Deformations of Galois Representations

Fernando Q. Gouvêa

Contents

IAS/Park City Mathematics Series
Volume 00, 0000

Deformations of Galois Representations

Fernando Q. Gouvêa

Introduction

These notes were prepared for a short graduate course which I was invited to teach at the Park City Mathematics Institute in the summer of 1999. The topic is the theory of deformations of Galois representations, which was created by Barry Mazur and has become, especially after Wiles' fundamental work on the modularity conjecture for elliptic curves, an important number-theoretical tool. This revised version of the notes includes several supplements, including three appendices. The first, by Mark Dickinson, gives a proof of the existence of the universal deformation that works directly from Grothendieck's theorem instead of using the Schlessinger criteria. The second, by Tom Weston, gives a detailed account of how to prove a theorem of Flach mentioned in Lecture 5. The third, by Matthew Emerton, gives an introduction to the theory of p-adic modular forms which is sketched very briefly in Lecture 7. The second and third appendix are write-ups of talks given at PCMI. I am grateful to all three authors for their permission to include their work in these notes.

I have tried to sprinkle problems throughout the write-up. These are of various kinds. Some simply ask the reader to fill in the details of an argument or to supply the proof of a theorem; these are mostly straightforward, but keep in mind that the notion of "straightforward" is highly dependent on each person's background. Other problems ask the reader to work out a specific example; I hope these will be helpful in understanding the material. Some problems are open-ended suggestions that the reader might want to investigate. A few problems ask questions whose answers I do not know, but which seemed natural to me as I was preparing these notes. Some of these are bound to be embarrassingly easy, while others may be quite hard.

[1] Department of Mathematics, Colby College, Waterville, Maine, USA 04901.
E-mail address: `fqgouvea@colby.edu`.

1991 *Mathematics Subject Classification.* Primary 11F80; Secondary 11F03, 11F33, 11R39, 14D15.

Key words and phrases. Galois representations, deformations, universal deformations.

Project sponsored by the National Security Agency under Grant number MDA904-98-1-0012. The United States Government is authorized to reproduce and distribute reprints.

The first six lectures form the core material for the course on deformations of Galois representations. I have tried to make them a useful introduction to the subject. The last two lectures have a different character. The seventh is a broad survey of how deformation theory interacts with the theory of modular forms, with special focus on (various forms of) the issue of deciding which deformations are modular. The last lecture is a brief account of the material in [**100**] and [**62**], describing the construction of the "infinite fern" in the deformation space of a modular residual representation. Due to this difference in style, the problems disappear when we get to the last two lectures. (The last two lectures are in fact a sort of meta-problem: fill in the details in this account.)

Several other accounts of the basic theory are available, and all were enormously useful (and at times intimidating!) as I was preparing my own notes. First of all, one should mention Mazur's three excellent (and very different) accounts of the theory: his original paper [**97**], his more elementary account in [**100**] (which puts special emphasis on modular representations), and the expository account [**101**] in the proceedings [**33**] of the Boston University conference on Fermat's Last Theorem. In the same volume one also finds the articles [**36**], which gives an alternative construction of the universal deformation, and [**32**], which looks closely at the theory of flat deformations. Another extended account of the theory is the unpublished book [**42**] by Charles Doran and Siman Wong. Shorter surveys include Mazur's own summary in [**100**], my survey in [**60**], and portions of the various surveys of the proof of Fermat's Last Theorem, such as [**119**], [**34**], and [**114**]. These are all worth a look.

I would like to thank the organizers of the 1999 Park City Mathematics Institute for their invitation to teach this course. Thanks also to the several people who made suggestions for this revised version, including Brian Conrad, Ralph Greenberg, Armand Brumer, and Blair Kelly. Special thanks to Mark Dickinson, my teaching assistant at PCMI, who made extensive suggestions, handled problem sessions, and saw to it that copies were made and distributed.

Fernando Gouvêa
Department of Mathematics
Colby College
Waterville, ME 04901
fqgouvea@colby.edu

LECTURE 1
Galois Groups and Their Representations

Our main concern throughout these lectures will be to study the representations of the absolute Galois group of a field. More often than not, the field in question will be $\mathbb{Q}$, the field of rational numbers, but we will also need to consider number fields (finite extensions of $\mathbb{Q}$), their various completions, and finite fields. So we'll start by saying some general things about Galois groups of infinite extensions, then quickly specialize to the cases that interest us. We will then try to collect what is known about the groups we want to study. This will give this lecture something of the nature of a survey; we have tried to add details only when they are not easily found in the standard references.

Galois groups of infinite algebraic extensions

In this section we give a very brief sketch of the Galois theory of infinite algebraic extensions. References for this material are [**112**, Chapter 1] and [**109**, Chapter IV].

Let K be a perfect field, and let F be a (finite or infinite) normal extension of K. The Galois group

$$G(F/K) = \mathrm{Gal}(F/K)$$

is defined, as usual, to consist of all automorphisms of F which induce the identity on K. When F/K is infinite, this is an infinite group. In the special case in which $F = \overline{K}$ is an algebraic closure of K, we will call this the *absolute Galois group of* K and we will denote it by G_K.

For finite extensions, the Galois correspondence nicely matches up subgroups of the Galois group with subextensions. The main difficulty with which we have to deal in this section is the fact that the naïve generalization of this correspondence does not work for infinite extensions. The easiest way to see this is to consider an example that will keep coming up: the absolute Galois group of a finite field.

Example. Let p be a prime number, $K = \mathbb{F}_p$ be the finite field with p elements, and let $F = \overline{\mathbb{F}}_p$ be an algebraic closure. The *Frobenius automorphism* $\phi = \phi_p : F \longrightarrow F$ (which we will often simply call "the Frobenius" or "the Frobenius at p") is defined by $\phi(x) = x^p$. Recall that for each n there is only one extension of $\mathbb{F}_p$ inside $\mathbb{F}$ which has degree n; this extension is fixed by ϕ^n. Let $Z \subset G_{\mathbb{F}_p}$ be the subgroup generated by ϕ. It is easy to see that Z is an infinite cyclic group and that its fixed subfield is $\mathbb{F}_p$. Galois theory for extensions of finite degree would then lead us to expect that $Z = G_{\mathbb{F}_p}$, but this is very far from being true in our case. To see this, choose any sequence of integers a_n such that we have $a_n \equiv a_m \pmod{m}$ whenever $m|n$. We define an automorphism ψ of $\mathbb{F}$ by requiring $\psi|_{\mathbb{F}_{p^n}} = \phi^{a_n}$. The conditions on the sequence a_n mean that these definitions are compatible. Since every element of $\mathbb{F}$ belongs to some subfield of finite degree over $\mathbb{F}_p$, we see that this defines an automorphism $\psi \in G_{\mathbb{F}_p}$. But $\psi \in Z$ if and only if the sequence a_n is constant, i.e., if there is some integer a such that $a_n = a$ for all n. Since there are many non-constant sequences of this type, we have shown that $G_{\mathbb{F}_p}$ is in fact much larger than Z.

Problem 1.1. Check the details in this construction. Specifically, show that many non-constant sequences $\{a_n\}$ exist and that the conditions defining ψ are indeed compatible.

Problem 1.2. How big *is* $G_{\mathbb{F}_p}$? For example, is it a countable set?

As usual, the way to fix the problem is to introduce a topology on our infinite Galois group, and then to show that the Galois correspondence will work provided we work only with *closed* subgroups. The definition of the topology is quite natural; basically, we need something that reduces to the discrete topology when F/K is finite and which selects the right subgroups for the correspondence to work.

Here is the formal definition. First of all, we'll say an extension of fields is a *Galois extension* if it is algebraic, normal, and separable. (In what follows, we will always assume that the base field K is perfect, so that we need not worry about separability.)

Definition 1.1. Let F/K be a (finite or infinite) Galois extension. For each *finite* Galois subextension K'/K, consider the Galois group $G(K'/K)$, and whenever we have two finite subextensions $K' \subset K''$ consider the homomorphism

$$G(K''/K) \longrightarrow G(K'/K)$$

given by restriction. This whole package defines an inverse system of groups, and we define the Galois group of F over K to be the inverse limit[1]

$$G(F/K) = \varprojlim_{K'/K} G(K'/K)$$

with its natural profinite topology.

This definition nicely generalizes the example above; after all, an element of the inverse limit is exactly given by an indexed set $\{\sigma_{K'}\}$ such that $\sigma_{K'} \in G(K'/K)$ and such that the various $\sigma_{K'}$ are compatible under restriction, which is very similar to the way we constructed the sequence $\{a_n\}$. The main thing we have added is the topology. It's worth noticing that if F is actually a finite extension of K, then the group is finite and the topology is just the discrete topology.

Problem 1.3. Let F/K be an infinite Galois extension, and let G be the group of automorphisms of F which induce the identity on K. For each *finite* Galois subextension K'/K, let $G(F/K')$ denote the normal subgroup of G consisting of all automorphisms which induce the identity on K'. Define a topology on G by defining a basis of neighborhoods of each $\sigma \in G$ to be the set of all cosets $\sigma G(F/K')$, where K' runs

[1] See the complements to lecture 1 for a quick overview of inverse limits.

through all finite Galois extensions of K. Show that this yields the same group and the same topology as in the definition above.

Problem 1.4. Show that $G(F/K)$ is Hausdorff, compact, and totally disconnected.

Problem 1.5. Let G be a topological group. Show that all open subgroups of G are also closed. If G is compact, show that all open subgroups are of finite index in G. Conversely, show that a closed subgroup of finite index in a topological group G is open.

Groups that are inverse limits of a projective system of finite groups are called *profinite*. As this discussion suggests, they appear quite naturally in arithmetical and algebraic contexts. One can often show that profinite groups have properties that are very close to the properties of finite groups. There will be several examples in what follows. A quick overview of profinite groups appears in the complements to lecture 1. For more information, see [**109**, Appendix C], [**139**] (or its English translation [**140**]), and [**143**].)

The immediate result of topologizing our group is that we now get a good Galois correspondence:

Theorem 1.1. *Let F/K be a (finite or infinite) Galois extension. The map*

$$K' \mapsto G(F/K')$$

defines a bijective inclusion-reversing correspondence between subextensions K'/K and closed subgroups of $G(F/K)$. The inverse correspondence is given by

$$H \mapsto F^H,$$

where, as usual, F^H denotes the subfield of F consisting of those elements which are fixed by every element of H.

In particular, the open subgroups (which are also closed and of finite index, see above) correspond to the finite subextensions.

Problem 1.6. Prove the theorem.

Problem 1.7. Consider the field $F \subset \overline{\mathbb{Q}}$ which is the compositum of all quadratic extensions of $\mathbb{Q}$. Describe the Galois group $G(F/\mathbb{Q})$ in as much detail as you can. Show that it has many subgroups of finite index which are not closed. (In fact, *most* of its subgroups of finite index are not closed!)

Problem 1.8. Let G_1 and G_2 be profinite groups. Show that a continuous injective homomorphism $G_1 \longrightarrow G_2$ is an isomorphism from G_1 onto a closed subgroup of G_2.

Problem 1.9. Show that every profinite group arises as a Galois group for some Galois extension.

Let's reconsider our initial example, and then take a brief look at the other main examples we'll need to consider.

Example. Let $\mathbb{F}$ be an algebraic closure of $\mathbb{F}_p$, and consider subfields $K \subset \mathbb{F}$. Then we know that there is a unique finite subextension $K/\mathbb{F}_p$ of degree n, which we call $\mathbb{F}_{p^n}$, and we have

$$G(\mathbb{F}_{p^n}/\mathbb{F}_p) \cong \mathbb{Z}/n\mathbb{Z},$$

where the isomorphism is obtained by mapping the Frobenius element ϕ to 1. The argument above shows directly that $G_{\mathbb{F}_p} \cong \hat{\mathbb{Z}}$, where $\hat{\mathbb{Z}}$ is the procyclic group

$$\hat{\mathbb{Z}} = \varprojlim_n \mathbb{Z}/n\mathbb{Z},$$

where the homomorphisms $\mathbb{Z}/n\mathbb{Z} \longrightarrow \mathbb{Z}/m\mathbb{Z}$ used to obtain the limit are defined, whenever $m|n$, to be simply reduction modulo m. The set Z which we considered above, consisting of all the integral powers of ϕ, is a dense subset of $G_{\mathbb{F}_p}$. Hence we say that ϕ *topologically generates* $G_{\mathbb{F}_p}$.

Problem 1.10. Show that the natural map $\hat{\mathbb{Z}} \longrightarrow \prod_p \mathbb{Z}_p$ is an isomorphism.

Problem 1.11. Show that $\mathbb{Z}_p$ (thought of as an additive group) is also a profinite group topologically generated by a single element, but is not isomorphic to $\hat{\mathbb{Z}}$.

The example of the absolute Galois group of a finite field is one for which we can get a precise description. This is far from being the case for the other two groups that will be at the center of our attention: the absolute Galois groups of $\mathbb{Q}_p$ and of $\mathbb{Q}$. We will consider both groups more carefully in the next section.

Before we go on to that, let's set up one more bit of the abstract theory.

Definition 1.2. Let G be a topological group. We define a *G-module* to be an abelian topological group M together with a *continuous* map

$$G \times M \longrightarrow M$$
$$(\sigma, m) \mapsto \sigma m$$

which satisfies the following conditions (where we write the operation on M additively):

i. $1m = m$, for all $m \in M$,
ii. $\sigma(m+n) = \sigma m + \sigma n$, for all $\sigma \in G$ and $m, n \in M$, and
iii. $(\sigma\tau)m = \sigma(\tau m)$, for all $\sigma, \tau \in G$ and all $m \in M$.

The most common situation is when G is a profinite group and M has the discrete topology. In this case, the condition of continuity can be translated to a simple group-theoretic condition:

Problem 1.12. Suppose G is profinite and M has the discrete topology. For each subgroup $H \subset G$, write M^H for the set of elements of M which are fixed by every element of H. Show that the map $G \times M \longrightarrow M$ is continuous if and only if we have

$$M = \bigcup_H M^H,$$

where H runs through all the *open* subgroups of G.

Problem 1.13. Let A be a topological ring whose undelying abelian group is profinite. We say that A is a *profinite ring*.

i. Prove that the natural map of topological rings

$$A \longrightarrow \varprojlim_I A/I$$

is an isomorphism, where I runs through the closed *ideals* of finite index in A. (To begin with, you need to show that if $A \neq 0$ such proper closed ideals $I \subset A$ of finite index do exist.)

ii. Let A be a complete noetherian local ring, and give A its "natural" topology, that is, the topology defined by the powers of its maximal ideal. Show that A is profinite if and only if its residue field is finite.

iii. Give an example of a profinite local ring which is not noetherian.

Problem 1.14. Let A be a profinite ring and let M be a finite free A-module with a continuous action of a profinite group G. For each open subgroup $H \subset G$, let M^H be defined as above. Show that the natural map

$$M \longrightarrow \varprojlim_H M^H$$

is a topological isomorphism. How does this relate to problem 1.12?

Given a G-module M, we can define the cohomology groups $\mathrm{H}^i(G, M)$ as in [**152**]. The case of interest for us, of course, will be when G is a Galois group. In addition to [**152**], see also [**156**] for a short introduction to Galois Cohomology. The books [**135**], [**139**], [**140**], and [**111**] contain more extensive treatments, the last being especially complete.

We conclude this section with two problems that deal with an idea that will later be very important for us: the notion of a *pro-p-group.*

Problem 1.15. Let p be a prime. A profinite group G is called a *pro-p-group* if every finite quotient of G is a p-group. (For example, $\mathbb{Z}_p$, as an additive group, is a pro-p-group.) Let $\Gamma_2(\mathbb{Z}_p)$ denote the kernel of the reduction mod p map

$$\mathrm{GL}_2(\mathbb{Z}_p) \longrightarrow \mathrm{GL}_2(\mathbb{F}_p).$$

Show that $\Gamma_2(\mathbb{Z}_p)$ is a pro-p-group.

Problem 1.16. Let G be a profinite group and p be a prime. Define another profinite group $G^{(p)}$ by

$$G^{(p)} = \varprojlim_H G/H,$$

where H runs through the open normal subgroups of G whose index in G is a power of p.

i. Show that there is a canonical continuous group homomorphism $\pi : G \longrightarrow G^{(p)}$ and that any continuous group homomorphism from G to a finite discrete p-group factors through π.

ii. Let $G = \hat{\mathbb{Z}}$. What is $G^{(p)}$?

iii. Is $G^{(p)}$ a quotient of G?

iv. Formulate and prove a universal property of $G \longrightarrow G^{(p)}$ in the category of profinite groups.

The Galois group of $\mathbb{Q}$

The Galois group which will mainly concern us is $G_{\mathbb{Q}} = G(\overline{\mathbb{Q}}/\mathbb{Q})$, the absolute Galois group of $\mathbb{Q}$. In this section we gather together some basic information about this group (and also try to point out that there is much about it that is still quite mysterious). Much of what we do would also apply to the Galois group of a general number field. All of this is to be found in standard references on algebraic number theory; our summary is inspired by the material in [**101**], [**34**, Section 2.1] and [**112**, Chapter 1].

Let's begin with what one might call the "local structure" of $G_{\mathbb{Q}}$. For each prime number p, there is a canonical inclusion of $\mathbb{Q}$ into its completion $\mathbb{Q}_p$. When we go to algebraic closures, however, there are many different inclusions $\overline{\mathbb{Q}} \longrightarrow \overline{\mathbb{Q}}_p$ (this is equivalent to saying that there are many ways to extend the p-adic valuation on $\mathbb{Q}$ to $\overline{\mathbb{Q}}$). Once we choose such an embedding, we get an inclusion of Galois groups $G_{\mathbb{Q}_p} \hookrightarrow G_{\mathbb{Q}}$. Changing the embedding changes this inclusion by conjugation. The image of $G_{\mathbb{Q}_p}$ is called a *decomposition group at p*. We will usually want to identify $G_{\mathbb{Q}_p}$ with its image in $G_{\mathbb{Q}}$, but when we do so we will have to bear in mind that the picture we have in mind is unique only up to conjugation.

We know quite a bit about the structure of algebraic extensions of $\mathbb{Q}_p$. First of all, there is a maximal unramified extension $\mathbb{Q}_p^{\mathrm{ur}}$, and we know that

$$G(\mathbb{Q}_p^{\mathrm{ur}}/\mathbb{Q}_p) \cong G(\overline{\mathbb{F}}_p/\mathbb{F}_p),$$

where $\overline{\mathbb{F}}_p$ is the residue field of the valuation ring of $\mathbb{Q}_p^{\mathrm{ur}}$ (which is an algebraic closure of $\mathbb{F}_p$). The restriction map then gives a surjective homomorphism

$$G(\overline{\mathbb{Q}}_p/\mathbb{Q}_p) \longrightarrow G(\overline{\mathbb{F}}_p/\mathbb{F}_p).$$

The kernel of this homomorphism is called the *inertia group* (at p) and we will denote it by I_p. Recall that $G(\overline{\mathbb{F}}_p/\mathbb{F}_p)$ is topologically generated by the Frobenius automorphism ϕ_p. We will call any lift of ϕ_p to $G_{\mathbb{Q}_p}$ a Frobenius automorphism, and we will confuse things even further by using the same notation ϕ_p for any such element. (To be fair, we usually do this in contexts where we are looking at the image of $G_{\mathbb{Q}_p}$ via a map whose kernel contains I_p. In this case, ϕ_p is any of many elements in a coset of I_p, but the image of ϕ_p is well-defined.)

The structure of the inertia group I_p is somewhat more complicated. It has a large normal Sylow pro-p-subgroup, which we denote by W_p and which is known as the *wild inertia group*. The quotient I_p/W_p is sometimes called the *tame inertia group*, and it is the better understood part of I_p. In fact, there is a (non-canonical) isomorphism

$$I_p/W_p \cong \prod_{\ell \neq p} \mathbb{Z}_\ell,$$

and if ϕ_p is any Frobenius element and $\bar{\sigma} \in I_p/W_p$, we have $\phi_p \bar{\sigma} \phi_p^{-1} = \bar{\sigma}^p$.

Problem 1.17. The group I_p/W_p corresponds to an extension of $\mathbb{Q}_p^{\mathrm{ur}}$. Describe that extension and the map from its Galois group to $\prod_{\ell \neq p} \mathbb{Z}_\ell$. (You will need to make a choice of a compatible sequence of ℓ^n-th roots of unity, which is why the isomorphism is non-canonical.)

The names of these subgroups reflect their origins in the theory of algebraic number fields. In fact, a Galois extension K of $\mathbb{Q}_p$ corresponds to a surjective homomorphism $G_{\mathbb{Q}_p} \longrightarrow G(K/\mathbb{Q}_p)$, and the extension will be called *unramified* if the image of I_p under this map is trivial. Similarly, we'll say the extension is *tamely ramified* if the image of W_p is trivial, and *wildly ramified* if not.

This analysis of the structure of $G_{\mathbb{Q}_p}$ can be continued, producing still smaller subgroups of W_p known as the "higher ramification groups." See, for example, [**135**] for details of all this.

Problem 1.18. Above we worked with a Galois extension $K/\mathbb{Q}_p$. How do we handle an extension which is not Galois? What changes?

Putting together the whole picture of what this says about the full Galois group $G_{\mathbb{Q}}$, we see that for each prime number p we have a complex package of information: a set of subgroups

$$W_p \subset I_p \subset G_{\mathbb{Q}_p} \subset G_{\mathbb{Q}}$$

(with both W_p and I_p normal in $G_{\mathbb{Q}_p}$) and a set of Frobenius elements at p. The last inclusion depends on the choice of the embedding of $\overline{\mathbb{Q}}$ into $\overline{\mathbb{Q}}_p$, and hence the whole picture is only determined up to conjugation.

As before, we can translate this group-theoretical picture in terms of algebraic number fields. A Galois extension $K/\mathbb{Q}$ corresponds to a surjective homomorphism $G_{\mathbb{Q}} \to G(K/\mathbb{Q})$, and the extension $K/\mathbb{Q}$ will be unramified at p when the images of all the inertia groups at p are trivial. Note that in this case there is a well-defined (up to conjugation) image of the Frobenius element ϕ_p in $G(K/\mathbb{Q})$. In finite extensions, all but a finite number of primes will be unramified, and if $K \neq \mathbb{Q}$ at least one prime will be ramified:

Theorem 1.2. *If $K/\mathbb{Q}$ is a finite extension, then K is ramified at finitely many primes (to be specific, they are the primes dividing the discriminant of $K/\mathbb{Q}$). Every non-trivial extension of $\mathbb{Q}$ is ramified at at least one prime.*

The first part of this theorem is relatively easy (and true over a general number field). The second is a theorem of Minkowski (and *not* true over a general number field).

Problem 1.19. Above, we described the "local picture" only for non-archimedean primes. We also need to consider the prime at infinity. Make the proper definitions. In particular, show that there is a well-defined conjugacy class of elements of order two in $G_{\mathbb{Q}}$; we call any element of this conjugacy class a "complex conjugation."

Problem 1.20. Describe all this for G_K, where K is a number field. Does anything significant change?

An important class of extensions of $\mathbb{Q}$ are the *cyclotomic* extensions obtained by adjoining roots of unity to $\mathbb{Q}$. Let m be a positive integer and let ζ_m be a primitive m-th root of unity. Then we know that $G(\mathbb{Q}(\zeta_m)/\mathbb{Q})$ is abelian, and in fact isomorphic to the group of units of $\mathbb{Z}/m\mathbb{Z}$. If we take the union K_ℓ of all $\mathbb{Q}(\zeta_m)$ as m ranges over all the powers of a prime ℓ, these isomorphisms compile to give an isomorphism between $G(K_\ell/\mathbb{Q})$ and $\mathbb{Z}_\ell^\times$. Composing this with the surjective map $G_{\mathbb{Q}} \longrightarrow G(K_\ell/\mathbb{Q})$ gives a homomorphism

$$\epsilon_\ell : G_{\mathbb{Q}} \longrightarrow \mathbb{Z}_\ell^\times.$$

This is called the ℓ-adic *cyclotomic character*; it can be described by saying that, for any ℓ-power root of unity ζ and any element $\sigma \in G_{\mathbb{Q}}$ we have

$$\sigma(\zeta) = \zeta^{\epsilon_\ell(\sigma)}.$$

The extension K_ℓ (or, equivalently, the ℓ-adic cyclotomic character) is ramified only at ℓ and infinity. If $p \neq \ell$, then, it makes sense to talk about the image of a Frobenius element at p under ϵ_ℓ.

Problem 1.21. Prove that if $p \neq \ell$ we have $\epsilon_\ell(\phi_p) = p$.

The Galois group of any cyclotomic extension, and hence also of any Galois subextension, is abelian. The *Kronecker-Weber Theorem* asserts that all abelian extensions of $\mathbb{Q}$ are of this kind. In keeping with our main theme, we can restate this in terms of characters of the Galois group. Given any group G, write G^{ab} for its abelianization, i.e., its (unique) maximal continuous[2] abelian quotient, i.e., the

[2] In general, we will consider only quotients that are quotients in the category of profinite groups, i.e., quotients by closed subgroups. When we want to emphasize this, we will speak of "continuous quotients."

quotient of G by the closed subgroup topologically generated by the commutators of G. Then we have:

Theorem 1.3 (Kronecker-Weber). *For each prime p, let ϵ_p denote the p-adic cyclotomic character. The product of all the ϵ_p, which maps $G_{\mathbb{Q}}$ to the product of all $\mathbb{Z}_p^\times$, induces an isomorphism*

$$(\prod \epsilon_p) : G_{\mathbb{Q}}^{ab} \xrightarrow{\cong} \prod_p \mathbb{Z}_p^\times \cong \hat{\mathbb{Z}}^\times.$$

Problem 1.22. Check that this isomorphism is equivalent to the theorem as it is usually stated: all abelian extensions of $\mathbb{Q}$ are contained in some cyclotomic field.

