A Lean proof of Fermat's Last Theorem

Kevin Buzzard, Richard Taylor

May 16, 2024

Chapter 1

Introduction

Fermat's Last Theorem is the statement that if a, b, c, n are positive whole numbers with $n \geq 3$, then $a^n + b^n \neq c^n$. It is thus a statement about a family of *Diophantine equations* ($a^3 + b^3 = c^3, a^4 + b^4 = c^4, \ldots$). Diophantus was a Greek mathematician who lived around 1800 years ago, and he would have been able to understand the statement of the theorem (he knew about positive integers, addition and multiplication).

Fermat's Last Theorem was explicitly raised by Fermat in 1637, and was proved by Wiles (with the proof completed in joint work with Taylor) in 1994. There are now several proofs but all of them go broadly in the same direction, using elliptic curves and modular forms.

Explaining a proof of Fermat's Last Theorem to Lean is in some sense like explaining the proof to Diophantus; for example, the proof starts by observing that before we go any further it's convenient to first invent/discover zero and negative numbers, and one can point explicitly at places in Lean's source code here and here where these things happen. However we will adopt a more efficient approach: we will assume all of the theorems both in Lean and in its mathematics library `mathlib`, and proceed from there. To give some idea of what this entails: `mathlib` at the time of writing contains most of an undergraduate mathematics degree and parts of several relevant Masters level courses (for example, the definitions and basic properties of elliptic curves and modular forms are in `mathlib`). Thus our task can be likened to teaching a graduate level course on Fermat's Last Theorem to a computer.

The proof explained in these notes was constructed by Taylor, taking into account Buzzard's comments on what would be easy or hard to do in Lean. The proof uses refinements of the original Taylor-Wiles method by Diamond/Fujiwara, Khare-Wintenberger, Skinner-Wiles, Kisin, Taylor and others – one could call it a 21st century proof of the theorem. We shall explain more about the exact path we're taking in Chapter 4. But before we go into those technical details, we can enjoy some of the more basic arguments at the start of the proof. And the proof starts, as every known proof does, with some basic reductions and the introduction of a certain elliptic curve. We explain this in the next chapter.

Chapter 2

First reductions of the problem

2.1 Overview

The proof of Fermat's Last Theorem is by contradiction. We assume that we have a counterexample $a^n + b^n = c^n$, and manipulate it until it satisfies the axioms of a "Frey package". From the Frey package we build a Frey curve – an elliptic curve defined over the rationals. We then look at a certain representation of a Galois group coming from this elliptic curve, and finally using two very deep and independent theorems (one due to Mazur, the other due to Wiles) we show that this representation is neither reducible or irreducible, a contradiction.

2.2 Reduction to $n \geq 5$ and prime

Lemma 2.1. *If there is a counterexample to Fermat's Last Theorem, then there is a counterexample $a^p + b^p = c^p$ with p an odd prime.*

Proof. Note: this proof is in mathlib already; we run through it for completeness' sake.

Say $a^n + b^n = c^n$ is a counterexample to Fermat's Last Theorem. Every positive integer is either a power of 2 or has an odd prime factor. If $n = kp$ has an odd prime factor p then $(a^k)^p + (b^k)^p = (c^k)^p$ is the counterexample we seek. It remains to deal with the case where n is a power of 2, so let's assume this. We have $3 \leq n$ by assumption, so $n = 4k$ must be a multiple of 4, and thus $(a^k)^4 = (b^k)^4 = (c^k)^4$, giving us a counterexample to Fermat's Last Theorem for $n = 4$. However an old result of Fermat himself (proved as `fermatLastTheoremFour` in `mathlib`) says that $x^4 + y^4 = z^4$ has no nontrivial solutions. □

Euler proved Fermat's Last Theorem for $p = 3$; at the time of writing this is not in mathlib.

Lemma 2.2. *There are no solutions in positive integers to $a^3 + b^3 = c^3$.*

Proof. A proof has been formalised in Lean in the FLT-regular project here. Another proof has been formalised in Lean in the FLT3 project here by a team from the Lean For the Curious Mathematician conference held in Luminy in March 2024 (its dependency graph can be visualised here). To get this node green, this latter proof needs to be upstreamed to mathlib. This is currently work in progress by the same team. □

Corollary 2.3. *If there is a counterexample to Fermat's Last Theorem, then there is a counterexample $a^p + b^p = c^p$ with p prime and $p \geq 5$.*

Proof. Follows from the previous two lemmas. □

2.3 Frey packages

For convenience we make the following definition.

Definition 2.4. *A* Frey package (a, b, c, p) *is three pairwise-coprime nonzero integers a, b, c, with $a \equiv 3 \pmod 4$ and $b \equiv 0 \pmod 2$, and a prime $p \geq 5$, such that $a^p + b^p = c^p$.*

Our next reduction is as follows:

Lemma 2.5. *If Fermat's Last Theorem is false for $p \geq 5$ and prime, then there exists a Frey package.*

Proof. Suppose we have a counterexample $a^p + b^p = c^p$ for the given p; we now build a Frey package from this data.

If the greatest common divisor of a, b, c is d then $a^p + b^p = c^p$ implies $(a/d)^p + (b/d)^p = (c/d)^p$. Dividing through, we can thus assume that no prime divides all of a, b, c. Under this assumption we must have that a, b, c are pairwise coprime, as if some prime divides two of the integers a, b, c then by $a^p + b^p = c^p$ and unique factorization it must divide all three of them. In particular we may assume that not all of a, b, c are even, and now reducing modulo 2 shows that precisely one of them must be even.

Next we show that we can find a counterexample with b even. If a is the even one then we can just switch a and b. If c is the even one then we can replace c by $-b$ and b by $-c$ (using that p is odd).

The last thing to ensure is that a is 3 mod 4. Because b is even, we know that a is odd, so it is either 1 or 3 mod 4. If a is 3 mod 4 then we are home; if however a is 1 mod 4 we replace a, b, c by their negatives and this is the Frey package we seek. □

2.4 Galois representations and elliptic curves

To continue, we need some of the theory of elliptic curves over $\mathbb{Q}$. So let $f(X)$ denote any monic cubic polynomial with rational coefficients and whose three complex roots are distinct, and let us consider the equation $E : Y^2 = f(X)$, which defines a curve in the (X, Y) plane. This curve (or strictly speaking its projectivisation) is a so-called elliptic curve (or an elliptic curve over $\mathbb{Q}$ if we want to keep track of the field where the coefficients of $f(X)$ lie). More generally if k is any field then there is a concept of an elliptic curve over k, again defined by a (slightly more general) plane cubic curve $F(X, Y) = 0$.

If E is an elliptic curve over a field k, and if K is any field which is a k-algebra, then we write $E(K)$ for the set of solutions to $y^2 = f(x)$ with $x, y \in K$, together with an additional "point at infinity". It is an extraordinary fact, and not at all obvious, that $E(K)$ naturally has the structure of an additive abelian group, with the point at infinity being the zero element (the identity). Fortunately this fact is already in `mathlib`. We shall use $+$ to denote the group law. This group structure has the property that three distinct points $P, Q, R \in K^2$ which are in $E(K)$ will sum to zero if and only if they are collinear.

The group structure behaves well under change of field.

Lemma 2.6. *If E is an elliptic curve over a field k, and if K and L are two fields which are k-algebras, and if $f : K \to L$ is a k-algebra homomorphism, the map from $E(K)$ to $E(L)$ induced by f is an additive group homomorphism.*

Proof. The equations defining the group law are ratios of polynomials with coefficients in k, and such things behave well under k-algebra homomorphisms. □

This construction is functorial (it sends the identity to the identity, and compositions to compositions).

Lemma 2.7. *The group homomorphism $E(K) \to E(K)$ induced by the identity map $K \to K$ is the identity group homomorphism.*

Proof. An easy calculation. □

Lemma 2.8. *If $K \to L \to M$ are k-algebra homomorphisms then the group homomorphism $E(K) \to E(M)$ induced by the map $K \to M$ is the composite of the map $E(K) \to E(L)$ induced by $K \to L$ and the map $E(L) \to E(M)$ induced by the map $L \to M$.*

Proof. Another easy calculation. □

Thus if $f : K \to L$ is an isomorphism of fields, the induced map $E(K) \to E(L)$ is an isomorphism of groups, with the inverse isomorphism being the map $E(L) \to E(K)$ induced by f^{-1}. This construction thus gives us an action of the multiplicative group $\mathrm{Aut}_k(K)$ of automorphisms of the field K on the additive abelian group $E(K)$.

Definition 2.9. *If E is an elliptic curve over a field k and K is a field and a k-algebra, then the group of k-automorphisms of K acts on the additive abelian group $E(K)$.*

In particular, if $\overline{\mathbb{Q}}$ denotes an algebraic closure of the rationals (for example, the algebraic numbers in $\mathbb{C}$) and if $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ denotes the group of field isomorphisms $\overline{\mathbb{Q}} \to \overline{\mathbb{Q}}$, then for any elliptic curve E over $\mathbb{Q}$ we have an action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on the additive abelian group $E(\overline{\mathbb{Q}})$.

We need a variant of this construction where we only consider the n-torsion of the curve, for n a positive integer. Recall that if A is any additive abelian group, and if n is a positive integer, then we can consider the subgroup $A[n]$ of elements a such that $na = 0$. If a group G acts on A via additive group isomorphisms, then there will be an induced action of G on $A[n]$.

Definition 2.10. *If E is an elliptic curve over a field k and K is a field and a k-algebra, and if n is a natural number, then the group of k-automorphisms of K acts on the additive abelian group $E(K)[n]$ of n-torsion points on the curve.*

If furthermore $n = p$ is prime, then $A[p]$ is naturally a vector space over the field $\mathbb{Z}/p\mathbb{Z}$, and thus it inherits the structure of a mod p representation of G. Applying this to the above situation, we deduce that if E is an elliptic curve over $\mathbb{Q}$ then $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ acts on $E(\overline{\mathbb{Q}})[p]$ and this is the *mod p Galois representation* attached to the curve E.

In the next section we apply this theory to an elliptic curve coming from a counterexample to Fermat's Last theorem.

2.5 The Frey curve

Definition 2.11 (Frey)**.** *Given a Frey package (a, b, c, p), the corresponding* Frey curve *(considered by Frey and, before him, Hellegouarch) is the elliptic curve E defined by the equation $Y^2 = X(X - a^p)(X + b^p)$.*

Note that the roots of the cubic $X(X-a^p)(X+b^p)$ are distinct because a, b, c are nonzero and $a^p + b^p = c^p$.

Given a Frey package (a, b, c, p) with corresponding Frey curve E, the mod p Galois representation associated to this package is the representation of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on $E(\overline{\mathbb{Q}})[p]$. Frey's observation is that this mod p Galois representation has some very surprising properties. We will make this remark more explicit in the next chapter. Here we shall show how these properties can be used to finish the job.

2.6 Reduction to two big theorems.

Recall that a representation of a group G on a vector space W is said to be *irreducible* if there are precisely two G-stable subspaces of W, namely 0 and W. The representation is said to be *reducible* otherwise.

Now say Fermat's Last Theorem is false, and hence by Lemma 2.5 a Frey package (a, b, c, p) exists. Consider the mod p representation of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ coming from the p-torsion in the Frey curve $Y^2 = X(x - a^p)(X + b^p)$ associated to the package. Let's call this representation ρ. Is it reducible or irreducible?

Theorem 2.12 (Mazur)**.** *ρ cannot be reducible.*

Proof. This follows from a profound result of Mazur [9] from 1979, namely the fact that the torsion subgroup of an elliptic curve over $\mathbb{Q}$ can have size at most 16. In fact a fair amount of work still needs to be done to deduce the theorem from Mazur's result. We will have more to say about this result later. □

Theorem 2.13 (Wiles,Taylor–Wiles, Ribet,...)**.** *ρ cannot be irreducible either.*

Proof. This is the main content of Wiles' magnum opus. We omit the argument for now, although later on in this project we will have a lot to say about a proof of this. □

Corollary 2.14. *There is no Frey package.*

Proof. Follows immediately from the previous two theorems 2.12 and 2.13. □

We deduce

Corollary 2.15. *Fermat's Last Theorem is true.*

Proof. Assume there is a there is a counterexample $a^n + b^n = c^n$. By Corollary 2.3 we may assume that there is also a counterexample $a^p + b^p = c^p$ with $p \geq 5$ and prime. Then there is a Frey package (a, b, c, p) by 2.5, contradicting Corollary 2.14. □

The structure of the rest of this (highly incomplete, for now) document is as follows. In Chapter 3 we develop some of the basic theory of elliptic curves and the Galois representations attached to their p-torsion subgroups. We then apply this theory to the Frey curve, deducing in particular how Mazur's result on torsion subgroups of elliptic curves implies Theorem 2.12, the assertion that ρ cannot be reducible. In Chapter 4 we give a high-level overview of our strategy to prove that ρ cannot be irreducible, which diverges from the original approach taken by Wiles; one key difference is that we work with the p-torsion directly rather than switching to the 3-torsion. We also give a precise statement of the modularity lifting theorem which we will use. Finally, in Chapter 7 we give a collection of theorem statements which we shall need in order to push our strategy through. All of these results were known in the 1980s or before. This chapter is incoherent in the sense that it is just a big list of apparently unrelated results. As our exposition of the proof expands, the results of this chapter will slowly move to more appropriate places. The chapter is merely there to give some kind of idea of the magnitude of the project.

Chapter 3

Elliptic curves, and the Frey Curve

3.1 Overview

In the last chapter we explained how, given a counterexample to Fermat's Last Theorem, we could construct a Frey package and thus a Frey curve, which is an elliptic curve with some interesting properties. In this chapter we start with an overview of parts of the theory of the arithmetic of elliptic curves. Following this we sketch proofs of the two main results of this chapter: firstly that the p-torsion ρ in the Frey curve is "hardly ramified", and secondly that Mazur's result on the possible torsion of elliptic curves implies that ρ must be irreducible. Everything here follows from standard results about elliptic curves, however almost none of these results are in `mathlib` as I am writing this, so there is plenty to be done here.

