CWI Quarterly Volume 9 (1 & 2) 1996, pp. 61 – 67

Simplicity of Frobenius Eigenvalues in the Galois Representations Associated to Modular Forms

Bas Edixhoven

Institut Mathématique, Campus de Beaulieu, 35042 Rennes, France

e-mail: edix@campagnarde.univ-rennes1.fr

This text is a written version of the talk that I gave at the occasion of the 50th anniversary of the SMC. It describes some results of an article in preparation with R.F. Coleman (University of California at Berkeley). I thank Rutger Noot for discussions that led me to a proof of a special case of Theorem 3.2, and Richard Taylor for a discussion concerning the general case.

1. Ramanujan's τ-function

Let Δ be the formal power series with integer coefficients defined by the product:

$$\Delta = \sum_{n\geq 1} \tau(n) q^n = q \cdot \prod_{n\geq 1} (1-q^n)^{24}.$$

The function $n \mapsto \tau(n)$ thus defined is the famous Ramanujan τ-function. We can interpret q as the function $z \mapsto \exp(2\pi i z)$ from the complex upper half plane $\mathbb{H} = \{z \in \mathbb{C} \mid \Im(z) > 0\}$ to $\mathbb{C}$. Then Δ defines an analytic function on $\mathbb{H}$, sending z to $\Delta(\exp(2\pi i z))$. This function Δ has a lot of symmetry. Recall that the group $\mathrm{SL}_2(\mathbb{R})$ acts on $\mathbb{H}$ by fractional linear transformations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z = \frac{az+b}{cz+d}.$$

The function Δ is then "invariant" for the action of the subgroup $\mathrm{SL}_2(\mathbb{Z})$ of $\mathrm{SL}_2(\mathbb{R})$ in the sense that for all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in $\mathrm{SL}_2(\mathbb{Z})$ one has:

$$\Delta\left(\frac{az+b}{cz+d}\right) = (cz+d)^{12}\Delta(z).$$

This formula means in fact that the expression $\Delta(z)(dz)^6$ is a $\mathrm{SL}_2(\mathbb{Z})$-invariant section of the 6th tensor power of the line bundle of holomorphic differentials on $\mathbb{H}$, which makes it by definition a modular form of weight 12 for $\mathrm{SL}_2(\mathbb{Z})$. The fact that the constant term in the formal power series giving Δ equals zero means that Δ is a cusp form. The complex vector space of cusp forms of weight 12 is of dimension one, hence Δ is an eigenform for certain operators that are naturally defined on the vector spaces of modular forms. This implies that the Dirichlet series associated to Δ has the following Euler product expansion over all prime numbers p:

$$\sum_{n\geq 1} \frac{\tau(n)}{n^s} = \prod_p \frac{1}{1-\tau(p)p^{-s}+p^{11}p^{-2s}},$$

valid for s in $\mathbb{C}$ with $\Re(s)$ big enough. Ramanujan conjectured that for all prime numbers p one has $|\tau(p)| \leq 2p^{11/2}$. This was proved, in two steps, by Deligne. The first step (1968) is the construction, for every prime number l, of an l-adic Galois representation $\rho_{\Delta,l}\colon \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{Q}_l)$ which is continuous, unramified at all primes $p \neq l$ and has the property that for all $p \neq l$ the Frobenius element $\rho_{\Delta,l}(\mathrm{Frob}_p)$ has trace $\tau(p)$ and determinant p^{11}. The representation $\rho_{\Delta,l}$ occurs in fact in the dual of the l-adic cohomology of a motive over $\mathbb{Q}$ with good reduction at all primes. The second step (1974) is the proof of the conjecture of Weil implying that the eigenvalues of such a $\rho_{\Delta,l}(\mathrm{Frob}_p)$ are algebraic numbers all of whose archimedean absolute values are equal to $p^{11/2}$ (the exponent is half of the degree of the cohomology group in which the dual of $\rho_{\Delta,l}$ occurs). Since $\tau(p)$ is the sum of the two eigenvalues of $\rho_{\Delta,l}(\mathrm{Frob}_p)$ (take any $l \neq p$), it follows indeed that $|\tau(p)| \leq 2p^{11/2}$.