There is also a local Kronecker-Weber theorem, and it too can be restated in terms of characters. Let $\pi : G_{\mathbb{Q}_p} \longrightarrow G_{\mathbb{F}_p}$ be the standard projection and let ϵ_p be the p-adic cyclotomic character. Then we have:

Theorem 1.4 (Local Kronecker-Weber). *The map $\pi \times \epsilon_p$ induces an isomorphism*

$$\pi \times \epsilon_p : G_{\mathbb{Q}_p}^{ab} \xrightarrow{\cong} G_{\mathbb{F}_p} \times \mathbb{Z}_p^\times.$$

The Kronecker-Weber theorem is the first piece of Class Field Theory over $\mathbb{Q}$. In general, Class Field Theory provides a detailed description of the abelian extensions of any number field, and so it can be used to study Galois groups and their representations. The theory is too complex to summarize here, so we refer readers to the literature for more details. There are several accounts of Class Field Theory available; one of the more accessible ones is [**112**].

There are many other questions to ask about the Galois group of $\mathbb{Q}$. For example, the following is a famous conjecture:

Conjecture. Any finite group can be obtained as a discrete quotient of $G_{\mathbb{Q}}$.

Much work has been done in the direction of this conjecture (for example, see [**138**]), but the full conjecture remains very much open. One reason for mentioning it here is to point out that it implies that $G_{\mathbb{Q}}$ must be quite complicated!

Various attempts have been made to come up with conjectural descriptions of $G_{\mathbb{Q}}$; one of the most interesting is Grothendieck's theory of "dessins d'enfants" (see [**126**] for details).

Restricting the ramification

As we saw above, a finite extension $K/\mathbb{Q}$ can be only be ramified at a finite number of primes. This is not true for infinite extensions (consider the example in Problem 1.7, for example), but there are good reasons to expect that the "natural" Galois representations (more precision later) are all finitely ramified. This section considers the Galois theory with bounded ramification.

We'll fix a finite set S of primes, including the prime at infinity. (This need not be done in general, but for our purposes we'll always want to allow ramification at infinity.) We want to consider extensions $K/\mathbb{Q}$ which are ramified only at the primes belonging to S; we describe these as "unramified outside S." Putting all such K together gives $\mathbb{Q}_S$, the maximal extension of $\mathbb{Q}$ which is unramified outside S. This is easily checked to be a Galois extension of $\mathbb{Q}$; we want to study the group

$$G_{\mathbb{Q},S} = G(\mathbb{Q}_S/\mathbb{Q}).$$

This is a quotient of the full $G_{\mathbb{Q}}$, of course, but in many ways it is much easier to understand.

Before we consider the main results, let's point out that we can make the same definition for a general number field K, and a set of primes[3] S of K (again, including all the archimedean primes), yielding a group that we will call $G_{K,S}$. Notice that in both cases we are putting no restrictions on the ramification at infinity.

Problem 1.23. Suppose S is a set of primes in $\mathbb{Q}$ and S_1 is the set of primes in K lying over the primes in S. Is there a simple description of the relation between G_{K,S_1} and $G_{\mathbb{Q},S}$?

Problem 1.24. Show that any open subgroup of $G_{K,S}$ is again of the form G_{K_1,S_1}, for some finite extension K_1/K.

Problem 1.25. What would change if we decided to restrict the ramification at the archimedean primes?

The first important fact about $G_{K,S}$ is the following finiteness result:

Theorem 1.5 (Hermite-Minkowski). *Let K be a finite extension of $\mathbb{Q}$, let S be a finite set of primes and let d be a positive integer. There are only finitely many extensions F/K of degree d which are unramified outside S.*

An important consequence of this theorem is the fact that the set

$$\mathrm{Hom}_{\mathrm{cont}}(G_{K,S}, \mathbb{Z}/p\mathbb{Z})$$

is finite, since each nontrivial continuous homomorphism corresponds to an extension of degree p, unramified outside S. Putting this together with Problem 1.24 gives the following crucial (for us) result:

Theorem 1.6. *Let p be a prime number, K a number field, and S a finite set of (non-archimedean) primes. Let $G \subset G_{K,S}$ be an open subgroup. Then there exist only a finite number of continuous homomorphisms from G to $\mathbb{Z}/p\mathbb{Z}$.*

Mazur calls this the *p-finiteness condition*, and we will use it in an essential way to understand the deformations of a Galois representation.

Problem 1.26. Show that the p-finiteness condition also holds for any of the $G_{\mathbb{Q}_\ell}$. (Is anything special about the case $\ell = p$?)

How big are the groups we are considering? As we saw, the absolute Galois group of a finite field is topologically generated by one element, the Frobenius. For a local field, one can also show a finite generation result:

Theorem 1.7. *If K is a finite extension of $\mathbb{Q}_p$, then G_K is topologically finitely generated.*

For this and much more about the Galois group of a local field, see [**76**], [**77**], and [**161**], which together give a detailed description of G_K in this case.

The situation for the $G_{K,S}$ is much more complicated. Shafarevich conjectured[4] that this group is also topologically finitely generated, but so far this remains an

[3] We actually get a choice here. We could keep S as a set of primes in $\mathbb{Q}$, and say that an extension of K is "unramified outside S" if it is unramified at all primes of K that do not lie above a prime belonging to S. This point of view has its advantages, but of course choosing a set of primes in K is more general.

[4] The reference is [**142**], but we should note that there Shafarevich simply asks whether it is the case that $G_{K,S}$ is finitely generated for any number field (and whether the number of generators can be bounded in terms of the number of elements of S). His main reason for posing the question is that the analogous statement is true for function fields over $\mathbb{C}$.

open question. (The p-finiteness property, which would follow from finite generation, is a sort of replacement for this still-unknown result.) We do know that $G_{K,S}$ is topologically *countably* generated.

Problem 1.27. Prove this. In other words, show that there exists a countable set of elements that generate a dense subgroup of $G_{K,S}$. (This is actually quite easy.)

Problem 1.28. (A test situation; as far as I know, this is an open problem.) For each elliptic curve E defined over $\mathbb{Q}$, with good reduction outside 2, let $\mathbb{Q}(T_2(E))$ be the extension of $\mathbb{Q}$ obtained by adjoining the coordinates of the 2^n-division points for all n (equivalently, it is the field fixed by the kernel of the 2-adic representation attached to E). Let K be the compositum of the $\mathbb{Q}(T_2(E))$ as E runs through all such elliptic curves. Is $G(K/\mathbb{Q})$ topologically finitely generated?

For every prime p, we can carry the local picture from the previous section over to $G_{\mathbb{Q},S}$ and get homomorphisms $G_{\mathbb{Q}_p} \longrightarrow G_{\mathbb{Q},S}$. When $p \notin S$, the image of the inertia group I_p is trivial, and therefore there is a well-defined Frobenius element ϕ_p in $G_{\mathbb{Q},S}$ which generates the image of $G_{\mathbb{Q}_p}$. It seems natural to conjecture[5] that the image of $G_{\mathbb{Q}_p}$ in $G_{\mathbb{Q},S}$ is as large as possible:

Conjecture. With the notations above, we have:

i. If $p \in S$, the map $G_{\mathbb{Q}_p} \longrightarrow G_{\mathbb{Q},S}$ is an inclusion.

ii. If $p \notin S$, the kernel of the map $G_{\mathbb{Q}_p} \longrightarrow G_{\mathbb{Q},S}$ is exactly I_p, so that we get an inclusion $G_{\mathbb{Q}_p}/I_p \hookrightarrow G_{\mathbb{Q},S}$.

This very natural conjecture seems to be quite difficult to prove.

As pointed out above, for each $p \notin S$ we have a well-defined Frobenius element ϕ_p in the image of $G_{\mathbb{Q}_p}$. The whole local picture, however, is defined only up to conjugation, so if we do not want to fix the homomorphisms $G_{\mathbb{Q}_p} \longrightarrow G_{\mathbb{Q},S}$ we should think of the Frobenius element as a conjugacy class of elements of $G_{\mathbb{Q},S}$. One of the most significant results about the set of all these conjugacy classes is the following density result.

Theorem 1.8 (Chebotarev). *Let $K/\mathbb{Q}$ be a Galois extension that is unramified outside a finite set S of primes. Let T be a finite set of primes containing S. For each prime $p \notin T$, there is a well-defined Frobenius conjugacy class $[\phi_p] \subset G(K/\mathbb{Q})$. The union of all these Frobenius conjugacy classes is dense in $G(K/\mathbb{Q})$.*

Problem 1.29. What does this say when K is a *finite* extension of $\mathbb{Q}$? (Easy question, but the fact is worth noting.)

Problem 1.30. Let ζ_m be a primitive m-th root of unity, and let $K = \mathbb{Q}(\zeta_m)$. In this case the Galois group is known completely explicitly, and we also know what the Frobenius elements are. What does the Chebotarev theorem tell us in this situation?

Problem 1.31. Is the set of (topological) generators given by the Chebotarev theorem countable?

Finally, the abelianization of $G_{\mathbb{Q},S}$ is easily understood by using the Kronecker-Weber theorem.

Problem 1.32. Show that $G^{\mathrm{ab}}_{\mathbb{Q},S}$ is isomorphic to $\prod_{p\in S} \mathbb{Z}_p^\times$.

As before, we can use Class Field Theory to understand the abelianization in the case of a number field.

[5] Thanks to Ralph Greenberg for mentioning this issue to me.

Galois representations

Why consider the representations of $G_{\mathbb{Q},S}$? One of the reasons is simply that such representations arise naturally, for example from the theory of elliptic curves and modular forms. Another reason, as Mazur has pointed out, is the fact that the whole picture we want to study, which includes not only $G_{\mathbb{Q},S}$ but also all the maps $G_{\mathbb{Q}_p} \longrightarrow G_{\mathbb{Q},S}$, is only defined up to conjugation. Group representations are well-suited to this situation. For example, for $p \notin S$ the Frobenius elements ϕ_p are only defined up to conjugation, but the characteristic polynomial of the image of ϕ_p under a representation is well-defined (and therefore so are the trace and determinant of the image).

Let's make the formal definition:

Definition 1.3. A Galois representation (defined over A, unramified outside S) is a continuous homomorphism

$$\rho : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_n(A),$$

where A is some topological ring and n is a positive integer. Two Galois representations ρ_1 and ρ_2 are *equivalent* if there is a matrix $P \in \mathrm{GL}_n(A)$ such that $P^{-1}\rho_1 P = \rho_2$.

Given such a thing, we can consider the free A-module of rank n and give it a continuous action of $G_{\mathbb{Q},S}$ by defining $g \cdot m = \rho(g)m$. Conversely, given a finite free A-module M of rank n with a continuous action of $G_{\mathbb{Q},S}$, we can get a representation ρ as above by choosing a basis for M. Changing the basis changes ρ into an equivalent representation.

If we have a finite free A-module M with a continuous action of a profinite group G such that

$$M = \varinjlim_{H} M^H$$

as H runs through the open normal subgroups of G, then we can canonically make M into a module over the completed group ring $A[[G]]$, defined as

$$A[[G]] = \varprojlim_{H} A[G/H],$$

where H runs through the open normal subgroups of G and $A[G/H]$ is the usual group ring of the finite group G/H over A. In problem 1.14, we checked that the condition on M is automatically verified when A is a profinite ring. Hence, giving (up to equivalence) a representation of G defined over a profinite ring A is the same as giving a continuous $A[[G]]$-module M which is finite and free as an A-module. This point of view is also occasionally useful.

There is one final point of view which is occasionally useful. Given a representation

$$\rho : G \longrightarrow \mathrm{GL}_n(A)$$

defined over a profinite ring A, we can extend it by linearity to the completed group ring $A[[G]]$, to get a continuous homomorphism of A-algebras

$$A[[G]] \longrightarrow \mathrm{M}_n(A).$$

Conversely, the restriction to G of any such homomorphism gives a representation in the usual sense.

In what follows, we will mostly stick to the first point of view, but every once in a while, when A is a profinite ring (as it will often be), it will be convenient to switch to the other versions.

The standard choices for the ring A are the following:

i. $A = \mathbb{C}$. These are known as "Artin representations," and are the most classical. Because of the topologies involved, the image of $G_{\mathbb{Q},S}$ in $\mathrm{GL}_n(A)$ must be *finite.*

ii. A is a finite field. These representations arise naturally from elliptic curves and modular forms, and they are the ones that Serre's conjecture tries to describe.

iii. $A = \mathbb{Z}_p$ or $\mathbb{Q}_p$ or finite extensions thereof. These also arise from elliptic curves and modular forms. Since $\mathbb{Z}_p$ (or a finite extension) also carries a profinite topology, this case gives the best "match" in topologies. In particular, the image of $G_{\mathbb{Q},S}$ is not necessarily finite in this case.

Our main interest is in the last two cases, and in the relation between them, so we will choose to work with rings A that are generalizations of those two situations. Specifically, we will assume A is a complete noetherian local ring with finite residue field. Note, by problem 1.13, that such an A is automatically a profinite ring.

In a sense, we are interested in trying to understand all the (finitely ramified) Galois representations into $\mathrm{GL}_n(A)$. For $n = 1$, this is essentially already done, since describing all such representations amounts to describing the abelianizations G_K^{ab}, and this is basically what Class Field Theory does. (If $K \neq \mathbb{Q}$ this is not quite true, since one runs into such difficulties as Leopoldt's Conjecture, but one still has quite good control of the situation.) Hence, we'll focus on $n \geq 2$. In fact, things are already so interesting for $n = 2$ that we'll often restrict ourselves to that case, which is also the case where Serre's conjecture applies and where we get representations from elliptic curves and modular forms.

Studying "all the representations" is far too vague to serve as a guideline for investigation, however, and so we have to come up with a more specific program. The point of view we will take, then, is to start with a given representation into $\mathrm{GL}_n(\mathsf{k})$, where k is a finite field, and then to consider all the representations which "lift" this representation to $\mathrm{GL}_n(A)$, where A runs through all complete noetherian local rings with residue field k. It turns out that we can make this into a well-defined question and (even better!) that the question has an interesting answer.

Complements to Lecture 1

The first lecture makes intensive use of both the notion of an inverse limit and the theory of profinite groups. We give brief summaries of each of these, with some references.

Inverse limits

Inverse limits make sense for various kinds of mathematical objects. We could phrase everything in the language of categories (as in the next lecture), but we avoid that for now. Hence, we discuss inverse limits of sets, of groups, and of rings; the reader should note, however, that we do not really use many specific properties of these objects.

We start with a partially ordered set I, which we want to be a *directed set.* This just means that given $i, j \in I$ there exists $k \in I$ such that $i \leq k$ and $j \leq k$. For example, I could be the set of all positive integers with the usual order, or the set of positive integers ordered by divisibility.

To give an *inverse system* we must specify:

- a directed set I,
- for each $i \in I$, a group (or ring, or set) G_i,
- for each pair $i, j \in I$ such that $i \leq j$, a group (or ring, or set) homomorphism $\phi_{ij} : G_j \longrightarrow G_i$.

(Of course, a "set homomorphism" is just an arbitrary function.) We could add requirements of continuity if our objects carried natural topologies, and so on. We require that this data satisfy the "obvious" conditions:

- ϕ_{ii} is the identity, and
- "all triangles commute," that is, if $i \leq j \leq k$ then $\phi_{ik} = \phi_{ij} \circ \phi_{jk}$.

Here are some famous examples:

Problem 1.33. Let I be the set of positive integers with the usual order, let p be a fixed prime number, and for each $n \in I$ let G_n be the ring $\mathbb{Z}/p^n\mathbb{Z}$. If $n \leq m$, let ϕ_{nm} be "reduction mod p^n." Check that this defines an inverse system.

Problem 1.34. Let I be the set of positive integers ordered by divisibility. For each $n \in I$, get G_n be the ring $\mathbb{Z}/n\mathbb{Z}$, and whenever $n|m$ let ϕ_{nm} be "reduction modulo n." Check that this defines an inverse system.

Problem 1.35. Let F/K be a field extension, and let I be the set of finite Galois subextensions, ordered by inclusion. For each $K' \in I$, let $G_{K'} = G(K'/K)$, and if $K' \subset K''$ let $\phi_{K'K''}$ be the restriction map $G(K''/K) \longrightarrow G(K'/K)$. Check that this defines an inverse system.

Of course, we can also make a trivial inverse system by making all the G_i be the same and taking all the maps to be the identity map.

Now we can define an inverse limit. Given an inverse system of groups (or rings, or sets) as above, we'll say a group (or ring, or set) G is the *inverse limit* of the system if it satisfies two conditions:

- G comes equipped with homomorphisms $\psi_i : G \longrightarrow G_i$ for every $i \in I$ making all triangles commute: if $i \leq j$, then $\psi_i = \phi_{ij} \circ \psi_j$.
- G is "universal" among groups (rings, sets) with this property, i.e., given any other group (ring, set) G' with such a set of homomorphisms, there exists a unique homomorphism $G' \longrightarrow G$ through which they all factor.

In this case, we write

$$G = \varprojlim_i G_i.$$

(This notation is somewhat abusive, since it doesn't sufficiently specify the inverse system. In most cases, however, it's easy to figure out which inverse system is meant.)

Of course, this kind of definition is useless without some kind of concrete construction to show that such things exist. So here is a constructive description of G (that can also serve as proof that inverse limits of groups, rings, and sets exist).

Consider the product

$$P = \prod_{i \in I} G_i.$$

Elements of P are sequences indexed by I, which we write as $(g_i)_{i\in I}$, where $g_i \in G_i$. We get the inverse limit G by considering the piece of this product set which contains the "coherent" sequences, i.e.,

$$G = \{(g_i)_{i\in I} | \phi_{ij}(g_j) = g_i \text{ whenever } i \leq j\}.$$

Problem 1.36. Check that this works, that is, that this G is a group (ring, set) and is the inverse limit.

Problem 1.37. Check that if each of the G_i is a topological group and the ϕ_{ij} are all continuous, then G is the inverse limit of the G_i as topological groups.

In most of the situations we work with, the G_i are *finite* groups (rings, sets). If we give them the discrete topology, they are then compact as topological spaces. The product P, with the product topology, is then compact, and it's easy to see that G is a closed subset of P. Hence G carries a natural compact Hausdorff topology.

The inverse limits of the three inverse systems we considered above are the ring $\mathbb{Z}_p$ of p-adic integers, the ring $\hat{\mathbb{Z}}$, and the Galois group $G(F/K)$. The first of these is a good example on which to test the theory, particularly if you think of $\mathbb{Z}_p$ in terms of p-adic expansions. (See [**59**, Sections 1.2 and 3.3] for a very elementary treatment and [**133**, Chapter 2] for a more sophisticated version.)

Finally, two problems to give you a chance to play with these ideas and extend them at the same time:

Problem 1.38. Define an "exact sequence of inverse systems" and decide whether inverse limits preserve exactness. (Not easy!)

Problem 1.39. Define "direct systems" and "direct limits" by turning all the arrows around. As an example, let I be the positive integers ordered by divisibility, and for each n let $G_n = \mathbb{Z}/n\mathbb{Z}$ (thought of as an additive group). Whenever $n|m$, define $\phi_{mn} : \mathbb{Z}/n\mathbb{Z} \longrightarrow \mathbb{Z}/m\mathbb{Z}$ by mapping 1 to m/n. Can you describe the direct limit?

Profinite groups

The definition of a profinite group is a direct application of the ideas we have just discussed:

Definition 1.4. A *profinite group* is a topological group which can be represented as the inverse limit of an inverse system of finite groups (thought of as carrying the discrete topology).

It follows that profinite groups are compact, by the discussion above. Some of the topological properties of profinite groups we discussed (or set as problems) in the lecture. Here are three others: suppose G is a profinite group, so that

$$G = \varprojlim_i G_i,$$

and let $K_i = \operatorname{Ker}(\psi_i : G \longrightarrow G_i)$. Then, since G_i is discrete, K_i is an open subgroup of G.

Problem 1.40. Show that the K_i, $i \in I$, form a basis of open neighborhoods of the identity in G.

Problem 1.41. Show that a subgroup of G is open if and only if it is closed and of finite index.

Problem 1.42. Show that any closed subgroup of a profinite group is the intersection of all open subgroups containing it.

As we pointed out in the lecture, profinite groups are *totally disconnected*, that is, the connected component of any point is the singleton set consisting only of that point. It turns out that this is equivalent to being profinite:

Theorem 1.9. *Let G be a (Hausdorff and) compact topological group. The following are equivalent:*

i. G is profinite,
ii. G is totally disconnected,
iii. G has a set of open normal subgroups which is a full system of neighborhoods of the identity.

See [**143**] for a proof.

The property of being profinite is preserved under taking closed subgroups, quotients (by closed subgroups), arbitrary direct products, and inverse limits. As the lecture suggests, many of the properties of finite subgroups still make sense for profinite subgroups. For example, it makes sense to talk about pro-p-groups, meaning profinite groups all of whose finite discrete quotients are p-groups.

For more on profinite groups, see [**109**, Appendix C], [**139**] (or its English translation [**140**]), and [**143**].)

LECTURE 2
Deformations of Representations

The basic situation we want to study is as follows. We are given either a number field K and a finite set of primes S, or a local field F, and we are given a representation of either $G_{K,S}$ or G_F into $\mathrm{GL}_n(\mathsf{k})$, where k is a finite field. We want to try to understand all possible lifts of this representation to $\mathrm{GL}_n(A)$, where A is a complete noetherian local ring with residue field k. It is not exactly clear, of course, what "understand all possible lifts" means, and so the main goal of this lecture is to make our question precise. We begin by discussing some of the historical motivation for the theory, then develop (the simplest form of) the precise deformation problem we want to study.

Why deform Galois representations?

Nowadays, the obvious reason to study deformations of Galois representations is that they played a crucial role in the proof of the modularity conjecture for elliptic curves over $\mathbb{Q}$ (work of Wiles, Taylor, Diamond, Breuil, and Conrad). However, the theory predates that work, and so the original motivation was different.

Historically, the first (p-adic) Galois representations to be carefully studied were those coming from elliptic curves. Every elliptic curve defined over $\mathbb{Q}$ with good reduction outside a set of primes S gives us representations of $G_{\mathbb{Q},S}$ into both $\mathrm{GL}_2(\mathbb{Z}_p)$ and $\mathrm{GL}_2(\mathbb{F}_p)$, and (usually) the representation over $\mathbb{Z}_p$ is a "lift" of the representation over $\mathbb{F}_p$. This already gives us a first example of a residual representation and one of its deformations.

The second classical source of Galois representations are modular forms, and once again one sees the same pattern: one gets a pair of representations, in characteristic zero and in characteristic p, which are (usually[1]) an example of a residual representation and a deformation.

The real push towards a careful study of such deformations, however, seems to have been inspired by Hida's results on the theory of ordinary p-adic modular forms, which yielded representations which were described at the time as "very large." In particular, specializing Hida's large representations in different ways produced a

[1] We say "usually" because of the following problem. We'll often want to work with a *semisimple* residual representation, which means we'll sometimes have to pass from a representation to its semisimplification. This means that the residual representation coming, for example, from a modular form may fail to be the reduction of the representation in characteristic zero..

large number of deformations of a residual representation, some of which did not look much like any of the "usual" representations. This seems to have led Mazur, in his seminal paper [**97**], to pose the question of understanding the deformations in general.

To explain this more fully, we give a very loose description of Hida's work on ordinary p-adic modular forms (see [**71**], [**70**] or [**72**], for example), focusing only on the features that are relevant to our theme. Our summary assumes the reader is familiar with the standard theory of modular forms; see [**13**].

Fix a prime $p \geq 5$, an integer N, not divisible by p, and an integer $k \geq 2$. Let $M_k(\mathrm{N}, \mathbb{Z}_p)$ be the $\mathbb{Z}_p$-module of modular forms of weight k on $\Gamma_1(\mathrm{N}) \cap \Gamma_0(p)$ with coefficients in $\mathbb{Z}_p$, and consider the submodule $M_k(\mathrm{N}, \mathbb{Z}_p)^0$ of "ordinary" modular forms, i.e., the submodule spanned by the eigenforms for the $\mathrm{U} = \mathrm{U}_p$ operator whose eigenvalues are p-adic units. (We could consider finite extensions of $\mathbb{Z}_p$ in exactly the same way.) Hida constructed a Hecke algebra $\mathbf{T}^0 = \mathbf{T}^0(\mathrm{N})$ attached to the whole family of spaces $M_k(\mathrm{N}, \mathbb{Z}_p)^0$, for $k \geq 2$. If we let $\Gamma = 1 + p\mathbb{Z}_p$ and $\Lambda = \mathbb{Z}_p[[\Gamma]]$ be the usual Iwasawa algebra, $\mathbf{T}^0$ is a finite flat Λ-algebra, and any Hecke eigenform in any of our spaces corresponds to a homomorphism $\mathbf{T}^0 \longrightarrow \mathbb{Z}_p$.