3.2 The arithmetic of elliptic curves

We give an overview of the results we need, citing the literature for proofs. Everything here is standard, and most of it dates back to the 1970s or before.

Theorem 3.1. *Let n be a positive integer, let F be a separably closed field with n nonzero in F, and let E be an elliptic curve over F. Then the n-torsion $E(K)[n]$ in the F-points of E is a finite group of size n^2.*

Proof. There are several proofs in the textbooks. The proof being worked on uses the theory of division polynomials; the formalisation is ongoing work of David Angdinata, and it will be part of his PhD thesis. □

This theorem actually tells us the structure of the n-torsion, because of the following purely group-theoretic result:

Lemma 3.2. *Say n is a positive integer, r is a natural, and A is an abelian group. Assume that for all $d \mid n$, the d-torsion $A[d]$ of A has size d^r. Then $A[n] \cong (\mathbb{Z}/n\mathbb{Z})^r$.*

Proof. The result is obvious if $n = 1$, so we may assume $n > 1$. One proof would be to write A as $\prod_{i=1}^{t}(\mathbb{Z}/a_i\mathbb{Z})$ with $a_i \mid a_{i+1}$ (this is possible by the structure theorem for finite abelian groups), and then to apply our hypothesis firstly with $d = a_1$ to deduce $t = r$ and then with $d = a_t$ to deduce $a_1 = a_t$. □

Corollary 3.3. *Let n be a positive integer, let F be a separably closed field with n nonzero in F, and let E be an elliptic curve over F. Then the n-torsion $E(F)[n]$ in the F-points of E is a finite group isomorphic to $(\mathbb{Z}/n\mathbb{Z})^2$.*

Proof. This follows from the previous group-theoretic lemma 3.2 and theorem 3.1. □

We saw in section 2.4 that if E is an elliptic curve over a field k and if k^{sep} is a separable closure of k, then the group $\mathrm{Gal}(k^{\mathrm{sep}}/k)$ acts on $E(k^{\mathrm{sep}})[n]$. Now let n be a positive integer which is nonzero in k. We have just seen that $E(k^{\mathrm{sep}})[n]$ is isomorphic to $(\mathbb{Z}/n\mathbb{Z})^2$, and it inherits an action of $\mathrm{Gal}(k^{\mathrm{sep}}/k)$. If we fix an isomorphism $E(k^{\mathrm{sep}})[n] \cong (\mathbb{Z}/n\mathbb{Z})^2$ then we get a representation $\mathrm{Gal}(k^{\mathrm{sep}}/k) \to \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$. A fundamental fact about this Galois representation is that its determinant is the cyclotomic character.

Theorem 3.4. *If E is an elliptic curve over a field k, and n is a positive integer which is nonzero in k, then the determinant of the 2-dimensional representation of* $\mathrm{Gal}(k^{\mathrm{sep}}/k)$ *on $E(k^{\mathrm{sep}})[n]$ is the mod n cyclotomic character.*

Proof. This presumably should be done via the Weil pairing. I have not yet put any thought into a feasible way to formalise this. □

3.3 Good reduction

We give a brief overview of the theory of good and multiplicative reduction of elliptic curves. For more details one can consult the standard sources such as [14]. We stick with the low-level approach, thinking of elliptic curves as plane cubics; whilst we cannot do this forever, it will suffice for these initial results.

Definition 3.5. *Let E be an elliptic curve over the field of fractions K of a valuation ring R with maximal ideal $\mathfrak{m}$. We say E has* good reduction *over R if E has a model with coefficients in R and the reduction mod $\mathfrak{m}$ is still non-singular. If E is an elliptic curve over a number field N and P is a maximal ideal of its integer ring $\mathcal{O}_N$, then one says that E has* good reduction *at P if E has good reduction over the $\mathcal{O}_{N,P}$, the localisation of $\mathcal{O}_N$ at P.*

Remark 3.6. *From this point on, our Frey curves and Frey packages will use notation (a, b, c, ℓ), with $\ell \geq 5$ a prime number, rather than p. This frees up p for use as another prime.*

Lemma 3.7. *If E is the Frey curve $Y^2 = X(X - a^\ell)(X + b^\ell)$ associated to a Frey package (a, b, c, ℓ), and if p is a prime not dividing abc (and in particular if $p > 2$), then E has good reduction at p.*

Proof. The reduction mod p of the equation defining the Frey curve is still a smooth plane cubic, because the three roots 0, a^ℓ and $-b^\ell$ are distinct modulo p (note that the difference between a^ℓ and $-b^\ell$ is c^ℓ). □

If E is an elliptic curve over a number field N and if ρ is the representation of $\mathrm{Gal}(\overline{N}/N)$ on the n-torsion of E then ρ is continuous and its image is finite, so by the fundamental theorem of (infinite) Galois theory the representation factors through an injection $\mathrm{Gal}(L/N) \to \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$ where L/N is a finite Galois extension of number fields. One says that ρ is *unramified* at a maximal ideal P of $\mathcal{O}_N$ if the extension L/N is unramified at P (or in other words, if the factorization of $P\mathcal{O}_L$ into prime ideals is squarefree).

At some point we will need a theory of finite flat group schemes over an affine base. Here is a working definition.

Definition 3.8. *If R is a commutative ring, then a* finite flat group scheme *over R is the spectrum of a commutative Hopf algebra H/R which is finite and flat as an R-module.*

(Probably this is not the correct definition in the non-Noetherian case; one should instead ask for locally free, which is equivalent in the Noetherian case and enables you to reduce to the Noetherian case in general)

Some facts we will need are:

Theorem 3.9. *If E is an elliptic curve over a number field N and E has good reduction at a maximal ideal P of $\mathcal{O}_N$, and if furthermore $n \notin P$, then the Galois representation on the n-torsion of E is unramified.*

Proof. One approach would be by showing that the n-torsion in the integral model of E over $\mathcal{O}_{N,P}$ is an etale finite flat group scheme. There might be simpler approaches however. It's worth looking to see what Silverman does. □

Theorem 3.10. *If E is an elliptic curve over a number field N and E has good reduction at a maximal ideal P of $\mathcal{O}_N$ containing the prime number p, then the Galois representation on the p-torsion of E comes from a finite flat group scheme over the localisation $\mathcal{O}_{N,P}$.*

Proof. Indeed, the kernel of the p-torsion on a good integral model is finite and flat. Checking this claim formally will probably involve a fair amount of work. □

3.4 Multiplicative reduction

Definition 3.11. *Let E be an elliptic curve over the field of fractions K of a valuation ring R with maximal ideal $\mathfrak{m}$. We say E has* multiplicative reduction *over R if E has a model with coefficients in R and which reduces mod $R/\mathfrak{m}$ to a plane cubic with one singularity, which is an ordinary double point. We say that the reduction is* split *if the two tangent lines at the ordinary double point are both defined over $R/\mathfrak{m}$, and* non-split *otherwise.*

If E is an elliptic curve over a number field N and P is a maximal ideal of its integer ring $\mathcal{O}_N$, then one says that E has *multiplicative reduction at P* if E has multiplicative reduction over the $\mathcal{O}_{N,P}$, the localisation of $\mathcal{O}_N$ at P.

Lemma 3.12. *If E is the Frey curve $Y^2 = X(X - a^\ell)(X + b^\ell)$ associated to a Frey package (a, b, c, ℓ), and if p is an odd prime which divides abc, then E has multiplicative reduction at p.*

Proof. The hypothesis $p \mid abc$ implies that precisely two of the three roots 0, a^ℓ and $-b^\ell$ of the cubic are equal mod p. Call $x \in \mathbb{Z}/p\mathbb{Z}$ this common value. Then the reduction mod p of the curve is smooth away from the point $(x, 0)$, and has an ordinary double point at $(x, 0)$. Hence the Frey curve has multiplicative reduction at p. □

Remark 3.13. *If the third root reduces mod p to $y \neq x$, then the reduction is split multiplicative iff $x - y$ is a square mod p. We shall not need this fact.*

Lemma 3.14. *If E is the Frey curve $Y^2 = X(X - a^\ell)(X + b^\ell)$ associated to a Frey package (a, b, c, ℓ) then E has multiplicative reduction at 2.*

Proof. Indeed, the change of variables $X = 4X'$ and $Y = 8Y' + 4X'$ transforms the equation to $64Y'^2 + 64X'Y' = 64X'^3 + 16X'^2(b^\ell - a^\ell - 1) - 4X'a^\ell b^\ell$ and, because $\ell \geq 5$, b is even and $a = 3 \bmod 4$, we see that the 64s cancel, giving an equation over $\mathbb{Z}$ which reduces mod 2 to $Y'^2 + X'Y' = X'^3 + cX'^2$ for some $c \in \{0, 1\}$. This cubic is smooth away from an ordinary double point at $(0, 0)$. Hence the Frey curve has multiplicative reduction at 2. □

Remark 3.15. *Note that E has split multiplicative reduction iff $c = 0$, which happens iff $a^\ell = 7$ mod 8. We shall not need this fact.*

In particular, the Frey curve associated to a Frey package is *semistable* – it has good or multiplicative reduction at all primes.

The main thing we need about elliptic curves with multiplicative reduction over nonarchimedean local fields is the *uniformisation theorem*, originally due to Tate.

Theorem 3.16. *If E is an elliptic curve over a field complete with respect to a nontrivial nonarchimedean (real-valued) norm K and if E has split multiplicative reduction, then there is a Galois-equivariant injection $(K^{\mathrm{sep}})^\times/q^{\mathbb{Z}} \to E(K^{\mathrm{sep}})$, where $q \in K^\times$ satisfies $|q| = |j(E)|^{-1}$.*

Proof. See [13], Theorems V.3.1, Remark V.3.1.2 (we don't need surjectivity), and Theorem V.5.3. This is a lot of work and is a good target for breaking down into many smaller lemmas. □

Corollary 3.17. *If E is an elliptic curve over a field K complete with respect to a nontrivial nonarchimedean (real-valued) norm and with perfect residue field, and if E has multiplicative reduction, then there's an unramified character χ of $\mathrm{Gal}(K^{\mathrm{sep}}/K)$ whose square is 1, such that for all positive integers n with $n \neq 0$ in K, the n-torsion $E(K^{\mathrm{sep}})[n]$ is an extension of χ by $\epsilon\chi$, where ϵ is the cyclotomic character. Furthermore, the element of $K^\times/(K^\times)^\ell$ corresponding to this extension is given by the q-invariant of the curve.*

Proof. After a quadratic twist we may assume that E has split multiplicative reduction. The result then follows from the uniformisation theorem and an explicit computation. Note that even if we do not prove surjectivity of Tate's uniformisation, we still know that it's surjective on the n-torsion, because we know that there are n^2 points in the n-torsion of E over K^{sep}, and they are all accounted for by the n-torsion in $(K^{\mathrm{sep}})^\times/q^{\mathbb{Z}}$. □

3.5 Hardly ramified representations

We make the following definition; this is not in the literature but it is a useful concept for us.

Definition 3.18. *Let $\ell \geq 5$ be a prime and let V be a 2-dimensional vector space over $\mathbb{Z}/\ell\mathbb{Z}$. A representation $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}(V)$ is said to be* hardly ramified *if it satisfies the following four axioms:*

1. *$\det(\rho)$ is the mod ℓ cyclotomic character;*
2. *ρ is unramified outside 2ℓ;*
3. *The semisimplification of the restriction of ρ to $\mathrm{Gal}(\overline{\mathbb{Q}}_2/\mathbb{Q}_2)$ is unramified;*
4. *The restriction of ρ to $\mathrm{Gal}(\overline{\mathbb{Q}}_\ell/\mathbb{Q}_\ell)$ comes from a finite flat group scheme.*

We are interested in hardly ramified representations for several reasons. One is that by using some deep theorems, we will be able to prove that all hardly ramified representations are *potentially automorphic*, which will give us our first foothold into the world of modular forms. We shall come back to these ideas later. In the next section we shall be concerned with the following rather simpler result, namely that the ℓ-torsion in the Frey curve associated to a Frey package (a, b, c, ℓ) is hardly ramified. The proof is standard; for another reference, see Theorem 2.15 of [5].

3.6 The ℓ-torsion in the Frey curve is hardly ramified.

Let (a, b, c, ℓ) be a Frey package, with associated Frey curve E and mod ℓ Galois representation $\rho = E[\ell]$. We now work through a proof that ρ is hardly ramified.

Theorem 3.19. *If $p \neq \ell$ is a prime not dividing abc then ρ is unramified at p.*

Proof. Indeed, E has good reduction at p, and hence ρ is unramified at p by 3.9. □

If however p divides abc then E has multiplicative reduction at p, and we can use the theory of the Tate curve to analyse ρ at p.

Theorem 3.20. *If (a, b, c, ℓ) is a Frey package then the j-invariant of the corresponding Frey curve is $2^8(C^2 - AB)^3/A^2B^2C^2$, where $A = a^\ell$, $B = b^\ell$ and $C = c^\ell$.*

Proof. Apply the explicit formula (presumably already in mathlib) □

Corollary 3.21. *If (a, b, c, ℓ) is a Frey package and the j-invariant of the corresponding Frey curve is j, and if $2 < p \mid abc$, then the p-adic valuation $v_p(j)$ of j is a multiple of ℓ.*

Proof. Indeed p does not divide 2^8 as $p > 2$, and (using the notation of the previous theorem) p does not divide $C^2 - AB$ either, because it divides precisely one of A, B and C. Hence $v_p(j) = -2v_p(a^\ell b^\ell c^\ell) = -2\ell v_p(abc)$ is a multiple of ℓ. □

Corollary 3.22. *If (a, b, c, ℓ) is a Frey package, if $2 < p \mid abc$ is a prime with $p \neq \ell$, then the ℓ-torsion in the Frey curve is unramified at p.*

Proof. After an unramified quadratic twist we may assume the curve is split at p. The theory of the Tate curve tells us that the extension of $\mathbb{Q}_p$ cut out by the ℓ-torsion of the Frey curve is $\mathbb{Q}_p(\mu_\ell, \sqrt[\ell]{q})$, with μ_ℓ the set of ℓth roots of unity in $\overline{\mathbb{Q}}_p$. Because $\ell \neq p$ the extension $\mathbb{Q}_p(\mu_\ell)$ is unramified at p. And because $p \neq 2$ divides abc, theorem 3.21 shows us that the j-invariant of the Frey curve has p-adic valuation a multiple of ℓ. Thus the extension can be written $\mathbb{Q}_p(\mu_\ell, \sqrt[\ell]{u})$, where $u \in \mathbb{Q}_p^\times$ is a unit. The extension is hence unramified (because, for example, Hensel's Lemma shows that the ℓth root of u is in the maximal unramified extension of $\mathbb{Q}_p$). □

Corollary 3.23. *If (a, b, c, ℓ) is a Frey package, then the ℓ-torsion in the Frey curve is unramified at all primes $p \neq 2, \ell$.*

Proof. Follows from 3.19 and 3.22. □

This analysis needs to be slightly modified if $p = 2$, because the j-invariant of the Frey curve may not have 2-adic valuation a multiple of ℓ. We obtain the following weaker result.