2. More general eigenforms

Deligne showed in fact that the analog of Ramanujan's conjecture for arbitrary cuspidal eigenforms is true. Let $N \geq 1$ and k be integers and let $\varepsilon\colon (\mathbb{Z}/N\mathbb{Z})^* \to \mathbb{C}^*$ be a character. A modular form of level N, weight k and character ϵ is then a holomorphic function $f\colon \mathbb{H} \to \mathbb{C}$ such that

$$f\left(\frac{az+b}{cz+d}\right) = \varepsilon(d)(cz+d)^k f(z)$$

for all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in $\mathrm{SL}_2(\mathbb{Z})$ with N dividing c, and such that for every $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in $\mathrm{SL}_2(\mathbb{Z})$ the function

$$\mathbb{H} \to \mathbb{C}, \quad z \mapsto (cz+d)^{-k} f\left(\frac{az+b}{cz+d}\right)$$

has a limit for $|\Im(z)|$ tending to infinity. Such a modular form is called a cusp form if all these limits are zero. For a modular form f one has $f(z+1) = f(z)$, implying that f can be written as a power series in q:

$$f = \sum_{n \geq 0} a_n q^n.$$

The set $M(N, k, \varepsilon)$ of modular forms with fixed N, k and ϵ is a $\mathbb{C}$-vector space of finite dimension (this follows from an interpretation of it as the space of global sections of some holomorphic line bundle on some compact Riemann surface). The dimension of $M(N, k, \varepsilon)$ can be calculated by the Riemann–Roch formula, except when $k = 1$; for $k < 0$ it is zero. The $M(N, k, \varepsilon)$ are equipped with certain operators T_n ($n \geq 1$), called Hecke operators, defined in terms of the action of $\mathrm{SL}_2(\mathbb{Q})$ on $\mathbb{H}$. These T_n commute with each other, so it makes sense to look at their common eigenspaces. There is a simple relation between the eigenvalues of a non-zero common eigenform f and its Fourier expansion $f = \sum_{n \geq 0} a_n q^n$: one has $a_1 T_n(f) = a_n f$. This relation implies that a_1 is non-zero and that the common eigenspaces are of dimension one. An eigenform is called normalized if $a_1 = 1$.

Suppose now that f is a normalized cuspidal eigenform of some level N, weight k and character ϵ. Then $T_n(f) = a_n f$. The Dirichlet series associated to f has the Euler product expansion:

$$\sum_{n \geq 1} \frac{a_n}{n^s} = \prod_p \frac{1}{1 - a_p p^{-s} + \varepsilon(p) p^{k-1} p^{-2s}},$$

valid for s in $\mathbb{C}$ with $\Re(s)$ big enough (for $p|N$ one defines $\varepsilon(p) := 0$). It can be proved, for example by using the theory of moduli spaces for elliptic curves, i.e., modular curves, that the a_n are algebraic integers generating a finite field extension K of $\mathbb{Q}$. For every prime number l one has a continuous representation $\rho_{f,l} \colon \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(K \otimes \mathbb{Q}_l)$ which is unramified at all primes p not dividing lN and has the property that for such primes $\rho_{f,l}(\mathrm{Frob}_p)$ has trace a_p and determinant $\varepsilon(p) p^{k-1}$. The eigenvalues of $\rho_{f,l}(\mathrm{Frob}_p)$, i.e., the roots in $\mathbb{C}$ of the polynomial $x^2 - a_p x + \varepsilon(p) p^{k-1}$, have absolute value equal to $p^{(k-1)/2}$, hence we have $|a_p| \leq 2p^{(k-1)/2}$. Let us note that we have $|a_p| = 2p^{(k-1)/2}$ if and only if the polynomial $x^2 - a_p x + \varepsilon(p) p^{k-1}$ has a double root.

3. The problem we want to solve

The kind of question we ask ourselves can now be easily formulated: can it happen that $|a_p| = 2p^{(k-1)/2}$, for some cuspidal normalized eigenform $f =$

$\sum a_n q^n$ and a prime number p not dividing the level of f? Stated like this, the answer is yes. Let f be a cuspidal normalized eigenform of weight one. Then the $\rho_{f,l}$ have finite image, and Chebotarev's density theorem implies that there exist infinitely many prime numbers p such that $\rho_{f,l}(\mathrm{Frob}_p)$ is the identity element. If we consider only forms of weight at least two, the situation is very different. Of course, for the modular form Δ the problem is trivial: $|\tau(p)|$ is an integer, hence it can not be equal to the irrational number $2p^{11/2}$. For a general cuspidal normalized eigenform $f = \sum a_n q^n$ with character ε this type of argument does not work: there can be prime numbers p such that $\varepsilon(p)p^{k-1}$ is a square in the field K generated by the a_n. Only very little seems to be known about these fields K. Douglas Ulmer obtained the following result as a kind of by-product in his article "A construction of local points on elliptic curves over modular curves", International Mathematics Research Notices 1995, No. 7.