Suppose we have such an eigenform $f \in M_k(\mathrm{N}, \mathbb{Z}_p)^0$. Write $\bar{f}$ for the reduction of f modulo p, which we can think of as a modular form of weight k over the finite field $\mathbb{F}_p$; the fact that f is ordinary translates, modulo p, into the assertion that $\bar{f}$ is not in the kernel of U. Because it is an eigenform, $\bar{f}$ corresponds to a homomorphism $\mathbf{T}^0 \longrightarrow \mathbb{F}_p$. The kernel of this homomorphism is a maximal ideal $\mathfrak{m} = \mathfrak{m}(f) \subset \mathbf{T}^0$. Let $R(f) = \mathbf{T}^0_{\mathfrak{m}}$ be the completion of $\mathbf{T}^0$ at the ideal $\mathfrak{m}$. Then $R(f)$ is a complete local Λ-algebra, and is finite and flat over Λ.

Now we bring in Galois representations. One of the fundamental facts in the theory is that every time we have a Hecke eigenform, it gives rise to a two-dimensional Galois representation. Since our f is an eigenform, we get a representation

$$\rho_f : G_{\mathbb{Q}} \longrightarrow \mathrm{GL}_2(\mathbb{Z}_p).$$

Reducing modulo p (and taking the semi-simplification if necessary, see below) gives a representation

$$\overline{\rho}_f : G_{\mathbb{Q}} \longrightarrow \mathrm{GL}_2(\mathbb{F}_p).$$

Hida's work showed that there exists a "big" representation

$$\rho_H : G_{\mathbb{Q}} \longrightarrow \mathrm{GL}_2(R(f))$$

which "interpolates" all the representations ρ_g coming from ordinary eigenforms $g \in M_{k'}(\mathrm{N}, \mathbb{Z}_p)^0$ for various weights k' such that the q-expansions of f and g coincide modulo p.

More precisely, suppose we find an eigenform $g \in M_{k'}(\mathrm{N}, \mathbb{Z}_p)^0$ such that $\bar{f} = \bar{g}$, where bars indicate reduction modulo p. Then it turns out that the "residual representations" $\overline{\rho}_f$ and $\overline{\rho}_g$ are the same (up to equivalence). Hida's theorem says that there exists a homomorphism $R(f) \longrightarrow \mathbb{Z}_p$ such that composing ρ_H with this homomorphism gives ρ_g. Thus, our big representation is somehow parametrizing all lifts of the residual representation $\overline{\rho}_f$ which are of a certain type. We can think of ρ_H as an *analytic family* of Galois representations, all of which have the same reduction modulo p. It is natural then, to ask about such families in general, and that question leads at once to the deformation theory.

Furthermore, it quickly became clear that there are other homomorphisms $R(f) \longrightarrow \mathbb{Z}_p$, ones that *do not* correspond to modular forms (or at least not to the usual kind of modular forms). Specializing the representation via one of these homomorphisms can produce rather strange representations (for example, see the final sections of [**106**]). This suggests, once again, that it makes sense to try to study "all the deformations."

The deformation functor

Mazur created the theory of deformations of Galois representations in his paper [**97**], which is one of the fundamental references for this section (and for much of what follows also). See also Mazur's notes from the Boston University conference, [**101**], and the notes [**42**] by Doran and Wong.

What we want to do is imitate the situation in Hida's theory, in maximal generality. So we'll start with a profinite group Π (which will later be a Galois group of some kind) and a representation of Π into matrices over a finite field. The basic question will be: can we describe all lifts of this representation to (appropriate) p-adically complete rings?

Let k be a finite[2] field of characteristic p. For this section, we make no assumptions on p. We will want to start with a representation into $\mathrm{GL}_n(\mathsf{k})$, and consider its lifts.

Let Π be a profinite group. In order for the theory to work, we need to know that Π satisfies the finiteness condition which we considered, in the case of Galois groups, in our first lecture.

Condition Φ_p: *For every open subgroup of finite index $\Pi_0 \subset \Pi$ there exist only a finite number of continuous homomorphisms $\Pi_0 \longrightarrow \mathbb{F}_p$.*

We already know, by Theorem 1.6 and Problem 1.26, that the Condition Φ_p holds for $G_{\mathbb{Q}_\ell}$ and for $G_{\mathbb{Q},S}$, where ℓ is a prime and S is a finite set of primes. It's worth pointing out that the condition can be stated in several equivalent forms.

Let's first set up some notation. First, the *pro-p-completion* of the profinite group Π is

$$\Pi^{(p)} = \varprojlim_{N} \Pi/N,$$

where we take the limit over all closed normal subgroups whose index is (finite and) a power of p. Second, the *p-Frattini quotient* of Π is the maximal continuous abelian quotient of Π which is of exponent p.

Problem 2.1. Show that the p-Frattini quotient exists and that it is the image of a surjective continuous homomorphism from $\Pi^{(p)}$.

The following lemma gives several equivalent ways of stating condition Φ_p.

Lemma 2.1. *Let Π_0 be a profinite group. The following conditions are equivalent:*

i. the pro-p-completion of Π_0 is topologically finitely generated,

ii. the abelianization of the pro-p-completion of Π_0 is a $\mathbb{Z}_p$-module of finite rank,

iii. the p-Frattini quotient of Π_0 is finite,

iv. the set of continuous homomorphisms from Π_0 to $\mathbb{F}_p$ is finite.

[2] The construction works just as well for something like the algebraic closure of $\mathbb{F}_p$, but the case of a finite field is the most significant for us, and we'll simplify things by restricting ourselves to this case.

Proof. Clearly a set of topological generators of the pro-p-completion becomes a set of generators over $\mathbb{Z}_p$ in the abelianization, and becomes a basis of the p-Frattini-quotient as a vector space over $\mathbb{F}_p$. Hence, it's clear that (1) $\Rightarrow$ (2) $\Rightarrow$ (3). Since any homomorphism $\Pi_0 \longrightarrow \mathbb{F}_p$ must factor through the p-Frattini-quotient, (3) and (4) are equivalent. To conclude the proof, we use the profinite version of the Burnside Basis Theorem, which says that if the image in the p-Frattini quotient of a set $\{g_1, \ldots, g_r\}$ of elements of the pro-p-group $\Pi_0^{(p)}$ is a basis for the quotient as a vector space over $\mathbb{F}_p$ then $g_1, \ldots, g_r$ topologically generate $\Pi_0^{(p)}$. It follows that (3) $\Rightarrow$ (1), and we're done. □

Problem 2.2. (The pro-p version of the Burnside Basis Theorem) Let G be a pro-p-group, i.e., an inverse limit of finite p-groups, and let $\mathrm{Fr}(G)$ be its pro-p-Frattini quotient. Prove that any lifting to G of a basis of $\mathrm{Fr}(G)$ as a vector space over $\mathbb{F}_p$ is a set of topological generators for G.

As we saw above, (finitely ramified) Galois groups satisfy condition Φ_p. This can be vastly generalized; see [**84**].

Having stated our crucial assumption about the profinite group Π, let's go on to talk about its representations, according to the "program" we have outlined. We want to start, then, with a homomorphism

$$\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k}),$$

and we want to consider lifts of $\overline{\rho}$, that is, homomorphisms

$$\rho : \Pi \longrightarrow \mathrm{GL}_n(R)$$

where R is a ring together with a homomorphism $\pi : R \longrightarrow \mathsf{k}$ such that the image of ρ under the homomorphism $\mathrm{GL}_n(R) \longrightarrow \mathrm{GL}_n(\mathsf{k})$ induced by π is our residual representation $\overline{\rho}$, i.e., the diagram

$$\begin{tikzcd}
 & \mathrm{GL}_n(R) \arrow[d, "\pi"] \\
\Pi \arrow[ur, "\rho"] \arrow[r, "\overline{\rho}"] & \mathrm{GL}_n(\mathsf{k})
\end{tikzcd}$$

is commutative. If we want to do this correctly, however, we need to be a bit more precise about the rings we will be considering. So we set this up in the language of categories.

Choose and fix a finite field k of characteristic p. Let $\mathcal{C}$ denote the category whose objects are complete noetherian local rings with residue field k and whose morphisms are local homomorphisms $R_1 \longrightarrow R_2$ of complete noetherian local rings which induce the identity on k.[3] In particular, this means that if $\mathfrak{m}$ is the maximal ideal of R, then we are requiring that, first, $R/\mathfrak{m} = \mathsf{k}$, and second,

$$R = \varprojlim_j R/\mathfrak{m}^j.$$

(It is useful to recall that by the Krull intersection theorem the intersection of the $\mathfrak{m}^j$ is 0, which is equivalent to saying that the natural topology on a noetherian local ring is always Hausdorff.)

[3] To be absolutely precise, we need to make our objects be complete noetherian local rings $(R, \mathfrak{m})$ together with a *fixed* isomorphism $R/\mathfrak{m} \cong ks$, but we'll refrain from being picky about this.

Sometimes it is also convenient to consider the full subcategory $\mathcal{C}^0$ whose objects are artinian local rings with residue field k. (Notice that the maximal ideal of an artinian local ring is always nilpotent, and hence such rings are automatically complete and noetherian.) Following Mazur, we will call the objects of $\mathcal{C}$ "coefficient rings" and we will call the morphisms "coefficient ring homomorphisms."

Notice that (as is implicit from our use of the word "complete") all coefficient rings carry a natural topology, in which the powers of the maximal ideal are a basis of neighborhoods of 0. Coefficient ring homomorphisms are continuous with respect to this topology.

Problem 2.3. Prove that objects of $\mathcal{C}$ are pro-objects of $\mathcal{C}^0$. Specifically, prove that if R is a complete noetherian local ring with maximal ideal $\mathfrak{m}$, then for every n the quotient $R/\mathfrak{m}^n$ is an object of $\mathcal{C}^0$, and R is the inverse limit of the $R/\mathfrak{m}^n$.

Problem 2.4. With the notations in the previous problem, show that the topology on $R/\mathfrak{m}^n$ is discrete, and that the topology on R is the inverse limit topology.

Problem 2.5. How serious are the restrictions on "coefficient ring homomorphisms?" Find examples of ring homomorphisms between objects of $\mathcal{C}$ which are not "coefficient ring homomorphisms."

The "simplest" example of a (non-artinian) element of $\mathcal{C}$ is the ring $W(\mathsf{k})$ of Witt vectors. Since k is finite, this is simply the (unique) unramified extension of $\mathbb{Z}_p$ whose residue field is k. When $\mathsf{k} = \mathbb{F}_p$, then $W(\mathsf{k})$ is $\mathbb{Z}_p$ itself. See [**135**] for more information on rings of Witt vectors.

Problem 2.6. Show that any coefficient ring R in $\mathcal{C}$ carries a canonical $W(\mathsf{k})$ algebra structure. (That is, show that every such R has a unique coefficient ring homomorphism $W(\mathsf{k}) \longrightarrow R$.)

Problem 2.7. Show that in fact every coefficient ring is a quotient of a power series ring in several variables with coefficients in $W(\mathsf{k})$.

As the last two problems show, our coefficient rings are automatically $W(\mathsf{k})$-algebras. It often happens, however, that we want to modify this somewhat. For example, suppose that we start the game with a representation that comes from a modular form. Then we have at hand not only a residual representation defined over a finite field k, but also a particular lift to a discrete valuation ring $\mathcal{O}$ that may very well not be $W(\mathsf{k})$. In such a situation, we may decide that we want to restrict the whole game to coefficient rings that are $\mathcal{O}$-algebras. This amounts to working in a slightly different category.

Let Λ be an object of $\mathcal{C}$, that is a complete noetherian local ring with residue field k. We'll define $\mathcal{C}_\Lambda$ to be the category whose objects are complete noetherian local Λ-algebras with residue field k and whose morphisms are coefficient-ring homomorphisms which are also Λ-algebra homomorphisms. As before we let $\mathcal{C}^0_\Lambda$ be the full subcategory of artinian Λ-algebras with residue field k. Of course, $\mathcal{C}$ is the same as $\mathcal{C}_{W(\mathsf{k})}$.

Problem 2.8. Is it true that every element of $\mathcal{C}_\Lambda$ is a quotient of a power series ring in several variables over Λ?

Given a coefficient ring R (i.e. an object of $\mathcal{C}$, or of $\mathcal{C}_\Lambda$ if we have fixed a different base ring), we will write π for the canonical projection $R \longrightarrow \mathsf{k}$ and also, by abuse of language, for the map it induces from $\mathrm{GL}_n(R)$ to $\mathrm{GL}_n(\mathsf{k})$. Finally, we let

$$\Gamma_n(R) = \operatorname{Ker}\left(\mathrm{GL}_n(R) \xrightarrow{\pi} \mathrm{GL}_n(\mathsf{k})\right).$$

Definition 2.1. Let R be a coefficient ring. We say two homomorphisms

$$\rho_1, \rho_2 : \Pi \longrightarrow \mathrm{GL}_n(R)$$

are *strictly equivalent* if there exists $M \in \Gamma_n(R)$ such that $\rho_1 = M^{-1}\rho_2 M$.

The idea, of course, is that strictly equivalent homomorphisms give the same homomorphism when we compose them with $\pi : R \longrightarrow \mathsf{k}$. This will give the right notion of equivalence for the theory we have in mind.

Suppose now that we start with a representation (i.e., a continuous group homomorphism)

$$\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k}).$$

We will call this a *residual representation.* For the rest of this section we will assume that we have chosen and fixed such a residual representation. We are finally ready to make the crucial definition.[4]

Definition 2.2. Let $\overline{\rho}$ be a residual representation and let R be a coefficient ring. A *deformation* of $\overline{\rho}$ to R is a strict equivalence class of continuous homomorphisms

$$\rho : \Pi \longrightarrow \mathrm{GL}_n(R)$$

which reduce to $\overline{\rho}$ via the projection π, that is, such that $\pi \circ \rho = \overline{\rho}$.

$$\begin{tikzcd} & \mathrm{GL}_n(R) \arrow[d, "\pi"] \\ \Pi \arrow[ur, "\rho"] \arrow[r, "\overline{\rho}"'] & \mathrm{GL}_n(\mathsf{k}) \end{tikzcd}$$

If we want to be precise, we really have to say that we require $\pi \circ \varphi = \overline{\rho}$ for any homomorphism φ in the strict equivalence class of ρ. Of course, this will be true for all φ in the strict equivalence class if and only if it is true for one of them, so this quibble isn't really very serious. We will, in fact, routinely confuse a homomorphism with its strict equivalence class, and deal with the possible confusions this will generate as they arise.

We think of this as defining a functor

$$\mathbf{D} = \mathbf{D}_{\overline{\rho}} : \mathcal{C} \rightsquigarrow \underline{\mathrm{Sets}}$$

where

$$\mathbf{D}_{\overline{\rho}}(R) = \{\text{deformations of } \overline{\rho} \text{ to } R\}.$$

Similarly, we define the functor $\mathbf{D}_{\overline{\rho},\Lambda}$ by restricting to the subcategory $\mathcal{C}_\Lambda$. We will often drop the $\overline{\rho}$ from the notation, since our residual representation will typically be fixed for the whole discussion.

Lemma 2.2. $\mathbf{D}$ *and* $\mathbf{D}_\Lambda$ *are functors.*

[4] People who are familiar with deformation theory in a geometric context should note that we are really talking of "infinitesimal" or "formal" deformations here.

Problem 2.9. Prove the lemma. (The main things here are to recall what it means to be a functor and to remember that deformations are strict equivalence classes of homomorphisms.)

Problem 2.10. Translate this into the language of free R-modules with a continuous action of Π. A residual representation becomes a k-vector space with a continuous action of Π and a deformation must be some kind of free R-module with a continuous action of Π. Make the appropriate definition and compare the resulting functor with the one(s) we have just defined.

Recall that the categories $\mathcal{C}$ and $\mathcal{C}^0$ are related because objects of $\mathcal{C}$ are pro-objects of $\mathcal{C}^0$. Specifically, if R is a complete noetherian local ring with residue field k, then, if $\mathfrak{m}$ is the maximal ideal of R, we have

$$R = \varprojlim_k R/\mathfrak{m}^k.$$

Now, if we have a functor $\mathbf{F}$ on $\mathcal{C}$, the sets $\mathbf{F}(R/\mathfrak{m}^k)$ will form an inverse system, and there will be compatible morphisms $\mathbf{F}(R) \longrightarrow \mathbf{F}(R/\mathfrak{m}^k)$. These compile to give a canonical morphism

$$\mathbf{F}(R) \longrightarrow \varprojlim_k \mathbf{F}(R/\mathfrak{m}^k).$$

Definition 2.3. We say a functor $\mathbf{F}$ on $\mathcal{C}$ is *continuous* when the canonical morphism

$$\mathbf{F}(R) \longrightarrow \varprojlim_k \mathbf{F}(R/\mathfrak{m}^k)$$

is an isomorphism.

Lemma 2.3. $\mathbf{D}$ *and* $\mathbf{D}_\Lambda$ *are continuous functors.*

Proof. We work with $\mathbf{D}$; the proof for $\mathbf{D}_\Lambda$ is the same.

Recall, first, that

$$\mathrm{GL}_n(R) = \varprojlim_k \mathrm{GL}_n(R/\mathfrak{m}^k)$$

and

$$\Gamma_n(R) = \varprojlim_k \Gamma_n(R/\mathfrak{m}^k).$$

Furthermore, note that the maps

$$\mathrm{GL}_n(R/\mathfrak{m}^{k+1}) \longrightarrow \mathrm{GL}_n(R/\mathfrak{m}^k)$$

and

$$\Gamma_n(R/\mathfrak{m}^{k+1}) \longrightarrow \Gamma_n(R/\mathfrak{m}^k)$$

are all surjective.

If deformations were simply homomorphisms, the continuity would now follow at once. However, deformations are strict equivalence classes of homomorphisms, and so we have to be a bit more careful.

The canonical map

$$\mathbf{D}(R) \longrightarrow \varprojlim_k \mathbf{D}(R/\mathfrak{m}^k)$$

maps a deformation $\rho = \rho_R$ of $\overline{\rho}$ to R to the coherent sequence $\{\rho_k\}$, where ρ_k is the deformation to $R/\mathfrak{m}^k$ obtained by reducing (any homomorphism representing) ρ modulo $\mathfrak{m}^k$.

To show that the canonical map is surjective, we need to show that any coherent sequence $\{\rho_k\}$ comes from a deformation ρ to R. For this, it is enough to show that we can choose the homomorphisms representing ρ_k so as to have a coherent sequence *of homomorphisms.* For $k = 1$, we must have $\rho_1 = \overline{\rho}$, so there is no choice at this level. Assume we have chosen homomorphisms $r_1, \dots, r_k$ representing the deformations $\rho_1, \dots, \rho_k$ and forming a coherent sequence. Let r' be any homomorphism representing ρ_{k+1}. The assumption that the sequence $\{\rho_k\}$ is coherent means that there exists $M_k \in \Gamma(R/\mathfrak{m}^k)$ such that $M_k^{-1}(r' \pmod{\mathfrak{m}^k})M_k = r_k$. Choosing a lift M_{k+1} of M_k to $\Gamma(R/\mathfrak{m}^{k+1})$ and setting $r_{k+1} = M_{k+1}^{-1} r' M_{k+1}$ extends the coherent sequence to level $k + 1$. By induction, we get a coherent sequence $\{r_k\}$ of homomorphisms $\Pi \to \mathrm{GL}_n(R/\mathfrak{m}^k)$. Taking the inverse limit of these homomorphisms then gives a deformation $\rho : \Pi \longrightarrow \mathrm{GL}_n(R)$ whose reduction modulo $\mathfrak{m}^k$ is ρ_k. This proves the canonical map is surjective.

To show that the canonical map is injective, we need to show that if ρ and ρ' are homomorphisms $\Pi \longrightarrow \mathrm{GL}_n(R)$ such that $\rho_k = \rho \pmod{\mathfrak{m}^k}$ and $\rho'_k = \rho' \pmod{\mathfrak{m}^k}$ are strictly equivalent for all k, then ρ and ρ' are strictly equivalent. The assumption is that for all k we can find $M_k \in \Gamma(R/\mathfrak{m}^k)$ such that

$$\rho_k = M_k^{-1} \rho'_k M_k.$$

It is clear that we can choose the M_k such that $M_{k+1} \equiv M_k \pmod{\mathfrak{m}^k}$, giving a coherent sequence and therefore and element of $\Gamma(R)$ such that $\rho = M^{-1}\rho' M$, as desired. This proves the canonical map is injective. □

The continuity of our functor is an important technical tool: basically it shows that $\mathbf{D}$ is completely determined by its values on the full subcategory $\mathcal{C}^0$. We will use this in a crucial way later, when we use the Schlessinger criteria for representability, which apply to functors on artinian rings.

We should note a final variation on the basic idea. Suppose we have a lift ρ_A of $\overline{\rho}$ to a coefficient ring A. Then it makes sense to look only at those deformations which are actually deformations of our fixed lift to A, that is, deformations to coefficient rings R with a map to A such that the induced deformation is ρ_A. This leads to a slightly modified functor again, for which we need to make two changes:

i. First, we work with the category whose objects are coefficient rings (or coefficient Λ-algebras) that come with an "A-augmentation," that is, a coefficient ring (Λ-algebra) homomorphism $R \longrightarrow A$. In [**101**], Mazur calls these "A-augmented coefficient rings (or Λ-algebras)." We call this category $\mathcal{C}(A)$ (or $\mathcal{C}_\Lambda(A)$).

ii. Second, we change the definition of strict equivalence to allow conjugation only by matrices in the kernel of the map $\mathrm{GL}_n(R) \longrightarrow \mathrm{GL}_n(A)$ induced by the augmentation.

As before, one needs to also consider the full subcategory $\mathcal{C}^0_\Lambda(A)$ of A-augmented artinian local Λ-algebras with residue field k. See Mazur's discussion in [**101**] for more about this "relative" version of the theory.

Universal deformations: why representable functors are nice

The question we want to ask about our deformation functor is whether it is *representable.* This means the following. Given any coefficient ring R, we can define a set-valued functor $\mathbf{h}_R$ on $\mathcal{C}$ by setting, for each coefficient ring S,

$$\mathbf{h}_R(S) = \mathrm{Hom}(R, S),$$

where of course Hom indicates coefficient ring homomorphisms, and where the action on homomorphisms $S_1 \longrightarrow S_2$ defined by composition. We say that a functor $\mathbf{F}$ is *representable* if it is isomorphic to the functor $\mathbf{h}_{\mathcal{R}}$ for some coefficient ring $\mathcal{R}$. Hence, to ask whether $\mathbf{D}$ is representable is to ask whether there exists a coefficient ring (or coefficient Λ-algebra) $\mathcal{R}_{\overline{\rho}}$ such that we have

$$\mathbf{D}_{\overline{\rho}}(R) = \mathrm{Hom}(\mathcal{R}_{\overline{\rho}}, R)$$

for every coefficient ring R and this identification is "functorial," i.e., transforms well under homomorphisms. Let's explore that idea a little in this section.

Assume, then, that the ring $\mathcal{R} = \mathcal{R}_{\overline{\rho}}$ exists. First of all, consider the case $R = \mathcal{R}$. Since the identity is a homomorphism from $\mathcal{R}$ to itself, it corresponds to a deformation

$$\boldsymbol{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathcal{R}).$$

This will turn out to deserve to be called the "universal" deformation. To see why, consider any deformation ρ to a coefficient ring R. By our assumption that the functor is represented by $\mathcal{R}$, this deformation must correspond to a (better: exactly one) coefficient ring homomorphism $\varphi : \mathcal{R} \longrightarrow R$, and the morphism mapping $\boldsymbol{\rho}$ to ρ must be "composition with φ." In other words, given any deformation ρ to a coefficient ring R, there is a coefficient ring homomorphism $\varphi : \mathcal{R} \longrightarrow R$ such that $\rho = \varphi \circ \boldsymbol{\rho}$. Thus, the ring $\mathcal{R}$ parametrizes all possible deformations, and the deformation $\boldsymbol{\rho}$ is "universal," because every deformation is derived from it.

The upshot, then, is that if we can show that our functor is representable we will get a large ring, which we will call the *universal deformation ring of* $\overline{\rho}$ and denote by $\mathcal{R}_{\overline{\rho}}$, and a representation

$$\boldsymbol{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathcal{R}_{\overline{\rho}}),$$

which we will call the *universal deformation of* $\overline{\rho}$.

Problem 2.11. Show that any representable functor is continuous.

As we noted above, we sometimes want to work not with the functor $\mathbf{D}$ but rather with the functor $\mathbf{D}_\Lambda$ which is the result of restricting our attention to coefficient rings which are also Λ-algebras. As we pointed out above, $\mathbf{D}$ is the same as $\mathbf{D}_{W(\mathsf{k})}$, so it would seem that we should work directly with the more general case. It turns out, however, that the moving from one case to another is quite easy:

Theorem 2.4. *If* $\mathbf{D}$ *is representable by a coefficient-ring* $\mathcal{R}$*, then* $\mathbf{D}_\Lambda$ *is representable by* $\mathcal{R}_\Lambda = \mathcal{R}\hat{\otimes}_{W(\mathsf{k})}\Lambda$.

Proof. This is essentially clear if one understands what a completed tensor product is. First of all, note that $\mathcal{R}_\Lambda$ is a coefficient ring and a Λ-algebra. (This is why we need a *completed* tensor product: the tensor product of two coefficient rings need not be complete, so we must pass to the completion to obtain a coefficient ring again.) Next, there is a canonical coefficient-ring homomorphism $\mathcal{R} \longrightarrow \mathcal{R}_\Lambda$, which induces a deformation $\boldsymbol{\rho}_\Lambda$ of $\overline{\rho}$ to $\mathcal{R}_\Lambda$. We claim that this is the universal deformation to coefficient Λ-algebras. To see this, just note that any map from $\mathcal{R}$ to a coefficient Λ-algebra A factors through the canonical map $\mathcal{R} \longrightarrow \mathcal{R}_\Lambda$. □

Problem 2.12. Show that the completed tensor product (over $W(\mathsf{k})$) of two coefficient rings is a coefficient ring. In fact, show that completed tensor product of R_1 and R_2 is the inverse limit over pairs (i,j) of the tensor products $R_1/\mathfrak{m}_{R_1}^i$ tensor $R_2/\mathfrak{m}_{R_2}^j$.

