The Power of 2: Small Primes in Number Theory

Jack A. Thorne (University of Cambridge, UK)

From Euclid to Gauss

The first proposition in Euclid's *Elements* gives the construction, with ruler and compass, of the equilateral triangle. Later, Euclid shows how to construct a regular n-gon for $n = 5$ and $n = 15$, and how to pass from a construction of the regular n-gon to a construction of the regular $2n$-gon. For which other values of n does a construction of the regular n-gon using ruler and compass exist?

The answer to this ancient question was given roughly 2000 years later, by Gauss. In 1796, Gauss showed that the regular 17-gon could be constructed, and eventually showed in *Disquisitiones Arithmeticae* that the n-gon is constructible when n is of the form $n = 2^k p_1 \dots p_r$, where $p_1, \dots, p_r$ are distinct *Fermat primes*, i.e., prime numbers of the form $F_m = 2^{2^m} + 1$.

Gauss' success in making progress on this question rested on the fact that the Fermat number $F_2 = 17$ is a prime. Legend has it that it was Gauss' pleasure in proving this that made his mind up to pursue mathematics as a career. The next Fermat number, $F_3 = 65537$, is also prime, but no other Fermat primes are known. Today's readers of the *Disquisitiones* can be thankful that the list of Fermat primes does not end at $F_1 = 5$!

My research is in algebraic number theory, and in particular the Langlands program, which aims to give the ultimate non-abelian generalisation of class field theory, a topic which has its roots in topics treated in *Disquisitiones* (consider quadratic reciprocity, the ideal class group and the reduction theory of binary quadratic forms, to name but a few).

Much research in this part of number theory pulls in ideas from many other parts of mathematics (geometry, representation theory, analysis, ...) – anything that will help reach the final goal. And in many cases, a lucky numerical coincidence helps to push us over the finishing line. My aim in this article is to introduce the reader to some of the remarkable recent research on the subject (as well as some of my own work), with a particular eye for the supporting roles played by small primes.

Cyclotomic fields

Let us first explain the modern point of view on the ideas behind Gauss' construction. First, identifying the plane with $\mathbb{C}$, one sees that it is enough to construct the n^{th} roots of unity. Second, one sees that the complex numbers which are constructible are precisely those that can be seen inside a tower of quadratic field extensions of the field $\mathbb{Q}$ of rational numbers. The challenge, therefore, is to explain when $e^{2\pi i/n}$ is contained in such a field extension.

This is precisely the kind of question that Galois theory is equipped to answer. Indeed, the number $e^{2\pi i/n}$ generates the n^{th} cyclotomic field $K_n = \mathbb{Q}(e^{2\pi i/n})$. The *Galois correspondence* states that there are as many subfields of K_n as there are subgroups of its Galois group $\mathrm{Gal}(K_n/\mathbb{Q})$, the group of all automorphisms of K_n. Galois theory shows that K_n can be obtained by iterated quadratic extensions precisely when its Galois group has order a power of 2.

One can show that there is an isomorphism $\mathrm{Gal}(K_n/\mathbb{Q}) \cong (\mathbb{Z}/n\mathbb{Z})^\times$, so we see that the n-gon is constructible precisely when the value $\phi(n)$ of Euler's totient function is a power of 2. Elementary number theory shows the equivalence of this condition with the criterion given by Gauss.

Class field theory

Cyclotomic fields are the most basic examples of abelian extensions of number fields. A number field is a field extension $L/\mathbb{Q}$ which can be obtained by adjoining finitely many algebraic numbers. We say that $L/\mathbb{Q}$ is Galois if L can be obtained by adjoining all (not just some) of the roots of a fixed polynomial with rational coefficients. If this condition is satisfied, then the Galois group $\mathrm{Gal}(L/\mathbb{Q})$ of automorphisms of L acts on L (and permutes these roots). We say that $L/\mathbb{Q}$ is abelian when it is Galois and its Galois group $\mathrm{Gal}(L/\mathbb{Q})$ is abelian. As we have seen, this class of number fields includes the cyclotomic fields K_n.

A fundamental additional structure carried by the Galois group of a number field is the presence, for each prime p, of a conjugacy class of subgroups $D_p \subset \mathrm{Gal}(L/\mathbb{Q})$. We call D_p the decomposition group at the prime p; it may be defined as the subgroup of automorphisms of L which are continuous with respect to the topology given by an absolute value on L extending the p-adic absolute value $|\cdot|_p$ on $\mathbb{Q}$ (the completion of which gives the field $\mathbb{Q}_p$ of p-adic rational numbers). Much of the charm of algebraic number theory comes from the interaction between global phenomena (e.g., the arithmetic of the field L) and local phenomena (e.g., the structure of the group D_p and the arithmetic of the completions of L with respect to its p-adic absolute values).

