Draft: March 25, 2012

DEFORMATIONS OF GALOIS REPRESENTATIONS AND APPLICATIONS — TALK AT THE IHP CONFERENCE ON THE OCCASION OF THE GALOIS BICENTENNIAL

MATTHEW EMERTON

1. Introduction

The theory of Galois representations with finite field coefficients begins, as far as I know, with Galois himself. In modern terms, he showed that an irreducible polynomial $f(x) \in \mathbb{Q}[x]$ of prime degree p is solvable by radicals if and only if the Galois group G of $f(x)$ admits an embedding

$$G \hookrightarrow \begin{pmatrix} \mathbb{F}_p^\times & \mathbb{F}_p \\ 0 & 1 \end{pmatrix} \subset \mathrm{GL}_2(\mathbb{F}_p).$$

(Here the group $\begin{pmatrix} \mathbb{F}_p^\times & \mathbb{F}_p \\ 0 & 1 \end{pmatrix}$ is thought of a subgroup of $\mathrm{Sym}(\mathbb{F}_p)$, the full permutation group of $\mathbb{F}_p$, by its affine-linear action on $\mathbb{F}_p$, i.e. via the action $\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \cdot x = ax + b$. The claimed embedding of G is then obtained by choosing an appropriate identification between $\mathbb{F}_p$ and the p roots of $f(x)$.) Given such a polynomial $f(x)$, thinking of its Galois group G as a quotient of the absolute Galois group $G_{\mathbb{Q}}$ of $\mathbb{Q}$, one obtains a representation $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_p)$.[1] This is an example of a (two-dimensional, mod p) Galois representation.

The basic objective of the theory of deformations of Galois representations is to study liftings of representations $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_n(\mathbb{F}_p)$ to representations $\rho_n : G_{\mathbb{Q}} \to \mathrm{GL}_n(\mathbb{Z}/p^n)$, and ultimately to p-adic representations $\rho : G_{\mathbb{Q}} \to \mathrm{GL}(\mathbb{Z}_p)$.[2]

Somewhat more precisely, given $\overline{\rho}$, one would like to describe the collection of all possible liftings, perhaps satisfying some conditions. Typically there are many liftings (although it is not always obvious *a priori* that there are *any* liftings at all), and one thinks of them as lying in a "space", the *Galois deformation space*; two liftings ρ and ρ' are considered close if they are p-adically close, i.e. if the mod p^n reductions ρ_n and ρ'_n coincide for some power p^n (and the larger the value of n, the closer are the two liftings).

The theory of deformations has another, grander, objective, though, which provided one of the primary motivations for its introduction (by Barry Mazur), and

[1] The overline in the notation is chosen to indicate that $\overline{\rho}$ has coefficients in characteristic p, rather than characteristic zero.

[2] There is a technical distinction between liftings — where one considers the literal homomorphism into a matrix group — and deformations, where one considers such homomorphisms up to conjugation. However, I will suppress this distinction in this talk. Essentially, if one understands the liftings of $\overline{\rho}$, then one passes to the deformations of $\overline{\rho}$ by forming an appropriate quotient. Also, $\mathbb{F}_p$ can be replaced by a more general finite field k, and $\mathbb{Z}_p$ by a the ring of integers in a finite extension of $\mathbb{Z}_p$ — and in fact it is technically important to consider these more general coefficients, although I will also suppress that point in this talk. Finally, the field $\mathbb{Q}$ whose absolute Galois group we are considering could be replaced by any number field, or also a local field.

which remains the primary motivation for its study. Namely, as Michael Harris explained in his talk, one expects there to be a reciprocity between (certain — namely algebraic) automorphic forms (for the group GL_n, say) and (certain — namely geometric, in the sense of Fontaine and Mazur) n-dimensional p-adic representations of $G_{\mathbb{Q}}$. A basic strategy for proving such a result is as follows:

- Construct Galois representations attached to the automorphic forms of interest, and identify which Galois deformation space(s) they lie in.
- Show that there are many automorphic forms, so that they contribute many points to the relevant Galois deformation space(s).
- Bound the size of the relevant Galois deformation space(s) from above.
- Combining the previous two points, conclude that every p-adic Galois representation satisfying the appropriate conditions arises from one of the automorphic form under consideration.