Problem 2.13. Suppose $\mathcal{R} = W(\mathsf{k})[[X_1, X_2, \dots, X_k]]/I$, where I is the closed ideal generated by $f_1, f_2, \dots, f_s$. Describe $\mathcal{R}_\Lambda$.

Problem 2.14. More generally, suppose two coefficient rings R_1 and R_2 are given explicitly as quotients of power series rings in several variables over $W(\mathsf{k})$. Describe $R_1 \hat{\otimes}_{W(\mathsf{k})} R_2$.

Problem 2.15. It is clearly not always possible to recover $\mathcal{R}$ from $\mathcal{R}_\Lambda$ (for example, consider $\Lambda = \mathsf{k}$). Is it ever possible to "descend" from $\mathcal{R}_\Lambda$ to $\mathcal{R}$?

As we learn from algebraic geometry, one can associate to a ring such as $\mathcal{R}$ a geometric object $\operatorname{Spec}\mathcal{R}$, and an "$A$-valued point on $\operatorname{Spec}\mathcal{R}$" is the same as a ring homomorphism $\mathcal{R} \longrightarrow A$. Since coefficient-ring homomorphisms $\mathcal{R} \longrightarrow A$ correspond to deformations, this suggests that we should call $\operatorname{Spec}\mathcal{R}$ the "universal deformation space" of $\overline{\rho}$. There is something to be careful of here, however: the "A-valued points of $\operatorname{Spec}\mathcal{R}$" include *all* ring-homomorphisms $\mathcal{R} \longrightarrow A$, and of course not all such homomorphisms will induce a deformation of $\overline{\rho}$ to A (one will always get a representation $\Pi \longrightarrow \mathrm{GL}_n(A)$, but it need not be continuous nor, even if continuous, need it be a lift of $\overline{\rho}$). A better way to obtain a "deformation space" from the universal deformation ring $\mathcal{R}$ is to consider its formal spectrum $\operatorname{Spf}\mathcal{R}$ as a formal scheme over $\operatorname{Spf} W(\mathsf{k})$ or the associated rigid-analytic space $\operatorname{Spf}\mathcal{R}^{\mathrm{rig}}$ (which, however, is not quasi-caompact; see [**35**] for the properties of the functor $(-)^{\mathrm{rig}}$).

Suppose, for example, that $\mathsf{k} = \mathbb{F}_p$ and $\mathcal{R} = \mathbb{Z}_p[[X_1, X_2, X_3]]$ (as will actually be the case in one of our examples). Then we want to think of the associated space as a three-dimensional space over $\mathbb{Z}_p$, with three parameters corresponding to the three variables. But the dimension of $\operatorname{Spec}\mathcal{R}$ (which is the same as the Krull dimension of the ring) is four, not three. On the other hand, the relative dimension of $\operatorname{Spf}\mathcal{R}$ over $\operatorname{Spf}\mathbb{Z}_p$ is indeed three.

Representable functors and fiber products

Suppose we are working in some category, and we are given objects A, B, and C and maps $\alpha : A \longrightarrow C$ and $\beta : B \longrightarrow C$. Visualize this as the beginning of a commutative diagram

$$\begin{tikzcd}
A \arrow[dr, "\alpha"'] & & B \arrow[dl, "\beta"] \\
& C &
\end{tikzcd}$$

which we want to complete to a commutative "diamond." If a "universal" solution to this problem exists, we call it the *fiber product of A and B over C*, which we denote by $A \times_C B$. This comes with maps $p : A \times_C B \longrightarrow A$ and $q : A \times_C B \longrightarrow B$

such that the diagram

$$\begin{tikzcd} & A \times_C B \arrow[dl, "p"'] \arrow[dr, "q"] & \\ A \arrow[dr, "\alpha"'] & & B \arrow[dl, "\beta"] \\ & C & \end{tikzcd}$$

commutes, and is the universal such object in the following sense. Suppose we have another ring D and maps $D \longrightarrow A$ and $D \longrightarrow B$ such that the diagram

$$\begin{tikzcd} & D \arrow[dl] \arrow[dr] & \\ A \arrow[dr, "\alpha"'] & & B \arrow[dl, "\beta"] \\ & C & \end{tikzcd}$$

is commutative. Then there exists a unique map $D \longrightarrow A \times_C B$ through which both the maps $D \longrightarrow A$ and $D \longrightarrow B$ factor. It is easy to see that if the fiber product exists it is unique up to unique isomorphism.

In the category of sets, the fiber product is given by

$$A \times_C B = \{(a, b) \in A \times B \mid \alpha(a) = \beta(b)\}.$$

Problem 2.16. Check that this set does have the universal property described above.

Now let's consider this in the context of representable functors. We continue to work in a general category. If R and S are objects of our category, we write $\mathrm{Hom}(R, S)$ for the set of morphisms from R to S. Then we can translate the universal property of the fiber product into the statement that for any object D we have

$$\mathrm{Hom}(D, A \times_C B) = \mathrm{Hom}(D, A) \times_{\mathrm{Hom}(D,C)} \mathrm{Hom}(D, B),$$

since the set on the right consists exactly of the pairs of morphisms $D \longrightarrow A$ and $D \longrightarrow B$ that make the diagram commute. If we think of D as the object representing a functor from our category to sets given by $\mathbf{F}(R) = \mathrm{Hom}(D, R)$, we can read this statement as saying that "representable functors commute with fiber products." In other words, if $\mathbf{F}$ is a representable functor from a category where fiber products exist to the category of sets, then we have

$$\mathbf{F}(A \times_C B) = \mathbf{F}(A) \times_{\mathbf{F}(C)} \mathbf{F}(B),$$

where of course the object on the right is a fiber product of sets. Following Mazur, we will call this property the *Mayer-Vietoris* property of representable functors.

For general functors, all we know is that $\mathbf{F}(A \times_C B)$ fits into the diagram

$$\begin{tikzcd} & \mathbf{F}(A \times_C B) \arrow[dl, "\mathbf{F}(p)"'] \arrow[dr, "\mathbf{F}(q)"] & \\ \mathbf{F}(A) \arrow[dr, "\mathbf{F}(\alpha)"'] & & \mathbf{F}(B) \arrow[dl, "\mathbf{F}(\beta)"] \\ & \mathbf{F}(C) & \end{tikzcd}$$

Hence, by the universal property of fiber products, we get a map

$$\mathbf{F}(A \times_C B) \longrightarrow \mathbf{F}(A) \times_{\mathbf{F}(C)} \mathbf{F}(B),$$

but a priori there is no reason to expect this map to have any special properties. On the other hand, when $\mathbf{F}$ is representable, this function will be a bijection.

The upshot of this discussion is the following: when we are working in a category in which fiber products exist, the Mayer-Vietoris property is a *necessary* condition for a functor to be representable. As we will see below, it is very close to being also sufficient.

How can we apply this in our situation? It turns out that we *cannot* apply it to $\mathcal{C}$, the category of all coefficient rings (i.e., complete noetherian local rings with residue field k), but we can use it if we work with the smaller category $\mathcal{C}^0$ of artinian local rings with residue field k; this is the main reason to bring up $\mathcal{C}^0$ in the first place. The reason is this: if A, B, and C are commutative rings and α and β are ring homomorphisms, then $A \times_C B$ has a natural ring structure that makes it the fiber product in the category of rings. The property of being local is preserved, and the property of having residue field k is also preserved. But the fiber product of noetherian rings doesn't need to be noetherian.

Problem 2.17. Let $A = \mathsf{k}[[X,Y]]$, $B = \mathsf{k}$, $C = \mathsf{k}[[X]]$. Let $\alpha : A \longrightarrow C$ be the map that sends Y to 0 and let $\beta : B \longrightarrow C$ be the inclusion of k in $\mathsf{k}[[X]]$. Note that A, B, and C are objects of $\mathcal{C}$ and that α and β are morphisms in $\mathcal{C}$. Show that the ring $A \times_C B$ is not noetherian, and hence is not an object of $\mathcal{C}$.

Problem 2.18. Show that if both α and β are surjective, then $A \times_C B$ is an object of $\mathcal{C}$, i.e., a complete noetherian local ring with residue field k, and is the fiber product of A and B over C in $\mathcal{C}$.

Problem 2.19. Show that if A, B, and C are in $\mathcal{C}^0$, i.e., are artinian rings with residue field k, then $A \times_C B$ is an object of $\mathcal{C}^0$.

Problem 2.20. Show that the same is true in the categories $\mathcal{C}^0_\Lambda$ and $\mathcal{C}^0_\Lambda(A)$.

Problem 2.21. Suppose we work in some subcategory $\mathcal{Z}$ of the category of commutative rings. Suppose we are given objects A, B, and C and morphisms $A \longrightarrow C$ and $B \longrightarrow C$. Let $A \times_C B$ be the ring-theoretical fiber product, i.e., the ring defined by

$$A \times_C B = \{(a,b) \in A \times B \mid \alpha(a) = \beta(b)\}$$

with the natural operations. Is it true that if $A \times_C B$ is an object of $\mathcal{Z}$ then it is the fiber product in $\mathcal{Z}$? Is it true that if $A \times_C B$ is not an object of $\mathcal{Z}$ then there is no fiber product of A and B over C in $\mathcal{Z}$?

Recall that objects of $\mathcal{C}$ are "pro-objects" of $\mathcal{C}^0$, that is, that any object of $\mathcal{C}$ is an inverse limit of objects of $\mathcal{C}^0$. To be specific, if R is a complete noetherian local ring with maximal ideal $\mathfrak{m}$, then $R/\mathfrak{m}^n$ is artinian and we have

$$R = \varprojlim_n R/\mathfrak{m}^n.$$

Suppose that our functor is *continuous*, which, as we explained above, means that

$$\mathbf{F}(R) = \varprojlim_n \mathbf{F}(R/\mathfrak{m}^n).$$

(As we noted above, the deformation functors do have this property.) Then $\mathbf{F}$ is completely determined by its values on the smaller category $\mathcal{C}^0$.

Furthermore, it may happen that $\mathbf{F}$ is not representable as a functor on $\mathcal{C}^0$, but that there exists an object $\mathcal{R}$ of the larger category $\mathcal{C}$ such that we have

$$\mathbf{F}(A) = \operatorname{Hom}(\mathcal{R}, A)$$

for every artinian coefficient ring A. In this case, we say that the functor $\mathbf{F}$ on the category $\mathcal{C}^0$ is *pro-representable.*

Problem 2.22. Show that if $\mathbf{F}$ is continuous, then it is pro-representable as a functor on $\mathcal{C}^0$ if and only if it is representable as a functor on $\mathcal{C}$.

It is easy to see that *pro-representable functors have the Mayer-Vietoris property.* In fact, this is quite close to being a *sufficient* condition for a functor on $\mathcal{C}^0$ to be pro-representable.

To state the exact theorem, let's introduce the coefficient ring of "dual numbers,"

$$\mathsf{k}[\varepsilon] = \mathsf{k}[X]/(X^2)$$

where $\varepsilon = X \pmod{X^2}$. If we are working with coefficient Λ-algebras, we make $\mathsf{k}[\varepsilon]$ into a Λ-algebra via the map

$$\Lambda \longrightarrow \Lambda/\mathfrak{m}_\Lambda = \mathsf{k} \hookrightarrow \mathsf{k}[\varepsilon].$$

Then we have:

Theorem 2.5 (Grothendieck). *Let*

$$\mathbf{F} : \mathcal{C}^0_\Lambda \rightsquigarrow \underline{\mathit{Sets}}$$

be a (covariant) functor such that $\mathbf{F}(\mathsf{k})$ *consists of a single element. Then* $\mathbf{F}$ *is pro-representable if and only if*

i. $\mathbf{F}$ *satisfies the Mayer-Vietoris property, and*
ii. $\mathbf{F}(\mathsf{k}[\varepsilon])$ *is a finite set.*

For the proof, see [**67**].

The statement of this theorem is one of the few places where we are *really* using our assumption that k is a finite field, but we are using it only to state this result before having explained why $\mathbf{F}(\mathsf{k}[\varepsilon])$ is a k-vector space. Once we know that, we can replace the finiteness assumption above with finite dimensionality over k.

As Mazur says in [**101**], this result "is easy to prove (it is a good exercise) but ... is difficult to use because its hypothesis is hard to check." The problem of course, is that the Mayer-Vietoris condition involves checking something for all diagrams

$$\begin{tikzcd} A \arrow[dr, "\alpha"'] & & B \arrow[dl, "\beta"] \\ & C & \end{tikzcd}$$

That is clearly hard to do in general. Schlessinger's theorem (to be discussed in the next lecture) should be viewed as basically a simplification of this result. On the other hand, see the complements to Lecture 3 for a proof that proceeds directly from Grothendieck's theorem.

The finiteness (or finite-dimensionality) condition is there to guarantee that the representing object is noetherian. If we were willing to work in a larger category (e.g., the category of all Λ-algebras which are inverse limits of objects of $\mathcal{C}_\Lambda^0$), then we could drop this finiteness assumption (which in fact is not in Grothendieck's original theorem).

Problem 2.23. (Some category theory needed.) Show that the first condition in the theorem, together with the condition that $\mathbf{F}(\mathsf{k})$ is a singleton, is equivalent to saying that $\mathbf{F}$ preserves all finite limits. (Grothendieck calls functors which preserve all finite limits *left exact*.)

The tangent space

For this section, we fix a coefficient ring Λ and work in the category $\mathcal{C}_\Lambda$ of coefficient Λ-algebras. We let $\mathfrak{m}_\Lambda$ denote the maximal ideal of Λ. Let R be a coefficient Λ-algebra, and let $\mathfrak{m}_R$ be its maximal ideal. The *Zariski cotangent space* of R is defined to be

$$t_R^* = \mathfrak{m}_R/(\mathfrak{m}_R^2, \mathfrak{m}_\Lambda),$$

where

$$(\mathfrak{m}_R^2, \mathfrak{m}_\Lambda) = \mathfrak{m}_R^2 + (\text{image of } \mathfrak{m}_\Lambda)R$$

is the ideal of R generated by the square $\mathfrak{m}_R^2$ of the maximal ideal of R and the image in R of the maximal ideal of Λ. Notice that t_R^* is a module over $\Lambda/\mathfrak{m}_\Lambda$, that is, it is a k-vector space.

The *Zariski tangent space* of R is, of course, the dual of the cotangent space:

$$t_R = \mathrm{Hom}_{\mathsf{k}}(\mathfrak{m}_R/(\mathfrak{m}_R^2, \mathfrak{m}_\Lambda), \mathsf{k}).$$

Since R is noetherian, t_R^* is a finite-dimensional vector space, so that there is no problem with the duality here.

Problem 2.24. Let R be a noetherian local ring with residue field k and define the tangent space t_R as above. Prove that t_R is a finite-dimensional vector space over $R/\mathfrak{m}_R$. Is the converse true? (Well, first of all, what would the converse say?)

Problem 2.25. Let $f : B \longrightarrow A$ be a morphism in $\mathcal{C}_\Lambda$. Show that f induces a k-linear transformation $f_* : t_B^* \longrightarrow t_A^*$ of cotangent spaces. Show that f is surjective if and only if f_* is surjective. (This is Lemma 1.1 in **[125]**.)

Problem 2.26. Use the duality between the tangent and cotangent spaces to reinterpret the previous problem in terms of tangent spaces.

Suppose we have a functor $\mathbf{F}$ as above which is represented by R. We'd like to reinterpret this construction in terms of the functor. The crucial observation is the following.

Lemma 2.6. *If* $\mathbf{F}$ *is a functor which is represented by R, there is a natural bijection*

$$\mathrm{Hom}_{\mathsf{k}}(\mathfrak{m}_R/(\mathfrak{m}_R^2, \mathfrak{m}_\Lambda), \mathsf{k}) \cong \mathrm{Hom}_\Lambda(R, \mathsf{k}[\varepsilon]),$$

where Hom_k *means* k*-vector space homomorphisms and* Hom_Λ *means homomorphisms of coefficient* Λ*-algebras.*

The basic point is that a homomorphism of coefficient Λ-algebras

$$R \longrightarrow \mathsf{k}[\varepsilon],$$

because it must induce the identity on residue fields, must have the form

$$r \mapsto \overline{r} + \varphi(r)\varepsilon,$$

where $\overline{r} = r \pmod{\mathfrak{m}}$ denotes the image of r in the residue field k, and $\varphi(r) \in \mathsf{k}$. Furthermore, since the map must be a homomorphism of Λ-algebras, φ is completely determined by its values on elements $r \in \mathfrak{m}_R$. Working out what φ must look like yields the Lemma.

Problem 2.27. Fill in the details to give a proof of the Lemma.

We have shown, then, that if a functor $\mathbf{F}$ is represented by a coefficient Λ-algebra R, then $\mathbf{F}(\mathsf{k}[\varepsilon]) = t_R$, at least as sets. To make this really work, we have to explain how to think of $\mathbf{F}(\mathsf{k}[\varepsilon])$, which a priori is just a set, as a k-vector space. Of course, we want to do that in such a way as to make the bijection in the Lemma be k-linear, and therefore an isomorphism of k-vector spaces.

It turns out that there is a natural vector space structure on $\mathbf{F}(\mathsf{k}[\varepsilon])$ that arises simply from the fact that $\mathbf{F}$ is a reasonably nice functor. One part of this is easy: an element α of k gives an automorphism of $\mathsf{k}[\varepsilon]$ by

$$a + b\varepsilon \mapsto a + \alpha b\varepsilon,$$

(yes, this is actually a ring homomorphism!) and therefore, by functoriality, gives an automorphism of $\mathbf{F}(\mathsf{k}[\varepsilon])$. This gives a scalar multiplication by k.

The addition is a bit harder. Since we are assuming that $\mathbf{F}$ is representable, we know it has the Mayer-Vietoris property. We apply it to the diagram

$$\begin{tikzcd}
\mathsf{k}[\varepsilon] \arrow[dr] & & \mathsf{k}[\varepsilon] \arrow[dl] \\
& \mathsf{k} &
\end{tikzcd}$$

where both arrows are the canonical projection onto the residue field. Since there is only one coefficient Λ-algebra homomorphism $R \longrightarrow \mathsf{k}$, $\mathbf{F}(\mathsf{k})$ consists of only one element, and therefore the fiber product $\mathbf{F}(\mathsf{k}[\varepsilon]) \times_{\mathbf{F}(\mathsf{k})} \mathbf{F}(\mathsf{k}[\varepsilon])$ is just a product. Hence the Mayer-Vietoris property says, in this situation, that

$$\mathbf{F}\left(\mathsf{k}[\varepsilon] \times_{\mathsf{k}} \mathsf{k}[\varepsilon]\right) \cong \mathbf{F}(\mathsf{k}[\varepsilon]) \times \mathbf{F}(\mathsf{k}[\varepsilon]).$$

Now, we have a homomorphism of coefficient Λ-algebras

$$\mathfrak{p} : \mathsf{k}[\varepsilon] \times_k \mathsf{k}[\varepsilon] \longrightarrow \mathsf{k}[\varepsilon]$$

defined by

$$\mathfrak{p}(x + y_1\varepsilon, x + y_2\varepsilon) = x + (y_1 + y_2)\varepsilon.$$

(The notation $\mathfrak{p}$ is meant to recall "plus," or perhaps the abbreviation of "piu" used by the early Italian algebraists.) Then we put all this together to define the addition: the composition

$$\mathbf{F}(\mathsf{k}[\varepsilon]) \times \mathbf{F}(\mathsf{k}[\varepsilon]) \cong \mathbf{F}\left(\mathsf{k}[\varepsilon] \times_{\mathsf{k}} \mathsf{k}[\varepsilon]\right) \xrightarrow{\mathbf{F}(\mathfrak{p})} \mathbf{F}(\mathsf{k}[\varepsilon])$$

gives the vector addition.

Problem 2.28. Check everything! In particular, check that if $\mathbf{F}$ is represented by R, then

i. $\mathfrak{p}$ is indeed a homomorphism of coefficient Λ-algebras,

ii. these two operations do make $\mathbf{F}$ a vector space over k, *iii.* with these definitions the natural bijection in Lemma 2.6 is in fact an isomorphism of k-vector spaces.

The reason to go through this effort of translation is the following: we can now try to define the "tangent space of a functor" by following this template. Of course, we did need to use the Mayer-Vietoris property, but we used it only for the specific diagram above. So we have in fact proved the following:

Proposition 2.7. *Let*

$$\mathbf{F} : \mathcal{C}_\Lambda^0 \rightsquigarrow \underline{Sets}$$

be a (covariant) functor such that $\mathbf{F}(\mathsf{k})$ *consists of a single element. Suppose that the natural map*

$$\mathbf{F}(\mathsf{k}[\varepsilon] \times_{\mathsf{k}} \mathsf{k}[\varepsilon]) \longrightarrow \mathbf{F}(\mathsf{k}[\varepsilon]) \times \mathbf{F}(\mathsf{k}[\varepsilon])$$

is a bijection. Then $\mathbf{F}(\mathsf{k}[\varepsilon])$ *has a natural vector space structure over* k.

Problem 2.29. In the discussion above, we were assuming that $\mathbf{F}$ was represented by R. So to prove the proposition we need to check that the only properties of $\mathbf{F}$ that we really used are the ones listed in the statement of the proposition. Do that.

We will refer to the assumption that the map

$$\mathbf{F}(\mathsf{k}[\varepsilon] \times_{\mathsf{k}} \mathsf{k}[\varepsilon]) \longrightarrow \mathbf{F}(\mathsf{k}[\varepsilon]) \times \mathbf{F}(\mathsf{k}[\varepsilon])$$

is a bijection as the *tangent space hypothesis over* k. When it is satisfied, we will write

$$t_{\mathbf{F}} = \mathbf{F}(\mathsf{k}[\varepsilon])$$

and call this the *tangent space of the functor* $\mathbf{F}$.

In [**101**], Mazur suggests that we say the functor $\mathbf{F}$ is *nearly representable* if it satisfies the tangent space hypothesis and the tangent space $t_{\mathbf{F}}$ is finite-dimensional over k. See Mazur's article for further discussion, and also for a discussion of how to adapt this to the "relative" case in which the category is $\mathcal{C}_\Lambda(A)$.

Finally, we note in passing another interpretation of the tangent space:

Problem 2.30. Show that t_R is naturally isomorphic to $\mathrm{Der}_\Lambda(R, \mathsf{k})$, the k-vector space of Λ-algebra derivations from R to k. (This is another reason to think of t_R as the tangent space.)

Complements to lecture 2

The language of categories and functors is a particularly convenient way to think about the deformation theory. Basically, category theory tries to make precise the idea that in a mathematical "universe of discourse" there is typically a collection of objects which we study (e.g., sets, groups, rings, topological spaces, complete noetherian local rings with residue field k, etc.), and for each such collection of objects we have a "correct" notion of function between our objects (e.g., for the list above, they would be: functions in general, group homomorphisms, ring homomorphisms, continuous functions, local homomorphisms inducing the identity on residue fields). Such a "universe of discourse" is called a *category.*

Functors connect different categories, transforming the objects of one to objects of the other and doing the same to the functions, while preserving some obvious structure (the identity function and compositions of functions). We could even speak of a "category of categories," in which the appropriate functions would be the functors. The functors we are interested in are set-valued, that is, they associate a set to each coefficient rings. Mathematics is full of functors, but the most famous ones are certainly the various functors that attach algebraic objects to various geometric objects (homology and cohomology, etc.).

Making this impressionistic description precise is the business of category theory, of which we need only a small amount. The basic notions are discussed in most algebra textbooks; for example, see [**90**, I, §11]. For more information, a good reference is MacLane's [**92**], which contains much more material than we have used (or will use).

LECTURE 3
The Universal Deformation: Existence

Our goal for this lecture is to prove that, under suitable hypotheses, the deformation functor is indeed (pro-)representable. Our proof will be very similar to Mazur's original proof (in [**97**]; see also [**8**], [**10**], [**9**], [**115**]), which is based on Schlessinger's criteria for pro-representability of a functor on a category of artinian rings first given in [**125**].

Since the publication of [**97**], several other approaches to proving the representability of the deformation functor (or, equivalently, the existence of a universal deformation) have been found. One is a "direct" approach that constructs the universal deformation ring by generators and relations from what is known about the Galois group. Constructions in this style have been given by Faltings (see [**34**]) and by Lenstra and de Smit (see [**36**]). We will see a little bit of this point of view when we discuss "explicit" deformations.

Another approach is based on the notion of a "pseudo-character," which is basically a function that "looks like" the character of a representation. This has been studied by Nyssen [**113**] and Rouquier [**121**], who find conditions for a pseudo-character to be the trace of a representation and use them to construct the universal deformation.

As before, k will denote a finite field of characteristic p and Π will denote a profinite group satisfying hypothesis Φ_p. We will assume we are given a residual representation

$$\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$$

whose lifts we want to understand.