The decomposition group comes with a normal subgroup I_p, the inertia group, and a canonical generator for the cyclic quotient D_p/I_p, called the Frobenius element Frob_p. For all but finitely many primes p (which are said to be unramified in L) the inertia group is trivial, and we obtain a well-defined conjugacy class of elements of $\mathrm{Gal}(L/\mathbb{Q})$. There is a similar story when $\mathbb{Q}$ is replaced by any base number field K.

We can now explain the importance of abelian extensions L/K of number fields. When the Galois group is abelian, the Frobenius elements are well-defined (not just up to conjugacy). This is the mechanism by which class field theory, one of the great achievements of mathematics in the first half of the 20^{th} century, describes all abelian extensions of a given

number field: it gives a canonical surjection from a generalised ideal class group of K to the group $\mathrm{Gal}(L/K)$, uniquely characterised by the requirement that the class of a prime ideal of the ring of integers of K is sent to the corresponding Frobenius element.

Serre's conjecture

The Langlands program should include, as a special case, a non-abelian generalisation of class field theory. By duality, we can think of class field theory as giving a correspondence between the 1-dimensional representations of Galois groups of number fields and the irreducible representations of generalised ideal class groups. The Langlands conjectures would describe n-dimensional representations of Galois groups in terms of automorphic representations, which play the role of characters of ideal class groups.

As a window into this circle of ideas, we are now going to describe Serre's conjecture, which aims to give an "automorphic" parameterisation of 2-dimensional representations of Galois groups over $\mathbb{Q}$ in characteristic p. Serre published his conjecture in 1987 [12]. It is closely related to, but not equivalent to, the Langlands conjectures, and has had a tremendous influence on their study.

We introduce some necessary notation. Let $\overline{\mathbb{Q}}$ be an algebraic closure of $\mathbb{Q}$, and let $G_{\mathbb{Q}} = \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ be the absolute Galois group, equipped with its Krull topology. If $L/\mathbb{Q}$ is any Galois number field (contained in $\overline{\mathbb{Q}}$) then there is a continuous surjection $G_{\mathbb{Q}} \to \mathrm{Gal}(L/\mathbb{Q})$.

Let p be a prime. The first class of objects we consider in Serre's conjecture consists of continuous representations $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$, with coefficients in the algebraic closure of the finite field $\mathbb{F}_p$ of p elements. We say that $\overline{\rho}$ is of S-type if it is irreducible and if $\det \overline{\rho}(c) = -1$, where $c \in G_{\mathbb{Q}}$ is complex conjugation. A typical source of such representations is in the p-torsion subgroups of elliptic curves. We will discuss this example in more detail below.

The second class of objects appears inside the cohomology groups of arithmetic groups, which play a role analogous to that of the generalised ideal class groups in class field theory. If $N \geq 1$ is an integer, then we define $\Gamma_1(N)$ to be the subgroup of matrices

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

satisfying the congruence conditions $c \equiv 0 \bmod N$, $a \equiv d \equiv 1 \bmod N$. The (group) cohomology groups $H^*(\Gamma_1(N), \overline{\mathbb{F}}_p)$ are finite-dimensional vector spaces. More generally, if we are given a pair of integers k, t with $2 \leq k \leq p+1$ and $0 \leq t \leq p-2$, then we can consider $V_{k,t} = \mathrm{Sym}^{k-2} \overline{\mathbb{F}}_\ell^2 \otimes \det^t$, a finite-dimensional, irreducible representation of $\mathrm{GL}_2(\mathbb{F}_p)$, hence the cohomology groups $H^*(\Gamma_1(N), V_{k,t})$.

The group $\Gamma_1(N)$ is a *congruence subgroup* of the group $\mathrm{GL}_2(\mathbb{Q})$. As a consequence, its cohomology groups receive additional symmetries, in the form of an action of the *Hecke algebra* of $\mathrm{GL}_2(\mathbb{Q})$. For each prime $\ell \nmid Np$, there is a distinguished Hecke operator T_ℓ which acts on the vector space $H^*(\Gamma_1(N), V_{k,t})$, and can be thought of as playing the role of the ideal class of a prime ideal in class field theory. If ℓ_1, ℓ_2 are two such primes then the operators T_{ℓ_1}, T_{ℓ_2} commute. Consequently, there exist elements $v \in H^*(\Gamma_1(N), V_{k,t})$ which are simultaneous eigenvectors for all of the Hecke operators T_ℓ ($\ell \nmid Np$). We call the collection $(a_\ell)_{\ell \nmid Np}$ of eigenvalues in $\overline{\mathbb{F}}_p$ a system of Hecke eigenvalues of level N and weight (k, t).