This was the strategy introduced by Wiles in his proof (completed by him together with Richard Taylor) of the modularity theorem for semistable elliptic curves over $\mathbb{Q}$, and hence of Fermat's Last Theorem, and has remained the basic strategy in the proof of the many subsequent results in the theory of automorphic forms and Galois representations, such as the proof of the full modularity theorem for elliptic curves over $\mathbb{Q}$ (by Breuil, Conrad, Diamond, and Taylor) and the proof of the Sato–Tate conjecture for elliptic curves over $\mathbb{Q}$ (by Clozel, Harris, Shepherd-Barron, Shin, and Taylor).

2. An example

I would like to illustrate Galois deformation theory through an example. My goal is to give a simple, but precise, statement in one particular case, which will give a feel for the kind of statements that are proved in the general theory. Along the way, I hope to illustrate by example some of the conditions that arise in deformation theory, as well as the relationship of the theory to Diophantine problems and to automorphic forms.

Example 1. In fact, I will give two examples. The first goes back in some sense before Galois, to Gauss. Namely, if p^n is a prime, then for any commutative ring A, we write

$$\mu_{p^n}(A) := \{a \in A \mid a^{p^n} - 1\}.$$

This is a functor of A; it functorially assigns an abelian group to each ring A. Since the assignment is made by solving an equation in A, it is a *group scheme* (in fact a *finite flat* group scheme) over $\operatorname{Spec}\mathbb{Z}$.

In particular, since it is a functor, $G_{\mathbb{Q}}$ acts on $\mu_{p^n}(\overline{\mathbb{Q}})$. This latter group is known to be cyclic of order p^n (it coincides with $\mu_{p^n}(\mathbb{C})$), and so we obtain a character $\chi_n : G_{\mathbb{Q}} \to \mathrm{GL}_1(\mathbb{Z}/p^n) = (\mathbb{Z}/p^n)^\times$, known as the *mod p^n cyclotomic character*. These characters are successive liftings of the mod p character χ_1, which we will also denote by $\overline{\chi}$, and taken together they give rise to a character

$$\chi : G_{\mathbb{Q}} \to \mathrm{GL}_1(\mathbb{Z}_p) = \mathbb{Z}_p^\times,$$

the p-adic cyclotomic character.

Since each χ_n arises by evaluating the finite flat group scheme μ_{p^n} on $\overline{\mathbb{Q}}$, we say that χ_n is finite flat over $\operatorname{Spec}\mathbb{Z}$. If p is odd, then one can furthermore characterize each χ_n as being the unique lift of $\overline{\chi}$ to a representation with coefficients in $\mathbb{Z}/p^n$

that is finite flat over $\operatorname{Spec}\mathbb{Z}$. (If $p = 2$, then $\overline{\chi}$ is the trivial character, which admits the trivial character as a lift, in addition to the χ_n.)

Example 2a. Let us return to the context of Galois mentioned at the beginning of the introduction, and consider the simplest, and one of the most celebrated, examples: namely let $f(x) = x^2 + 1$. This is certainly an irreducible quadratic equation over $\mathbb{Q}$, and Galois's theorem suggests that we consider the Galois representation

$$\overline{\rho} : G_{\mathbb{Q}} \to \begin{pmatrix} 1 & \mathbb{F}_2 \\ 0 & 1 \end{pmatrix},$$

which factors through the quotient $\operatorname{Gal}(\mathbb{Q}(i)/\mathbb{Q}) = \{1, c\}$ (with c denoting complex conjugation), and is defined by mapping c to the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

Now this Galois representation has a certain very nice property. Namely, just as with the cyclotomic characters χ_n, the representation $\overline{\rho}$ is finite flat, i.e. arises by evaluating a finite flat group scheme on $\overline{\mathbb{Q}}$.