We let $\mathcal{C}$ stand for the category whose objects are complete noetherian local rings with residue field k and whose morphisms are local homomorphisms which induce the identity on k. We use the shorthand expression "coefficient ring" for an object of $\mathcal{C}$. We let $\mathcal{C}^0$ be the full subcategory of $\mathcal{C}$ whose objects are artinian coefficient rings. If Λ is an object of $\mathcal{C}$, we write $\mathcal{C}_\Lambda$ for the category whose objects are complete noetherian local Λ-algebras with residue field k and whose morphisms are local Λ-algebra homomorphisms which induce the identity on residue fields. We use the shorthand expression "coefficient Λ-algebra" for an object of $\mathcal{C}_\Lambda$. Finally, we write $\mathcal{C}^0_\Lambda$ for the full subcategory of artinian local Λ-algebras with residue field k.

If R is a coefficient ring, we write $\Gamma_n(R)$ for the kernel of the map $\mathrm{GL}_n(R) \longrightarrow \mathrm{GL}_n(\mathsf{k})$ given by reduction modulo the maximal ideal, so that $\Gamma_n(R) = 1 + \mathrm{M}_n(\mathfrak{m})$, i.e., it consists of matrices whose off-diagonal elements are in $\mathfrak{m}$ and whose diagonal elements belong to $1 + \mathfrak{m}$.

Given the residual representation $\overline{\rho}$, we have defined set-valued covariant functors $\mathbf{D}$ and $\mathbf{D}_\Lambda$ whose value on a coefficient ring (resp, a coefficient Λ-algebra) R is the set of deformations of $\overline{\rho}$ to R. The functors depend on $\overline{\rho}$, of course, but since $\overline{\rho}$ will usually be fixed throughout we omit it from the notation; if it is necessary to emphasize the dependence on $\overline{\rho}$, we will write $\mathbf{D}_{\overline{\rho}}$ and $\mathbf{D}_{\overline{\rho},\Lambda}$, respectively.

Schlessinger's criteria

As we saw above, we can think of the deformation functor $\mathbf{D}$ (or $\mathbf{D}_\Lambda$) as a functor on the category $\mathcal{C}^0$ (or $\mathcal{C}^0_\Lambda$) of artinian coefficient rings (or Λ-algebras). From this point of view, our goal is to show that these functions are pro-representable. As we saw from Grothendieck's theorem, pro-representability is closely related to the Mayer-Vietoris property, that is, to how our functors act on fiber products. In [**125**], Schlessinger obtained a set of criteria for pro-representability of functors on categories of artinian rings which are much easier to apply. In this section, we recall Schlessinger's criteria in preparation for using them, in the next section, to prove that the deformation functors are pro-representable.

Let $\mathbf{F}$ be a covariant functor

$$\mathbf{F} : \mathcal{C}^0_\Lambda \rightsquigarrow \underline{\text{Sets}},$$

and assume that $\mathbf{F}(\mathsf{k})$ consists of one element. We want to give sufficient conditions for $\mathbf{F}$ to be pro-representable by a ring $\mathcal{R}$ in $\mathcal{C}_\Lambda$.

In general, if R is an artinian coefficient Λ-algebra, we write $\mathfrak{m}_R$ for the maximal ideal in R. If R and S are two coefficient Λ-algebras, we say a homomorphism

$$\phi : R \longrightarrow S$$

is *small* if it is surjective and if $\operatorname{Ker}(\phi)$ is principal and is annihilated by $\mathfrak{m}_R$.

Problem 3.1. Show that any surjective homomorphism in $\mathcal{C}^0_\Lambda$ factors as the composition of small homomorphisms.

The prototypical example of a small homomorphism, which will be of great importance in what follows, is the homomorphism

$$\mathsf{k}[\varepsilon] \longrightarrow \mathsf{k},$$

where $\mathsf{k}[\varepsilon]$, as above, is the ring of dual numbers.

To set up the Schlessinger criteria, consider rings R_0, R_1, and R_2 in $\mathcal{C}^0_\Lambda$, and suppose we have morphisms

$$\begin{tikzcd}
R_1 \arrow[dr, "\phi_1"'] & & R_2 \arrow[dl, "\phi_2"] \\
& R_0 &
\end{tikzcd}$$

Let

$$R_3 = R_1 \times_{R_0} R_2 = \{(r_1, r_2) \in R_1 \times R_2 \mid \phi_1(r_1) = \phi_2(r_2)\}$$

be the fiber product of R_1 and R_2 over R_0, which is again an artinian coefficient Λ-algebra. Since $\mathbf{F}$ is a functor, we get a map

$$(*) \qquad \mathbf{F}(R_3) \longrightarrow \mathbf{F}(R_1) \times_{\mathbf{F}(R_0)} \mathbf{F}(R_2)$$

If $\mathbf{F}$ is representable, we know (this is just the Mayer-Vietoris condition again) that the map $(*)$ is a bijection. We also know that if $(*)$ is a bijection in the case where $R_1 = R_2 = \mathsf{k}[\varepsilon]$ and $R_0 = \mathsf{k}$, then $\mathbf{F}(\mathsf{k}[\varepsilon])$ has a natural k-vector space structure.

Now we can state the Schlessinger conditions, which we label as **H1**, **H2**, **H3**, and **H4** (the "H" stands for hull; see below). They are basically a weakened form of the conditions in Grothendieck's theorem. The first two specify that the map $(*)$ should be nice when the map $R_2 \longrightarrow R_0$ is particularly simple.

H1: If the map $R_2 \longrightarrow R_0$ is small, then $(*)$ is surjective.

H2: If $R_0 = \mathsf{k}$ and $R_2 = \mathsf{k}[\varepsilon]$, then $(*)$ is bijective.

If **H2** holds, applying it to the case when $R_1 = R_2 = \mathsf{k}[\varepsilon]$ shows that the tangent space hypothesis over k is satisfied, and hence we can think of $t_{\mathbf{F}} = \mathbf{F}(\mathsf{k}[\varepsilon])$ as a k-vector space; as before, we call this the tangent space of $\mathbf{F}$. Schlessinger's third condition is:

H3: The vector space $t_{\mathbf{F}} = \mathbf{F}(\mathsf{k}[\varepsilon])$ is finite-dimensional.

The fourth condition is another Mayer-Vietoris variant:

H4: If $R_1 = R_2$, the maps $R_i \longrightarrow R_0$ are the same, and $R_i \longrightarrow R_0$ is small, then $(*)$ is bijective.

Theorem 3.1 (Schlessinger). *Let* $\mathbf{F}$ *be a set-valued covariant functor on* $\mathcal{C}^0_\Lambda$ *such that* $\mathbf{F}(\mathsf{k})$ *has exactly one element. If* $\mathbf{F}$ *satisfies conditions **H1** to **H4**, then* $\mathbf{F}$ *is pro-representable. In particular, there exists an object* $\mathcal{R}$ *of* $\mathcal{C}_\Lambda$ *such that* $\mathbf{F}(A) = \operatorname{Hom}(\mathcal{R}, A)$ *for every* A *in* $\mathcal{C}^0_\Lambda$.

Schlessinger's theorem is in fact more general: he shows that if $\mathbf{F}$ satisfies only conditions **H1** to **H3** then it has a "hull" which satisfies some of the properties one would expect the representing object to have. (See [**125**], [**101**], and the problems at the end of this section for more discussion of what this means.) This is the reason for the otherwise rather peculiar ordering of the four conditions.

It's also worth noting that since representable functors *do* satisfy all four conditions, the theorem gives necessary and sufficient conditions for representability. Similarly, **H1** to **H3** are necessary and sufficient conditions for the existence of a hull.

Let $d = \dim_{\mathsf{k}} t_{\mathbf{F}}$. The proof constructs $\mathcal{R}$ as an inverse limit of quotients of $\Lambda[[X_1, X_2, \dots, X_d]]$. See [**125**] for the details. It is probably worth pointing out that the proof does not give much information about the resulting ring beyond the fact that it is a quotient of $\Lambda[[X_1, X_2, \dots, X_d]]$.

We'll often want to apply Schlessinger's criteria to a subfunctor of a functor which we already know is representable. It turns out to be quite easy to do this. We say a set-valued functor $\mathbf{F}_1$ on $\mathcal{C}^0_\Lambda$ is a *subfunctor* of $\mathbf{F}$ if, for every coefficient Λ-algebra R, we have $\mathbf{F}_1(R) \subset \mathbf{F}(R)$. (If we want to be more precise, we'd have to say that there exists a natural transformation $\mathbf{F}_1 \longrightarrow \mathbf{F}$ which induces the inclusions $\mathbf{F}_1(R) \subset \mathbf{F}(R)$ for every R.)

Proposition 3.2. *Let $\mathbf{F}_1$ be a subfunctor of $\mathbf{F}$ such that $\mathbf{F}_1(\mathsf{k}) = \mathbf{F}(\mathsf{k})$ is a singleton, and suppose $\mathbf{F}$ is pro-representable, and so satisfies conditions **H1** to **H4**. If $\mathbf{F}_1$ satisfies condition **H1**, then $\mathbf{F}_1$ satisfies the other three conditions, and therefore is also pro-representable.*

Problem 3.2. Prove the proposition. (The main point is that the restriction of an injective homomorphism is automatically injective.)

Problem 3.3. Check that the proposition is also true if we replace "is pro-representable" by "has a hull" (equivalently, if we omit property **H4**) in both hypothesis and conclusion.

Problem 3.4. In the situation of the proposition, let $\mathcal{R}$ be the coefficient Λ-algebra that represents $\mathbf{F}$. Prove that the object representing $\mathbf{F}_1$ is a quotient of $\mathcal{R}$.

An alternative approach is suggested by Mazur in [**101**], which does not depend on knowing that $\mathbf{F}$ is pro-representable. We need one bit of language first: in any category where fiber products exist, let's say that a diagram

$$\begin{tikzcd}
& D \arrow[dl] \arrow[dr] & \\
A \arrow[dr, "\alpha"'] & & B \arrow[dl, "\beta"] \\
& C &
\end{tikzcd}$$

is *Cartesian* if the induced map $D \longrightarrow A \times_C B$ is an isomorphism. Now suppose $\mathbf{F}_1$ is a subfunctor of a covariant set-valued functor on $\mathcal{C}^0_\Lambda$, and suppose that $\mathbf{F}_1(\mathsf{k}) = \mathbf{F}(\mathsf{k})$. Given a diagram in $\mathcal{C}^0_\Lambda$

$$\begin{tikzcd}
A \arrow[dr, "\alpha"'] & & B \arrow[dl, "\beta"] \\
& C &
\end{tikzcd}$$

we consider the commutative square

$$\begin{tikzcd}
\mathbf{F}_1(A \times_C B) \arrow[r] \arrow[d] & \mathbf{F}_1(A) \times_{\mathbf{F}_1(C)} \mathbf{F}_1(B) \arrow[d] \\
\mathbf{F}(A \times_C B) \arrow[r] & \mathbf{F}(A) \times_{\mathbf{F}(C)} \mathbf{F}(B)
\end{tikzcd}$$

in which the vertical arrows are inclusions and the horizontal arrows are obtained as in our previous discussion. If every such diagram is Cartesian, Mazur says that $\mathbf{F}_1 \subset \mathbf{F}$ is *relatively representable.*

Problem 3.5. Show that if $\mathbf{F}_1 \subset \mathbf{F}$ is relatively representable, then, for each i, $\mathbf{F}_1$ satisfies **H**i if $\mathbf{F}$ does, and similarly for the tangent space hypothesis.

The point, then, is that Mazur's condition allows one to transfer each property separately, while our previous theorem only applies when we already know that $\mathbf{F}$ is pro-representable. Nevertheless, in most of the cases with which we will be working the deformation functor will indeed be pro-representable, so that Proposition 3.2 is good enough.

Universal Deformations exist

We now apply Schlessinger's theorem to the deformation functor

$$\mathbf{D}_\Lambda : \mathcal{C}_\Lambda \rightsquigarrow \underline{\text{Sets}}$$

given by

$$\mathbf{D}_\Lambda(R) = \{\text{deformations of } \overline{\rho} \text{ to } R\}.$$

As we will see, the first three conditions will always hold, but the fourth will depend on what $\overline{\rho}$ is.

Definition 3.1. Let $\overline{\rho}$ be a residual representation. We let

$$C(\overline{\rho}) = \mathrm{Hom}_\Pi(\mathsf{k}^n, \mathsf{k}^n) = \{P \in \mathrm{M}_n(\mathsf{k}) \mid P\overline{\rho}(g) = \overline{\rho}(g)P \text{ for all } g \in \Pi\}.$$

As the definition suggests, we can think of k^n as a Π-module via $\overline{\rho}$, and then $C(\overline{\rho})$ is its ring of Π-module endomorphisms. More generally, if A is a coefficient Λ-algebra and ρ is a deformation of $\overline{\rho}$ to A, we can use ρ to make A^n a Π-module, and then make the analogous definition:

Definition 3.2. Let $\overline{\rho}$ be a residual representation, and let ρ be a deformation of $\overline{\rho}$ to a coefficient Λ-algebra A. We define

$$C_A(\rho) = \mathrm{Hom}_\Pi(A^n, A^n) = \{P \in \mathrm{M}_n(A) \mid P\rho(g) = \rho(g)P \text{ for all } g \in \Pi\}.$$

In particular $C(\overline{\rho}) = C_{\mathsf{k}}(\overline{\rho})$. We will be especially interested in the case where $C(\overline{\rho}) = \mathsf{k}$, that is, where the only matrices in $\mathrm{M}_n(\mathsf{k})$ that commute with the image of $\overline{\rho}$ are the scalar matrices.

Theorem 3.3 (Mazur, Ramakrishna). *Suppose Π is a profinite group that satisfies property Φ_p, $\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$ is a continuous representation, and Λ is a complete noetherian ring with residue field k. Then the deformation functor $\mathbf{D}_\Lambda$ always satisfies properties **H1**, **H2**, and **H3**. Furthermore, if $C(\overline{\rho}) = \mathsf{k}$, then $\mathbf{D}_\Lambda$ also satisfies property **H4**.*

Mazur essentially proved this theorem in [**97**], except that he showed property **H4** under the assumption that $\overline{\rho}$ is absolutely irreducible. Ramakrishna pointed out in [**115**] that the hypothesis could be weakened as above. Note that by Schur's Lemma (see below) we do know that $C(\overline{\rho}) = \mathsf{k}$ when $\overline{\rho}$ is absolutely irreducible.

We will prove the theorem by a series of lemmas. We fix the following notation throughout.

Let R_0, R_1 and R_2 be artinian coefficient Λ-algebras, and suppose we are given $\phi_1 : R_1 \longrightarrow R_0$ and $\phi_2 : R_2 \longrightarrow R_0$ as above. Let

$$E_i = \mathrm{Hom}_{\overline{\rho}}(\Pi, \mathrm{GL}_n(R_i))$$

be the set of homomorphisms from Π to $\mathrm{GL}_n(R_i)$ which reduce to $\overline{\rho}$ modulo the maximal ideal. Then $\Gamma_n(R_i)$ acts on E_i by conjugation and (because deformations are strict equivalence classes of homomorphisms) we have

$$\mathbf{D}_\Lambda(R_i) = E_i/\Gamma_n(R_i).$$

The only difficulties in the proof arise in passing to the quotient from E_i to $\mathbf{D}_\Lambda(R_i)$.

The map in $(*)$ is

$$b : E_3/\Gamma_n(R_3) \longrightarrow E_1/\Gamma_n(R_1) \times_{E_0/\Gamma_n(R_0)} E_2/\Gamma_n(R_2).$$

If $R_2 \longrightarrow R_0$ is surjective, then $\Gamma_n(R_2) \longrightarrow \Gamma_n(R_0)$ is also surjective (see problem 3.24).

Lemma 3.4. *Property **H1** is true.*

Proof. Suppose $R_2 \longrightarrow R_0$ is small (in fact, all we need to know is that it is surjective). We want to show that if we have a pair (ρ_1, ρ_2) of deformations to R_1 and R_2 which induce the same deformation to R_0, we can paste them together to get a deformation to R_3. This is clear for homomorphisms (i.e., for elements of E_i), so we need to check that we can pick representatives for the strict equivalence classes so that they "match" when projected down to R_0.

To do this, we pick any two representatives ϕ_1 and ϕ_2. The assumption is that their images in R_0 are strictly equivalent, so that there is an element $\overline{M}$ of $\Gamma_n(R_0)$ such that conjugating the image of ϕ_2 by $\overline{M}$ gives the image of ϕ_1. Since $R_2 \longrightarrow R_0$ is surjective, so is $\Gamma_n(R_2) \longrightarrow \Gamma_n(R_0)$; hence, we can lift $\overline{M}$ to $M \in \Gamma_n(R_2)$. Then ϕ_1 and $M^{-1}\phi_2 M$ are group homomorphisms that have the same image in $\mathrm{GL}_n(R_0)$, and hence they define a homomorphism $\phi_3 \in E_3$. The strict equivalence class of ϕ_3 maps to (ρ_1, ρ_2), and so we have proved that $(*)$ is surjective. □

This settles **H1**, but we still need to consider when it is that b is injective. For this, we call on two lemmas. Let $\phi_2 \in E_2$ and let $\phi_0 \in E_0$ be its image. Set

$$G_i(\phi_i) = \{g \in \Gamma_n(R_i) \mid g \text{ commutes with the image of } \phi_i \text{ in } \mathrm{GL}_n(R_i)\}.$$

(Note that this is similar but not identical to $C_{R_i}(\phi_i)$ defined above. In particular, this is a subgroup of $\Gamma_n(R_i)$ while the other is a ring.) The first result is:

Lemma 3.5. *If for all $\phi_2 \in E_2$ the map*

$$G_2(\phi_2) \longrightarrow G_0(\phi_0)$$

is surjective, then the map b is injective.

Proof. Suppose ϕ and ψ are elements of E_3 that induce elements ϕ_i and ψ_i in E_i for each $i = 0, 1, 2$. Saying that ϕ and ψ have the same image under $(*)$ means that for each $i = 1, 2$ there is an $M_i \in \Gamma_n(R_i)$ such that $\phi_i = M_i^{-1}\psi_i M_i$. Mapping down to E_0 we see that

$$\phi_0 = \overline{M}_1^{-1}\psi_0\overline{M}_1 = \overline{M}_2^{-1}\psi_0\overline{M}_2,$$

and so that $\overline{M}_2\overline{M}_1^{-1}$ commutes with the image of ϕ_0, i.e., $\overline{M}_2\overline{M}_1^{-1} \in G_0(\phi_0)$.

Now use the surjectivity assumption to find $N \in G_2(\phi_2)$ which maps to $\overline{M}_2\overline{M}_1^{-1}$. Let $N_2 = N^{-1}M_2$. Then we have

$$N_2^{-1}\psi_2 N_2 = M_2^{-1}N\psi_2 N^{-1}M_2 = M_2^{-1}\psi_2 M_2 = \phi_2.$$

On the other hand, the image of N_2 in $\Gamma_n(R_0)$ is

$$\overline{N}_2 = (\overline{M}_2\overline{M}_1^{-1})^{-1}\overline{M}_2 = \overline{M}_1.$$

Since M_1 and N_2 have the same image in $\Gamma_n(R_0)$, the pair (M_1, N_2) defines an element $M \in \Gamma_n(R_3)$ and we have $M^{-1}\psi M = \phi$. Thus, ϕ and ψ are strictly equivalent, and we are done. □

Lemma 3.6. *Property **H2** is true.*

Proof. If $R_0 = \mathsf{k}$ and $R_2 = \mathsf{k}[\varepsilon]$, then we already know $(*)$ is surjective by **H1**. Injectivity will follow if we know that the map $G_2(\phi_2) \longrightarrow G_0(\phi_0)$ is always surjective. But when $R_0 = \mathsf{k}$, $G_0 = \Gamma_n(R_0)$ consists only of the identity matrix, and $G_0(\phi_0)$, which is a subgroup, is again just the identity. So the surjectivity holds (trivially) and we are done. □

Lemma 3.7. *Property **H3** is true.*

Proof. Let $\Pi_0 = \operatorname{Ker}\overline{\rho}$ and let ρ be a lift of $\overline{\rho}$ to $\mathsf{k}[\varepsilon]$. If $x \in \Pi_0$, we have $\overline{\rho}(x) = 1$, and hence $\rho(x) \in \Gamma_n(\mathsf{k}[\varepsilon])$. Hence, ρ determines a map from $\Pi_0 = \operatorname{Ker}\overline{\rho}$ to $\Gamma_n(\mathsf{k}[\varepsilon])$. Two lifts that determine the same map must be identical. Now, Π_0 is an open subgroup of Π and by problem 3.23, we know $\Gamma_n(\mathsf{k}[\varepsilon])$ is a (finite) p-elementary abelian group. By property Φ_p, there are finitely many maps $\Pi_0 = \operatorname{Ker}\overline{\rho}$ to $\Gamma_n(\mathsf{k}[\varepsilon])$. Hence, $\mathbf{D}_\Lambda(\mathsf{k}[\varepsilon])$ is a finite set, and we're done. □

Problem 3.6. This proof relies, once again, on the fact that k is a *finite* field. Is it possible to modify it to cover the case of an infinite field of characteristic p?

Lemma 3.8. *If $C(\overline{\rho}) = \mathsf{k}$, then for any i the group $G_i(\phi_i) \subset R_i$, i.e., $G_i(\phi_i)$ consists of the scalar matrices in $\Gamma_n(R_i)$.*

Proof. We prove the stronger assertion that for any deformation ρ of $\overline{\rho}$ to an artinian coefficient ring A we have $C_A(\rho) = A$. Our argument follows the one given in [**42**].

Since the map $A \longrightarrow \mathsf{k}$ is surjective, it factors as a sequence of small extensions. Since we know that $C_{\mathsf{k}}(\overline{\rho}) = \mathsf{k}$, the lemma will follow, by induction, from the claim that if $C_B(\rho_B) = B$ and $A \longrightarrow B$ is small, then $C_A(\rho_A) = A$.

To prove this, take $c \in C_A(\rho_A)$. By our assumption, the image of c in $\mathrm{M}_n(B)$ is a scalar matrix. Suppose $c \mapsto \overline{r}$, where the scalar $\overline{r} \in B$ is the image of $r \in A$. Then we can write $c = r + tM$ where t is a generator of the kernel of $A \longrightarrow B$ and $M \in \mathrm{M}_n(A)$.

Now, c commutes with the image of ρ_A, so that we must have, for every $g \in \Pi$,

$$(r + tM)\rho_A(g) = \rho_A(g)(r + tM),$$

which, since scalars commute with everything, boils down to

$$M\rho_A(g) = \rho_A(g)M.$$

Reducing modulo the maximal ideal $\mathfrak{m}_A$ and using the fact that $C(\overline{\rho}) = \mathsf{k}$, we see that M must be of the form $M = s + M_1$, where $s \in A$ is a scalar and all the entries of M_1 belong to $\mathfrak{m}_A$. But, since $A \longrightarrow B$ is small, we have $t\mathfrak{m}_A = 0$, from which it follows that $M = r + ts$ is a scalar. □

Problem 3.7. Show that if $C(\overline{\rho}) = \mathsf{k}$ then two lifts ρ and ρ' of $\overline{\rho}$ to a coefficient Λ-algebra A are equivalent if and only if they are strictly equivalent.

Lemma 3.9. *Suppose $C(\overline{\rho}) = \mathsf{k}$. Then property **H4** is true.*

Proof. From the previous lemma, $G_i(\phi_i)$ consists only of scalars (of the form $1 + \mathfrak{m}_{R_i}$, in fact), and the lemma follows. □

The upshot is:

Theorem 3.10 (Mazur, Ramakrishna). *Suppose Π is a profinite group satisfying condition Φ_p and*

$$\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$$

is a continuous representation such that $C(\overline{\rho}) = \mathsf{k}$. Then there exists a ring $\mathcal{R} = \mathcal{R}(\Pi, \mathsf{k}, \overline{\rho})$ in $\mathcal{C}_\Lambda$ and a deformation $\boldsymbol{\rho}$ of $\overline{\rho}$ to $\mathcal{R}$,

$$\boldsymbol{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathcal{R})$$

such that any deformation of $\overline{\rho}$ to a coefficient Λ-algebra A is obtained from $\boldsymbol{\rho}$ via a unique morphism $\mathcal{R} \longrightarrow A$.

We call $\mathcal{R}$ the *universal deformation ring* and $\boldsymbol{\rho}$ the *universal deformation* of $\overline{\rho}$. The ring $\mathcal{R}(\Pi, \mathsf{k}, \overline{\rho})$ is unique in the following strong sense.

Theorem 3.11 (Mazur). *Suppose*

$$\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$$

is a continuous representation such that $C(\overline{\rho}) = \mathsf{k}$. If $\overline{\rho}\,'$ is a representation equivalent to $\overline{\rho} \otimes \chi$, where χ is a representation of dimension one, then there is a canonical isomorphism

$$r(\overline{\rho}\,', \overline{\rho}) : \mathcal{R}(\Pi, \mathsf{k}, \overline{\rho}) \longrightarrow \mathcal{R}(\Pi, \mathsf{k}, \overline{\rho}\,')$$

mapping the universal deformation of $\overline{\rho}$ to the universal deformation of $\overline{\rho}\,'$. This system of canonical isomorphisms satisfies the natural compatibility conditions.

Proof. This is basically immediate from the definition. See the next lecture for some hints about what is involved, and [**97**] for the details. □

Absolutely irreducible representations

Given the important role of the hypothesis that $C(\overline{\rho}) = \mathsf{k}$ in the theorem, it's important to ask which representations have this property. The most important part of the answer is *Schur's Lemma*, which says absolutely irreducible representations satisfy the condition $C(\overline{\rho}) = \mathsf{k}$.