We are now ready to state Serre's conjecture.[1] The first approximation is that for any S-type representation, there exists $N, (k, t)$ and a system of Hecke eigenvalues of level N and weight (k, t) such that for all $\ell \nmid Np$,

$$\mathrm{tr}\, \overline{\rho}(\mathrm{Frob}_\ell) = a_\ell.$$

The full conjecture asserts that the smallest possible N with this property should be equal to the conductor $N(\overline{\rho})$, an integer which measures the ramification of the representation $\overline{\rho}$, and that the possible (k, t)'s for which this system of Hecke eigenvalues appears can be described explicitly using a recipe which depends only on the restriction of $\overline{\rho}$ to the inertia group $I_p \subset G_{\mathbb{Q}}$.

This conjecture is remarkable on many levels. It implies that the set of isomorphism classes of S-type representations $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ of bounded conductor is finite (since the cohomology groups on the automorphic side are finite-dimensional). This elementary statement has no known proof that does not rely on automorphic forms. More generally, the cohomology groups can be computed algorithmically, and this information used to give precise information on the existence (or otherwise) of specific Galois representations. An amazing application of this is the Cremona database, which lists all elliptic curves E over $\mathbb{Q}$ of conductor $N(E) \leq 5 \times 10^5$ [13].

Also significant for the development of the subject has been the paradigm suggested by the conjecture: the most optimistic form of a local-global principle relating the relative position of decomposition groups in the Galois groups of number fields to the cohomology of arithmetic groups. From this point of view one can explain the existence of Ribet's level-raising and level-lowering congruences [10,11] between modular forms using a simple computation with Galois representations. The recipe for the set of weights (k, t) which should give rise to $\overline{\rho}$ suggests the existence of a close relationship between the representation theory of the decomposition group $D_p \subset G_{\mathbb{Q}}$ and that of the group $\mathrm{GL}_2(\mathbb{Z}_p)$. This theme that has been made precise in the Breuil–Mézard conjecture [3] and has seen its ultimate expression in the formulation of the p-adic Langlands correspondence for $\mathrm{GL}_2(\mathbb{Q}_p)$ [2].

Fermat's Last Theorem

Another reason for the great interest of Serre's conjecture is that it implies Fermat's Last Theorem, following a famous gambit using the Frey curve associated to a putative non-trivial solution to the Fermat equation. Indeed, suppose given a solution

$$a^p + b^p = c^p$$

[1] In fact, Serre's formulation uses the reduction modulo p of certain spaces of automorphic forms in the place of the cohomology of arithmetic groups. We have followed Buzzard–Diamond–Jarvis [4] in using cohomology, since it is both easier to describe and more amenable to generalisation, for example to base fields other than $\mathbb{Q}$. The existence of the Eichler–Shimura isomorphism implies that the systems of Hecke eigenvalues are the same in either case.

to the Fermat equation, where $p \geq 5$ is a prime and a, b, c are coprime integers such that $abc \neq 0$. (It is enough to consider this case, since the non-existence of solutions in exponents 3 and 4 was already proved by Euler and Fermat, respectively.) To this solution one associates the elliptic curve E given by the equation

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

Thus E is an algebraic curve of genus one, defined over $\mathbb{Q}$, which therefore admits a structure of commutative group variety, with identity element given by the unique point at infinity. The complex points $E(\mathbb{C})$ are isomorphic, as a Lie group, to $S^1 \times S^1$; consequently, if $n \geq 1$ is an integer, then the n-torsion points $E[n](\mathbb{C})$ form a finite abelian group abstractly isomorphic to $(\mathbb{Z}/n\mathbb{Z})^2$. Since the group operations of E are defined over $\mathbb{Q}$, these points are defined over the subfield $\overline{\mathbb{Q}} \subset \mathbb{C}$ of algebraic numbers. We write $L_n/\mathbb{Q}$ for the number field generated by the x, y co-ordinates of the non-trivial n-torsion points.

Then $L_n/\mathbb{Q}$ is a Galois extension, and its Galois group $\mathrm{Gal}(L_n/\mathbb{Q})$ acts on $E[n](\mathbb{C})$. Choosing a basis for this free $\mathbb{Z}/n\mathbb{Z}$-module gives a Galois representation

$$\overline{\rho}_{E,n} : \mathrm{Gal}(L_n/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z}).$$

Our discussion up to this point would apply equally well for any elliptic curve over $\mathbb{Q}$. However, something very special happens for our curve, associated to a non-trivial solution to the Fermat equation in degree p. Indeed, in this case the number field L_p associated to the p-torsion points turns out to have very little ramification, essentially because the discriminant $\Delta(E)$ of E is (up to powers of 2) a p^{th} power.