Here is the relevant group scheme: it corresponds to the functor

$$\mathcal{G}_1 : A \mapsto \{a_0 + a_1 I \in A[I] \,|\, (a_0 + a_1 I)^2 = 1, a_0 a_1 = 0\}.$$

Here A is any commutative ring, and I is a formal variable adjoined to A which satisfies the relation $I^2 = -1$. The group law is just given by multiplication in the ring $A[I]$.

If we compute the values of $\mathcal{G}_1$ on $\overline{\mathbb{Q}}$, we find that $\mathcal{G}_1(\overline{\mathbb{Q}})$ is the two-dimensional $\mathbb{F}_2$-vector space spanned by $-1 + 0 \cdot I$ and $0 + i \cdot I$. (Here $i \in \overline{\mathbb{Q}}$ denotes the usual element of square -1.) Since $\mathcal{G}_1$ is a functor, $G_{\mathbb{Q}}$ acts on $\mathcal{G}_1(\overline{\mathbb{Q}})$, and one immediately sees that it does so via the representation $\overline{\rho}$. Thus $\overline{\rho}$ does indeed arise from the finite flat group scheme $\mathcal{G}_1$ over $\operatorname{Spec}\mathbb{Z}$, and so $\overline{\rho}$ is finite flat over $\operatorname{Spec}\mathbb{Z}$.

We can easily write down some liftings of $\overline{\rho}$: For each $n \geq 1$, define the finite flat group scheme

$$\mathcal{G}_n : A \mapsto \big\{a_0 + a_1 Z_n + \cdots + a_{2^n-1} Z_n^{2^n-1} \in A[Z_n] \quad | \\ (a_0 + a_1 Z_n + \cdots + a_{2^n-1} Z_n^{2^n-1})^{2^n} = 1, a_i a_j = 0 \text{ for all } i \neq j\big\},$$

where again A is any commutative ring, and now Z_n is a formal variable adjoined to A which satisfies the relation $Z_n^{2^n} = -1$. Squaring, and identifying Z_n^2 with Z_{n-1} (and setting $Z_1 = I$), we obtain maps

$$\mathcal{G}_n \to \mathcal{G}_{n-1} \to \cdots \to \mathcal{G}_1. \tag{1}$$

Furthermore, if we choose $\zeta_n \in \overline{\mathbb{Q}}$ such that $\zeta_n^{2^n} = -1$ (so ζ_n is a primitive 2^{n+1}st root of 1), then one computes that $\mathcal{G}_n(\overline{\mathbb{Q}})$ is the free $\mathbb{Z}/2^n$-module of rank two spanned by $\zeta_n^2 + 0 \cdot Z_n + \cdots + 0 \cdot Z_n^{2^n-1}$ and $0 + \zeta_n \cdot Z_n + \cdots 0 \cdot Z_n^{2^n-1}$. Thus the action of $G_{\mathbb{Q}}$ on $\mathcal{G}_n(\overline{\mathbb{Q}})$ gives rise to a representation $\rho_n : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}/2^n)$, and a consideration of (1) shows that the ρ_n are successive lifts of $\overline{\rho}$. Again, each of them is (by construction) finite flat over $\operatorname{Spec}\mathbb{Z}$.

We may combine all the ρ_n into a 2-adic lift $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_2)$ of $\overline{\rho}$, and using class field theory (more or less), one can then prove the following result:

ρ is the unique deformation of $\overline{\rho}$ which has determinant equal to χ (the 2-adic cyclotomic character), and for which each reduction ρ_n is finite flat over $\operatorname{Spec}\mathbb{Z}$.

Example 2b. The preceding result is a typical "upper bound" on the size of a deformation space, albeit in a very special (and perhaps not so interesting) context. However, we can make the context more interesting by introducing an elliptic curve, as we will now do.