Corollary 3.24. *If (a, b, c, ℓ) is a Frey package, then the semisimplification of the restriction of the ℓ-torsion ρ in the associated Frey curve to $\mathrm{Gal}(\overline{\mathbb{Q}}_2/\mathbb{Q}_2)$ is unramified.*

Proof. After a quadratic twist to make the curve have split multiplicative reduction, the theory of the Tate curve shows us that ρ is an extension of the trivial character by the cyclotomic character. Hence the semisimplification of this representation is the direct sum of two unramified characters and is hence unramified. □

Theorem 3.25. *Let ρ be the ℓ-torsion in the Frey curve associated to a Frey package (a, b, c, ℓ). Then the restriction of ρ to $\mathrm{Gal}(\overline{\mathbb{Q}}_\ell/\mathbb{Q}_\ell)$ comes from a finite flat group scheme.*

Proof. The Frey curve either has good reduction at ℓ (case 1 of FLT) or multiplicative reduction at ℓ (case 2 of FLT). In the first case the ℓ-torsion is finite and flat at ℓ by theorem 3.10. In the second case the theory of the Tate curve shows that the ℓ-torsion is (up to quadratic twist) an extension of the trivial character by the cyclotomic character corresponding (via Hilbert 90) to the ℓth power of an ℓ-adic unit. This extension is known to be finite and flat; see for example Proposition 8.2 of [6]. Note that the proof in [6] uses fppf cohomology, although one can write down a much more elementary proof of this using arguments in [8]. □

We have now proved the first main result of this chapter.

Theorem 3.26. *Let ρ be the Galois representation on the ℓ-torsion of the Frey curve coming from a Frey package (a, b, c, ℓ). Then ρ is hardly ramified.*

Proof. This follows from the results above. The fact that $\ell \geq 5$ follows from the definition of a Frey package. The first condition is theorem 3.4, and the second is theorem 3.23. The third condition is theorem 3.24, and the fourth is theorem 3.25. □

3.7 The ℓ-torsion in the Frey curve is irreducible.

We finish this chapter by showing that Mazur's theorem implies that the ℓ-torsion in the Frey curve is irreducible. We start by stating Mazur's theorem.

Theorem 3.27. *Let E be an elliptic curve over $\mathbb{Q}$. Then the torsion subgroup of E has size at most 16.*

Proof. This is the main theorem of [9]. Formalising this result will be a highly non-trivial project; note that this theorem is used in all known proofs of FLT, so there seems to be no way around it. □

Let (a, b, c, ℓ) be a Frey package, with associated Frey curve E and mod ℓ Galois representation $\rho = E[\ell]$. We know that ρ is 2-dimensional; let's suppose for a contradiction that that it is reducible, so in particular its semisimplification is the direct sum of two characters α and β.

The next two results are Lemme 6 on p307 of [11].

Theorem 3.28. *With notation as above, the characters α and β are unramified at p for all primes $p \neq \ell$.*

Proof. We have seen in theorem 3.23 that ρ is unramified at all primes $p \neq 2, \ell$, so the characters α and β are unramified at all such primes. If $p = 2$ then the semisimplification of the restriction of ρ to $\mathrm{Gal}(\overline{\mathbb{Q}}_2/\mathbb{Q}_2)$ is unramified by corollary 3.24, so α and β are unramified at 2. □

Remark 3.29. *Does this innocuous-looking proof above use some form of the Brauer-Nesbitt theorem?*

Theorem 3.30. *One of α and β is unramified at ℓ.*

Proof. In the multiplicative case this follows immediately from the theory of the Tate curve. In the good reduction case, the ℓ-torsion is finite and flat at ℓ by theorem 3.25, so we now need to understand what such representations look like. If the reduction is supersingular, then ρ is necessarily irreducible, contradicting our assumption. If however the reduction is ordinary, then the theory of the canonical subgroup shows that the ℓ-torsion is an extension of an unramified character by an unramified twist of the cyclotomic character (see Proposition 11 on p273 of [11]). □

Corollary 3.31. *One of α and β is trivial.*

Proof. The previous two theorems show that one of α and β is a character unramified at all primes, and hence cuts out an extension unramified at all primes, so by Minkowski's theorem this character is trivial. □

To summarise, we have shown the following.

Theorem 3.32. *If ρ is reducible, then either ρ has a trivial 1-dimensional submodule or a trivial 1-dimensional quotient (here "trivial" means that the Galois group* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ *acts trivially).*

Proof. Follows from the above. □

We now split into two cases, depending on whether ρ has a trivial submodule or a trivial quotient.

Lemma 3.33. *If ρ has a trivial 1-dimensional submodule then the Frey curve has a non-trivial point of order ℓ.*

Proof. Indeed, the trivial 1-dimensional submodule is a Galois-invariant subgroup of $E[\ell]$, so it corresponds to a Galois-stable point of order ℓ. □

Corollary 3.34. *ρ cannot have a trivial 1-dimensional submodule.*

Proof. We have just seen that in this case, the Frey curve has a point of order ℓ. It also has three points of order 2, meaning that its torsion subgroup has order at least $4\ell \geq 20$, contradicting Mazur's theorem 3.27. □

It remains to rule out the case where ρ is reducible and has a trivial quotient. To do this, we need to quotient out ρ by its 1-dimensional Galois-stable submodule.

Theorem 3.35. *If p is a prime and if E is an elliptic curve over a field K of characteristic not equal to p, and if $C \subseteq E(K^{\mathrm{sep}})[p]$ is a Galois-stable subgroup of order p, then there's an elliptic curve $E' :=$"E/C" over K and an isogeny of elliptic curves $E \to E'$ over K inducing a Galois-equivariant surjection $E(K^{\mathrm{sep}}) \to E'(K^{\mathrm{sep}})$ with kernel precisely C.*

Proof. Brian Conrad suggested the following approach, applicable as well for abelian schemes $A \to S$ over a base. Let G be a finite locally free S-subgroup of A, say G with constant rank $n > 0$ by working locally on the base, so G is contained in $A[n]$. Then $n : A \to A$ is the fppf quotient of the source by $A[n]$, so it expresses A as an $A[n]$-torsor over itself. The problem of building A/G as an abelian scheme is then seen to be the "same" as that of constructing the quotient of this $A[n]$-torsor by the G-action.

In other words, the problem them becomes one having nothing specific to do with abelian schemes, at the cost of working over a base (such as the original target A) even when S was the spectrum of a field in the application. The question is now: for a finite locally free commutative S-group H and a closed locally free S-subgroup G, build a reasonable quotient H/G. One approach is to look at the Cartier dual $H^\vee \to G^\vee$, show that it's faithfully flat, and then deduce that the Cartier dual of the kernel of this map does the job. Note that one input for this proof is the claim that inclusions of Hopf algebras over fields are flat, proved nicely in Waterhouse's book. □

I suspect that the proof above is a no-go right now; there will presumably be a much easier proof of this result in Silverman though. Note also that this approach will not give us a plane cubic, but rather a smooth proper group scheme so we would need Riemann-Roch to turn it into a plane cubic, although it's unlikely that one will be able to prove Mazur's theorem without developing all of this machinery and much more.

Corollary 3.36. *ρ cannot have a trivial 1-dimensional quotient.*

Proof. ρ has a Galois-stable submodule C. The quotient curve E/C now has a trivial 1-dimensional submodule, and also three points of order 2 (the images of the three 2-torsion points in E). Hence the torsion subgroup of E/C has order at least $4\ell \geq 20$, again contradicting Mazur's theorem. □

Theorem 3.37. *The ℓ-torsion in the Frey curve associated to a Frey package (a, b, c, ℓ) is irreducible.*

Proof. Follows from theorem 3.32, corollary 3.34 and corollary 3.36. □

Chapter 4

An overview of the proof

So far we have seen that, modulo Mazur's theorem (and various other things which will still take some work to formalise but which are much easier), Fermat's Last Theorem can be reduced to the statement that there is no prime $\ell \geq 5$ and hardly-ramified irreducible 2-dimensional Galois representation $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$.

In this chapter we give an overview of our strategy for proving this, and collect various results which we will need along the way. Note that we no longer need to assume that ρ comes from the ℓ-torsion in an elliptic curve.

4.1 Potential modularity.

We will only speak about modularity for 2-dimensional representations of the absolute Galois group of a totally real field F of even degree over $\mathbb{Q}$, and in particular we will never say that a representation of the absolute Galois group of $\mathbb{Q}$ is modular! What we will mean by "modular" is "associated to an automorphic representation of the units of the totally definite quaternion algebra over F ramified at no finite places". We can furthermore even demand that the infinity type is trivial, as these are the only forms we shall need for FLT.

Assume we have a hardly-ramified representation ρ as above. Let K be the number field corresponding to the kernel of ρ. Our first claim is that there is some totally real field F of even degree, Galois over $\mathbb{Q}$, unramified at ℓ, and disjoint from K, such that $\rho|G_F$ is modular. The proof of this is very long, and uses a host of machinery. For example:

- Moret–Bailly's result [10] on points on curves with prescribed local behaviour;
- several nontrivial results in global class field theory;
- the Jacquet–Langlands correspondence;
- The assertion that irreducible 2-dimensional mod p representations induced from a character are modular (this follows from converse theorems);
- A modularity lifting theorem.

Almost everything here dates back to the the 1980s or before. The exception is the modularity lifting theorem, which we now state explicitly.

4.2 A modularity lifting theorem

Suppose $\ell \geq 5$ is a prime, that F is a totally real field of even degree in which ℓ is unramified, and that S is a finite set of finite places of F not dividing ℓ. Write G_F for the absolute Galois group of F.

If $v \in S$ then let F_v denote the completion of F at v, fix an inclusion $\overline{F} \to \overline{F_v}$, let $\mathcal{O}_v$ denote the integers of F_v and $k(v)$ the residue field. Let $I_v \subset G_F$ denote the inertia subgroup at v. Local class field theory (or a more elementary approach) gives a map $I_v \to \mathcal{O}_{F_v}^\times$ and hence a map $I_v \to k(v)^\times$. Let J_v denote the kernel of this map.

Let R be a complete local Noetherian $\mathbb{Z}_\ell$-algebra with finite residue field of characteristic ℓ. We will be interested in representations $\rho : G_F \to \mathrm{GL}_2(R)$ with the following four properties.

- $\det(\rho)$ is the cyclotomic character;
- ρ is unramified outside $S \cup \{\ell\}$;
- If $v \in S$ then $\rho(g)$ has trace equal to 2 for all $g \in J_v$;
- If $v \mid \ell$ is a place of F then ρ is flat at v.

In the last bullet point, “flat” means “projective limit of representations arising from finite flat group schemes”. Let us use the lousy temporary notation “S-good” to denote representations with these four properties.

Say k is a finite extension of $\mathbb{Z}/\ell\mathbb{Z}$ and $\overline{\rho} : G_F \to \mathrm{GL}_2(k)$ is continuous, absolutely irreducible when restricted to $F(\zeta_\ell)$, and S-good. One can check that the functor representing S-good lifts of $\overline{\rho}$ is representable.

Theorem 4.1. *If $\overline{\rho}$ is modular of level $\Gamma_1(S)$ and $\rho : G_F \to \mathrm{GL}_2(\mathcal{O})$ is an S-good lift of $\overline{\rho}$ to $\mathcal{O}$, the integers of a finite extension of $\mathbb{Q}_\ell$, then ρ is also modular of level $\Gamma_1(S)$.*

Right now we are very far from even stating this theorem in Lean.

I am not entirely sure where to find a proof of this in the literature, although it has certainly been known to the experts for some time. Theorem 3.3 of [15] comes close, although it assumes that ℓ is totally split in F rather than just unramified. Another near-reference is Theorem 5.2 of [7], although this assumes the slightly stronger assumption that the image of ρ contains $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ (however it is well-known to the experts that this can be weakened to give the result we need). One reference for the proof is Richard Taylor’s 2018 Stanford course.

Proof. (Sketch)

The proof is a two-stage procedure and has a nontrivial analytic input. First one uses the Skinner–Wiles trick to reduce to the “minimal case”, and this needs cyclic base change for GL(2) and also a characterisation of the image of the base change construction; this seems to need a multiplicity one result, which (because of our definition of “modular”) will need Jacquet–Langlands as well.

In the minimal case, the argument is the usual Taylor–Wiles trick, using refinements due to Kisin and others. □

Given this modularity lifting theorem, the strategy to show potential modularity of ρ is to use Moret–Bailly to find an appropriate totally real field F, an auxiliary prime p, and an auxiliary elliptic curve over F whose mod ℓ Galois representation is ρ and whose mod p

Galois representation is induced from a character. By converse theorems (for example) the mod p Galois representation is associated to an automorphic representation of GL_2/F and hence by Jacquet–Langlands it is modular. Now we use the modularity lifting theorem to deduce the modularity of the curve over F and hence the modularity of the ℓ-torsion.