We can make this precise by computing the invariants N, (k, t) attached to the representation $\overline{\rho}_{E,p}$ (which is of S-type). Assuming, as we may, that $a \equiv 3 \bmod 4$ and $b \equiv 0 \bmod 2$, we find that $N = 2$, $(k, t) = (2, 0)$. Serre's conjecture implies that the representation $\overline{\rho}_{E,p}$ should be associated to a system of Hecke eigenvalues occurring in $H^1(\Gamma_1(2), \overline{\mathbb{F}}_p)$. This is a contradiction! Indeed, the space $H^1(\Gamma_1(2), \overline{\mathbb{F}}_p)$ is 1-dimensional, and the unique system of Hecke eigenvalues it carries is not associated to a representation of S-type (in fact, it is associated to a reducible Galois representation).

Of course, Fermat's Last Theorem was proved first by Wiles in 1993, more than 10 years before the proof by Khare and Wintenberger of Serre's conjecture. Wiles' proof introduced a vast number of new ways to study the relation between Galois representations and automorphic forms, many of which appear again in an essential way in the work of Khare–Wintenberger. However, the route that Wiles followed to Fermat's Last Theorem is essentially the one we have outlined above: he proved the modularity of the elliptic curve E, hence of the representation $\overline{\rho}_{E,p}$, at level $N = N(E)$. Ribet's level-lowering results, alluded to earlier, imply the modularity of $\overline{\rho}_{E,p}$ at level $N = N(\overline{\rho}_{E,p}) = 2$, leading to a contradiction.

This is an appropriate moment to explain what it means for a general elliptic curve E over $\mathbb{Q}$ to be modular. In fact, it is simpler from our point of view to explain what it means for a Galois representation

$$\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$$

with coefficients in the algebraic closure of the field $\mathbb{Q}_p$ of p-adic rationals to be modular. Exactly as in the case of $\overline{\mathbb{F}}_p$-coefficients, the cohomology groups $H^*(\Gamma_1(N), \overline{\mathbb{Q}}_p)$ receive actions of the pairwise commuting Hecke operators T_ℓ, defined for each prime $\ell \nmid Np$. We say that ρ is modular of level N and weight[2] $(2, 0)$ if there exists a simultaneous eigenvector $v \in H^1(\Gamma_1(N), \overline{\mathbb{Q}}_p)$ for the Hecke operators T_ℓ such that for all $\ell \nmid Np$, ρ is unramified at ℓ and we have the equality

$$\mathrm{tr}\,\rho(\mathrm{Frob}_\ell) = \text{eigenvalue of } T_\ell \text{ on } v.$$

One appealing feature here is that these cohomology groups, together with their Hecke operators, are defined over $\mathbb{Q}$: they arise by base extension from the vector space $H^1(\Gamma_1(N), \mathbb{Q})$. If $v \in H^1(\Gamma_1(N), \overline{\mathbb{Q}})$ is a simultaneous eigenvector for all of the Hecke operators T_ℓ ($\ell \nmid N$), then the eigenvalues generate a number field $K = \mathbb{Q}(\{a_\ell\}_{\ell \nmid N})$. Associated to v is a *compatible system* of Galois representations $\rho_\lambda : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$, one for each prime p and choice of embedding $\lambda : K \to \overline{\mathbb{Q}}_p$.

If E is an elliptic curve, then for any prime p we can glue the representations $\overline{\rho}_{E,p^n} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}/p^n\mathbb{Z})$ together into a representation $\rho_{E,p^\infty} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Q}_p)$. These representations form a compatible system, and we say that E is modular if one (equivalently, all) of them is modular in the above sense.

Modularity lifting theorems

The most important innovation in Wiles' work is probably the concept of the modularity lifting theorem. To explain this, we first recall that the topology on $\overline{\mathbb{Q}}_p$ is defined by the p-adic absolute value $|\cdot|_p$. The set of elements of absolute value at most 1 is a subring, denoted $\overline{\mathbb{Z}}_p$, and the set of elements of absolute value strictly less than 1 is an ideal in this subring, denoted $\mathfrak{m}_{\overline{\mathbb{Z}}_p}$. The quotient $\overline{\mathbb{Z}}_p/\mathfrak{m}_{\overline{\mathbb{Z}}_p}$ may be identified with $\overline{\mathbb{F}}_p$.