Consider the elliptic curve $\mathcal{E}$ cut out by (the projectivization of) the equation

$$y^2 + xy + y = x^3 - x^2 - x - 14.$$

This elliptic curve is often denoted $X_0(17)$, because it is isomorphic (over $\mathbb{C}$) to (the completion of) the quotient of the complex upper half-plane by the discrete group

$$\Gamma_0(17) := \left\{ \begin{pmatrix} a & b \\ 17c & d \end{pmatrix} \mid a, b, c, d \in \mathbb{Z}, ad - 17bc = 1 \right\}$$

of Mobius transformations.

The usual chord–tangent law makes $\mathcal{E}$ an algebraic group, and we may consider its subgroup scheme of 2^n-torsion points $\mathcal{E}[2^n]$, for any $n \geq 1$. If we take the $\overline{\mathbb{Q}}$-points, then $\mathcal{E}[2^n](\overline{\mathbb{Q}})$ is free of rank two over $\mathbb{Z}/2^n$, and so the $G_{\mathbb{Q}}$-action on it gives a representation $\psi_n : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}/n)$. As usual, we write $\overline{\psi} := \psi_1$. The ψ_n are successive lifts one of the other, and we may combine them into a 2-adic lift $\psi : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_2)$ of $\overline{\psi}$, the so-called 2-*adic Tate module* of $\mathcal{E}$.

We can compute the 2-torsion points explicitly: they are the points at which the tangent line to the curve is vertical, i.e. at which

$$2y + x + 1 = 0.$$

One finds the three 2-torsion points

$$P = (11/4, -15/8),\ Q = (-1 - 2i, i),\ P + Q = R = (-1 + 2i, -i),$$

as well, of course, as the point O at infinity, which is the identity for the group law. Clearly the action of $G_{\mathbb{Q}}$ on these points factors through $\mathrm{Gal}(\mathbb{Q}(i)/\mathbb{Q})$, and one sees immediately that $c(P) = P$, while $c(Q) = R = P + Q$. Thus in fact $\overline{\psi} \cong \overline{\rho}$.

We can go further and compute $\mathcal{E}[4](\overline{\mathbb{Q}})$, and doing this, one discovers that ψ_2 is isomorphic to ρ_2. At this point one might wonder if ψ_n is isomorphic to ρ_n for all n, but general principles show that this is not possible: if $\mathcal{E}[2^n]$ were finite flat for every n, then a theorem of Grothendieck would show that $\mathcal{E}$ has good reduction at every prime, but the curve $\mathcal{E}$ becomes a nodal cubic modulo 17 (it acquires a node at the point $(7, 13)$).

In fact, the group schemes $\mathcal{E}[2^n]$ giving rise to the Galois representations ψ_n are all finite flat over $\mathrm{Spec}\,\mathbb{Z}[1/17]$ (since $\mathcal{E}$ has good reduction away from 17), and so we say that the representations ψ_n are finite flat over $\mathrm{Spec}\,\mathbb{Z}[1/17]$. But the ψ_n have only a weaker property locally at the prime 17: they are *semistable* at 17. (This choice of terminology reflects the use of the term semistable to describe nodal singularities of curves.)

One can then prove the following result:

ρ and ψ are the only two 2-adic deformations of $\overline{\rho}$ whose determinants are equal to χ the 2-adic cyclotomic character*, and whose reductions modulo 2^n are finite flat over* $\mathrm{Spec}\,\mathbb{Z}[1/17]$ *and semistable at* 17*, for each $n \geq 1$.*

The elliptic curve $\mathcal{E}$ corresponds (in the sense of the modularity theorem for elliptic curves over $\mathbb{Q}$) to the unique normalized cuspform of weight 2 on $\Gamma_0(17)$,

$$f_{\mathcal{E}} := q - q^2 - q^4 - 2q^5 + 4q^7 + \cdots,$$

and this statement is a typical result relating a certain deformation space to a space of automorphic forms.