4.3 Compatible families, and reduction at 3

We now use Khare–Wintenberger to lift ρ to a potentially modular ℓ-adic Galois representation of conductor 2, and put it into an ℓ-adic family using the Brauer's theorem trick in [1]. Finally we look at the 3-adic specialisation of this family. Reducing mod 3 we get a representation which is flat at 3 and tame at 2, so must be reducible because of the techniques introduced in Fontaine's paper on abelian varieties over $\mathbb{Z}$ (an irreducible representation would cut out a number field whose discriminant violates the Odlyzko bounds). One can now go on to deduce that the 3-adic representation must be reducible, which contradicts the irreducibility of ρ.

We apologise for the sketchiness of what is here, however at the time of writing it is so far from what we are even able to *state* in Lean that there seems to be little point right now in fleshing out the argument further. As this document grows, we will add a far more detailed discussion of what is going on here. Note in particular that stating the modularity lifting theorem in Lean is the first target.

Chapter 5

An example of an automorphic form

5.1 Introduction

The key ingredient in Wiles' proof of Fermat's Last Theorem is a *modularity lifting theorem*, sometimes called an $R = T$ theorem. For Wiles, the R came from elliptic curves, the T came from classical modular forms, and the fact that they're equal is basically the Shimura–Taniyama–Weil conjecture, now known as the Breuil–Conrad–Diamond–Taylor modularity theorem: any elliptic curve over the rationals is modular.

At the heart of the proof we shall formalise is also an $R = T$ theorem, however the T which we shall use will be associated not to classical modular forms, but to spaces of more general automorphic forms called quaternionic modular forms. Those of you who know something about classical modular forms might well know that the groups $\mathrm{SL}_2(\mathbb{R})$ and $\mathrm{SL}_2(\mathbb{Z})$ are intimately involved; these are the norm 1 units in the matrix rings $M_2(\mathbb{R})$ and $M_2(\mathbb{Z})$. In the theory of quaternionic modular forms, the analogous groups are the norm 1 units in rings such as Hamilton's quaternions $\mathbb{R} \oplus \mathbb{R}i \oplus \mathbb{R}j \oplus \mathbb{R}k$, and subrings such as $\mathbb{Z} \oplus \mathbb{Z}i \oplus \mathbb{Z}j \oplus \mathbb{Z}k$.

One of the main goals of the FLT project at the time of writing this sentence, is formalising the *statement* of the modularity lifting theorem which we shall use. So we are going to need to develop the theory of quaternionic modular forms, which is rather different to the theory of classical modular forms (for example, in the cases we need, the definition is completely algebraic; there are no holomorphic functions in sight, and the analogue of the upper half plane in the quaternionic theory is a finite set of points).

We could just launch into the general theory over totally real fields, which will be the generality which we'll need. But when I was a PhD student, I learnt about these objects by playing with explicit examples. So, whilst not logically necessary for the proof, I thought it would be fun, and perhaps also instructional, to compute a concrete example of a space of quaternionic modular forms. The process of constructing the example might even inform what kind of machinery we should be developing in general. Let's begin by discussing the quaternion algebra we shall use.

5.2 A quaternion algebra

Let's define D to be the quaternion algebra $\mathbb{Q} \oplus \mathbb{Q}i \oplus \mathbb{Q}j \oplus \mathbb{Q}k$. As a vector space, D is 4-dimensional over $\mathbb{Q}$ with $[1, i, j, k]$ giving a basis. It has a (non-commutative)ring structure, with multiplication satisfying the usual quaternion algebra relations $i^2 = j^2 = k^2 = ijk = -1$. You can think of D as an analogue of 2×2 matrices with rational coefficients, hence its units $D^\times$ are an analogue of the group $\mathrm{GL}_2(\mathbb{Q})$.

We will also need an analogue of the group $\mathrm{GL}_2(\mathbb{Z})$, which will come from an integral structure on D. We choose the Hurwitz order, namely the subring $\mathcal{O} := \mathbb{Z} \oplus \mathbb{Z}i \oplus \mathbb{Z}j \oplus \mathbb{Z}\omega$, where $\omega = \frac{-1+(i+j+k)}{2}$, a cube root of unity, as $(i + j + k)^2 = -3$. The simplest way to understand $\mathcal{O}$ is that it's quaternions $a + bi + cj + dk$ where either a, b, c, d are all integers or are all in $\frac{1}{2} + \mathbb{Z}$.

Note that $\mathcal{O}$ is a maximal order and a Euclidean domain, which is why we prefer it over the more obvious sublattice $\mathbb{Z} \oplus \mathbb{Z}i \oplus \mathbb{Z}j \oplus \mathbb{Z}k$.

In this chapter, we are going to compute a complex vector space which could be called something like the "weight 2 level 2 modular forms for $D^\times$". The main result will be that this space is 1-dimensional.

Note that mathlib has modular forms, but it doesn't have enough complex analysis to deduce that the space of modular forms of a given weight and level is finite-dimensional. If all the 'sorry's in this chapter are completed before mathlib gets the necessary complex analysis, then the first nonzero space of modular forms to be proved finite-dimensional in Lean will be a space of quaternionic modular forms.

We will use a modern "adelic" definition of our modular forms, so the first thing we need to do is to talk about profinite completions.

5.3 $\widehat{\mathbb{Z}}$

Classically automorphic forms were defined as functions on symmetric spaces (like the upper half plane) which transformed well under the action of certain discrete groups (for example $\mathrm{SL}_2(\mathbb{Z})$). However such definitions became combinatorially problematic when generalised to number fields with nontrivial class group, because the classical theory needed a *number* p to define the Hecke operator T_p, and in the case where p was a non-principal prime ideal in a number field, there was no appropriate number. One fix is to take disjoint unions of symmetric spaces indexed by the ideal class group of the field in question, but it is easier to work adelically, which is morally what we shall do. However we are able to avoid introducing the adeles explicitly; we can work instead with the conceptually simpler object $\widehat{\mathbb{Z}}$, the profinite completion of $\mathbb{Z}$. So what is $\widehat{\mathbb{Z}}$? We offer a low-level definition of this object.

Given an integer z, we can reduce it mod N for every positive natural number and get elements $z_N = \overline{z} \in \mathbb{Z}/N\mathbb{Z}$. These elements are not completely arbitrary though – they must satisfy some compatibility conditions. For example there can be no positive integer z such that $z_{10} = 6$ and $z_2 = 1$, because $z_{10} = 6$ tells us that z ends in a 6 when written in base 10, and in particular it's even, so z_2 must be 0. The general rule: if $D \mid N$ then z_D must be equal to image of z_N under the natural ring homomorphism from $\mathbb{Z}/N\mathbb{Z}$ to $\mathbb{Z}/D\mathbb{Z}$. We say that a collection of elements $z_N \in \mathbb{Z}/N\mathbb{Z}$ is *compatible* if it satisfies this rule.

Definition 5.1. *The profinite completion $\widehat{\mathbb{Z}}$ of $\mathbb{Z}$ is the set of all compatible collections $c = (c_N)_N$ of elements of $\mathbb{Z}/N\mathbb{Z}$ indexed by $\mathbb{N}^+ := \{1, 2, 3, ...\}$. A collection is said to be* compatible *if for all positive integers $D \mid N$, we have c_N mod D equals c_D.*

Lemma 5.2. *$\widehat{\mathbb{Z}}$ is a subring of $\prod_{N\geq 1}(Z/N\mathbb{Z})$ and in particular is a ring.*

Proof. Follow your nose. □

Examples of elements of $\widehat{\mathbb{Z}}$ are are given by integers, where we define z_N to be z mod N for all N. This gives us a natural map from $\mathbb{Z}$ to $\widehat{\mathbb{Z}}$. In particular we can talk about $0 \in \widehat{\mathbb{Z}}$ and $1 \in \widehat{\mathbb{Z}}$.

Lemma 5.3. $0 \neq 1$ *in* $\widehat{\mathbb{Z}}$.

Proof. Recall that you can evaluate an element of $\widehat{\mathbb{Z}}$ at a positive integer. Evaluating 0 at 2 gives 0, and evaluating 1 at 2 gives 1, and these are distinct elements of $\mathbb{Z}/2\mathbb{Z}$, so $0 \neq 1$ in $\widehat{\mathbb{Z}}$. □

Lemma 5.4. *The map from the naturals into $\widehat{\mathbb{Z}}$ sending n to n is injective.*

Proof. Generalise the above idea. Feel free to write up a LaTeX proof and PR it. □

Note that it follows easily that that the map from the integers to $\widehat{\mathbb{Z}}$ is injective.

But $\widehat{\mathbb{Z}}$ is *much* larger than $\mathbb{Z}$; it has the same cardinality as the reals in fact. Let's write down an explicit example of an element of $\widehat{\mathbb{Z}}$ which isn't obviously in $\mathbb{Z}$.

Definition 5.5. *The infinite sum $0! + 1! + 2! + 3! + 4! + 5! + \cdots$ looks like it makes no sense at all; it is the sum of an infinite series of larger and larger positive numbers. However, the sum is* finite *modulo N for every positive integer N, because all the terms from $N!$ onwards are multiples of N and thus are zero in $\mathbb{Z}/N\mathbb{Z}$. Thus it makes sense to define e_N to be the value of the finite sum modulo N. Explicitly, $e_N = 0! + 1! + \cdots + (N-1)!$ modulo N.*

Lemma 5.6. *The collection $(e_N)_N$ is an element of $\widehat{\mathbb{Z}}$.*

Proof. This boils down to checking that $D! + (D+1)! + \cdots + (N-1)!$ is a multiple of D. □

Lemma 5.7. *The element $(e_N)_N$ of $\widehat{\mathbb{Z}}$ is not in $\mathbb{Z}$.*

Proof. First imagine that $e = n$ with $n \in \mathbb{Z}$ and $0 \leq n$. In this case, choose j such that $0! + 1! + 2! + \cdots + j! > n$ and check also that the sum is less than $(j+1)!$. Now set $N = (j+1)!$ and let's compare e_N and $n_N = n$. The trick is that e_N must be $0! + 1! + \cdots + j!$ mod N, because all the terms beyond this are multiples not just of $(j+1)$ but of $(j+1)! = N$. Thus mod N we have $0 \leq n < e_N < N$ so $n \neq e$.

Now we deal with $n = -t < 0$; choose j large such that $(j+1)! - (0! + 1! + \cdots + j!) > t$ (possible because the sum is at most $2 \times j!$) and then set $N = (j+1)!$ and we have $0 < e_N < N - t < N$ so we cannot have $e_N = -t$ in $\mathbb{Z}/N\mathbb{Z}$, so again $e \neq n$. □

Let's prove some more basic facts about $\widehat{\mathbb{Z}}$.

Lemma 5.8. *If $0 < N$ is an integer then multiplication by N is injective on $\widehat{\mathbb{Z}}$.*

Proof. Suppose that $(z_i)_i \in \widehat{\mathbb{Z}}$ and $Nz = 0$. This means that $Nz_i = 0 \in \mathbb{Z}/i\mathbb{Z}$ for all i. Let us fix an arbitrary positive integer j; we need to prove that $z_j = 0 \in \mathbb{Z}/j\mathbb{Z}$. Consider the element $z_{Nj} \in \mathbb{Z}/Nj\mathbb{Z}$. By assumption, we have $Nz_{Nj} = 0$, meaning that if we lift z_{Nj} to an integer, we have $Nj \mid Nz_{Nj}$, and thus $j \mid z_{Nj}$. Thus by the compatibility assumption on the z_i we have that $z_j \in \mathbb{Z}/j\mathbb{Z}$ is the mod j reduction of z_{Nj} and hence is zero. □

We will also need to understand exactly which elements of $\widehat{\mathbb{Z}}$ are multiples of N.

Lemma 5.9. *The multiples of N in $\widehat{\mathbb{Z}}$ are precisely the compatible collections $(z_i)_i \in \widehat{\mathbb{Z}}$ with $z_N = 0$.*

Proof. Clearly $z_N = 0$ is a necessary condition to be a multiple of N. To see it is sufficient, take a general $(z_i) \in \widehat{\mathbb{Z}}$ such that $z_N = 0$, and now define a new element $(y_j)_j$ of $\widehat{\mathbb{Z}}$ by $y_j = z_{Nj}/N$. Just to clarify what this means: $z_{Nj} \in \mathbb{Z}/Nj\mathbb{Z}$ reduces mod N to $z_N = 0$ by the compatibility assumption, so it is in the subgroup $N\mathbb{Z}/Nj\mathbb{Z}$ of $\mathbb{Z}/Nj\mathbb{Z}$, which is isomorphic (via "division by N") to the group $\mathbb{Z}/j\mathbb{Z}$; this is how we construct y_j. It is easily checked that the y_j are compatible and that $Ny = z$. □

5.4 More advanced remarks on $\widehat{\mathbb{Z}}$ versus $\mathbb{Q}$

This section can be skipped on first reading.

People who have seen some more advanced algebra might recognise this construction of $\widehat{\mathbb{Z}}$ as being the profinite completion of the additive abelian group $\mathbb{Z}$, so it is a fundamental object of mathematics in some sense. But usually, when building mathematics, after $\mathbb{Z}$ we go to $\mathbb{Q}$, a multiplicative localisation of $\mathbb{Z}$, and only complete after that (to get $\mathbb{R}$). The process of "completing before localising" gives us a far more arithmetic completion of $\mathbb{Z}$.

Even though $\mathbb{Q}$ is a divisible abelian group and hence its profinite completion vanishes, we can still attempt to "locally profinitely complete it" by defining $\widehat{\mathbb{Q}} := \mathbb{Q} \otimes_{\mathbb{Z}} \widehat{\mathbb{Z}}$. This object is more commonly known as the *finite adeles* of $\mathbb{Q}$. More generally if F is any number field then $F \otimes_{\mathbb{Z}} \widehat{\mathbb{Z}}$ is the ring of finite adeles of F. To get to the full ring of adeles of a number field F you need to take the product with the ring of infinite adeles of F, which is $F \otimes_{\mathbb{Q}} \mathbb{R}$: some kind of universal archimedean completion of F. I don't know a reference which develops the theory of adeles in this way, so this is what we shall do here.