3.1 THEOREM (ULMER). *Let p be a prime number. Suppose that the Birch–Swinnerton-Dyer conjecture for elliptic curves over function fields of characteristic p is true. Then for every cuspidal normalized eigenform $f = \sum a_n q^n$ of level prime to p and of weight 3, one has $|a_p| < 2p$.*

Ulmer also noted that one should be able to prove by his method that Tate's conjecture implies the analog of his result for all weights $k \geq 3$. (This conjecture claims that the dimension of the $\mathbb{Q}$-vector space of codimension r cycles on a smooth projective variety over finite field of characteristic p equals the order of its zeta function at r.) Indeed, we have the following result, obtained by a different method. This method can be described briefly by saying that it uses that the motive over $\mathbb{F}_p$ that one considers is actually the reduction modulo p of a motive over $\mathbb{Z}_p$. The existence of this unramified lift forces certain restrictions on its corresponding Hodge filtration on the crystalline cohomology.

3.2 THEOREM (COLEMAN–EDIXHOVEN). *Let p be a prime number. Let $f = \sum a_n q^n$ be a cuspidal normalized eigenform of weight $k \geq 2$ and of prime-to-p level. Suppose that the crystalline Frobenius at p is semi-simple. Then one has $|a_p| < 2p^{(k-1)/2}$.*

The crystalline Frobenius in this statement is given by the crystalline realization of the reduction modulo p of the rank two motive associated to f. This will become more explicit in the next two sections, when we discuss the proof. We remark that for f of weight two this crystalline Frobenius element is known to be semi-simple because the category of abelian varieties up to isogeny over a fixed finite field is semi-simple. Hence the following corollary.

3.3 COROLLARY. *Let $f = \sum a_n q^n$ be a cuspidal normalized eigenform of weight two and character ε. Let p be a prime number not dividing the level of f. Then the polynomial $x^2 - a_p x + \varepsilon(p)p$ has simple roots.*

For f of general weight $k \geq 2$ the semi-simplicity of the crystalline Frobenius element at a prime not dividing the level of f is a consequence of Tate's conjecture mentioned above: see the first three lines of Section 2 of Milne's article "Motives over finite fields", Proceedings of Symposia in Pure Mathematics, Volume 55, Part 1.

Theorem 3.2 has the following interesting consequence.

3.4 COROLLARY. *Let $N \geq 1$ and $k \geq 2$ be integers, with N cube free, i.e., N is not divisible by any third power of a prime number. If $k > 2$ suppose that Tate's conjecture mentioned above is true. Then the Hecke algebra of type (N, k), i.e., the sub-$\mathbb{Z}$-algebra generated by the Hecke operators and the diamond operators of the endomorphism algebra of the $\mathbb{C}$-vector space of modular forms of level N and weight k is reduced.*

This result implies that the discriminants of such Hecke algebras are non-zero. Abbes and Ullmo relate, for prime level p, weight two and trivial character, the discriminant of that Hecke algebra to the height of the modular curve $X_0(p)$.

Finally, according to Mazur, Theorem 3.2 sheds some light on a question that arises in the relation between p-adic modular forms and deformations of Galois representations.

4. AN ELEMENTARY PROOF IN THE CASE OF WEIGHT TWO

We will now sketch an elementary proof of Theorem 3.2 for forms f of weight two ("elementary" meaning elementary compared to the next section). So suppose that f is as in Theorem 3.2, of weight two, and that $x^2 - a_p x + \varepsilon(p)p$ has a double root λ in $\overline{\mathbb{Q}}$ for some prime number p not dividing the level of f. Then of course we have $\lambda^2 = \varepsilon(p)p$ and $2\lambda = a_p$. Let K be the finite extension of $\mathbb{Q}$ generated by the a_n, and let O_K be its ring of integers. A construction of Eichler and Shimura gives an abelian variety $A_{\mathbb{Q}}$ over $\mathbb{Q}$ of dimension $[K : \mathbb{Q}]$ and a morphism of rings $O_K \to \mathrm{End}(A_{\mathbb{Q}})$, such that the representations $\rho_{f,l}$ are realized by the l-adic Tate modules of $A_{\mathbb{Q}}$. This abelian variety has good reduction at p; let $A_{\mathbb{Z}_p}$ denote the corresponding abelian scheme over $\mathbb{Z}_p$. Let $M := \mathrm{H}^1_{\mathrm{DR}}(A_{\mathbb{Z}_p}/\mathbb{Z}_p)$ be the first algebraic de Rham cohomology group of this abelian scheme. It is a free $\mathbb{Z}_p$-module of rank $2[K : \mathbb{Q}]$, equipped with its Hodge filtration

$$M = \mathrm{Fil}^0 M \supset \mathrm{Fil}^1 M = \mathrm{H}^0(A_{\mathbb{Z}_p}, \Omega^1).$$