Definition 3.3. A representation $\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$ is called *reducible* if the representation space k^n (with the Π-action given by $\overline{\rho}$) has a proper subspace that is invariant under the action of Π. It is called *irreducible* if no such subspace exists. Finally, we say that $\overline{\rho}$ is *absolutely irreducible* if there is no extension k'/k such that $\overline{\rho} \otimes_{\mathsf{k}} \mathsf{k}'$ is reducible.

The idea of "absolute" irreducibility is just this: it's perfectly possible for a representation to have no invariant subspaces as given, but for the subspaces to appear once we move to a larger field. For example, suppose we send a cyclic group of order 4 to $\mathrm{GL}_2(\mathbb{R})$, representing the generator by a ninety degree rotation. Then this representation has no fixed lines (i.e., no real eigenvalues). But if we base-change to $\mathbb{C}$, two fixed lines will appear. Hence our representation was irreducible but not absolutely irreducible.

Problem 3.8. Let $\overline{\mathsf{k}}$ be an algebraic closure of k. Show that $\overline{\rho}$ is absolutely irreducible if and only if $\overline{\rho} \otimes_{\mathsf{k}} \overline{\mathsf{k}}$ is irreducible.

The main reason we like absolutely irreducible representations is the following result:

Theorem 3.12 (Schur's Lemma). *If $\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$ is absolutely irreducible, then $C(\overline{\rho}) = \mathsf{k}$.*

The proof can be found in any standard text on group representation theory. For example, see page 7 of [**55**].

Hence, absolutely irreducible representations have universal deformations. Of course, there are other representations that satisfy $C(\overline{\rho}) = \mathsf{k}$. The following problems ring some changes on these ideas.

Problem 3.9. Find an example of a reducible representation which nevertheless satisfies the condition $C(\overline{\rho}) = \mathsf{k}$.

Problem 3.10. Show that if $\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_2(\mathsf{k})$ is irreducible and its image contains an element of exact order 2 and determinant -1, then $C(\overline{\rho}) = \mathsf{k}$. (The case of characteristic 2 has to be considered separately.)

Problem 3.11. Show that $\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_2(\mathsf{k})$ is absolutely irreducible if and only if it is irreducible and $C(\overline{\rho}) = \mathsf{k}$. (The same is true for representations to $\mathrm{GL}_n(\mathsf{k})$, but it's slightly harder to prove.)

Problem 3.12. We can extend a representation $\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$ to a continuous homomorphism of k-algebras

$$f_{\overline{\rho}} : \mathsf{k}[[\Pi]] \longrightarrow \mathrm{M}_n(\mathsf{k}).$$

Show that $\overline{\rho}$ is absolutely irreducible if and only if $f_{\overline{\rho}}$ is onto. (Note that Schur's Lemma follows from this result.)

Example: the case $n = 1$

As we will see in the next section, it can often be quite difficult to get concrete information about the universal deformation ring. In this section, we look at one example in which it is possible to get a complete description.

Let's consider, then, what happens when $n = 1$. If we consider the uniqueness statement above, we see that the deformation ring in the one-dimensional case does not depend (up to canonical isomorphism) on the representation at all (of course, the universal deformation *does* depend on the representation). We will confirm that by constructing the ring directly.

So let's start with a one-dimensional representation, that is, a character

$$\overline{\rho} = \overline{\chi} : \Pi \longrightarrow \mathsf{k}^{\times} = \mathrm{GL}_1(\mathsf{k}),$$

and try to describe the deformation ring explicitly. The basic idea is as follows: there is a canonical ("Teichmüller") lift of $\overline{\chi}$ to the ring of Witt vectors $W(\mathsf{k})$, and thence to Λ (which is a $W(\mathsf{k})$-algebra. Once that lift is given, the "rest" of any lift must differ from this one by a homomorphism with values in an abelian pro-p-group, and these we can describe in a somewhat explicit manner.

So let $\Gamma = \Pi^{\mathrm{ab},(p)}$ be the abelianization of the pro-p-completion (see Lecture 2 for the definition) of Π, and let $\gamma : \Pi \longrightarrow \Gamma$ be the canonical projection. Note that any map from Π to an abelian pro-p-group must factor through Γ.

Problem 3.13. Prove that any homomorphism from Π to an abelian pro-p-group must factor through the abelianization of its pro-p-completion. (Easy, but helps you remember the definitions.)

Problem 3.14. Is there a difference between the abelianization of the pro-p-completion and the pro-p-completion of the abelianization?

Problem 3.15. Let $\Pi = G_{\mathbb{Q},S}$. Describe Γ. (Use the description of the abelianization of Π in Lecture 1. You'll need to distinguish the cases $p \in S$ and $p \notin S$.)

Problem 3.16. Let $\Pi = G_{\mathbb{Q}_\ell}$. Describe Γ. (Use the description of the abelianization of Π in Lecture 1. You'll need to distinguish the cases $\ell = p$ and $\ell \neq p$.)

As usual, we will let Λ be a coefficient ring and work in the category $\mathcal{C}_\Lambda$ of coefficient Λ-algebras. (If we want to work with all coefficient rings we just take

$\Lambda = W(\mathsf{k})$.) Let $\Lambda[[\Gamma]]$ be the completed group ring over Λ, that is,

$$\Lambda[[\Gamma]] = \varprojlim_{H} \Lambda[\Gamma/H]$$

where H ranges through the open normal subgroups of Γ and $\Lambda[\Gamma/H]$ is the usual group ring of the (finite) group Γ/H over Λ. If $u \in \Gamma$, we write $[u]$ for the corresponding element in the group ring.

Problem 3.17. Show that $\Lambda[[\Gamma]]$ is a coefficient Λ-algebra, i.e., an object of $\mathcal{C}_\Lambda$.

Since Λ is a complete noetherian ring, and hence is henselian, the units of Λ split (canonically) into a product:

$$\Lambda^\times \cong \mathsf{k}^\times \times (1 + \mathfrak{m}_\Lambda).$$

Using this splitting we get a canonical lift of $\overline{\chi}$ to Λ, which we will call $\chi_0 : \Pi \longrightarrow \Lambda$.

Now we can state the theorem. Recall that $\gamma : \Pi \longrightarrow \Gamma$ is the canonical projection. Then we have:

Proposition 3.13. *The universal deformation ring for a character $\overline{\chi} : \Pi \longrightarrow \mathsf{k}^\times$ is $\mathcal{R}(\Pi, \mathsf{k}, \overline{\chi}) = \Lambda[[\Gamma]]$ and the universal deformation is given by*

$$\boldsymbol{\chi}(x) = \chi_0(x)[\gamma(x)].$$

Proof. First of all, we know, by hypothesis Φ_p, that Γ is finitely generated as a $\mathbb{Z}_p$-module. If r is the number of generators, then we know that $\Lambda[[\Gamma]]$ is a quotient of the power series ring $\Lambda[[X_1, X_2, \ldots, X_r]]$, and hence is a coefficient Λ-algebra. It's also clear that $\boldsymbol{\chi}$ is a character. It remains to show, then, that this is indeed the universal deformation.

Consider, then, a lift $\chi : \Pi \longrightarrow A^\times$ to some coefficient Λ-algebra A. Let $\psi = \chi_0^{-1}\chi$. Then it's clear that ψ is a character of Π taking values in $1 + \mathfrak{m}_A$. Since $1 + \mathfrak{m}_A$ is an abelian pro-p-group, ψ must factor through the homomorphism $\gamma : \Pi \longrightarrow \Gamma$; this defines a map $f_\chi : \Gamma \longrightarrow 1 + \mathfrak{m}_A$ which extends to a homomorphism of Λ-algebras $f_\chi : \Lambda[[\Gamma]] \longrightarrow A$. We then have $\chi = f_\chi \circ \boldsymbol{\chi}$. Thus, $\Lambda[[\Gamma]]$ is the universal deformation ring and $\boldsymbol{\chi}$ is the universal deformation of χ. □

Note that, as we pointed out above, $\Lambda[[\Gamma]]$ is independent of χ.

Problem 3.18. Check that $1 + \mathfrak{m}_A$ is indeed a pro-p-group.

Problem 3.19. Given two characters $\overline{\chi}_1, \overline{\chi}_2 : \Pi \longrightarrow \mathsf{k}^\times$, Theorem 3.11 says that there must be an isomorphism

$$r(\overline{\chi}_1, \overline{\chi}_2) : \Lambda[[\Gamma]] \longrightarrow \Lambda[[\Gamma]].$$

Can you describe this isomorphism?

Here's an interesting consequence of this calculation. Suppose we have a residual representation

$$\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k}).$$

Then we can look at its determinant, $\det \overline{\rho}$, which is a one-dimensional representation. If ρ is a deformation of $\overline{\rho}$ to a ring R, then clearly $\det \rho$ is a deformation of $\det \overline{\rho}$ to R. In particular, it follows that $\det \boldsymbol{\rho}$ is a deformation of $\det \overline{\rho}$ to the universal ring $\mathcal{R}(\Pi, \mathsf{k}, \overline{\rho})$. By universality, it follows that there is a map

$$\Lambda[[\Gamma]] \longrightarrow \mathcal{R} = \mathcal{R}(\Pi, \mathsf{k}, \overline{\rho}).$$

This homomorphism, which we will call the determinant homomorphism, makes $\mathcal{R}$ a $\Lambda[[\Gamma]]$-algebra. This extra structure is sometimes important.

Problem 3.20. Let $K = \mathbb{Q}$, p be an odd prime, $S = \{p\}$, $\Lambda = \mathbb{Z}_p$, and $\Pi = G_{\mathbb{Q},S}$. Show that $\Gamma \cong 1 + p\mathbb{Z}_p$ and therefore that the deformation ring $\mathbb{Z}_p[[\Gamma]]$ is the usual "Iwasawa algebra."

Extra Problems

This section collects a few extra problems related to the material in this lecture. First, a case where the deformation theory is easy to compute.

Problem 3.21. Suppose Π is a *finite* group of order not divisible by p, and suppose that $\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$ is an inclusion. Show that there exists a lift of $\overline{\rho}$ to $\mathrm{GL}_n(W(\mathsf{k}))$ (and hence, via the canonical map $W(\mathsf{k}) \longrightarrow \Lambda$, to $\mathrm{GL}_n(\Lambda)$). Show that every deformation of $\overline{\rho}$ factors through this lift, and hence that the universal deformation ring is $\mathcal{R}_{\overline{\rho}} = \Lambda$.

For other subgroups of $\mathrm{GL}_n(\mathsf{k})$, the problem is much harder. For example:

Problem 3.22. Suppose $\Pi = \mathrm{SL}_n(\mathsf{k}) \subset \mathrm{GL}_n(\mathsf{k})$ and $\overline{\rho}$ is the inclusion. Can you determine the universal deformation ring?

The next two problems have to do with the group $\Gamma_n(R)$, i.e., the kernel of the reduction map from $\mathrm{GL}_n(R)$ to $\mathrm{GL}_n(\mathsf{k})$.

Problem 3.23. Show that $\Gamma_n(\mathsf{k}[\varepsilon])$ is a finite p-elementary abelian group.

Problem 3.24. Show that if $R \longrightarrow S$ is a surjective homomorphism in $\mathcal{C}^0_\Lambda$ then the induced homomorphism of groups

$$\Gamma_n(R) \longrightarrow \Gamma_n(S)$$

is also surjective.

Finally, a few problems related to the notion of a "pro-representable hull" (the definition appears below).

Problem 3.25. Suppose $\mathbf{F}$ and $\mathbf{G}$ are set-valued functors on $\mathcal{C}^0_\Lambda$ such that both $\mathbf{F}(\mathsf{k})$ and $\mathbf{G}(\mathsf{k})$ are singletons. We say that a morphism of functors $\xi : \mathbf{F} \longrightarrow \mathbf{G}$ is *smooth* if, given any surjective homomorphism $B \longrightarrow A$ of artinian coefficient Λ-algebras, any element $f \in \mathbf{F}(A)$ and any lifting of $\xi(f) \in \mathbf{G}(A)$ to an element $g \in \mathbf{G}(B)$, there exists a lifting $f' \in \mathbf{F}(B)$ of f such that $\xi(f') = g$.

i. Show that this condition is equivalent to saying that for any surjective homomorphism $B \longrightarrow A$ of artinian Λ-algebras the natural mapping

$$\mathbf{F}(B) \longrightarrow \mathbf{F}(A) \times_{\mathbf{G}(A)} \mathbf{G}(B)$$

is surjective.

ii. Show that we can replace $\mathcal{C}^0_\Lambda$ by $\mathcal{C}_\Lambda$ in the definition.

iii. Show that if $\mathbf{F} \longrightarrow \mathbf{G}$ is smooth, the map $\mathbf{F}(B) \longrightarrow \mathbf{G}(B)$ is surjective for every coefficient Λ-algebra B.

iv. Suppose $\mathbf{F}$ is represented by a coefficient Λ-algebra R and $\mathbf{G}$ is represented by a coefficient Λ-algebra S. Then a morphism $\mathbf{F} \longrightarrow \mathbf{G}$ corresponds to a homomorphism $S \longrightarrow R$. If $\mathbf{F} \longrightarrow \mathbf{G}$ is smooth, what does that tell you about the corresponding ring homomorphism?

If R is an object of $\mathcal{C}_\Lambda$, let $\mathbf{h}_R$ denote the functor $\mathrm{Hom}(R, \cdot\,)$ that maps each coefficient Λ-algebra A to the set $\mathrm{Hom}(R, A)$.

Problem 3.26. Suppose $\mathbf{F}$ is a functor on $\mathcal{C}^0_\Lambda$ such that $\mathbf{F}(\mathsf{k})$ is a singleton. Show that any element $\rho \in \mathbf{F}(R)$ induces a morphism of functors $\mathbf{h}_R \longrightarrow \mathbf{F}$.

Problem 3.27. Let R_1 and R_2 be coefficient Λ-algebras and let $\varphi : R_1 \longrightarrow R_2$ be a homomorphism. Show that the following statements are equivalent:

- The homomorphism $\varphi : R_1 \longrightarrow R_2$ is surjective.
- The corresponding morphism of functors $\mathbf{h}_{R_2} \longrightarrow \mathbf{h}_{R_2}$ is injective.

Problem 3.28. Suppose $\mathbf{F}$ is a functor on $\mathcal{C}^0_\Lambda$ such that $\mathbf{F}(\mathsf{k})$ is a singleton, R is a coefficient Λ-algebra, and $\rho \in \mathbf{F}(R)$. We say the pair (R, ρ) is a *pro-representable hull* for $\mathbf{F}$ if

- the map $\mathbf{h}_R \longrightarrow \mathbf{F}$ induced by ρ is smooth, and
- the induced map $t_R \longrightarrow t_{\mathbf{F}}$ of tangent spaces is an isomorphism.

Prove that any two pro-representable hulls for $\mathbf{F}$ are isomorphic.

Problem 3.29. In the situation of the previous problem, prove that if $\mathbf{F}$ is representable by $\mathcal{R}$, then $\mathcal{R}$ (together with the universal element $\boldsymbol{\rho} \in \mathbf{F}(\mathcal{R})$ corresponding to the identity $\mathcal{R} \longrightarrow \mathcal{R}$) is a hull of $\mathbf{F}$. Are there other hulls of $\mathbf{F}$?

Problem 3.30. We mentioned above that since the deformation functor $\mathbf{D}_{\overline{\rho}}$ always satisfies conditions **H1** to **H3**, it has a hull $(\mathcal{R}, \rho)$. The representation ρ is sometimes called a *versal deformation* of $\overline{\rho}$. Translate the definition of a hull into deformation-theoretical terms: what properties does a versal deformation have?

Finally a meta-problem: in this and the following lectures, how much of the theory survives if all we have is a hull (or a "versal deformation")?

Complements to lecture 3

Mark Dickinson has given an alternative proof of the main theorem in this lecture that sidesteps the Schlessinger conditions, working instead with Grothendieck's theorem from the previous chapter. See Appendix 1 for Dickinson's proof.

LECTURE 4
The Universal Deformation: Properties

Now that we have proved that (under certain conditions) universal deformations exist, we want to find out more about them and about the universal deformation rings. This turns out, of course, to be quite difficult. In this lecture we will look at what can be said in general about the deformation ring and its properties. We will continue to use the notations we established above. In particular, Π is a profinite group satisfying the hypothesis Φ_p. More and more, however, we want to think of Π as being either $G_{K,S}$ for some number field K and some set of primes S including the archimedean primes or the absolute Galois group of a local field.

Functorial properties

The simplest properties of the universal deformation rings might be described as "functorial" properties. Basically, they are derived from various constructions involving group representations, together with the universality properties of the universal deformation. These properties are worked out in detail in [**97**], and we won't spend too much time on them. For this whole section, assume that $\overline{\rho}$ satisfies the condition $C(\overline{\rho}) = \mathsf{k}$, so that the deformation functor $\mathbf{D}_\Lambda$ is representable.

The kind of properties we want to consider here are those which arise simply from the fact that the functor is representable, together with the fact that GL_n itself has functorial properties (specifically, it is an affine group scheme of finite type over $\mathbb{Z}$). One example of this kind of property already appeared in the previous lecture, when we considered the determinant function

$$\det : \mathrm{GL}_n \longrightarrow \mathrm{GL}_1 .$$

Since the determinant is a homomorphism of affine group schemes over Λ, it sends deformations of $\overline{\rho}$ into deformations of $\det \overline{\rho}$. By the universal property, this gives a homomorphism of coefficient Λ-algebras from the completed group ring $\Lambda[[\Gamma]]$ (where $\Gamma = \Pi^{ab,(p)}$, as defined in the previous lecture) to the universal deformation ring $\mathcal{R}(\overline{\rho})$.

Another example of the same thing is as follows. Suppose we have two residual representations $\overline{\rho}$ and $\overline{\rho}'$ which are equivalent. Then there is a matrix $x \in \mathrm{GL}_n(W(\mathsf{k}))$ (which we can think of as in $\mathrm{GL}_n(\Lambda)$) such that $\overline{\rho} = \overline{x}^{-1}\overline{\rho}'\overline{x}$. Conjugation by x is an isomorphism of group schemes over Λ,

$$\delta_x : \mathrm{GL}_n \longrightarrow \mathrm{GL}_n,$$

and it transforms deformations of $\overline{\rho}'$ into deformations of $\overline{\rho}$. Therefore, it determines an isomorphism of deformation rings

$$r(\delta_x) : \mathcal{R}(\overline{\rho}) \longrightarrow \mathcal{R}(\overline{\rho}').$$

Problem 4.1. Show that the isomorphism $r(\delta_x)$ is determined by the representations $\overline{\rho}$ and $\overline{\rho}'$. (In other words, it does not depend on our choice of the particular matrix $x \in \mathrm{GL}_n(W(\mathsf{k}))$.)

Since $r(\delta_x)$ depends only on the residual representations $\overline{\rho}$ and $\overline{\rho}'$, we can denote it by $r(\overline{\rho}', \overline{\rho})$. This proves part of the uniqueness assertion in Theorem 3.11.

Problem 4.2. Play the same game with the "transpose-inverse" automorphism of GL_n to show that the universal deformation ring of $\overline{\rho}$ and of its contragredient (i.e., the representation $g \mapsto (\overline{\rho}(g)^{-1})^t$) are canonically isomorphic.

Problem 4.3. Suppose $C(\overline{\rho}_1) = \mathsf{k}$, $C(\overline{\rho}_2) = \mathsf{k}$, and $C(\overline{\rho}_1 \otimes \overline{\rho}_2) = \mathsf{k}$, so that all three representations have universal deformation rings. Given a deformation ρ_1 of $\overline{\rho}_1$ to a ring A_1 and a deformation ρ_2 of $\overline{\rho}_2$ to a ring A_2, show that the tensor product $\rho_1 \otimes \rho_2$ is a deformation of $\overline{\rho}_1 \otimes \overline{\rho}_2$ to the ring $A_1 \hat{\otimes} A_2$, and deduce a natural homomorphism

$$\mathcal{R}(\overline{\rho}_1 \otimes \overline{\rho}_2) \longrightarrow \mathcal{R}(\overline{\rho}_1) \hat{\otimes}_\Lambda \mathcal{R}(\overline{\rho}_2).$$

Problem 4.4. Still in the situation of the previous problem, suppose we pick a lift ρ_1 of $\overline{\rho}_1$ to $\mathrm{GL}_n(\Lambda)$. By the universal property, this corresponds to a map $h_1 : \mathcal{R}(\overline{\rho}_1) \longrightarrow \Lambda$. Use this and the map defined in the previous problem to deduce that there exists a homomorphism

$$\mathcal{R}(\overline{\rho}_1 \otimes \overline{\rho}_2) \longrightarrow \mathcal{R}(\overline{\rho}_2).$$

This is called "contraction with the lift ρ_1."

Problem 4.5. In the situation of the previous problem, suppose $\overline{\rho}_1$ is one-dimensional, i.e., is a character. In this case we sometimes refer to the homomorphism obtained in that problem as the "twisting homomorphism" corresponding to ρ_1. Show that the twisting homomorphism is in fact an isomorphism and that it satisfies the obvious "homomorphic" property with respect to ρ_1. (This problem, together with what was done above, completes the proof of Theorem 3.11.)

We can continue in this vein to consider, for example, what happens when we change the group Π (e.g., by restricting to a subgroup) and what happens when we change the base field k. See [**97**] for a careful discussion of all this.

Tangent spaces and cohomology groups

Fix a residual representation $\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$ such that $C(\overline{\rho}) = \mathsf{k}$ and a coefficient ring Λ. For this section, let $\mathbf{D} = \mathbf{D}_{\overline{\rho},\Lambda}$, in order to make the notation lighter. We know, then, that the functor $\mathbf{D}$ is representable; denote the representing coefficient Λ-algebra by $\mathcal{R}$. We have already introduced the tangent space $t_{\mathbf{D}} = \mathbf{D}(\mathsf{k}[\varepsilon])$ of the functor $\mathbf{D}$. Since $\mathbf{D}$ is represented by $\mathcal{R}$, we know that

$$t_{\mathbf{D}} = \mathbf{D}(\mathsf{k}[\varepsilon]) = \mathrm{Hom}_\Lambda(\mathcal{R}, \mathsf{k}[\varepsilon]) = \mathrm{Hom}_{\mathsf{k}}(\mathfrak{m}_{\mathcal{R}}/(\mathfrak{m}_{\mathcal{R}}^2, \mathfrak{m}_\Lambda), \mathsf{k}).$$

This is true for any representable functor; for the deformation functor, we can say a bit more. Suppose we know $\overline{\rho}(g) = a$, where a is some matrix in $\mathrm{GL}_n(\mathsf{k})$, and suppose that ρ_1 is a deformation of $\overline{\rho}$ to $\mathsf{k}[\varepsilon]$. Then we must have $\rho_1(g) = (1 + b_g\varepsilon)a$ for some matrix $b_g \in \mathrm{M}_n(\mathsf{k})$. In other words, ρ_1 determines (and is determined by) a map $b : \Pi \longrightarrow \mathrm{M}_n(\mathsf{k})$ mapping g to the matrix b_g. (Equivalently, the point is that $\mathrm{GL}_n(\mathsf{k}[\varepsilon])$ is the semi-direct product of $1+\varepsilon\mathrm{M}_n(\mathsf{k})$ and $\mathrm{GL}_n(\mathsf{k})$, and the first group is isomorphic to the additive group $\mathrm{M}_n(\mathsf{k})$.)

Imposing the condition that ρ_1 be a homomorphism boils down to saying that the map

$$g \mapsto b_g$$

should be a cocycle with values in $\mathrm{M}_n(\mathsf{k})$, where we make Π act on $\mathrm{M}_n(\mathsf{k})$ via conjugation:

$$g \cdot b = \overline{\rho}(g) b \overline{\rho}(g)^{-1}.$$

The k-vector space $\mathrm{M}_n(\mathsf{k})$ with this action of Π is usually called the *adjoint representation* of $\overline{\rho}$, and denoted $\mathrm{Ad}(\overline{\rho})$. One checks, then, that this association gives an isomorphism

$$t_{\mathbf{D}} \cong \mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho})).$$

This gives a connection between the deformation theory and the cohomology of $\mathrm{Ad}(\overline{\rho})$ which is quite important and which we will continue to examine in the next section.

Problem 4.6. Check that the map $g \mapsto b_g$ is a cocycle and that the cocycles corresponding to strictly equivalent lifts differ by a coboundary.

Problem 4.7. Check that the map $t_{\mathbf{D}} \longrightarrow \mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho}))$ defined above is indeed an isomorphism of k-vector spaces.

Notice that when Π is a Galois group, this puts us in the realm of Galois cohomology, which means it puts at our disposal an enormous array of techniques and theorems. We already have a simple numerical consequence:

Corollary 4.1. *Retain the assumptions and notations of this section, so that, in particular, a universal deformation of $\overline{\rho}$ exists and the universal deformation ring is $\mathcal{R}$. Let $d_1 = \dim_{\mathsf{k}} \mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho}))$. Then $\mathcal{R}$ is a quotient of a power series ring in d_1 variables over Λ.*

In other words, $\mathcal{R}$ fits into an exact sequence

$$0 \longrightarrow I \longrightarrow \Lambda[[X_1, X_2, \dots, X_{d_1}]] \longrightarrow \mathcal{R} \longrightarrow 0.$$

One possible approach to understanding $\mathcal{R}$, then, is to try to determine the dimension d_1 and the ideal I. In many situations, in fact, the value of d_1 is the crucial piece of information.

Tangent spaces and extensions of modules

We have already given several different interpretations of the tangent space to the deformation functor. One often uses still another one, relating the tangent space to extensions of Π-modules. We'll work out the basic correspondence, and leave the details to the reader.