5.5 $\widehat{\mathbb{Q}}$ and tensor products.

The definition of $\widehat{\mathbb{Q}}$ is easy if you know about tensor products of additive abelian groups.

Definition 5.10. *The profinite completion $\widehat{\mathbb{Q}}$ of $\mathbb{Q}$ is the tensor product $\mathbb{Q} \otimes_{\mathbb{Z}} \widehat{\mathbb{Z}}$, or $\widehat{\mathbb{Q}} = \mathbb{Q} \otimes \widehat{\mathbb{Z}}$ for short.*

5.6 A crash course in tensor products

We've defined $\widehat{\mathbb{Q}}$ to be $\mathbb{Q} \otimes \widehat{\mathbb{Z}}$. Whatever does this mean? Well just to orient yourself, if A and B are additive abelian groups, then $A \otimes B$ is also an abelian group. And if A and B are commutative rings (as they are in our case), then $A \otimes B$ is also a commutative ring.

Even if A and B are completely concrete commutative rings, their tensor product $A \otimes B$ might be incomprehensible. For example $\mathbb{C} \otimes \mathbb{C}$ is completely incomprehensible (note that we are tensoring over the integers). It is not like the product of groups or the disjoint union of two sets, where you have a completely explicit unambiguous formula for each element.

In this sense, the theory of tensor products is a bit like the theory of continuous functions. Humanity started off studying concrete polynomial equations such as $x^2 + 1$ and then moved on to concrete analytic functions such as $\log(x)$ and $\sin(x)$, but eventually the abstract concept of a continuous function from the reals to the reals was born. There is no "formula" for a general continuous function, and continuous functions such as e^{-1/x^2} or $|x|$ have no power series. Even if there *were* a formula for a specific continuous function of interest, it is not clear in general how to make sense of the claim that it's the "best" formula. In other words, there is no "canonical form" for a general continuous function, and yet we prove things about them anyway. We shall adopt the same attitude for elements of $A \otimes B$.

The first thing to know about the tensor product $A \otimes B$ of two abelian groups A and B is a "constructor" for the type. In other words, how can we make elements $A \otimes B$? Well, it turns out that given elements $a \in A$ and $b \in B$, we can form the element $a \otimes_t b \in A \otimes B$.

Example 5.11. *Recall that the sum of all the factorials is an element $e \in \widehat{\mathbb{Z}}$, and $22/7$ is certainly a rational number, so we can make the element $\frac{22}{7} \otimes_t e \in \widehat{\mathbb{Q}}$.*

This example is in the Lean code.

Elements of the form $a \otimes_t b \in A \otimes B$ are known as *pure tensors*. In the literature, pure tensors are often written $a \otimes b$, but we shall follow `mathlib`'s convention in reserving the $\otimes$ symbol for *groups* like $A \otimes B$, and adorning it with a t when using it on *elements* of the groups (or, as Lean calls them, *terms*, which explains the notation).

Addition of pure tensors obeys the "distributivity" rules $a \otimes_t b_1 + a \otimes_t b_2 = a \otimes_t (b_1 + b_2)$ and $a_1 \otimes_t b + a_2 \otimes_t b = (a_1 + a_2) \otimes_t b$, but there is no rule which simplifies a general sum $a \otimes_t b + c \otimes_t d$ into a pure tensor. Indeed, in general it is *not* the case that every element of a tensor product $A \otimes B$ is of the form $a \otimes_t b$; there can be tensors which aren't pure. However every element of $A \otimes B$ is a finite sum of pure tensors, with the result that one can attempt to define additive maps from $A \otimes B$ by saying what they do on pure tensors, and then extending linearly.

Another thing worth understanding is that just like how rational numbers can be written as quotients of integers in several ways (for example $1/2 = 2/4 = 3/6 = \cdots$), a general pure tensor in $A \otimes B$ can be represented as $a \otimes_t b$ in many ways. For example, in $\widehat{\mathbb{Q}}$ we have $1 \otimes_t 2 = 2 \otimes_t 1$. A general rule for equality of pure tensors is that if $a \in A$ and $b \in B$ and $z \in \mathbb{Z}$, then $za \otimes_t b = a \otimes_t zb$; integers can move over the tensor symbol. But equality is hard: in general there may not be an algorithm to decide whether two pure tensors $a \otimes_t b$ and $c \otimes_t d$ are equal in $A \otimes B$.

Remark 5.12. *A summary of the situation: if A and B are abelian groups, then every element of $A \otimes B$ can be written in the form $\sum_{i=1}^{N} a_i \otimes_t b_i$. It's just that this representation is highly nonunique, and furthermore given explicit elements $a_1, a_2 \in A$ and $b_1, b_2 \in B$ it might be a hard problem to figure out if $a_1 \otimes_t b_1 = a_2 \otimes_t b_2$.*

For example, it turns out that $(\mathbb{Z}/2\mathbb{Z}) \otimes (\mathbb{Z}/3\mathbb{Z}) = 0$ and so in this tensor product all the $a \otimes_t b$ are equal to each other and to $0 \otimes 0$.

Having said all of that, one nice property of $\widehat{\mathbb{Q}}$ is that every tensor *is* pure; let's prove this now.

Lemma 5.13. *Every element of $\widehat{\mathbb{Q}} := \mathbb{Q} \otimes \widehat{\mathbb{Z}}$ can be written as $q \otimes_t z$ with $q \in \mathbb{Q}$ and $z \in \widehat{\mathbb{Z}}$. Furthermore one can even assume that $q = \frac{1}{N}$ for some positive integer N.*

Proof. A proof I would write on the board would look like the following. Take a general element of $\widehat{\mathbb{Q}}$; we know it can be expressed as a finite sum $\sum_i q_i \otimes_t z_i$ with $q_i \in \mathbb{Q}$ and $z_i \in \widehat{\mathbb{Z}}$. Now choose a large positive integer N, the lowest common multiple of all the denominators showing up in the q_i, and then rewrite $\sum_i q_i \otimes_t z_i$ as $\sum_i \frac{n_i}{N} \otimes z_i$ with $n_i \in \mathbb{Z}$. Now using the fundamental fact that $na \otimes_t b = a \otimes_t nb$ for $n \in \mathbb{Z}$, we can rewrite the sum as $\sum_i \frac{1}{N} \otimes_t n_i z_i$ which is equal to the pure tensor $\frac{1}{N} \otimes (\sum_i n_i z_i)$.

In Lean I would prove this using `TensorProduct.induction_on`, which quickly reduces us to the claim that the sum of two pure tensors is pure, which we can prove using the above technique whilst avoiding the general theory of finite sums. □

Be careful though: just because every element of $\widehat{\mathbb{Q}}$ can be written as $q \otimes z$, this representation may not be unique. For example $2 \otimes 1 = 1 \otimes 2$. However, writing $\frac{1}{N} \otimes_t z$ as z/N does tempt us into the following definition.

Definition 5.14. *If $N \in \mathbb{N}^+$ and $z \in \widehat{\mathbb{Z}}$ then we say that N and z are* coprime *if $z_N \in (\mathbb{Z}/N\mathbb{Z})^\times$. We write z/N as notation for the element $\frac{1}{N} \otimes_t z$.*

Lemma 5.15. *Every element of $\widehat{\mathbb{Q}}$ can be uniquely written as z/N with $z \in \widehat{\mathbb{Z}}$, $N \in \mathbb{N}^+$, and with N and z coprime.*

Proof. Existence: by the previous lemma, an arbitrary element can be written as z/N; let D be the greatest common divisor of N and z_N (lifted to a natural). If $D = 1$ then the fraction is by definition in lowest terms. However if $1 < D \mid N$ then z_D is the reduction of z_N and is hence 0. By lemma 5.9 we deduce that $z = Dy$ is a multiple of D, and hence $z/N = \frac{1}{N} \otimes_t Dy = \frac{1}{E} \otimes y$, where $E = N/D$. Now if a natural divided both y_E and E then this natural would divide both z_N/D and N/D, contradicting the fact that D is the greatest common divisors.

Uniqueness: if $z/N = w/M$, we deduce $1 \otimes_t Mz = 1 \otimes_t Nw$, and by injectivity of $\widehat{\mathbb{Z}} \to \widehat{\mathbb{Q}}$ we deduce that $Mz = Nw = y$. In particular, if L is the lowest common multiple of M and N then y_L is a multiple of both M and N and is hence zero, so $y = Lx$ is a multiple of L by 5.9, and we deduce from torsionfreeness that $z = (L/M)x$ and $w = (L/N)x$. If some prime divided L/M then it would have to divide N which means that z is not in lowest terms; similarly if some prime divided L/N then w/M would not be in lowest terms. We deduce that $L = M = N$ and hence $z = w$ by torsionfreeness. □

If A and B are additive abelian groups then $A \otimes B$ is also an additive abelian group. However if A and B are commutative rings, then $A \otimes B$ also inherits the structure of a commutative ring, with $0 = 0 \otimes_t 0$ and $1 = 1 \otimes_t 1$. Pure tensors multiply in the obvious way: the product of $a_1 \otimes_t b_1$ and $a_2 \otimes_t b_2$ is $a_1 a_2 \otimes_t b_1 b_2$. There are ring homomorphisms $A \to A \otimes B$ and $B \to A \otimes B$ sending a to $a \otimes_t 1$ and b to $1 \otimes_t b$. In general such maps are not injective, but in the case of $\widehat{\mathbb{Q}} = \mathbb{Q} \otimes \widehat{\mathbb{Z}}$ both maps from $\mathbb{Q}$ and $\widehat{\mathbb{Z}}$ are inclusions.

Lemma 5.16. *The ring homomorphism $\mathbb{Q} \to \widehat{\mathbb{Q}}$ sending q to $q \otimes_t 1$ is injective.*

Proof. We have seen that the map from $\mathbb{Z}$ to $\widehat{\mathbb{Z}}$ is injective. Now $\mathbb{Q}$ is a flat $\mathbb{Z}$-module, because it's torsion-free, so tensoring up we deduce that the map from $\mathbb{Q} = \mathbb{Q} \otimes \mathbb{Z}$ to $\widehat{\mathbb{Q}} = \mathbb{Q} \otimes \widehat{\mathbb{Z}}$ is also injective. There is no doubt a more elementary proof of this fact. □

Lemma 5.17. *The ring homomorphism $\widehat{\mathbb{Z}} \to \widehat{\mathbb{Q}}$ sending z to $1 \otimes_t z$ is injective.*

Proof. The map from $\mathbb{Z}$ to $\mathbb{Q}$ is injective, and we have seen that $\widehat{\mathbb{Z}}$ is a torsion-free and thus flat $\mathbb{Z}$-module, so the map from $\widehat{\mathbb{Z}}$ to $\widehat{\mathbb{Q}}$ is also injective. □

We can thus identify $\mathbb{Q} = \mathbb{Q} \otimes \mathbb{Z}$ and $\widehat{\mathbb{Z}} = \mathbb{Z} \otimes \widehat{\mathbb{Z}}$ with subrings of $\widehat{\mathbb{Q}} = \mathbb{Q} \otimes \widehat{\mathbb{Z}}$. Note that, being commutative rings, $\mathbb{Q}$ and $\widehat{\mathbb{Z}}$ both contain a copy of $\mathbb{Z}$ as a subring, and the corresponding copies of $\mathbb{Z}$ in $\widehat{\mathbb{Q}}$ are equal; this is because $1 \otimes a = a \otimes 1$ for all $a \in \mathbb{Z}$.

5.7 Additive structure of $\widehat{\mathbb{Q}}$.

Here we forget the ring structure on everything, and analyse $\widehat{\mathbb{Q}}$ as an additive abelian group, and in particular how the subgroups $\mathbb{Z}$, $\mathbb{Q}$ and $\widehat{\mathbb{Z}}$ sit within it.

The two results we prove in this section are that $\mathbb{Q} \cap \widehat{\mathbb{Z}} = \mathbb{Z}$ and that $\mathbb{Q} + \widehat{\mathbb{Z}} = \widehat{\mathbb{Q}}$. Using lattice-theoretic notation we could write these results as $\mathbb{Q} \sqcap \widehat{\mathbb{Z}} = \mathbb{Z}$ and $\mathbb{Q} \sqcup \widehat{\mathbb{Z}} = \widehat{\mathbb{Q}}$.

Lemma 5.18. *The intersection of $\mathbb{Q}$ and $\widehat{\mathbb{Z}}$ in $\widehat{\mathbb{Q}}$ is $\mathbb{Z}$.*

Proof. Clearly $\mathbb{Z} \subseteq \mathbb{Q} \cap \widehat{\mathbb{Z}}$. Now suppose that $x \in \mathbb{Q} \cap \widehat{\mathbb{Z}}$. Because x is rational we can write it as $\frac{A}{B} \otimes_t 1$ for some fraction A/B in lowest terms, and hence $x = A/B$ where now we regard $A \in \widehat{\mathbb{Z}}$ and note that A/B is still in lowest terms. However $x \in \widehat{\mathbb{Z}}$ implies that $x = x/1$ is in lowest terms, so we deduce that $B = 1$ and thus $x = A \in \mathbb{Z}$. □

Lemma 5.19. *The sum of $\mathbb{Q}$ and $\widehat{\mathbb{Z}}$ in $\widehat{\mathbb{Q}}$ is $\widehat{\mathbb{Q}}$. More precisely, every element of $\widehat{\mathbb{Q}}$ can be written as $q + z$ with $q \in \mathbb{Q}$ and $z \in \widehat{\mathbb{Z}}$, or more precisely as $q \otimes_t 1 + 1 \otimes_t z$.*

Proof. Write $x \in \widehat{\mathbb{Q}}$ as $x = z/N$ in lowest terms. Lift z_N to an integer t and observe that $(z-t)_N = 0$, hence $z - t = Ny$ for some $y \in \widehat{\mathbb{Z}}$. Now $x = t/N + y \in \mathbb{Q} + \widehat{\mathbb{Z}}$. □

5.8 Multiplicative structure of the units of $\widehat{\mathbb{Q}}$.

We now forget the additive structure on the commutative ring $\widehat{\mathbb{Q}}$ and consider the multiplicative structure of its group of units $\widehat{\mathbb{Q}}^\times$ (which I couldn't get into the section title). We have the obvious subgroups $\mathbb{Q}^\times$, $\mathbb{Z}^\times$ and $\widehat{\mathbb{Z}}^\times$.