The submodule $\mathrm{Fil}^1 M$ is a free of rank $[K:\mathbb{Q}]$ as $\mathbb{Z}_p$-module, and has the property that $\mathrm{Fil}^0 M/\mathrm{Fil}^1 M$ is torsion free. The double root λ of $x^2-a_p x+\varepsilon(p)p$ is in O_K, since it is integral and 2λ is in K. In the endomorphism ring of $A_{\mathbb{F}_p}$ we have the Eichler–Shimura relation:

$$0=(\mathrm{Frob}_p-\mathrm{Frob}_p)(\mathrm{Frob}_p-\mathrm{Frob}'_p)=\mathrm{Frob}_p^2-a_p\mathrm{Frob}_p+\varepsilon(p)p=(\mathrm{Frob}_p-\lambda)^2,$$

where Frob_p denotes the Frobenius endomorphism and Frob'_p the Verschiebung, multiplied by $\varepsilon(p)$. Now Frob_p is semi-simple, meaning that it satisfies an identity of the form $P(\mathrm{Frob}_p)=0$ with P a polynomial with coefficients in $\mathbb{Q}$ having simple roots. It follows that $\mathrm{Frob}_p=\lambda$ in $\mathrm{End}(A_{\mathbb{F}_p})$. Since $O_K\otimes\mathbb{Z}_p$ is a product of a finite number of discrete valuation rings, $\mathrm{Fil}^1 M$ is a locally free module over it; it is in fact free of rank one. It follows that λ does not annihilate $\mathrm{Fil}^1 M\otimes\mathbb{F}_p$, since we have $\lambda^2=\varepsilon(p)p$. But $\mathrm{Fil}^1 M\otimes\mathbb{F}_p$ is the same as $\mathrm{H}^0(A_{\mathbb{F}_p},\Omega^1)$, and on this module λ acts as Frob_p, hence it does annihilate. This contradiction finishes the proof.

5. The general case

In this last section we sketch the proof of Theorem 3.2. So let f be as in that theorem, and suppose that p is a prime number not dividing the level of f such that $x^2-a_p x+\varepsilon(p)p^{k-1}$ has a double root λ. Consider the representation $\rho_{f,p}\colon \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})\to \mathrm{GL}_2(K\otimes\mathbb{Q}_p)$. Fontaine has constructed the so-called "mysterious functor" $\mathrm{D}_{\mathrm{cris}}$ from the category of finite dimensional representations over $\mathbb{Q}_p$ of $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ to the category of filtered ϕ-modules. A filtered ϕ-module is a finite dimensional $\mathbb{Q}_p$-vector space M with a filtration Fil and an endomorphism ϕ. The morphisms are the obvious ones: the linear maps respecting Fil and ϕ. It is a theorem of Faltings (of which a special case was proved earlier by Fontaine and Messing) that for X a motive over $\mathbb{Q}_p$ with good reduction the filtered ϕ-module $\mathrm{D}_{\mathrm{cris}}(\mathrm{H}^i_{\mathrm{et}}(X_{\overline{\mathbb{Q}}_p},\mathbb{Q}_p))$ is functorially isomorphic to the crystalline cohomology group $\mathrm{H}^i_{\mathrm{cris}}(X,\mathbb{Q}_p)$ with its Hodge filtration and Frobenius endomorphism. Most important for us is the consequence of this theory that says that such filtered ϕ-modules $\mathrm{H}^i_{\mathrm{cris}}(X,\mathbb{Q}_p)$ are what is called "weakly admissible". To a filtered ϕ-module M one can associate two polygons: the Hodge polygon, depending only on the filtration, and the Newton polygon, depending only on ϕ. Weakly admissible means that the Newton polygon lies above the Hodge polygon, and that these polygons have the same endpoint. An equivalent formulation is the following. For M a filtered ϕ-module let $t_N(M)$ be the p-adic valuation of the determinant of ϕ, and let $t_H(M)$ be the maximal i such that $\mathrm{Fil}^i \det M\neq 0$. Then M is weakly admis-

sible if and only if 1: $t_N(M) = t_H(M)$ and 2: for all subobjects M' of M one has $t_H(M') \leq t_N(M')$.

Consider now the weakly admissible filtered ϕ-module $M := \mathrm{D}_{\mathrm{cris}}(\rho_{f,p})$. As before, we have $(\phi - \lambda)^2 = 0$ on M. Since we suppose that ϕ is semi-simple (i.e., ϕ is the crystalline Frobenius mentioned in the theorem), it follows that $\phi = \lambda$. Hence $\mathrm{Fil}^{k-1}M$ is a subobject of M. We have $t_H(\mathrm{Fil}^{k-1}M) = [K : \mathbb{Q}](k-1)$ and $t_N(\mathrm{Fil}^{k-1}M) = [K : \mathbb{Q}](k-1)/2$. Since $k \geq 2$, this contradicts the weak admissibility of M.