Let, then

$$\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$$

be a residual representation, $\mathbf{D}$ be the deformation functor, and $t_{\mathbf{D}} = \mathbf{D}(\mathsf{k}[\varepsilon])$ be its tangent space. We want to establish a correspondence between elements of $t_{\mathbf{D}}$

and extensions of $\overline{\rho}$ by $\overline{\rho}$, by which we mean k-vector spaces E with an action of Π such that there is an exact sequence in the category of $\mathsf{k}[[\Pi]]$-modules

$$0 \longrightarrow V_{\overline{\rho}} \longrightarrow E \longrightarrow V_{\overline{\rho}} \longrightarrow 0,$$

where by $V_{\overline{\rho}}$ we mean[1] k^n with the action of Π given by the residual representation $\overline{\rho}$.

Suppose, first, that we are given an element of $t_{\mathbf{D}}$, that is, a deformation ρ of $\overline{\rho}$ to $\mathsf{k}[\varepsilon]$. Let M be $\mathsf{k}[\varepsilon]^n$ with the action of Π given by ρ. Clearly M is of dimension $2n$ as a vector space over k. To see that M fits into an exact sequence as above, consider the submodule εM and the quotient $M/\varepsilon M$. These are both clearly n-dimensional, and in fact they are both isomorphic to $V_{\overline{\rho}}$ as Π-modules.

Problem 4.8. Check that both εM and $M/\varepsilon M$ are free k-modules of dimension n on which Π acts via $\overline{\rho}$. (If $\{e_1, e_2, \ldots, e_n\}$ is a basis of M over $\mathsf{k}[\varepsilon]$, check that

$$\{\varepsilon e_1, \varepsilon e_2, \ldots, \varepsilon e_n\}$$

is a basis of εM over k and that

$$\{(e_1 \bmod \varepsilon M), (e_2 \bmod \varepsilon M), \ldots, (e_n \bmod \varepsilon M)\}$$

is a basis of $M/\varepsilon M$ over k. Then check that the action of Π is correct.)

Thus, if we identify both εM and $M/\varepsilon M$ with $V_{\overline{\rho}}$, we have the exact sequence we wanted. This shows that every element of $t_{\mathbf{D}}$ determines an extension. For the converse, suppose we are given a $2n$-dimensional k-vector space E which fits into an exact sequence

$$0 \longrightarrow V_{\overline{\rho}} \xrightarrow{\alpha} E \xrightarrow{\beta} V_{\overline{\rho}} \longrightarrow 0.$$

We then make E into a $\mathsf{k}[\varepsilon]$-module by defining multiplication by ε to be

$$\alpha \circ \beta : E \xrightarrow{\beta} V_{\overline{\rho}} \xrightarrow{\alpha} E$$

(this reverse composition makes sense, since the image of β is the same as the domain of α). It's easy to see that $(\alpha \circ \beta)^2 = 0$, since $\beta \circ \alpha = 0$ by the exactness of our sequence. In addition, since both α and β are homomorphisms of Π-modules, this $\mathsf{k}[\varepsilon]$-module structure commutes with the action of Π. We can now check directly that this makes E into a free $\mathsf{k}[\varepsilon]$-module of rank n with an action of Π, and therefore defines a representation $\Pi \longrightarrow \mathrm{GL}_n(\mathsf{k}[\varepsilon])$, which is clearly a deformation of $\overline{\rho}$.

Problem 4.9. Check the details. (For example, why is E a free $\mathsf{k}[\varepsilon]$-module? Why is the representation defined by E a lifting of $\overline{\rho}$? Do isomorphic E's give strictly equivalent deformations?)

Problem 4.10. Check that strictly equivalent deformations correspond to isomorphic extensions and conversely.

There is a standard k-vector space structure on the set of isomorphism classes of extensions of $V_{\overline{\rho}}$ by $V_{\overline{\rho}}$. This vector space is denoted $\mathrm{Ext}^1_{\mathsf{k}[[\Pi]]}(V_{\overline{\rho}}, V_{\overline{\rho}})$, and what we have shown so far is that there is a bijection

$$\mathbf{D}(\mathsf{k}[\varepsilon]) \longrightarrow \mathrm{Ext}^1_{\mathsf{k}[[\Pi]]}(V_{\overline{\rho}}, V_{\overline{\rho}}).$$

[1] This is somewhat nonstandard notation, but the usual notation, which is to call this space $\overline{\rho}$ (thereby identifying the homomorphism with the Π-module it induces), is a bit too confusing to use here.

Problem 4.11. (For those who know the vector space structure on Ext.) Show that this bijection is in fact an isomorphism of k-vector spaces.

Another point of view here is to think in terms of matrices. If E is a $2n$-dimensional k-vector space on which Π acts, the existence of an exact sequence

$$0 \longrightarrow V_{\overline{\rho}} \xrightarrow{\alpha} E \xrightarrow{\beta} V_{\overline{\rho}} \longrightarrow 0$$

amounts to saying that the representation

$$\rho_E : \Pi \longrightarrow \mathrm{GL}_{2n}(\mathsf{k})$$

corresponding to E can be put into the block form

$$\rho_E(g) = \begin{pmatrix} \overline{\rho}(g) & A_g \\ 0 & \overline{\rho}(g) \end{pmatrix}$$

with $A_g \in \mathrm{M}_n(\mathsf{k})$.

Problem 4.12. Show that the correspondence $g \mapsto A_g\overline{\rho}(g)^{-1}$ is a 1-cocycle and therefore determines an element of $\mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho}))$. Show that the resulting map

$$\mathrm{Ext}^1_{\mathsf{k}[[\Pi]]}(V_{\overline{\rho}}, V_{\overline{\rho}}) \longrightarrow \mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho}))$$

is an isomorphism.

Thus, we have established canonical isomorphisms

$$t_{\mathbf{D}} = \mathbf{D}(\mathsf{k}[\varepsilon]) \cong \mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho})) \cong \mathrm{Ext}^1_{\mathsf{k}[[\Pi]]}(V_{\overline{\rho}}, V_{\overline{\rho}}).$$

This gives us one more way to get at the tangent space (and especially its dimension).

Obstructed and unobstructed deformation problems

The use of the notation d_1 for the dimension of the tangent space probably signals that there is a d_2 about to show up. This is indeed the case. We deepen the connection between deformation theory and cohomology by using a standard idea in deformation theory: we try to compute the obstruction to lifting a homomorphism.

Keep the notations and assumptions as above. Suppose we have rings R_1 and R_0 in $\mathcal{C}_\Lambda$, and a surjective coefficient Λ-algebra homomorphism $R_1 \longrightarrow R_0$ with kernel I satisfying $I \cdot \mathfrak{m}_{R_1} = 0$ (in particular, we could be working with a small homomorphism). Because of the last assumption, we can (and do) view I as a k-vector space. Suppose we are given a homomorphism $\rho : \Pi \longrightarrow \mathrm{GL}_n(R_0)$. What keeps us from finding a deformation to R_1?

Well, we can certainly find a set-theoretic lift, i.e., a function $\gamma : \Pi \longrightarrow \mathrm{GL}_n(R_1)$ that lifts ρ. To test whether this is a homomorphism, we would have to compute

$$c(g_1, g_2) = \gamma(g_1g_2)\gamma(g_2)^{-1}\gamma(g_1)^{-1}$$

for every g_1, $g_2 \in \Pi$. If γ were a homomorphism, we would have $c(g_1, g_2) = 1$. Since it *is* a homomorphism modulo I, we do know that

$$c(g_1, g_2) = 1 + d(g_1, g_2)$$

with $d(g_1, g_2) \in \mathrm{M}_n(I) \cong \mathrm{Ad}(\overline{\rho}) \otimes_{\mathsf{k}} I$. It now isn't too hard to check that $d(g_1, g_2)$ is a 2-cocycle with values in $\mathrm{Ad}(\overline{\rho}) \otimes_{\mathsf{k}} I$, and that replacing γ by a different lift changes this cocycle by a coboundary.

Problem 4.13. Check this!

Therefore, the cocycle $d(g_1, g_2)$ gives an element $\mathcal{O}(\rho_0)$ in the cohomology $\mathrm{H}^2(\Pi, \mathrm{Ad}(\overline{\rho}) \otimes_{\mathsf{k}} I) \cong \mathrm{H}^2(\Pi, \mathrm{Ad}(\overline{\rho})) \otimes_{\mathsf{k}} I$, and this element is trivial if and only if there exists a homomorphism $\Pi \longrightarrow \mathrm{GL}_n(R_1)$ lifting ρ_0. We might call $\mathcal{O}(\rho_0)$ the obstruction class of ρ_0 relative to $R_1 \longrightarrow R_0$.

In general, one can't readily compute obstruction classes. However, the fact that liftings exist exactly when $\mathcal{O}(\rho_0) = 0$ means that if $\mathrm{H}^2(\Pi, \mathrm{Ad}(\overline{\rho})) = 0$ the deformation problem should be especially simple. And this is indeed the case, as Mazur showed.

Theorem 4.2. *Suppose $C(\overline{\rho}) = \mathsf{k}$ and let $\mathcal{R} = \mathcal{R}(\Pi, \mathsf{k}, \overline{\rho})$ be the universal deformation ring representing the deformation functor $\mathbf{D}_\Lambda$. Let*

$$d_1 = \dim \mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho})) \qquad \textit{and} \qquad d_2 = \dim \mathrm{H}^2(\Pi, \mathrm{Ad}(\overline{\rho})).$$

Then we have

$$(**) \qquad \operatorname{Krull}\dim(\mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R}) \geq d_1 - d_2.$$

Furthermore, if $d_2 = 0$ we have equality in (**), and in fact

$$\mathcal{R} \cong \Lambda[[X_1, X_2, \ldots, X_{d_1}]].$$

Proof. We already know that there is a surjective homomorphism of coefficient Λ-algebras

$$\Lambda[[T_1, T_2, \ldots, T_{d_1}]] \longrightarrow \mathcal{R}$$

which induces an isomorphism on tangent spaces. Reducing modulo the maximal ideal gives a homomorphism $\mathsf{k}[[T_1, T_2, \ldots, T_{d_1}]] \longrightarrow \mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R}$ which still induces an isomorphism on tangent spaces, and therefore is still surjective. Let J be the kernel of this surjection. To save on notation, write $F = \mathsf{k}[[T_1, T_2, \ldots, T_{d_1}]]$ and let $\mathfrak{m}_F$ be its maximal ideal. We have an exact sequence

$$0 \longrightarrow J \longrightarrow F \longrightarrow \mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R} \longrightarrow 0$$

in which the surjective homomorphism $F \longrightarrow \mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R}$ induces an isomorphism of tangent spaces. What we need to prove is that the minimal number of generators for J is at most d_2.

Since $\mathfrak{m}_F J \subset J$, the sequence of k-vector spaces

$$0 \longrightarrow J/\mathfrak{m}_F J \longrightarrow F/\mathfrak{m}_F J \longrightarrow \mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R} \longrightarrow 0$$

is still exact and the map on the right still induces an isomorphism of tangent spaces. Hence, the Krull dimension of $\mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R}$ is at least $d_1 - \dim_{\mathsf{k}}(J/\mathfrak{m}_F J)$. (Equivalently, the minimal number of generators for J is at most $\dim_{\mathsf{k}}(J/\mathfrak{m}_F J)$.)

Let $\boldsymbol{\rho}_p$ be image of the universal deformation $\boldsymbol{\rho}$ under the quotient map $\mathcal{R} \longrightarrow \mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R}$. It is clear that $\boldsymbol{\rho}_p$ is universal among deformations of $\overline{\rho}$ to Λ-algebras killed by $\mathfrak{m}_\Lambda$ (equivalently, to k-algebras). The construction above gives a cohomology class

$$\mathcal{O}(\boldsymbol{\rho}_p) \in \mathrm{H}^2(\Pi, \mathrm{Ad}(\overline{\rho})) \otimes J/\mathfrak{m}_F J$$

which is the obstruction to lifting $\boldsymbol{\rho}_p$ to $F/\mathfrak{m}_F J$.

Consider the k-linear map

$$\mathrm{Hom}_{\mathsf{k}}(J/\mathfrak{m}_F J, \mathsf{k}) \xrightarrow{\alpha} \mathrm{H}^2(\Pi, \mathrm{Ad}(\overline{\rho}))$$

given by

$$f \mapsto (1 \otimes f)(\mathcal{O}(\boldsymbol{\rho}_p)).$$

If we can show α is injective, then we will have $\dim_{\mathsf{k}}(J/\mathfrak{m}_F J) \leq d_2$, whence we can conclude that the Krull dimension of $\mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R}$ is at least $d_1 - d_2$, as claimed.

To prove the injectivity of α, let f be a nonzero element in the kernel, let A be the quotient of $F/\mathfrak{m}_F J$ by the kernel of f and let I be the image of $J/\mathfrak{m}_F J$ in the quotient,so that $I = (J/\mathfrak{m}_F J)/\operatorname{Ker}(f) = \operatorname{Im}(f) = \mathsf{k}$. Then we get an exact sequence

$$0 \longrightarrow I \longrightarrow A \longrightarrow \mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R} \longrightarrow 0$$

where I is isomorphic to k and which still induces an isomorphism on tangent spaces (check!). But now the obstruction to lifting $\boldsymbol{\rho}_p$ to A vanishes. Thus we get a deformation of $\overline{\rho}$ to A lifting $\boldsymbol{\rho}_p$. But A is a k-algebra and $\boldsymbol{\rho}_p$ is universal among lifts to such rings, so this lift must be induced by a homomorphism $\mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R} \longrightarrow A$. This means that the sequence splits, but this and the fact that $I \neq 0$ contradict the fact that $A \longrightarrow \mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R}$ induces an isomorphism of tangent spaces. Thus, there cannot be a nonzero element f in the kernel of α, that is, α is injective as claimed. This proves the inequality.

The last assertion follows at once. If $d_2 = 0$, then the kernel of the surjective homomorphism of coefficient Λ-algebras

$$\Lambda[[T_1, T_2, \ldots, T_{d_1}]] \longrightarrow \mathcal{R}$$

has at most 0 generators. Hence, $\mathcal{R} \cong \Lambda[[T_1, T_2, \ldots, T_{d_1}]]$ in this case; in particular, equality holds in $(**)$. □

When we are in the situation where $d_2 = 0$, we say the lifting problem is *unobstructed.*

In the examples where the deformation ring of an absolutely irreducible $\overline{\rho}$ has been explicitly computed (see the next lecture for how one might do such a thing), the Krull dimension has always turned out to be equal to $d_1 - d_2$. Hence, one might conjecture that one always has the equality.

Conjecture. If $\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$ is an absolutely irreducible residual representation, and $\mathcal{R}$ is the universal deformation ring, we have

$$\operatorname{Krull\,dim}(\mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R}) = d_1 - d_2.$$

We refer to this as the "Dimension Conjecture." Böckle has shown in [**6**] that this conjecture does not hold in some cases where $\overline{\rho}$ is reducible but still has a universal deformation (because it satisfies $C(\overline{\rho}) = \mathsf{k}$). He has also been able to show that it is true in many cases.

Mazur points out in [**97**] that one should think of this conjecture as a generalization of Leopoldt's Conjecture. To see why will require computing d_1 and d_2 when Π is a global Galois group, and we will do this in the next section.

Galois representations

Let K be a number field, S a finite set of primes in K. We will assume S includes all the primes above p and also (as we have assumed from the beginning) the primes at infinity. Let $S_\infty \subset S$ be the set of primes at infinity. Finally, let $\Pi = G_{K,S}$ and let

$$\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$$

be a residual representation such that $C(\overline{\rho}) = \mathsf{k}$, and let $\mathcal{R}$ be its universal deformation ring. From our work in the previous section, we know a lower bound for the dimension of $\mathcal{R}$, expressed in terms of the dimensions of two cohomology groups. The goal of this section is to compute that bound somewhat more explicitly using known results about Galois cohomology. It turns out that we will be able to obtain a better formula for $d_1 - d_2$, but not for d_1 and d_2 separately.

The result we need from Galois cohomology is Tate's global Euler characteristic formula. Here's what it says. We take an extension $K/\mathbb{Q}$ of degree d, S a finite set of primes in K including all the infinite primes, M a *finite* $G_{K,S}$-module such that S contains all the primes that divide the order of M. (In our application, the order of M will be a power of p.) For each prime v of K, let K_v be the completion at v. In particular, if $v \in S_\infty$, K_v is either $\mathbb{R}$ or $\mathbb{C}$. Then the global Euler characteristic formula (see [**156**], [**68**], or [**107**]) says that

$$\frac{\#\mathrm{H}^0(G_{K,S}, M)\cdot\#\mathrm{H}^2(G_{K,S}, M)}{\#\mathrm{H}^1(G_{K,S}, M)} = \frac{1}{(\#M)^d}\prod_{v\in S_\infty}\#\mathrm{H}^0(G_{K_v}, M).$$

In our situation M will be $\mathrm{Ad}(\overline{\rho})$, which is a k-vector space and hence will have order a power of p and S will include all primes above p. Since the cohomology groups in this case will also be k-vector spaces, all of the groups in the formula have order a power of p, and we can translate the formula into a statement about dimensions:

$$\begin{aligned}\dim\mathrm{H}^0(G_{K,S}, M) - \dim\mathrm{H}^1(G_{K,S}, M) + \dim\mathrm{H}^2(G_{K,S}, M) &= \\ = \sum_{v\in S_\infty}\dim\mathrm{H}^0(G_{K_v}, M) - d\dim M&,\end{aligned}$$

where all of the dimensions are dimensions over k.

Now let $M = \mathrm{Ad}(\overline{\rho})$, and write $d_i = \dim\mathrm{H}^i(G_{K,S}, \mathrm{Ad}(\overline{\rho}))$ as before. Then the formula becomes

$$d_0 - d_1 + d_2 = \sum_{v\in S_\infty}\dim\mathrm{H}^0(G_{K_v}, \mathrm{Ad}(\overline{\rho})) - dn^2,$$

and hence

$$d_1 - d_2 = d_0 + dn^2 - \sum_{v\in S_\infty}\dim\mathrm{H}^0(G_{K_v}, \mathrm{Ad}(\overline{\rho})).$$

But d_0 we can compute:

$$\mathrm{H}^0(G_{K,S}, \mathrm{Ad}(\overline{\rho})) = (\mathrm{Ad}(\overline{\rho}))^{G_{K,S}}$$

is the set of matrices in $\mathrm{M}_n(\mathsf{k})$ fixed by the conjugation action of $G_{K,S}$, i.e., it is $C(\overline{\rho}) = \mathsf{k}$. So $d_0 = 1$. The upshot, then, is the following.

Proposition 4.3. *Let K be a number field of degree d over $\mathbb{Q}$, let $\overline{\rho} : G_{K,S} \longrightarrow \mathrm{GL}_n(\mathsf{k})$ be a residual representation such that $C(\overline{\rho}) = \mathsf{k}$, and let $\mathcal{R}$ be its universal deformation ring. Then*

$$\operatorname{Krull}\dim\mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R} \geq 1 + dn^2 - \sum_{v\in S_\infty}\dim\mathrm{H}^0(G_{K_v}, \mathrm{Ad}(\overline{\rho})).$$

The advantages of this formula are two. First, it only refers to the 0-th cohomology groups, which are just the fixed points under the Galois action and so are relatively easy to compute. Second, the groups acting are the G_{K_v} for v an archimedean prime, so that G_{K_v} has order two if v is real and order one if v is complex.

Let's work out what the formula gives in the two most interesting cases. First, let K be a number field and $\overline{\rho}$ be a character (i.e., a one-dimensional representation). As we saw above, the universal deformation does not depend on $\overline{\rho}$ in this case; in fact, we showed that

$$\mathcal{R} = \Lambda[[G_{K,S}^{ab,(p)}]].$$

Now notice that

$$\mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R} = \mathsf{k}[[G_{K,S}^{ab,(p)}]],$$

so the Krull dimension of this ring is equal to the rank of $G_{K,S}^{ab,(p)}$ as a $\mathbb{Z}_p$-module, or, equivalently, to the rank of $\operatorname{Hom}(G_{K,S}, \mathbb{Z}_p)$ as a $\mathbb{Z}_p$-module.

On the other hand, the formula above says that the Krull dimension is at least $1 + r_2$, where r_2 is the number of complex primes of K. So we have shown that

$$\operatorname{rank}_{\mathbb{Z}_p} \operatorname{Hom}(G_{K,S}, \mathbb{Z}_p) \geq 1 + r_2,$$

which is in fact a well-known result. The assertion that these two numbers are equal is equivalent to the *Leopoldt Conjecture* for the field K. This is why we said that the general dimension conjecture should be viewed as a vastly generalized Leopoldt Conjecture.

The Leopoldt Conjecture is known to be true for abelian extensions of $\mathbb{Q}$, and for abelian extensions of quadratic imaginary fields, so in that case the deformation ring has the expected dimension. The general case seems very elusive.

Problem 4.14. Check these computations.

Problem 4.15. Look up the classical statement of the Leopoldt Conjecture and explain why it is equivalent to

$$\operatorname{rank}_{\mathbb{Z}_p} \operatorname{Hom}(G_{K,S}, \mathbb{Z}_p) = 1 + r_2.$$

The next case in which we want to go through the computation is the case that is related to modular forms and elliptic curves: $n = 2$, p an odd prime, $K = \mathbb{Q}$, S containing p and ∞. In this situation there is only one infinite prime, and G_∞ is a group of order two generated by the complex conjugation σ. Since $\sigma^2 = 1$ and p is odd, $\overline{\rho}(\sigma)$ is a matrix of order 2 in $\mathrm{GL}_2(\mathsf{k})$, and hence we must have

$$\overline{\rho}(\sigma) \sim \pm\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \qquad \text{or} \qquad \overline{\rho}(\sigma) \sim \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}.$$

In the first case, $\det\overline{\rho}(\sigma) = 1$, and we call $\overline{\rho}$ an *even* representation. In the second, $\det\overline{\rho}(\sigma) = -1$ and we say $\overline{\rho}$ is *odd.*

Now it's easy to compute the dimension d_0 of $\mathrm{H}^0(G_\infty, \mathrm{Ad}(\overline{\rho}))$. If $\overline{\rho}$ is even, then $\overline{\rho}(\sigma)$ is a scalar matrix, and hence the action of G_∞ on $\mathrm{Ad}(\overline{\rho})$ is trivial, so $d_0 = 4$. If $\overline{\rho}$ is odd, then an easy computation shows that $d_0 = 2$. Plugging all this into the formula gives:

Proposition 4.4. *Let p be an odd prime, let S be a set of rational primes including p and ∞, let $\overline{\rho} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathsf{k})$ be a residual representation satisfying $C(\overline{\rho}) = \mathsf{k}$, and let $\mathcal{R}$ be the universal deformation ring of $\overline{\rho}$. Then:*

- *if $\overline{\rho}$ is even, then* $\operatorname{Krull}\dim \mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R} \geq 1$, *and*
- *if $\overline{\rho}$ is odd, then* $\operatorname{Krull}\dim \mathcal{R}/\mathfrak{m}_\Lambda\mathcal{R} \geq 3$.

One conjectures, at least when $\overline{\rho}$ is absolutely irreducible, that both of these inequalities are in fact equalities, i.e., that the cohomological constraints on the dimension are the only constraints. This is indeed the case in many cases that have been computed explicitly.

It is worth noting that the representations coming from elliptic curves and from modular forms are always odd.

Problem 4.16. What happens if $p = 2$? In this case, the distinction between $\det \overline{\rho}(\sigma) = 1$ and $\det \overline{\rho}(\sigma) = -1$ vanishes; can the computation still be done?

Problem 4.17. (Hard) Work out $d_1 - d_2$ when $\Pi = G_{\mathbb{Q}_p}$.

LECTURE 5
Explicit Deformations

So far, we have developed an elaborate theory about the universal deformation, but (except for the case of GL_1) we have not been able to get our hands on one in any concrete way. In this lecture, we want to discuss a point of view, going back to the work of Boston in [**8**], [**9**], and [**10**], which allows us, in many cases, to get a rather explicit description of the universal deformation ring of a Galois representation. We will try to describe the basic idea, and then give some sample theorems that have been proved using this method.

The basic setup

The first thing to do is to "see" our residual representation $\overline{\rho}$ in terms of field extensions. Take, as before, S to be a finite set of primes in $\mathbb{Q}$ including p and ∞, $\mathbb{Q}_S$ the maximal extension of $\mathbb{Q}$ unramified outside S, $\Pi = G_{\mathbb{Q},S} = G(\mathbb{Q}_S/\mathbb{Q})$, and suppose $\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathsf{k})$ is absolutely irreducible or, more generally, assume that $C(\overline{\rho}) = \mathsf{k}$. (This is mostly for convenience; the method also allows us to compute "versal deformations" in cases where the deformation functor is not representable.) We want to understand the deformation theory of $\overline{\rho}$.

$$\begin{tikzcd}
\mathbb{Q}_S \arrow[d, dash, "G(\mathbb{Q}_S/K)\cong\Pi_0=\mathrm{Ker}(\overline{\rho})"] \\
K \arrow[d, dash, "G(K/\mathbb{Q})=H\cong\mathrm{Im}(\overline{\rho})"] \\
\mathbb{Q}
\end{tikzcd}$$

Let $\Pi_0 = \mathrm{Ker}(\overline{\rho})$, and let K be the fixed field of Π_0, so that we have a tower of fields $\mathbb{Q} \subset K \subset \mathbb{Q}_S$, with Galois groups $H = G(K/\mathbb{Q}) \cong \mathrm{Im}(\overline{\rho})$ and $G(\mathbb{Q}_S/K) \cong \Pi_0$. Let S_1 be the set of primes of K which lie above S.