Lemma 5.20. *The intersection of $\mathbb{Q}^\times$ and $\widehat{\mathbb{Z}}^\times$ in $\widehat{\mathbb{Q}}^\times$ is $\mathbb{Z}^\times$.*

Proof. Clearly the intersection is contained within $\mathbb{Q} \cap \widehat{\mathbb{Z}} = \mathbb{Z}$. If $n \in \mathbb{Z}$ is in $\widehat{\mathbb{Z}}^\times$ then $n \neq 0$ and its inverse $1/n = \pm 1/|n|$ is in lowest terms but also in $\widehat{\mathbb{Z}}$, and hence $|n| = 1$ by uniqueness of lowest term representation. □

Lemma 5.21. *The product of $\mathbb{Q}^\times$ and $\widehat{\mathbb{Z}}^\times$ in $\widehat{\mathbb{Q}}^\times$ is all of $\widehat{\mathbb{Q}}^\times$. More precisely, every element of $\widehat{\mathbb{Q}}^\times$ can be written as qz with $q \in \mathbb{Q}^\times$ and $z \in \widehat{\mathbb{Z}}^\times$.*

Note that by the previous lemma, this representation will be unique up to sign.

Proof. We already know that a general element of $\widehat{\mathbb{Q}}^\times$ can be written as x/N with N positive, so this reduces us to proving that a general element $x \in \widehat{\mathbb{Z}}$ which is invertible in $\widehat{\mathbb{Q}}^\times$ can be written as qz with $q \in \mathbb{Q}^\times$ and $z \in \widehat{\mathbb{Z}}^\times$.

We know $1/x$ can be written in lowest terms as y/M, and multiplying up we deduce that $xy = M$, and hence x divides a positive integer. If $i : \mathbb{Z} \to \widehat{\mathbb{Z}}$ denotes the inclusion, then we've just seen that the preimage of the principal ideal (x), namely, $J := i^{-1}(x\widehat{\mathbb{Z}})$ is nonzero, as it contains M. Let $g \in J$ be the smallest positive integer; it's well-known that $J = (g)$.

I claim that it suffices to show that $x\widehat{\mathbb{Z}} = g\widehat{\mathbb{Z}}$. Because knowing $g = yx$ and $x = gz$ for some $y, z \in \widehat{\mathbb{Z}}$ tells us that $g(1 - yz) = 0$, and we know that multiplication by g is injective, hence $yz = 1$, so z is a unit and we have written $x = gz$ with $g \in \mathbb{Q}^\times$ and $z \in \widehat{\mathbb{Z}}^\times$.

It remains to prove the claim. By definition $g \in J \subseteq x\widehat{\mathbb{Z}}$ so this is one inclusion. For the other, it suffices to prove that $x_g = 0$. However if $0 < x_g < g$ lifts x_g to the naturals then I claim that $x_g \in J$, for $x_g - x$ is a multiple of g and hence of x, and this contradicts minimality of g. □

We are nearly ready to embark upon the multiplicative adelic theory for quaternion algebras. However before we do this, we need to develop the theory of the Hurwitz quaternions a bit more formally.

5.9 The Hurwitz quaternions

Definition 5.22. *The Hurwitz quaternions are the set $\mathcal{O} := \mathbb{Z} \oplus \mathbb{Z}\omega \oplus \mathbb{Z}i \oplus \mathbb{Z}i\omega$ (as an abstract abelian group or as a subgroup of the usual quaternions). Here $\omega = \frac{-1+(i+j+k)}{2}$ and note that $(i+j+k)^2 = -3$. We have $\overline{\omega} = \omega^2 = -(\omega+1)$. A general quaternion $a+bi+cj+dk$ is a Hurwitz quaternion if either $a, b, c, d \in \mathbb{Z}$ or $a, b, c, d \in \mathbb{Z} + \frac{1}{2}$.*

Lemma 5.23. *The Hurwitz quaternions form a ring.*

Proof. Follow your nose. □

This ring is isomorphic to $\mathbb{Z}^4$ as an additive group, and $\mathcal{O} \otimes_{\mathbb{Z}} \mathbb{R} = \mathbb{R} \oplus \mathbb{R}i \oplus \mathbb{R}j \oplus \mathbb{R}\omega$ is the usual Hamilton quaternions.

Definition 5.24. *There's a conjugation map (which we'll call "star") from the Hurwitz quaternions to themselves, sending integers to themselves and purely imaginary elements like $2\omega + 1$ to minus themselves. It satisfies $(x^*)^* = x$, $(xy)^* = y^*x^*$ and $(x+y)^* = x^* + y^*$. In particular, the Hurwitz quaternions are a "star ring" in the sense of mathlib.*

Definition 5.25. *The Hurwitz quaternions come equipped with an integer-valued norm, which is $a^2 + b^2 + c^2 + d^2$ on $a + bi + cj + dk$ but needs to be modified a bit to deal with ω.*

Lemma 5.26. *We have $N(x) = x\overline{x}$.*

Proof. Easy calculation. □

Lemma 5.27. *The norm of* 0 *is* 0*.*

Proof. A calculation. □

Lemma 5.28. *The norm of* 1 *is* 1*.*

Proof. A calculation. □

Lemma 5.29. *The norm of a product is the product of the norms.*

Proof. A calculation. □

Lemma 5.30. *The norm of an element is nonnegative.*

Proof. It's a sum of rational squares. □

Lemma 5.31. *The norm of an element is zero if and only if the element is zero.*

Proof. It's a sum of rational squares. □

Lemma 5.32. *Given two Hurwitz quaternions a and b with b nonzero, there exists q and r such that $a = qb + r$ and $N(r) < N(b)$.*

Proof. Let q be the nearest Hurwitz quaternion to a/b; one can check that $N(a/b - q) < 1$ and now everything follows. □

Corollary 5.33. *All left ideals of $\mathcal{O}$ are principal.*

Proof. Choose a nonzero element of smallest norm. □

Remark 5.34. *All right ideals are principal too, because there's another version of Euclid saying $a = bq + r$.*

5.10 Profinite completion of the Hurwitz quaternions

We define $\widehat{\mathcal{O}}$ to be $\mathcal{O}\otimes\widehat{\mathbb{Z}}$, so it's elements $a+bi+cj+d\omega$ with $a,b,c,d\in\widehat{\mathbb{Z}}$. The basic thing we need is this:

Theorem 5.35. *If N is a positive natural then the obvious map $\mathcal{O}\to\widehat{\mathcal{O}}/N\widehat{\mathcal{O}}$ is surjective.*

Proof. This is just four copies of the surjection $\mathbb{Z}\to\widehat{\mathbb{Z}}/N\widehat{\mathbb{Z}}$. Note that this latter map is surjective because $\mathbb{Z}\to\mathbb{Z}/N\mathbb{Z}$ is surjective, hence given $z\in\widehat{\mathbb{Z}}$ you can subtract an integer w such that $(z-w)_N=0$, so $z-w$ is a multiple of N. □

We define $D:=\mathbb{Q}\otimes\mathcal{O}=\mathbb{Q}\oplus\mathbb{Q}i\oplus\mathbb{Q}j\oplus\mathbb{Q}\omega=\mathbb{Q}\oplus\mathbb{Q}i\oplus\mathbb{Q}j\oplus\mathbb{Q}k$. Finally, we define $\widehat{D}:=D\otimes\widehat{\mathbb{Z}}$. Just as with $\widehat{\mathbb{Q}}$ we have

Lemma 5.36. *Every element of $\widehat{D}$ can be written as z/N with $z\in\widehat{\mathcal{O}}$ and $N\in\mathbb{N}^+$.*

Proof. Same as the proof for $\widehat{\mathbb{Q}}$. □

It is not hard to check that $\widehat{D}$ contains $\widehat{\mathcal{O}}$ and D as subrings, and that as additive abelian groups we have $\widehat{\mathcal{O}}\cap D=\mathcal{O}$ and $\widehat{\mathcal{O}}+D=\widehat{D}$. This is because $\mathcal{O}$ is just four copies of $\mathbb{Z}$ and we've proved the analogous result for $\mathbb{Z}$.

However the multiplicative structure is more interesting, especially as D is not commutative. For a general quaternion algebra it is *not* true that $(\widehat{D})^\times=D^\times(\widehat{\mathcal{O}})^\times$, because there are "class group obstructions". The double coset space is some kind of non-commutative analogue of a class group. However for our particular choice of D and $\mathcal{O}$ the result is true.

Theorem 5.37. *The group of units of $\widehat{D}$ is $D^\times\widehat{\mathcal{O}}^\times$. More precisely, every element of $\widehat{D}^\times$ can be written as a product δu with $\delta\in D^\times$ and $u\in\widehat{\mathcal{O}}^\times$.*

Proof. Given an element x of $\widehat{D}^\times$, we can use lemma 5.36 to write it as z/N with N a positive integer and $z\in\widehat{\mathcal{O}}$. Note that N is central and in $D^\times$. Similarly, we can write x^{-1} as y/M with M a positive integer and $y\in\widehat{\mathcal{O}}$. Then $1=xx^{-1}=zy/NM$ and so $zy=NM=MN$, and $1=x^{-1}x=yz/MN$ so $yz=MN$ too. In particular y both left and right divides a positive integer.

Now consider the left ideal $\widehat{\mathcal{O}}y$ generated by y. We've just seen that this ideal has nontrivial intersection with $\mathcal{O}$, because it contains $MN>0$. Hence its intersection with $\mathcal{O}$ is a nonzero left ideal of $\mathcal{O}$, which is hence principal by corollary 5.33. Write it as $\mathcal{O}\alpha$ with $0\neq\alpha\in\mathcal{O}$.

It suffices to show that $\widehat{\mathcal{O}}\alpha=\widehat{\mathcal{O}}y$. For this would imply that $u\alpha=y$ and $vy=\alpha$ for some $u,v\in\widehat{\mathcal{O}}$ and thus $(vu-1)\alpha=0$ and $(uv-1)y=0$, and both α and y are left divisors of positive integers (the norm of α, and MN respectively), so now using the fact that $\widehat{\mathcal{O}}$ is $\mathbb{Z}$-torsion-free (is the tensor product of torsion-free abelian groups torsion-free? That would be a cheap way of doing it. Otherwise use $\mathcal{O}=\mathbb{Z}^4$) we deduce that u and v are units, and thus $x^{-1}=\frac{1}{M}u\alpha$ so $x=(M\alpha^{-1})v\in D^\times\widehat{\mathcal{O}}^\times$.

What remains is this. We have $y\in\widehat{\mathcal{O}}$ which left and right divides some positive integer. We've defined $0\neq\alpha\in\mathcal{O}$ such that $\mathcal{O}\alpha$ is the pullback of the abelian group $\widehat{\mathcal{O}}y$ along the map $\mathcal{O}\to\widehat{\mathcal{O}}$. We need to show that when we push this ideal $\mathcal{O}\alpha$ forwards to $\widehat{\mathcal{O}}$ we get $\widehat{\mathcal{O}}y$ again. The fact that $\widehat{\mathcal{O}}\alpha\subseteq\widehat{\mathcal{O}}y$ is easy, because $\alpha\in\widehat{\mathcal{O}}y$ by definition. So it remains to show that $y\in\widehat{\mathcal{O}}\alpha$.

Let's define T to be a positive integer which is both a left and right multiple of both y and α (for example $T=MN\alpha\overline{\alpha}$ will do). Now note that we have an isomorphism

$\mathcal{O}/T\mathcal{O} = \widehat{\mathcal{O}}/T\widehat{\mathcal{O}}$, so we can choose some $\beta \in \mathcal{O}$ such that $\beta - y \in T\widehat{\mathcal{O}}$ is a multiple of T. Next note that $\beta \in y + \widehat{\mathcal{O}}T \subset \widehat{\mathcal{O}}y$ is in $\widehat{\mathcal{O}}y \cap \mathcal{O} = \mathcal{O}\alpha$, meaning $\beta = \gamma\alpha$ for some $\gamma \in \mathcal{O}$. Hence $y \in \beta + \widehat{\mathcal{O}}T \subseteq \widehat{\mathcal{O}}\alpha$. □

Chapter 6

Stating the modularity lifting theorems

I think that a nice and accessible goal (which will maybe take a month or two) would be to *state* the modularity lifting theorems which we'll be formalising. There are in fact two; one (the ”minimal case”) is proved using an extension of the original Taylor–Wiles techniques, and the other is deduced from it using various more modern tricks which were developed later. This chapter (currently work in progress) will contain a detailed discussion of all the things involved in the statement of the theorem.

6.1 Automorphic forms and analysis

Modular forms were historically the first nontrivial examples of automorphic forms, but by the 1950s or so it was realised that they were special cases of a very general notion of an automorphic form, as were Dirichlet characters! Modular forms are holomorphic automorphic forms for the group $\mathrm{GL}_2/\mathbb{Q}$, and Dirichlet characters are automorphic forms for the group $\mathrm{GL}_1/\mathbb{Q}$. It's possible to make sense of the notion of an automorphic form for the group G/k. Here k is a "global field" – that is, a field which is either a finite extension of $\mathbb{Q}$ (a number field) or a finite extension of $(\mathbb{Z}/p\mathbb{Z})(T)$ (a function field), and G is a connected reductive group variety over k.