More precisely, the condition of finite flatness away from 17, the semistability at 17, and the determinant condition, taken together correspond to the condition of a modular form being of weight two and level $\Gamma_0(17)$. The two possible lifts then correspond to the two normalized Hecke eigenforms of weight two and level $\Gamma_0(17)$, namely the Eisenstein series

$$E_2 := \frac{2}{3} + q + 3q^2 + 4q^3 + 6q^5 + 12q^6 + 8q^7 + \cdots,$$

and the cuspform $f_{\mathcal{E}}$.[3]

3. The general theory

The computations of the particular deformation spaces discussed in the preceding section are relatively straightforward, requiring just class field theory and a little bit of the theory of congruences of modular forms (together with Wiles's miraculous *numerical criterion*!).[4] In general, the problem of computing all the deformations of a given $\overline{\rho}$, and of relating these lifts to automorphic forms, is difficult. Taylor and Wiles introduced a method for doing this, which again relies on class field theory, but in its more subtle manifestations through the theory of Galois cohomology.

It is not possible in the present talk to give many details about the Taylor–Wiles method, but I will try to give some idea of how it works. For this, let me return to the strategy described in the introduction, and elaborate a little on each of the steps.

Constructing Galois representations attached to automorphic forms. [Much of this may be covered in Michael Harris's talk.] For automorphic forms on GL_1, this is more-or-less class field theory. The two-dimensional Galois representations attached to classical modular forms (which are automorphic for GL_2) were constructed by Deligne (and Deligne-Serre in the case of weight one modular forms), building on earlier ideas of Eichler, Shimura, and Ihara, and following the conjecture of their existence by Serre. The generalization of Deligne's construction from classical modular forms to Hilbert modular forms was achieved by Carayol, Wiles, and Taylor.

The most general currently known results are due to many people, including Clozel, Kottwitz, Harris and Taylor, Labesse, Shin, and Chenevier. Collectively, their results attach Galois representations to regular algebraic essentially conjugate self-dual cuspidal automorphic representations on GL_n over a CM field (with n arbitrary). In order to know precisely which deformation space these Galois representations lie in (i.e. which conditions, such as finite flat or semistable, these Galois representations satisfy), it is important to know precise *local-global compatibility* statements about the Galois representations. In the case of $n = 1$ this is the compatibility of local and global Artin maps. In the case of $n = 2$ and Hilbert modular forms, it is due to Langlands, Deligne, Carayol, T. Saito, Taylor, Skinner, Kisin,

[3]The fact that our deformation problem captures *all* the eigenforms of weight 2 and level 17 reflects the fact that the unique cuspform $f_{\mathcal{E}}$ is in fact congruent to the Eisenstein series mod 2 (in fact even mod 4, if we ignore the constant term). If we replaced 17 by another prime, then the analogous deformation problem would only capture those modular forms which are congruent to the Eisenstein series mod 2.

[4]See Calegari and Emerton, *On the ramification of Hecke algebras at Eisenstein primes*.

and T. Liu. In the case of general n, it is due to Harris–Taylor, Taylor–Yoshida, Shin, Barnet-Lamb–Gee–Geraghty–Taylor, and Caraiani.

Showing that there are many automorphic forms, and hence many automorphic Galois representations. This is one of the key steps in the Taylor–Wiles method. One quantifies "sufficiently many" in the following way: one augments the *level* of the automorphic forms under consideration, and shows that the dimension of the space of automorphic forms grows in proportion to the increase in level. This is typically achieved by a cohomological argument, and necessitates a restriction on the nature of the automorphic forms under consideration — e.g. they (or rather the automorphic representations that they generate) should be discrete series at infinity.

There is another, very fundamental, issue that has to be confronted in this step, which is related to the fact that, in the sketch of the strategy, I have written Galois deformation space(s), rather than just Galois deformation space. What is intended by this is that one must identify the candidate $\overline{\rho}$ whose deformations are to be considered.

