, and R. Taylor, *An* $\hat{A}_4$ *extension of* $\mathbf{Q}$ *attached to a non-selfdual automorphic form on* GL(3), Math. Ann. **291** (1991) pp. 753–766.
[4] L. Clozel, *Motifs et formes automorphes: applications du principe de fonctorialité*, Automorphic Forms, Shimura Varieties, and L-functions, Proceedings of the Ann Arbor Conference (L. Clozel and J. S. Milne, eds.), Academic Press, New York, 1990, pp. 77–159.
[5] B. Conrad, F. Diamond, and R. Taylor, *Modularity of certain potentially Barsotti-Tate Galois representations*, Journal of the American Mathematical Society **12** (1999), pp. 521–568.
[6] H. Darmon, *A Proof of the Full Shimura-Taniyama-Weil Conjecture is Announced*, Notices of the AMS, **46** (1999), pp. 1397–1401.
[7] F. Diamond, *On deformation rings and Hecke rings*, Annals of Mathematics **144** (1996) pp. 137–166.
[8] M. Eichler, *Quaternäre quadratische Formen und die Riemannsche Vermutung für die Kongruenzzetafunktion*, Acta Math. **5** (1954) pp. 355–366.
[9] B. van Geeman, W. van der Kallen, J. Top, and A. Verberkmoes, *Hecke eigenforms in the cohomology of congruence subgroups of* $\mathrm{SL}(3, \mathbf{Z})$, Experiment. Math. **6** (1997) pp. 163–174.
[10] B. van Geeman and J. Top, *A non-selfdual automorphic representation of* GL_3 *and a Galois representation*, Inv. Math. **117** (1994) pp. 391–401.
[11] S. Gelbart, *Elliptic curves and automorphic representations*, Adv. in Math. **21** (1976) pp. 235–292.
[12] U. Jannsen, S. Kleiman, and J.-P. Serre, *Motives*, Proc. Symp. Pure Math. 55, Parts 1 and 2, Amer. Math. Soc., Providence, 1994.
[13] A. W. Knapp, *Elliptic Curves*, Mathematical Notes 40, Princeton University Press, Princeton, New Jersey, 1992.
[14] R. P. Langlands, *Some contemporary problems with origins in the Jugendtraum (Hilbert's problem 12)*, Hilbert Problems, Proc. Symp. Pure Math. 28, Part 2, Amer. Math. Soc., Providence, 1976, pp. 401–418.
[15] K. Ribet, *Galois Representations and Modular Forms*, Bulletin of the AMS **32** (1995), pp. 375–402.
[16] J.-P. Serre, *A Course in Arithmetic*, Graduate Text in Math. 7, Springer Verlag, New York, 1973.
[17] G. Shimura, *A non-solvable reciprocity law*, J. Reine Angew. Math **221** (1966), pp. 209–220.
[18] G. Shimura, *Introduction to the Arithmetic Theory of Automorphic Forms*, Princeton University Press, Princeton, 1971.
[19] J. Silverman, *The Arithmetic of Elliptic Curves*, Graduate Text in Math. 106, Springer Verlag, New York, 1986.

[20] J. Tate, *Problem 9: The general reciprocity law*, Hilbert Problems, Proc. Symp. Pure Math. 28, Part 2, Amer. Math. Soc., Providence, 1976, pp. 311–322.
[21] R. Taylor and A. Wiles, *Ring theoretic properties of certain Hecke algebras*, Annals of Mathematics **141** (1995), pp. 553–572.
[22] A. Weil, *Number Theory: An Approach Through History from Hammurapi to Legendre*, Birkhauser, Boston, 1984.
[23] A. Wiles, *Modular elliptic curves and Fermat's Last Theorem*, Annals of Mathematics **141** (1995), pp. 443–551.
[24] B. Wyman, *What is a Reciprocity Law?*, American Mathematical Monthly **79** (1972), pp. 571–586.

(Ohio State University) Columbus, OH 43210-1174

(Boston College) Chestnut Hill, MA 02467-3806
E-mail address: ash.4@osu.edu
E-mail address: gross@bc.edu