The reason that the definition of a modular form involves some analysis (they are holomorphic functions) is that if you quotient out the group $\mathrm{GL}_2(\mathbb{R})$ by its centre and the maximal compact subgroup $O_2(\mathbb{R})$, you get something which can be naturally identified with the upper half plane, a symmetric space with lots of interesting differential operators associated to it (for example a Casimir operator). However if you do the same thing with $\mathrm{GL}_1(\mathbb{R})$ then you get a one point set, which is why a Dirichlet character is just a combinatorial object; it's a group homomorphism $(\mathbb{Z}/N\mathbb{Z})^\times \to \mathbb{C}^\times$ where N is some positive integer. It turns out that there are many other connected reductive groups where the associated symmetric space is 0-dimensional, and in these cases the definition of an automorphic form is again combinatorial. An example would be the group variety associated to the units of a totally definite quaternion algebra over a totally real field. In this case, the analogue of $\mathrm{GL}_2(\mathbb{R})$ would be the units $\mathbb{H}^\times$ in the Hamilton quaternions, a maximal compact subgroup would be the quaternions of norm 1 (homeomorphic to the 3-sphere S^3) and quotienting out $\mathbb{H}^\times$ by its centre $\mathbb{R}^\times$ and S^3 again just gives you 1 point.

Before we talk about quaternion algebras, let's talk about central simple algebras.

6.2 Central simple algebras

Convention: in this section, fields are commutative, but algebras over a field may not be. An example of what we are considering below would be Hamilton's quaternions $\mathbb{R}\oplus\mathbb{R}i\oplus\mathbb{R}j\oplus\mathbb{R}k$ as an algebra over $\mathbb{R}$.

Definition 6.1. *A* central simple algebra *over a field K is a nonzero K-algebra D such that K is the centre of D and that D has no nontrivial two-sided ideals.*

Equivalently, every surjective ring homomorphism $D \twoheadrightarrow A$ to any non-commutative ring A is either an isomorphism, or the zero map to the zero ring. Note that this latter condition has nothing to do with K.

Lemma 6.2. *If $n \geq 1$ then the $n \times n$ matrices $M_n(K)$ are a central simple algebra over K.*

Proof. We prove more generally that matrices with coefficients in K and indexed by an arbitrary nonempty finite type are a central simple algebra over K.

They are clearly an algebra over K, with K embedded via scalar matrices as usual (the injectivity of the map from K comes from nonemptiness of the finite index type). The centre clearly contains K; to show that it equals K, we argue as follows. Let $e(i,j)$ be the matrix with a 1 in the ith row and jth column, and zeros everywhere else. An element $Z = (Z_{s,t})_{s,t}$ of the centre commutes with all matrices $e(i,j)$ for $i \neq j$ and these equations immediately imply that $Z_{i,j} = 0$ if $i \neq j$ and that $Z_{i,i} = Z_{j,j}$.

It suffices to prove that any nonzero two-sided ideal I is all of $M_n(K)$. So say $0 \neq M \in I$ and let's fix (i,j) such that $M_{i,j} \neq 0$. One easily checks that $M_{i,j}\mathrm{id} = \sum_k e(k,i) \times M \times e(j,k) \in I$ (where $\mathrm{id} \in M_n(K)$ is the identity matrix). Therefore, $\mathrm{id} \in I$, so $I = M_n(K)$.

The definition also requires that the ring be non-zero, but this follows from the index type being nonempty. □

Lemma 6.3. *If D is a central simple algebra over K and L/K is a field extension, then $L \otimes_K D$ is a central simple algebra over L.*

Proof. This is not too hard: it's lemma b of section 12.4 in Peirce's ”Associative algebras”. Will maybe write more on Saturday. □

Next: define trace and norm.

Chapter 7

Appendix: A collection of results which are needed in the proof.

In this (temporary, unorganised) appendix we list a whole host of definitions and theorems which were known to humanity by the end of the 1980s and which we shall need. These definitions and theorems will find their way into more relevant sections of the blueprint once I have written more details. Note that some of these things are straightforward; others are probably multi-year research projects. The purpose of this chapter right now is simply to give the community (and possibly AIs) some kind of idea of the task we face. Note also that many of the *definitions* here are yet to be formalised in Lean, and this needs to be done before we can start talking about formalising theorems.

7.1 Results from class field theory

We start with the local case. In fact we restrict to the p-adic case, but only for simplicity of exposition because it's all we'll need (and, to be frank, because I'm not 100 percent of what is true in the function field case).

Let K be a finite extension of $\mathbb{Q}_p$. We write $\widehat{\mathbb{Z}}$ for the profinite completion of $\mathbb{Z}$; it is isomorphic to $\prod_p \mathbb{Z}_p$ where $\mathbb{Z}_p$ is the p-adic integers and the product is over all primes.

Theorem 7.1. *The maximal unramified extension K^{un} in a given algebraic closure of K is Galois over K with Galois group "canonically" isomorphic to $\widehat{\mathbb{Z}}$ in two ways; one of these two isomorphisms identifies $1 \in \widehat{\mathbb{Z}}$ with an arithmetic Frobenius (the endomorphism inducing $x \mapsto x^q$ on the residue field of K^{un}, where q is the size of the residue field of K). The other identifies 1 with geometric Frobenius (defined to be the inverse of arithmetic Frobenius).*

It is impossible to say which of the two canonical isomorphisms is "the most canonical"; people working in different areas make different choices in order to locally minimise the number of minus signs in their results.

As a result, the absolute Galois group of K surjects onto $\widehat{\mathbb{Z}}$; its kernel is said to be the *inertia subgroup* of this Galois group. Now pull back this surjection along the continuous map from $\mathbb{Z}$ (with its discrete topology) to $\widehat{\mathbb{Z}}$, in the category of topological groups. We end up with a group containing the inertia group as an open normal subgroup, and with quotient isomorphic to the integers.

Definition 7.2. *The topological group described above is called the* Weil group *of K.*

The following theorem is nontrivial.

Theorem 7.3. *If K is a finite extension of $\mathbb{Q}_p$ then there are two "canonical" isomorphisms of topological abelian groups, between $K^\times$ and the abelianisation of the Weil group of K.*

Proof. This is the main theorem of local class field theory; see for example the relevant articles in [4] or many other places. □

Note that María Inés de Frutos Fernández and Filippo Nuccio are working on a formalisation of the proof of this using Lubin–Tate formal groups.

Now let M be an abelian group (with the discrete topology) equipped with a continuous action of G_K, the Galois group $\mathrm{Gal}(K^{\mathrm{sep}}/K)$ where we fix an algebraic closure $\overline{K}$ of K. Note that if one doesn't want to choose an algebraic closure of K one can instead think of M as being an etale sheaf of abelian groups on $\mathrm{Spec}(K)$.

Continuous group cohomology $H^i(G_K, M)$ in this setting can be defined using continuous cocycles and continuous coboundaries, or just as a colimit of usual group cohomology over the finite quotients of this absolute Galois group (or as etale cohomology, if you prefer). Here are some of the facts we will need about cohomology in this situation. A nice summary is that cohomology of a local Galois group behaves like the cohomology of a compact connected 2-manifold. All the theorems below will need extensive planning.

Theorem 7.4. *If M is finite then the cohomology groups $H^i(G_K, M)$ all finite.*

Proof. This is Proposition 14 in section 5.2 of [12]. □

Theorem 7.5 ("the dimension is 2")**.** *If M is torsion then $H^i(G_K, M) = 0$ if $i > 2$.*

Proof. This follows from Proposition 15 in section 5.3 of [12]. □

Theorem 7.6 ("top degree")**.** *$H^2(G_K, \mu_n)$ is "canonically" isomorphic to $\mathbb{Z}/n\mathbb{Z}$.*

Proof. This is also included in Lemma 2 of section 5.2 of [12] (Serre just writes that the groups are equal; he clearly is not a Lean user. I can see no explanation in his book of this use of the equality symbol. When the statement of this "theorem" is formalised in Lean it may well actually be a definition, giving the map). □

Theorem 7.7 ("Poincaré duality")**.** *If $\mu = \bigcup_{n\geq 1} \mu_n$ and $M' := \mathrm{Hom}(M, \mu)$ is the dual of M then for $0 \leq i \leq 2$ the cup product pairing $H^i(G_K, M) \times H^{2-i}(G_K, M') \to H^2(G_K, \mu) = \mathbb{Q}/\mathbb{Z}$ is perfect.*

Proof. This is Theorem 2 in section 5.2 in [12]. Note again the dubious (as far as Lean is concerned) use of the equality symbol. □

Theorem 7.8 ("Euler-Poincaré characteristic")**.** *If $h^i(M)$ denotes the order of $H^i(G_K, M)$ then $h^0(M) - h^1(M) + h^2(M) = 0$.*

If μ_∞ denotes the Galois module of all roots of unity in our fixed $\overline{K}$, then one can define the dual Galois module M' as $\mathrm{Hom}(M, \mu)$ with its obvious Galois action.

If $0 \leq i \leq 2$ then the cup product gives us a map $H^i(K, M) \times H^{2-i}(K, M') \to H^2(K, \mu_\infty)$.

Theorem 7.9 (Local Tate duality)**.** *(i) There is a "canonical" isomorphism $H^2(K, \mu_\infty) = \mathbb{Q}/\mathbb{Z}$; (ii) The pairing above is perfect.*

Proof. This is Theorem II.5.2 in [12]. □

We now move onto the global case. If N is a number field, that is, a finite extension of $\mathbb{Q}$, then let $\mathbb{A}_N^f := N \otimes_{\mathbb{Z}} \widehat{\mathbb{Z}}$ denote the finite adeles of N and let $N_\infty := N \otimes_{\mathbb{Q}} \mathbb{R}$ denote the product of the completions of N at the infinite places, so $\mathbb{A}_N := \mathbb{A}_N^f \times N_\infty$ is the ring of adeles of N.

Theorem 7.10. *If N is a finite extension of $\mathbb{Q}$ then there are two "canonical" isomorphisms of topological groups between the profinite abelian groups $\pi_0(\mathbb{A}_N^\times/N^\times)$ and $\mathrm{Gal}(\overline{N}/N)^{\mathrm{ab}}$; one sends local uniformisers to arithmetic Frobenii and the other to geometric Frobenii; each of the global isomorphisms is compatible with the local isomorphisms above.*

Proof. This is the main theorem of global class field theory; see for example Tate's article in [4]. □

We need the following consequence:

Theorem 7.11. *Let S be a finite set of places of a number field K . For each $v \in S$ let L_v/K_v be a finite Galois extension. Then there is a finite solvable Galois extension L/K such that if w is a place of L dividing $v \in S$, then L_w/K_v is isomorphic to L_v/K_v as K_v-algebra. Moreover, if K^{avoid}/K is any finite extension then we can choose L to be linearly disjoint from K^{avoid}.*

We also need Poitou-Tate duality; I'll refrain from writing it down for now, because we don't even have Galois cohomology in Lean yet (although we are very close to it).

7.2 Structures on the points of an affine variety.

All rings and algebras in this section are commutative with a 1, and all morphisms send 1 to 1.

Let $X = \mathrm{Spec}(A)$ be an affine scheme of finite type over a field K. For example X could be an affine algebraic variety; in fact we shall only be interested in smooth affine varieties in the applications, but the initial definition and theorem are fine for all finite type schemes.

If R is any K-algebra then one can talk about the R-points $X(R)$ of X, which in this case naturally bijects with the K-algebra homomorphisms from A to R.

Definition 7.12. *If X is an affine scheme of finite type over K, and if R is a K-algebra which is also a topological ring, then we define a topology on the R-points $X(R)$ of K by embedding the K-algebra homomorphisms from A to R into the set-theoretic maps from A to R with its product topology, and giving it the subspace topology.*

Theorem 7.13. *If X is as above and $X \to \mathbb{A}_K^n$ is a closed immersion, then the induced map from $X(R)$ with its topology as above to R^n is an embedding of topological spaces (that is, a homeomorphism onto its image).*

Proof. See Conrad's notes. □

We now specialise to the smooth case. I want to make the following conjectural "definition":

Definition 7.14. *Let K be a field equipped with an isomorphism to the reals, complexes, or a finite extension of the p-adic numbers. Let X be a smooth affine algebraic variety over K. Then the points $X(K)$ naturally inherit the structure of a manifold over K.*

Remark 7.15. *Probably this is fine for a broader class of fields K.*

Theorem 7.16. *If X is as in the previous definition and $X \to \mathbb{A}^n_K$ is a closed immersion, then the induced map from $X(K)$ with its manifold structure to K^n is an embedding of manifolds.*

Proof. I'm assuming this is standard, if true. □

Corollary 7.17. *If G is an affine algebraic group of finite type over $K = \mathbb{R}$ or $\mathbb{C}$ then $G(K)$ is naturally a real or complex Lie group.*

Remark 7.18. *The corollary, for sure, is true! And it's all we need. I have not yet made any serious effort to find a reference for the definition or independence, although there seem to be some ideas here. As a toy example, one can embed $\mathrm{GL}_n(\mathbb{R})$ into either $\mathbb{R}^{n^2+1}$ via $M \mapsto (M, \det(M)^{-1})$ or into $\mathbb{R}^{2n^2}$ via $M \mapsto (M, M^{-1})$ and the claim is that the two induced manifold structures are the same.*

7.3 Algebraic groups.

The concept of an affine algebraic group over a field K can be implemented in Lean as a commutative Hopf algebra over K, as a group object in the category of affine schemes over K, as a representable group functor on the category of affine schemes over K, or as a representable group functor on the category of schemes over K which is represented by an affine scheme. All of these are the same to mathematicians but different to Lean and some thought should go into which of these should be the actual definition, and which should be proved to be the same thing as the definition.

Definition 7.19. *An affine algebraic group G of finite type over a field k is said to be* connected *if it is connected as a scheme, and* reductive *if $G_{\overline{k}}$ has no nontrivial smooth connected unipotent normal k-subgroup.*

7.4 Automorphic forms and representations

This section needs a lot of work; I am just attempting to write down some approximation to the (well-known) definitions but in great generality (far greater than we need). Some definitions below are short on details; indeed there may even be errors or imprecisions right now (because we are working in more generality than I am used to). It will be a very interesting project to get these details down. One reference (which leaves a lot of exercises) is Borel-Casselman in [2]. Even *stating* these definitions will be a big challenge in Lean; indeed one of the motivations of the project is that it forces us to write down all the below properly.