Essentially, if one wants to employ this method to prove that a particular p-adic Galois representation, or class of Galois representations, arises from an automorphic form, then one must first show that its mod p reduction arises from an automorphic form (so as to know that there any automorphic points in the relevant deformation space at all). That this should be true goes under the general rubric of *Serre's conjecture* (since Serre proposed the first form of such a conjecture in the case of two-dimensional mod p representations and classical modular forms), and turns out to be one of the most difficult parts in implementing the strategy. In Wiles's original arguments, he got around the problem of not knowing Serre's conjecture by appealing to a deep theorem in automorphic forms due to Langlands and Tunnell, and combining it with his celebrated $3-5$ trick. Since then, the $3-5$ trick has been turned (particularly through the work of Taylor) into a basic method in the theory, and in its most refined form was used by Khare and Wintenberger, and Kisin, to prove Serre's original conjecture for two-dimensional mod p representations.

For higher dimensional Galois representations, one currently has a potential Serre's conjecture (developed in the work of Harris, Shepherd-Barron, and Taylor, and further developed since) — that is to say one has Serre's conjecture, but only after restricting $\overline{\rho}$ to a certain open index subgroup of the Galois group on which it is originally defined. Removing the "potential" from this result is one of the most important, but perhaps one of the most difficult, open problems in the theory.

Bounding the size of a deformation space from above. This is the heart of the Taylor–Wiles method. The basic problem is that it is hard to measure the size of a deformation space *a priori*. The tangent space to $\overline{\rho}$ in the deformation space can be computed via Galois cohomology, but it is typically large, and *a priori* one doesn't know if this is because the deformation space is actually of high dimension, or just very singular at $\overline{\rho}$.

What the Taylor–Wiles method does is very carefully add primes to the level of the automorphic forms under consideration, and simultaneously consider larger deformation spaces, in which the conditions at these primes are relaxed. I say "very carefully" because the heart of the method is that these primes are chosen in such a way that *the tangent space to $\overline{\rho}$ doesn't change* as the primes are added. On the other hand, by the previous step, one knows that the number of automorphic points

in deformation space is systematically increasing as the primes are added. Passing to an appropriate limit, one ends up in a situation where *all* the directions in the tangent space are accounted for by automorphic forms.

The details of how to choose these auxiliary "Taylor–Wiles" primes is one of the most delicate points of the theory, involving subtle points of finite group theory. At this point, and also in the previous step, one finds that it is in fact technically more difficult (even impossible, in general) to work with reducible $\overline{\rho}$, and so it is common to restrict attention to p-adic Galois representations for which the mod p representation $\overline{\rho}$ is irreducible (or satisfies an even stronger condition, known as "non-degeneracy", or "having big image").

There are many other subtleties that arise; for example, adding Taylor–Wiles primes lets one show that the correct *dimension* is achieved by automorphic Galois representations, but doesn't necessarily show that they fill out all the *irreducible components* of deformation space.[5] Kisin introduced a technique for dealing with this problem, and thus pushed the method further; hence people now speak of the "Taylor–Wiles–Kisin method" ....

Concluding. Once the Taylor–Wiles method has been successfully carried out, one knows that all the points in the Galois deformation space of interest, *but with the conditions at the auxiliary, Taylor–Wiles, primes relaxed*, are explained by automorphic forms. It is now easy to deduce that all the points in the original Galois deformation space are also explained by automorphic forms: one just reimposes the original conditions at the auxiliary primes (whatever they were), and takes into account local-global compatibility at these primes for the Galois representations attached to automorphic forms.

Applications. [Return to some recent results such as Sato–Tate, ...]

Mathematics Department, University of Chicago, 5734 S. University Ave., Chicago, IL 60637

E-mail address: emerton@math.uchicago.edu

[5]In fact the issue is more nuanced than this; it is really components in the *local deformation space at p* which are at issue, but I will elide this point.