If $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$ is a continuous representation, then we may conjugate ρ to take values in $\mathrm{GL}_2(\overline{\mathbb{Z}}_p)$, and reduce modulo the ideal $\mathfrak{m}_{\overline{\mathbb{Z}}_p}$ to obtain a representation $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$. The character of $\overline{\rho}$ is determined by that of ρ.

We have defined what it means for $\overline{\rho}$ to be modular, and also what it means for ρ to be modular. It is natural to ask how these concepts are related. One direction is easy: if ρ is modular, then so is $\overline{\rho}$, as can be shown by considering the reduction modulo p of classes in $H^1(\Gamma_1(N), \overline{\mathbb{Q}}_p)$. Much harder is to go in the opposite direction. In general, there are many more systems of Hecke eigenvalues occurring in $H^1(\Gamma_1(N), \overline{\mathbb{Q}}_p)$ than the analogous group with $\overline{\mathbb{F}}_p$-coefficients. This reflects the existence of plentiful congruences between modular forms, and is the source both of the difficulty of the problem and of the power of its solution.

Wiles proved the first modularity lifting theorem, stating that for a representation ρ satisfying some technical conditions, the modularity of $\overline{\rho}$ implies the modularity of ρ. (These technical conditions usually take the form of a global condition on $\overline{\rho}$, for example that it is irreducible, and some necessary local conditions on ρ, for example that $\rho|_{D_p}$ can be realised inside an abelian variety.)

To prove the modularity of an elliptic curve E using such a theorem, one needs to choose a prime p (with the aim of proving ρ_{E,p^∞} is modular) and verify the modularity of the residual

[2] One can consider other weights (k, t); this amounts to replacing $\overline{\mathbb{Q}}_p$ by a non-trivial coefficient system, just as we have done above in the mod p case.

representation $\overline{\rho}_{E,p}$. Wiles chooses $p = 3$, and makes use of the following two remarkable coincidences: first, that the reduction homomorphism $\mathrm{GL}_2(\mathbb{Z}_3) \to \mathrm{GL}_2(\mathbb{F}_3)$ has a splitting $s : \mathrm{GL}_2(\mathbb{F}_3) \to \mathrm{GL}_2(\mathbb{Z}_3)$; second, that the resulting representation $s \circ \overline{\rho}_{E,3} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_3)$ has soluble image, and so can be shown to be modular using earlier work of Langlands–Tunnell (in particular, Langlands' proof of cyclic base change and descent for automorphic forms on GL_2). Neither of these two facts holds for any prime $p > 3$. Combining these observations with his modularity lifting theorem, Wiles is able to prove the modularity of semistable elliptic curves E, provided that $\overline{\rho}_{E,3}$ is irreducible.

To treat the remaining case, Wiles introduces another famous technique, the "3–5 switch". To prove the modularity of an elliptic curve E such that $\overline{\rho}_{E,3}$ is reducible, he introduces an auxiliary elliptic curve A with the property that $\overline{\rho}_{A,5} \cong \overline{\rho}_{E,5}$ and $\overline{\rho}_{A,3}$ is irreducible. This is possible since $X(\overline{\rho}_{E,5})$, the modular curve which parameterises those elliptic curves A such that $\overline{\rho}_{A,5} \cong \overline{\rho}_{E,5}$, is isomorphic to $\mathbb{P}^1_{\mathbb{Q}}$, and therefore has infinitely many rational points (once again, this would be false if 5 was replaced here by any larger prime.) The modularity of A follows using the argument of the previous paragraph. Finally, applying the modularity lifting theorem with $p = 5$ we deduce the modularity of E.

The proof of Serre's conjecture

As evidenced by Wiles' proof, modularity lifting theorems become especially potent in the presence of compatible systems of Galois representations. In fact, this combination formed the basis of the proof of Serre's conjecture by Khare and Wintenberger, in which the "3–5 switch" becomes a "p–P"-switch, where p, P are primes which become arbitrarily large.

Let us sketch the earlier proof by Khare [7] of the $N = 1$ case of Serre's conjecture, i.e., the modularity of S-type representations $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ which are unramified outside p. The argument is by induction on the prime p; in fact, it is enough to show the truth of the conjecture for infinitely many primes.

The base cases of the induction is the case $p = 2$. As we saw earlier, $H^1(\Gamma_1(2), \overline{\mathbb{F}}_2)$ is essentially trivial, leading to the expectation that the conjecture is vacuously true in the case $p = 2$: there are no irreducible representations. This is true, and was proved directly by Tate by analysing the discriminant of the number field cut out by a putative irreducible representation $\overline{\rho}$. This is a more sophisticated version of the elementary deduction, from Minkowski's bound, that there is no number field $L/\mathbb{Q}$ which is unramified everywhere.