Let G be a connected reductive group over a number field N. We note that $G(\mathbb{A}_N^f)$ is a (locally profinite) topological space and $G(N_\infty)$ is a real Lie group; their product is $G(\mathbb{A}_N)$. If $g \in G(\mathbb{A}_N)$, write $g_f \in G(\mathbb{A}_N^f)$ for the finite part and $g_\infty \in G(N_\infty)$ for its infinite part.

For some reason, in the literature people seem to fix a choice of maximal compact subgroup U_∞ of $G(N_\infty)$. I believe that all such subgroups are conjugate, and probably this gives some route between the different definitions coming from the different choices.

Example: if $G = \mathrm{GL}_2$ and $N = \mathbb{Q}$ then $N_\infty = \mathbb{R}$ and $G(N_\infty)$ is just $\mathrm{GL}_2(\mathbb{R})$ with its usual Lie group structure and we can take U_∞ to be $O_2(\mathbb{R})$; $G(\mathbb{A}_N^f)$ is the restricted product of $\mathrm{GL}_2(\mathbb{Q}_p)$ over $\mathrm{GL}_2(\mathbb{Z}_p)$, for all primes p.

By assumption, $G(N_\infty)$ admits a finite-dimensional (algebraic) representation ρ with finite kernel. Consider ρ as taking values in $GL_N(\mathbb{C}) = \mathrm{Aut}_\mathbb{C}(V)$. Fix a hermitian sesquilinear form on V which is U_∞ invariant, and now define a norm $||g||_\rho$ on $G(N_\infty)$ by

$$||g||_\rho = (\mathrm{tr}\, \rho(g)^* \rho(g))^{1/2},$$

where the asterisk denotes adjoint with respect to the sesquilinear form. According to the article by Borel–Jacquet in [2] (p189), if ρ' is another such choice then there exists a positive real C and a positive integer n such that $||g||_{\rho'} \leq C||g||_\rho^n$ for all $g \in G(N_\infty)$.

Definition 7.20. *A function $f : G(N_\infty) \to \mathbb{C}$ is* slowly-increasing *if there exists some $C > 0$ and $n \geq 1$ such that $|f(x) \leq C||x||_\rho^n$.*

Theorem 7.21. *This is independent of the choice of ρ as above.*

Proof. Follows from the above. □

We can now give the definition of an automorphic form. For FLT we only need the definition for G being either an abelian algebraic group, or an inner form of $GL(2)$, but we have chosen to work in full generality here.

Definition 7.22. *An* automorphic form *is a function $\phi : G(\mathbb{A}_N) \to \mathbb{C}$ satisfying the following conditions:*

- *ϕ is locally constant on $G(\mathbb{A}_N^f)$ and C^∞ on $G(N_\infty)$. In other words, for every g_∞, $\phi(-, g_\infty)$ is locally constant, and for every g_f, $\phi(g_f, -)$ is smooth.*
- *ϕ is left-invariant under $G(N)$;*
- *ϕ is right-U_∞-finite (that is, the space spanned by $x \mapsto \phi(xu)$ as u varies over U_∞ is finite-dimensional);*
- *ϕ is right K_f-finite, where K_f is one (or equivalently all) compact open subgroups of $G(\mathbb{A}_N^f)$;*
- *ϕ is z-finite, where z is the centre of the universal enveloping algebra of the Lie algebra of $G(N_\infty)$, acting via differential operators. Equivalently ϕ is annihiliated by a finite index ideal of this centre, so morally ϕ satisfies lots of differential equations of a certain type;*
- *For all g_f, the function $g_\infty \mapsto \phi(g_f g \infty)$ is slowly-increasing in the sense above.*

Automorphic forms form a typically infinite-dimensional vector space.

Definition 7.23. *An automorphic form is* cuspidal *(or "a cusp form") if it furthermore satisfies $\int_{U(N)\backslash U(\mathbb{A}_N)} \phi(ux)du = 0$, where P runs through all the proper parabolic subgroups of G defined over N and U is the unipotent radical of P, and the integral is with respect to the measure coming from Haar measure.*

The cuspidal automorphic forms form a complex subspace of the space of automorphic forms.

Definition 7.24. *The group $G(\mathbb{A}_N)$ acts on itself on the right, and this induces a left action of its subgroup $G(\mathbb{A}_N^f) \times U_\infty$ on the spaces of automorphic forms and cusp forms. The Lie algebra $\mathfrak{g}$ of $G(N_\infty)$ also acts, via differential operators. Furthermore the actions of $\mathfrak{g}$ and U_∞ are compatible in the sense that the differential of the U_∞ action is the action of its Lie algebra considered as a subalgebra of $\mathfrak{g}$. We say that the spaces are $(G(\mathbb{A}_N^f) \times U_\infty, \mathfrak{g})$-modules.*

Theorem 7.25. *The cusp forms decompose as a (typically infinite) direct sum of irreducible $(G(\mathbb{A}_N^f) \times U_\infty, \mathfrak{g})$-modules.*

Definition 7.26. *A cuspidal automorphic representation is an irreducible $(G(\mathbb{A}_N^f) \times U_\infty, \mathfrak{g})$-module isomorphic to an irreducible summand of the space of cusp forms.*

For non-cuspidal representations, they do not decompose as a direct sum; there is a continuous spectrum which decomposes as a direct integral. We may not ever need these. As a result the definition of an automorphic representation has to be slightly modified in the non-cuspidal case.

Definition 7.27. *An automorphic representation is an irreducible $(G(\mathbb{A}_N^f) \times U_\infty, \mathfrak{g})$-module isomorphic to an irreducible subquotient of the space of automorphic forms.*

Admissibility is a finiteness condition on an irreducible representation of $(G(\mathbb{A}_N^f) \times U_\infty, \mathfrak{g})$; automorphic representations are admissible, and this seems to boil down to theorems of Godement and Harish-Chandra in the general case.

Theorem 7.28. *An irreducible admissible $(G(\mathbb{A}_N^f) \times U_\infty, \mathfrak{g})$-module is a restricted tensor product of irreducible representations π_v of $G(N_v)$ as v runs through the finite places of N, tensored with a tensor product of irreducible $(\mathfrak{g}_v, U_{\infty,v})$-modules as v runs through the infinite places of N. The representations π_v are unramified for all but finitely many v.*

Proof. See Flath's article in [3]. □

As mentioned above, we only need all of this for abelian algebraic groups and for inner forms of GL_2 over totally real fields, where everything can be made more concrete (and in particular where I can write down concrete definitions, although this still needs to be done). In particular, we don't strictly speaking need all of the above, we could just cheat and deal with $\mathrm{GL}_2(\mathbb{R})$ and $\mathbb{H}^\times$ separately.

The theorems I need are: Jacquet-Langlands for inner forms of GL_2 over totally real fields, and multiplicity 1 for these inner forms. We also need cyclic base change plus classification of image, all for totally definite quaternion algebras, and we need automorphic induction from $\mathrm{GL}_1(K)$ to $\mathrm{GL}_2(F)$ when K/F is a degree 2 totally imaginary extension. There seems to be little point formalising the statements of the theorems if we cannot yet even formalise the definition of an automorphic representation properly.

7.5 Galois representations

Ivan Farabella has formalised the definition of a compatible family of Galois representations, modulo the existence of Frobenius elements, which has been established by Jou Glasheen.

Definition 7.29. *Let N be a number field. A* compatible family of d-dimensional Galois representations over *N is a finite set of finite places S of N, a number field E, a monic degree d polynomial $F_\mathfrak{p}(X) \in E[X]$ for each finite place $\mathfrak{p}$ of K not in S and, for each prime* *number ℓ and field embedding $\phi : E \to \overline{\mathbb{Q}}_\ell$ (or essentially equivalently for each finite place of E), a continuous homomorphism $\rho : \mathrm{Gal}(K^{\mathrm{sep}}/K) \to \mathrm{GL}_2(\overline{\mathbb{Q}}_\ell)$ unramified outside S and the primes of K above ℓ, such that $\rho(\mathrm{Frob}_{\mathfrak{p}})$ has characteristic polynomial $P_\pi(X)$ if π lies above a prime number $p \neq \ell$ with $p \notin S$.*

The big theorem, which again we are far from even *stating* right now, is

Theorem 7.30. *Given an automorphic representation π for an inner form of GL_2 over a totally real field and with reflex field E, such that π is weight 2 discrete series at every infinite place, there exists a compatible family of 2-dimensional Galois representations associated to π, with S being the places at which π is ramified, and $F_{\mathfrak{p}}(X)$ being the monic polynomial with roots the two Satake parameters for π at $\mathfrak{p}$.*

7.6 Algebraic geometry

We have already mentioned Mazur's Theorem on torsion subgroups of elliptic curves (theorem 3.27). The proof of this is the main theorem of [9], 150 pages of subtle arithmetic geometry involving the bad reduction of modular curves, exotic cohomology theories (etale and more), and the consequences of this for the Neron models of their Jacobians. After a beautiful introductory chapter containing a history and examples, the convention is established that throughout the paper, N will denote a prime number which is at least 5. And then the first sentence of chapter 1 of the paper proper is "Consider quasi-finite separated commutative group schemes of finite presentation over the base $S := \operatorname{Spec} \mathbb{Z}$ which are finite flat group schemes over $S' := \operatorname{Spec}(Z[1/N])$.". At the time of writing (May 2024), Lean's algebraic geometry cannot get us through *the first sentence of Mazur's proof*, which occupies pages 43 to 172 of the paper (not including the appendix or references, that's just the proof). Anyone interested in formalising Mazur's paper should make a formalisation of its first sentence their first milestone.

Talking of modular curves, we also need the existence of Shimura curves and surfaces over totally real fields F (of degree greater than 2, so always compact). The curves are "modeles étranges" in the sense of Deligne, so we also need moduli spaces of unitary Shimura varieties over CM extensions. We need to decompose the first and second etale cohomology groups of these varieties into Galois representations, by understanding them in terms of automorphic representations.

Definition 7.31. *We need the definition of (the canonical model over F of) the Shimura curve attached to an inner form of GL_2 with precisely one split infinite place, and the same for the Shimura surface associated to an inner form split at two infinite places (and ramified elsewhere, so it's compact).*

We also need Moret-Bailly's theorem from [10]:

Theorem 7.32. *Let K^{avoid}/K be a Galois extension of number fields. Suppose also that S is a finite set of places of K. For $v \in S$ let L_v/K_v be a finite Galois extension. Suppose also that T/K is a smooth, geometrically connected curve and that for each $v \in S$ we are given a nonempty, $\mathrm{Gal}(L_v/K_v)$-invariant, open subset $\Omega_v \subseteq (L_v)$. Then there is a finite Galois extension L/K and a point $P \in T(L)$ such that*

- *L/K is Galois and linearly disjoint from K^{avoid} over K;*
- *if $v \in S$ and w is a prime of L above v then L_w/K_v is isomorphic to L_v/K_v;*

- *and $P \in \Omega_v \subseteq T(L_v) \cong (L_w)$ via one such K_v-algebra morphism (this makes sense as Ω_v is* $\mathrm{Gal}(L_v/Kv)$*-invariant).*

Note that we do not even have the definition of a curve over a field in Lean.

7.7 Algebra

We need the classification of finite subgroups of $\mathrm{PGL}_2(\overline{\mathbb{F}}_p)$. The answer is that they are all cyclic, dihedral, A_4, S_4, A_5, or isomorphic to $\mathrm{PSL}_2(k)$ or $\mathrm{PGL}_2(k)$ for some finite field of characteristic p. This should at least be easy to state!

Bibliography

[1] Thomas Barnet-Lamb, Toby Gee, David Geraghty, and Richard Taylor. Potential automorphy and change of weight. *Ann. of Math. (2)*, 179(2):501–609, 2014.

[2] Armand Borel and W. Casselman, editors. *Automorphic forms, representations and L-functions. Part 1*, volume XXXIII of *Proceedings of Symposia in Pure Mathematics.* American Mathematical Society, Providence, RI, 1979.

[3] Armand Borel and W. Casselman, editors. *Automorphic forms, representations, and L-functions. Part 2*, volume XXXIII of *Proceedings of Symposia in Pure Mathematics.* American Mathematical Society, Providence, RI, 1979.

[4] J. W. S. Cassels and A. Fröhlich, editors. *Algebraic number theory.* Academic Press, London; Thompson Book Co., Inc., Washington, DC, 1967.

[5] Henri Darmon, Fred Diamond, and Richard Taylor. Fermat's last theorem. In *Current developments in mathematics, 1995 (Cambridge, MA)*, pages 1–154. Int. Press, Cambridge, MA, 1994.

[6] Bas Edixhoven. The weight in Serre's conjectures on modular forms. *Invent. Math.*, 109(3):563–594, 1992.

[7] Toby Gee. Modularity lifting theorems. *Essent. Number Theory*, 1(1):73–126, 2022.

[8] Nicholas M. Katz and Barry Mazur. *Arithmetic moduli of elliptic curves*, volume 108 of *Annals of Mathematics Studies.* Princeton University Press, Princeton, NJ, 1985.

[9] B. Mazur. Modular curves and the Eisenstein ideal. *Inst. Hautes Études Sci. Publ. Math.*, (47):33–186, 1977. With an appendix by Mazur and M. Rapoport.

[10] Laurent Moret-Bailly. Groupes de Picard et problèmes de Skolem. I, II. *Ann. Sci. École Norm. Sup. (4)*, 22(2):161–179, 181–194, 1989.

[11] Jean-Pierre Serre. Propriétés galoisiennes des points d'ordre fini des courbes elliptiques. *Invent. Math.*, 15(4):259–331, 1972.

[12] Jean-Pierre Serre. *Galois cohomology.* Springer Monographs in Mathematics. Springer-Verlag, Berlin, english edition, 2002. Translated from the French by Patrick Ion and revised by the author.

[13] Joseph H. Silverman. *Advanced topics in the arithmetic of elliptic curves*, volume 151 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1994.

[14] Joseph H. Silverman. *The arithmetic of elliptic curves*, volume 106 of *Graduate Texts in Mathematics*. Springer, Dordrecht, second edition, 2009.

[15] Richard Taylor. On the meromorphic continuation of degree two L-functions. *Doc. Math.*, pages 729–779, 2006.