What about the induction step? Suppose that the conjecture is true for a given prime p, and fix a second prime $P > p$ and an irreducible representation $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_P)$. The first step is to lift $\overline{\rho}$ to a compatible system of representations $\rho_\lambda : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_\lambda)$. Taking a p-adic member ρ_p of this compatible system, one hopes to verify the residual modularity of $\overline{\rho}_p$ (by induction) and then apply a modularity lifting theorem to deduce the modularity of ρ_p. We can then appeal, as in the case of the compatible systems of Galois representations associated to elliptic curves, to the fact that the modularity of a single member of a compatible family is equivalent to the modularity of every member.

Carrying out this argument in practice is subtle because of the technical conditions imposed by modularity lifting theorems, and would not have been possible without important refinements of the modularity lifting theorems in Wiles' original in work by many authors. Another key ingredient is a technique for constructing lifts of $\overline{\mathbb{F}}_p$-representations $\overline{\rho}$ to $\overline{\mathbb{Q}}_p$-representations with prescribed local behaviour (for example, with the same conductor N as $\overline{\rho}$). The existence of such lifts is a key consequence of Serre's conjecture; Khare and Wintenberger turned this on its head and established it, using modularity lifting theorems, on the way to proving the full conjecture.

Symmetric power functoriality

We now discuss applications of these kinds of techniques to a different problem. Symmetric power functoriality refers to a special case of Langlands' functoriality conjectures, which suggest a beautiful set of relations between automorphic forms on different reductive groups. These relations should reflect, through Langlands duality, the relations between the Langlands dual groups. The most basic relations are those associated to the symmetric powers of the standard representation of GL_2.

Both the shape and the importance of these conjectures can be motivated by considering the case of modular elliptic curves over $\mathbb{Q}$: in this case, they are related to the now-proved Sato–Tate conjecture. If E is an elliptic curve over $\mathbb{Q}$, then for all primes $p \nmid N(E)$, the curve E has good reduction, and it makes sense to consider the set of $\mathbb{F}_p$-points of E. Write n_p for the cardinality of this set; then Hasse's theorem implies the estimate

$$|p + 1 - n_p| \leq 2\sqrt{p}.$$

The Sato–Tate conjecture[3] concerns the distribution of the normalised error terms $a_p = (p + 1 - n_p)/2\sqrt{p} \in [-1, 1]$ as the prime p varies: it states that the numbers $\{a_p \mid p < X\}$ become equidistributed as $X \to \infty$ with respect to the Sato–Tate measure $\frac{2}{\pi}\sqrt{1 - t^2}dt$.

Serre identified how one might hope to prove the Sato–Tate conjecture, using a method inspired at some level by the Hadamard–de la Vallée Poussin proof of the prime number theorem. The crux is to consider the so-called symmetric power L-functions associated to the elliptic curve E.

Recall first that the L-function associated to an elliptic curve E over $\mathbb{Q}$ is defined as an Euler product[4]

$$L(E, s) \doteq \prod_p (1 - a_p p^{-s} + p^{1-2s})^{-1},$$

where s is a complex variable. Hasse's theorem implies that this product converges absolutely in the right half-plane $\mathrm{Re}\, s > 3/2$. In fact, the resulting holomorphic function admits an analytic continuation to the whole complex plane; this is a consequence of the modularity of the elliptic curve E. The famous Birch–Swinnerton-Dyer conjecture relates the group $E(\mathbb{Q})$ of rational points of E to the leading coefficient in the Taylor expansion of this function at the point $s = 1$ [14].

[3] This statement is valid provided that E does not have complex multiplication (CM). When E does have CM, a different measure must be used, reflecting the existence of the curve's additional symmetries.

[4] Some care is needed to define the Euler factors at the primes $p|N(E)$; we elide this detail here.

For each $n \geq 1$, we may equally define the n^{th} symmetric power L-function as follows. Let $\mathrm{Sym}^n : \mathrm{GL}_2 \to \mathrm{GL}_{n+1}$ denote the n^{th} symmetric power of the standard (identity) representation of GL_2. Let $t_p \in \mathrm{GL}_2(\mathbb{C})$ be a matrix with characteristic polynomial $\det(X - t_p) = X^2 - a_p X + p$. Then we define

$$L(E, \mathrm{Sym}^n, s) \doteq \prod_p \det\left(1 - p^{-s}\,\mathrm{Sym}^n(t_p)\right)^{-1}.$$

(When $n = 1$, Sym^1 is the standard representation and $L(E, \mathrm{Sym}^1, s) = L(E, s)$.) Once again, this Euler product converges absolutely in a right-half plane $\mathrm{Re}\, s > 1 + n/2$. Serre's observation was that if all the symmetric power L-functions can be shown to admit an analytic continuation to the whole complex plane, non-vanishing on the line $\mathrm{Re}\, s = 1 + n/2$, then the Sato–Tate conjecture follows.

These properties of the symmetric power L-functions follow from the Langlands conjectures! Indeed, $L(E, \mathrm{Sym}^n, s)$ would be precisely the standard L-function associated to an automorphic representation of GL_{n+1}, which deserves to be called the symmetric power lifting of the automorphic representation of GL_2 associated to E. These standard L-functions are known to have the required analytic continuation and non-vanishing properties.

The Sato–Tate conjecture for elliptic curves has now been proved. It turns out that the necessary analytic properties of the symmetric power L-functions can be proved to follow from the *potential automorphy* of the functorial lifts. More precisely, it is enough to show that the symmetric power L-functions admit meromorphic continuation and are holomorphic and non-vanishing on the appropriate line; these properties follow from the automorphy of the Galois representations $\mathrm{Sym}^n \rho_{E,p^\infty}|_{G_{M_n}}$, for some (inexplicit) totally real number field $M_n/\mathbb{Q}$. This was established in a series of works (culminating in [1]) based on Taylor's technique of potential automorphy and relying on contributions to the Langlands program by many other mathematicians.

This year, James Newton and I proved the automorphy of the symmetric power L-functions of elliptic curves; this shows in particular that they have the expected analytic (as opposed to merely meromorphic) continuation to the entire complex plane. More generally, we showed that for any automorphic representation of GL_2 which contributes to the cohomology of congruence subgroups of $\mathrm{GL}_2(\mathbb{Q})$, all of the symmetric power lifts exist, as predicted by Langlands' conjectures [8, 9].

We now sketch the proof of this result in the essential case of automorphic representations of level 1 (equivalently, which contribute to the cohomology of $\mathrm{SL}_2(\mathbb{Z})$ for some choice of coefficient system); this includes the important case of the representation generated by Ramanujan's Δ-function

$$\Delta(q) = q \prod_{n=1}^{\infty} (1 - q^n)^{24}.$$

Our proof is based on the existence of the Coleman–Mazur eigencurve $\mathcal{E}_p$ [6], which can be defined for any fixed prime p, and is a kind of a universal p-adic family of systems of Hecke eigenvalues. The eigencurve $\mathcal{E}_p$ is a 1-dimensional p-adic rigid analytic space, which admits a morphism

$$\mathcal{E}_p \to \mathcal{W}$$

to the space $\mathcal{W} = \mathrm{Hom}(\mathbb{Z}_p^\times/\{\pm 1\}, \mathbb{G}_m)$, called weight space. The eigencurve $\mathcal{E}_p$ has a dense set of "classical points" corresponding to pairs $(\{a_\ell\}_{\ell \nmid p}, \alpha_p)$, where $\{a_\ell\}_\ell$ is a system of Hecke eigenvalues appearing in some group $H^1(\mathrm{SL}_2(\mathbb{Z}), \mathrm{Sym}^{k-2} \overline{\mathbb{Q}}^2)$ and α_p is a root of the Hecke polynomial $X^2 - a_p X + p^{k-1}$; the image of this classical point in weight space is the character $x \mapsto x^{k-2}$.

The density of these classical points is a reflection of the fact that systems of Hecke eigenvalues can be put in p-adic families, in which the Hecke eigenvalues vary continuously (in the p-adic topology) as a function of the weight k. The non-classical points of $\mathcal{E}_p$ can be interpreted as arising from the systems of Hecke eigenvalues appearing in p-adic "overconvergent" cohomology groups.

The first step in our proof is to show that (for fixed n) the automorphy of the n^{th} symmetric power lifting is a property which is constant on irreducible components of $\mathcal{E}_p$: put another way, we can analytically continue the functorial lift along irreducible components of the eigencurve. We are also able to establish the existence (using modularity lifting theorems) of *some* modular forms for which a given symmetric power lift exists.

Each irreducible component of the eigencurve contains infinitely many classical points, so this shows at least that for each n, infinitely many modular forms admit a symmetric power lifting. This does not yet solve the problem completely, since the irreducible components of the eigencurve, and its global geometry more generally, remain mysterious.

We have not yet specified a choice of prime p. We now choose $p = 2$. Buzzard and Kilford were able to compute a large part of the 2-adic eigencurve $\mathcal{E}_2$, namely the part 'close to the boundary of weight space' [5]. When $p = 2$ the group $\mathbb{Z}_p^\times/\{\pm 1\}$ is free and $\mathcal{W}$ may be identified with the open p-adic disc $\{|w| < 1\}$. By the boundary of weight space, we mean the annulus $\{1/8 < |w| < 1\}$. Buzzard–Kilford showed that above this boundary annulus, the geometry of the eigencurve in fact becomes very simple: a countably infinite collection of open annuli, each of which maps isomorphically to the boundary of weight space.

To finish the proof, we need only to show that each of this infinite collection of boundary annuli inside $\mathcal{E}_2$ meets an irreducible component over which the symmetric power lifting exists. This we can achieve by combining our freedom to analytically continue along components with the freedom to move between the two classical points corresponding to the two roots of the Hecke polynomial $X^2 - a_p X + p^{k-1}$: above the boundary of weight space, this has the effect of jumping between different boundary annuli in $\mathcal{E}_2$.

We conclude with a concrete numerical consequence of the Buzzard–Kilford theorem: if $n \geq 3$ and $\chi : (\mathbb{Z}/2^n\mathbb{Z})^\times \to \overline{\mathbb{Q}}_2^\times$ is a primitive character such that $\chi(-1) = 1$, then the space of cuspidal modular forms of level 2^n, weight 2, and character χ has dimension 2^{n-3}, and the 2-adic valuations of the eigenvalues of the U_2 operator are the numbers in $(\frac{1}{2^{n-3}}\mathbb{Z}) \cap (0, 1)$, each appearing with multiplicity 1.

This statement, a beautiful generalisation of the triviality of $H^1(\Gamma_1(2), \overline{\mathbb{Q}}_2)$ which underpins the proof of Fermat's Last Theorem, is the essential starting point for our proof of symmetric power functoriality for holomorphic modular forms.

Bibliography

[1] T. Barnet-Lamb, D. Geraghty, M. Harris, and R. Taylor. A family of Calabi-Yau varieties and potential automorphy II. *Publ. Res. Inst. Math. Sci.*, 47(1):29–98, 2011.

[2] C. Breuil. The emerging p-adic Langlands programme. In *Proceedings of the International Congress of Mathematicians. Volume II*, pages 203–230. Hindustan Book Agency, New Delhi, 2010.

[3] C. Breuil and A. Mézard. Multiplicités modulaires et représentations de $GL_2(\mathbf{Z}_p)$ et de $\mathrm{Gal}(\overline{\mathbf{Q}}_p/\mathbf{Q}_p)$ en $l = p$. *Duke Math. J.*, 115(2):205–310, 2002. With an appendix by Guy Henniart.

[4] K. Buzzard, F. Diamond, and F. Jarvis. On Serre's conjecture for mod ℓ Galois representations over totally real fields. *Duke Math. J.*, 155(1):105–161, 2010.

[5] K. Buzzard and L. J. P. Kilford. The 2-adic eigencurve at the boundary of weight space. *Compos. Math.*, 141(3):605–619, 2005.

[6] R. Coleman and B. Mazur. The eigencurve. In *Galois representations in arithmetic algebraic geometry (Durham, 1996)*, volume 254 of *London Math. Soc. Lecture Note Ser.*, pages 1–113. Cambridge Univ. Press, Cambridge, 1998.

[7] C. Khare. Serre's modularity conjecture: the level one case. *Duke Math. J.*, 134(3):557–589, 2006.

[8] J. Newton and J. A. Thorne. Symmetric power functoriality for holomorphic modular forms, 2020.

[9] J. Newton and J. A. Thorne. Symmetric power functoriality for holomorphic modular forms, II, 2020.

[10] K. A. Ribet. Congruence relations between modular forms. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983)*, pages 503–514. PWN, Warsaw, 1984.

[11] K. A. Ribet. On modular representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ arising from modular forms. *Invent. Math.*, 100(2):431–476, 1990.

[12] J.-P. Serre. Sur les représentations modulaires de degré 2 de $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. *Duke Math. J.*, 54(1):179–230, 1987.

[13] The LMFDB Collaboration. The L-functions and modular forms database. http://www.lmfdb.org, 2020. [Online; accessed 15 October 2020].

[14] A. Wiles. The Birch and Swinnerton-Dyer conjecture. In *The millennium prize problems*, pages 31–41. Clay Math. Inst., Cambridge, MA, 2006.

Jack Thorne [thorne@dpmms.cam.ac.uk] is Professor of Number Theory at the University of Cambridge. In 2020 he was elected Fellow of the Royal Society and received the EMS Prize.