Let $\boldsymbol{\rho} : \Pi \longrightarrow \mathrm{GL}_n(\mathcal{R})$ be the universal deformation of $\overline{\rho}$. As above, write $\Gamma_n(\mathcal{R})$ for the kernel of the natural projection $\mathrm{GL}_n(\mathcal{R}) \longrightarrow \mathrm{GL}_n(\mathsf{k})$. If $\gamma \in \Pi_0$, then $\overline{\rho}(\gamma) = 1$, and therefore $\boldsymbol{\rho}(\gamma) \in \Gamma_n(\mathcal{R})$. Thus, the restriction of $\boldsymbol{\rho}$ to Π_0 gives a homomorphism $\Pi_0 \longrightarrow \Gamma_n(\mathcal{R})$. We note, then, that

Lemma 5.1. *For any ring R in $\mathcal{C}$, $\Gamma_n(R)$ is a pro-p-group.*

Proof. This appeared as a problem in a previous lecture. The basic idea of the proof is this: one writes R as the inverse limit of the quotients $R/\mathfrak{m}^k$, where $\mathfrak{m}$ is the maximal ideal in R. Then

$$\Gamma_n(R) = \varprojlim_k \Gamma_n(R/\mathfrak{m}^k).$$

To prove the lemma, note first that $\Gamma_n(R/\mathfrak{m}) = \Gamma_n(\mathsf{k}) = \{1\}$ is a p-group, and then consider each of the transition homomorphisms

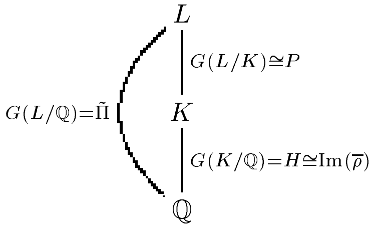
$$\Gamma_n(R/\mathfrak{m}^k) \longrightarrow \Gamma_n(R/\mathfrak{m}^{k-1}).$$

By induction, we know the image is a p-group. The kernel consists of those matrices whose off-diagonal entries are in the ideal $\mathfrak{m}^{k-1}/\mathfrak{m}^k$ and whose diagonal entries are in $1 + \mathfrak{m}^{k-1}/\mathfrak{m}^k$, i.e., it is

$$1 + \mathrm{M}_n(\mathfrak{m}^{k-1}/\mathfrak{m}^k).$$

One checks easily that this multiplicative group is isomorphic to the *additive* group $\mathrm{M}_n(\mathfrak{m}^{k-1}/\mathfrak{m}^k)$, which is easily seen to be a p-group. □

Problem 5.1. To complete the proof of the Lemma, check that the additive group $\mathrm{M}_n(\mathfrak{m}^{k-1}/\mathfrak{m}^k)$ is a p-group.


L
$G(L/K)\cong P$
$G(L/\mathbb{Q})=\tilde{\Pi}$
K
$G(K/\mathbb{Q})=H\cong\mathrm{Im}(\overline{\rho})$
$\mathbb{Q}$

So here's what we have: the universal deformation $\boldsymbol{\rho}$ induces a homomorphism $\Pi_0 \longrightarrow \Gamma_n(\mathcal{R})$, and $\Gamma_n(\mathcal{R})$ is a pro-p-group. Therefore the homomorphism $\Pi_0 \longrightarrow \Gamma_n(\mathcal{R})$ must factor through some pro-p quotient of Π_0. Any such quotient will be the Galois group of a pro-p-extension of K, so let L be the maximal pro-p-extension of K unramified outside S_1. Then $P = G(L/K)$ is a pro-p-group (in fact, it is the maximal continuous pro-p-quotient of Π_0), and we see that $\boldsymbol{\rho}$ must factor through $\tilde{\Pi} = G(L/\mathbb{Q})$. (Boston and Mazur call $\tilde{\Pi}$ the *p-completion of* Π *relative to* $\overline{\rho}$.)

We've shown, then, that the universal deformation $\boldsymbol{\rho}$ must factor through the quotient $\tilde{\Pi}$ of $G_{\mathbb{Q},S}$. It follows that all the deformations must factor through $\tilde{\Pi}$. Hence, the upshot of this discussion is that we can replace Π with $\tilde{\Pi}$ when studying the deformation theory of $\overline{\rho}$. The crucial feature of $\tilde{\Pi}$ is that it has a big normal subgroup P which is a pro-p-group, and the quotient $\tilde{\Pi}/P$ is isomorphic to the image of $\overline{\rho}$, so that the sequence

$$1 \longrightarrow P \longrightarrow \tilde{\Pi} \longrightarrow \mathrm{Im}(\overline{\rho}) \longrightarrow 1$$

is exact. The basic idea is now the following: to understand deformations of $\overline{\rho}$ to a coefficient Λ-algebra R, we need to understand all maps from P to $\Gamma_n(R)$ and then to consider how they may be extended to all of $\tilde{\Pi}$ in such a way as to be a deformation of $\overline{\rho}$. It turns out that we know enough about pro-p-extensions of number fields that in many cases this program can be pushed through to give a good description of $\mathcal{R}$ and often also of the universal deformation $\boldsymbol{\rho}$.

Rather than do this in the most general case, we will focus on the simpler case of "tame representations."

Definition 5.1. We say a residual representation $\overline{\rho}$ is *tame* if the order of $\mathrm{Im}(\overline{\rho})$ is not divisible by p.

Notice that in the tame case, the sequence

$$1 \longrightarrow P \longrightarrow \tilde{\Pi} \longrightarrow \mathrm{Im}(\overline{\rho}) \longrightarrow 1$$

tells us that $\tilde{\Pi}$ is a profinite group with a normal pro-p-Sylow subgroup, which allows us to get at the structure of $\tilde{\Pi}$ in a pretty explicit way.

Group theory

In this section, we summarize some group-theoretical results that will help us understand the universal deformation of a tame representation. The relevance of these results was first pointed out by Boston, whose exposition in [**9**] we follow.

As we saw above, when the residual representation $\overline{\rho}$ is tame, any deformation factors through a profinite group $\tilde{\Pi}$ which has a normal pro-p-Sylow subgroup P such that

$$\tilde{\Pi}/P \cong \mathrm{Im}(\overline{\rho}).$$

In this situation, we can apply the following theorem:

Theorem 5.2 (Schur-Zassenhaus). *Let G be a profinite group with normal pro-p-Sylow subgroup P of finite index in G. Let $\pi : G \longrightarrow G/P$ be the projection on the quotient. Then G contains a subgroup A such that π induces an isomorphism $A \xrightarrow{\cong} G/P$. Furthermore, any two subgroups with this property are conjugate by an element of P.*

(See [**120**, p. 246].)

As a consequence, G is the semi-direct product of P and A. We will exploit this later to define a homomorphism on G by defining it on A and on P in a compatible way.

Problem 5.2. Suppose G is as above, G' is a topological group, and we are given continuous homomorphisms $\alpha : A \longrightarrow G'$ and $\beta : P \longrightarrow G'$. Under what conditions do α and β together define a homomorphism $G \longrightarrow G'$?

The group P is a pro-p-group. If P is topologically finitely generated, it is a quotient of a free pro-p-group on finitely many generators. The minimal number of generators for P is called the *generator rank* of P and sometimes denoted $d(P)$. The kernel of the map from the free pro-p-group on $d(P)$ generators to P is itself finitely generated; the minimal number of generators for the kernel is called the *relation rank* of P, and sometimes denoted $r(P)$.

The Burnside Basis Theorem (Problem 2.2) provides a way to compute the generator rank. To make the notation simpler, let $\overline{P} = \mathrm{Fr}(P)$ be the Frattini quotient of P, i.e., the maximal p-elementary abelian (continuous) quotient of P. The Burnside Basis Theorem says that if $x_1, x_2, \dots, x_d$ are elements of P such that their image in $\overline{P}$ generates $\overline{P}$, then $x_1, x_2, \dots, x_d$ topologically generate P.

The following strengthening of the theorem is due to Boston (see [**9**]).

Theorem 5.3. *Let G be a profinite group with normal pro-p-Sylow subgroup P of finite index in G, and let A be a subgroup of G mapping isomorphically to G/P. Let A act on P and on $\overline{P}$ by conjugation. If $\overline{V}$ is an $\mathbb{F}_p[A]$-submodule of $\overline{P}$, then there exists an A-invariant subgroup V of P with $\dim_{\mathbb{F}_p} \overline{V}$ generators which maps onto $\overline{V}$ under π.*

For example, consider the case when $\overline{V}$ is one-dimensional. Then the assumption is that we have $\overline{x} \in \overline{P}$ such that $a^{-1}\overline{x}a = \overline{x}^{\phi(a)}$ for some character $\phi : A \longrightarrow \mathbb{F}_p^{\times}$. In this case, the theorem says that $\overline{x}$ can be lifted to an element $x \in P$ such that $a^{-1}xa = x^{\psi(a)}$, where $\psi : A \longrightarrow \mathbb{Z}_p^{\times}$ is a character lifting ϕ. (Since A is finite, this must be a character of finite order, and hence must be the Teichmüller lift of ϕ.)

One final bit of general group representation theory: suppose $H \subset \mathrm{GL}_n(\mathsf{k})$ is a subgroup whose order is prime to p, and let M be the adjoint representation of

H, that is $M = \mathrm{M}_n(\mathsf{k})$ with H acting by conjugation. Suppose V is some other finite-dimensional k-vector space with an action of H; as usual, we can think of M and V as modules over the group ring $\mathsf{k}[H]$. Since the order of H is not divisible by p, Maschke's Theorem says that every $\mathsf{k}[H]$-module can be written as a direct sum of irreducible $\mathsf{k}[H]$-modules.

Definition 5.2. We will say V is *prime to adjoint* if V and M have no irreducible sub-representations in common.

The "prime-to-adjoint" condition plays a big role in understanding certain deformation problems. See [**97**, section 1.12], [**9**, section 2], [**42**, section 3.6], and [**7**] for instances where this condition plays a significant role.

Problem 5.3. Show that the subspace of M consisting of matrices whose trace is zero is an H-invariant subspace. What is the complementary subspace? (You may want to assume the p does not divide n.)

Problem 5.4. Let R be a coefficient ring, and let

$$K_r = \mathrm{Ker}(\Gamma_n(R) \longrightarrow \Gamma_n(R/\mathfrak{m}^r)).$$

Suppose H is a finite subgroup of $\mathrm{GL}_n(\mathsf{k})$ of order prime to p. Show that K_{r-1}/K_r has a natural $\mathsf{k}[H]$-module structure, and that it is isomorphic to a multiple of the adjoint representation of H (that is, a direct sum of several copies of the adjoint representation).

Problem 5.5. Let R be a coefficient ring, let X be a topologically finitely generated closed subgroup of $\Gamma_n(R)$, and let H be a finite subgroup of $\mathrm{GL}_n(R)$ whose order is not divisible by p and which normalizes X. Suppose that the Frattini quotient $\bar{X}$, viewed as a $\mathsf{k}[H]$-module, is prime-to-adjoint. Show that X must be trivial. (Hint: Using the notation of the previous problem, let r be minimal such that X is not contained in K_r, and consider the image of X in K_{r-1}/K_r.)

The last two problems are taken from [**9**, section 2]. The last one is particularly significant: it highlights the importance of the adjoint representation and explains why the prime-to-adjoint condition is so significant.

Pro-p-extensions

In the picture above, we have a number field K, a set of primes S_1 including all primes above p and all archimedean primes, and an extension L/K which is the maximal pro-p-extension of K unramified outside S_1. Our goal in this section is to collect some of the known facts about the Galois group $P = G(L/K)$. We again follow Boston's exposition in [**9**].

The main result we want to quote gives the generator and relation ranks of P in terms of the arithmetic of K. (For a first hint about why this is possible, notice that the Burnside Basis Theorem allows one to reduce the question about the generator rank to class field theory, since $\overline{P}$ is abelian.)

Let r_2 denote the number of complex primes of K. As usual, if v is a (finite or infinite) prime of K, we write K_v for the completion of K at v. For any field F, let $\delta(F) = 1$ if F contains any primitive p-th roots of unity, and $\delta(F) = 0$ otherwise. Let $H = G(K/\mathbb{Q}) \cong \mathrm{Im}(\overline{\rho})$.

Let Z_S be the set of nonzero elements $x \in K$ such that the fractional ideal (x) generated by x is the p-th power of some ideal and such that x is a p-th power in each completion K_v for $v \in S_1$. Of course, if x is already a p-th power in K, then

$x \in Z_S$. Notice that both $(K^\times)^p$ and Z_S are stable under the Galois action of H. Let B_S denote the $\mathbb{F}_p[H]$-module $Z_S/(K^\times)^p$.

We can now (finally) state the theorem about P.

Theorem 5.4. *Let $d(P)$ and $r(P)$ denote, as above, the generator and relation ranks of P. Then*

$$r(P) = \left(\sum_{v \in S_1} \delta(K_v)\right) - \delta(K) + \dim_{\mathbb{F}_p} B_S,$$

and

$$d(P) = r_2 + 1 + r(P).$$

In particular, P is topologically finitely generated.

For what follows, we will need to know something about $\overline{P}$ as an H-module. To set up the theorem, let $\overline{E}$ denote the units of K modulo p-th powers, and let $\overline{E}_v$ denote the units of K_v modulo p-th powers. If the class number of K is prime to p, then we can deduce from global class field theory an exact sequence of $\mathbb{F}_p[H]$-modules

$$0 \longrightarrow B_S \longrightarrow \overline{E} \longrightarrow \bigoplus_{v \in S_1} \overline{E}_v \longrightarrow \overline{P} \longrightarrow 0.$$

For each rational prime ℓ, let H_ℓ be a decomposition subgroup of H at ℓ, and let H_∞ be the subgroup of H generated by a complex conjugation. Let $\mu_p(K)$ be the group of p-th roots of unity in K, which we think of as a module over H and over the H_ℓ. Note that it's perfectly possible for $\mu_p(K)$ to be trivial.

Theorem 5.5 (Boston-Mazur)**.** *Under the hypotheses above, if H has order prime to p, then we have the following isomorphisms of $\mathbb{F}_p[H]$-modules:*

$$\bigoplus_{v \in S_1} \overline{E}_v \cong \mathbb{F}_p[H] \oplus \left(\bigoplus_{\ell \in S} \operatorname{Ind}_{H_\ell}^H \mu_p\right)$$
$$\overline{E} \oplus \mathbb{F}_p \cong \mu_p \oplus \operatorname{Ind}_{H_\infty}^H \mathbb{F}_p$$

See [**10**] for the proof. The point is this: using this theorem and the exact sequence above, we can determine the decomposition of $\overline{P}$ into irreducible subrepresentations of H. Later, we'll want to look for homomorphisms from P to $\Gamma_n(R)$, where R is a coefficient ring. We will often be able to reduce this to looking for H-homomorphisms from $\overline{P}$ to appropriate quotients of submodules of $\Gamma_n(R)$. By the results in the previous section, the latter only involve irreducible subrepresentations contained in the adjoint representation of H. So the subrepresentations of $\overline{P}$ which are *not* contained in the adjoint representation will have to have trivial image.

Tame representations

Keep notations as above, and assume now that $\overline{\rho}$ is tame, that is, that the order of $H = \operatorname{Im}(\overline{\rho})$ is not divisible by p. Recall that in this case P is a normal pro-p-Sylow subgroup of $\tilde{\Pi}$. By the Schur-Zassenhaus theorem, $\tilde{\Pi}$ is the semidirect product of

P and a subgroup $A \cong G/P = H$. Because $\Gamma_n(W(\mathsf{k}))$ is pro-p, we can use Schur-Zassenhaus again to find a subgroup H_1 of $\mathrm{GL}_n(W(\mathsf{k}))$ which is isomorphic to H and therefore we can find a lift

$$\rho_1 : \tilde{\Pi} \longrightarrow \mathrm{GL}_n(W(\mathsf{k}))$$

inducing an isomorphism from A to H_1. We get, then, an induced inclusion $\sigma : A \longrightarrow \mathrm{GL}_n(W(\mathsf{k}))$, which we will fix from now on.

Problem 5.6. Show that the lift ρ_1 is unique up to strict equivalence, and conclude that any two choices of the inclusion σ are conjugate by an element of $\Gamma_n(W(\mathsf{k}))$.

For any coefficient ring R we have a canonical homomorphism $W(\mathsf{k}) \longrightarrow R$ and hence a homomorphism $\sigma_R : A \longrightarrow \mathrm{GL}_n(R)$. We let A act on $\Gamma_n(R)$ by conjugation via this homomorphism.

Given all this setup, recall that any deformation of $\overline{\rho}$ induces a homomorphism from $P = \operatorname{Ker} \overline{\rho}$ to $\Gamma_n(R)$. We can make this into a precise correspondence by taking into account the A-actions on both sides.

Define a set-valued covariant functor $\mathbf{E}_{\overline{\rho}}$ on $\mathcal{C}$ by defining, for each coefficient ring R,

$$\mathbf{E}_{\overline{\rho}}(R) = \mathrm{Hom}_A(P, \Gamma_n(R)),$$

where Hom_A denotes the set of continuous homomorphisms from P to $\Gamma_n(R)$ which commute with the A action. We want to compare this functor to the deformation functor $\mathbf{D}_{\overline{\rho}}$.

Notice that since $\tilde{\Pi}$ is the semidirect product of P and A (and we have been careful to take the A-action into account), any element $\phi \in \mathbf{E}_{\overline{\rho}}(R)$, together with the inclusion σ_R, defines a deformation of $\overline{\rho}$ to R. Hence, there is a natural morphism of functors $\mathbf{E}_{\overline{\rho}} \longrightarrow \mathbf{D}_{\overline{\rho}}$.

Theorem 5.6 (Boston). *The functor $\mathbf{E}_{\overline{\rho}}$ is always representable. Furthermore,*

i. If $C(\overline{\rho}) = \mathsf{k}$, the natural morphism of functors $\mathbf{E}_{\overline{\rho}} \longrightarrow \mathbf{D}_{\overline{\rho}}$ is an isomorphism.

ii. Otherwise, the morphism is smooth and induces an isomorphism on tangent spaces.

Proof. Let's first prove that we have an isomorphism when $C(\overline{\rho}) = \mathsf{k}$. To lighten the notation, write $\mathbf{E} = \mathbf{E}_{\overline{\rho}}$ and $\mathbf{D} = \mathbf{D}_{\overline{\rho}}$. Given a coefficient ring R, we claim that the induced map $\mathbf{E}(R) \longrightarrow \mathbf{D}(R)$ is a bijection.

To see that it is surjective, suppose ρ is a deformation of $\overline{\rho}$ to R. Then ρ induces a lift $A \longrightarrow \mathrm{GL}_n(R)$. Since all such lifts are conjugate by elements of $\Gamma_n(R)$ (see the problem above), we can choose a homomorphism ψ in the strict equivalence class of ρ such that $\psi|_A = \sigma_R$. Then $\psi|_P$ is an element of $E(R)$ which maps to the strict equivalence class of ψ, that is, to ρ.

To see that the map is injective, suppose ϕ_1 and ϕ_2 produce strictly equivalent lifts ψ_1 and ψ_2 of $\overline{\rho}$. Since both ψ_1 and ψ_2 induce σ on A, the matrix realizing the strict equivalence must be an element of $\Gamma_n(R)$ acting trivially on A by conjugation, i.e., commuting with the image of A. However, under our assumption that $C(\overline{\rho}) = \mathsf{k}$, the only elements of $\Gamma_n(R)$ commuting with the image of A are the scalars. Hence ψ_1 and ψ_2 differ by conjugation by a scalar, i.e., don't differ at all. In particular, their restrictions to P are the same, and hence $\phi_1 = \phi_2$.

We leave the case when $C(\overline{\rho}) \neq \mathsf{k}$ to the reader, and proceed to prove that $\mathbf{E}$ is representable. Choose generators $x_1, x_2, \ldots, x_d$ of P. The image of x_r in $\Gamma_n(R)$

is a matrix

$$\begin{pmatrix} 1+m_{11}^{(r)} & m_{12}^{(r)} & \dots & m_{1n}^{(r)} \\ m_{21}^{(r)} & 1+m_{22}^{(r)} & \dots & m_{2n}^{(r)} \\ \dots & & & \\ m_{n1}^{(r)} & m_{n2}^{(r)} & \dots & 1+m_{nn}^{(r)} \end{pmatrix},$$

where the $m_{ij}^{(r)}$ are in the maximal ideal of R.

We will construct the ring representing $\mathbf{E}$ as a quotient of the power series ring $W(\mathsf{k})[[T_{11}^{(1)},\dots,T_{nn}^{(d)}]]$ in dn^2 variables. Let $\mathcal{F}$ be the free pro-p-group on $x_1, x_2, \dots, x_d$, so that we get an exact sequence of groups

$$1 \longrightarrow N \longrightarrow \mathcal{F} \longrightarrow P \longrightarrow 1.$$

A homomorphism from P to $\Gamma_n(R)$ is exactly the same as a homomorphism from $\mathcal{F}$ to $\Gamma_n(R)$ such that N is in the kernel. We begin by defining a homomorphism from the free pro-p-group $\mathcal{F}$ to $\Gamma_n(W(\mathsf{k})[[T_{ij}^{(r)}]])$ such that that the image of x_r is the matrix

$$\begin{pmatrix} 1+T_{11}^{(r)} & T_{12}^{(r)} & \dots & T_{1n}^{(r)} \\ T_{21}^{(r)} & 1+T_{22}^{(r)} & \dots & T_{2n}^{(r)} \\ \dots & & & \\ T_{n1}^{(r)} & T_{n2}^{(r)} & \dots & 1+T_{nn}^{(r)} \end{pmatrix}.$$

Requiring that N be in the kernel amounts to requiring that certain equations involving the $T_{ij}^{(r)}$ hold. Requiring that the A-actions commute with the homomorphism imposes further equations. Let I be the ideal of $W(\mathsf{k})[[T_{ij}^{(r)}]]$ generated by all these equations. If we let $\mathcal{R} = W(\mathsf{k})[[T_{ij}^{(r)}]]/I$, we have produced a homomorphism $\phi : P \longrightarrow \Gamma_n(\mathcal{R})$. It is clear that this is the universal such homomorphism, and hence that $\mathcal{R}$ represents the functor $\mathbf{E}$. □

Problem 5.7. Check the remaining assertions in the theorem.

It's worth noting that this proof sheds some light on the issue of what it means for the functor $\mathbf{D}$ is representable, at least in the case when $\overline{\rho}$ is tame. As usual, the question of representability turns out to be connected to whether there are "extra automorphisms." In our situation, the question turns out to be whether the lift $A \longrightarrow \mathrm{GL}_n(R)$ is unique.

This result has been generalized by Böckle; see [**5**], for example.

Since the two functors $\mathbf{E}$ and $\mathbf{D}$ are isomorphic, so are their tangent spaces. Hence

$$\dim_{\mathsf{k}} t_{\mathbf{D}} = \dim_{\mathsf{k}} \mathrm{Hom}_A(P, \Gamma_n(\mathsf{k}[\varepsilon])).$$

Note, now that $\Gamma_n(\mathsf{k}[\varepsilon])$ is isomorphic to $\mathrm{Ad}(\overline{\rho})$ as an A-module; in particular, it is a p-elementary abelian group, so every homomorphism from P to $\Gamma_n(\mathsf{k}[\varepsilon])$ must factor through $\overline{P}$. Hence

$$\dim_{\mathsf{k}} t_{\mathbf{D}} = \dim_{\mathsf{k}} \mathrm{Hom}_A(\overline{P}, \mathrm{Ad}(\overline{\rho})).$$

Now, since the order of A is not divisible by p, both $\overline{P}$ and $\mathrm{Ad}(\overline{\rho})$ can be decomposed as a sum or irreducible A-modules. The dimension on the right, then, can be computed in terms of the number of irreducible A-modules which appear in the decomposition of both $\overline{P}$ and $\mathrm{Ad}(\overline{\rho})$. We are back to the situation described above:

if we can determine which sub-representations occur in $\overline{P}$, we can compute the dimension of the tangent space.

Suppose $C(\overline{\rho}) = \mathsf{k}$ and let $\mathcal{R}$ be the universal deformation ring. Note that if we can compute the dimension d of the tangent space, we know that there is a surjective homomorphism of coefficient rings

$$W(\mathsf{k})[[T_1, \dots, T_d]] \longrightarrow \mathcal{R}$$

inducing the identity on tangent spaces. Thus, when $\overline{\rho}$ is tame we can read the dimension off from the structure of $\overline{P}$ as an $\mathbb{F}_p[H]$-module. This crucial idea, together with the fact that what matters are the sub-representations which occur in the adjoint representation, has been dubbed the "prime-to-adjoint" principle by Böckle.

Here's a sample theorem:

Theorem 5.7 (Boston). *Let p be an odd prime. Suppose that $\overline{\rho} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathbb{F}_p)$ is odd and absolutely irreducible. Let $H = \mathrm{Im}(\overline{\rho})$, and suppose that p does not divide the order of H, so that $\overline{\rho}$ is tame. Let K be the field fixed by the kernel of $\overline{\rho}$, and let S_1 be the set of primes of K which lie above the primes in S. Let*

$$V = \operatorname{coker}\left(\mu_p(K) \longrightarrow \bigoplus_{v \in S_1} \mu_p(K_v)\right),$$

and let $B = B_S$ defined as above. Both V and B are $\mathbb{F}_p[H]$-modules. Suppose that the class number of K is not divisible by p and that both V and B are relatively prime to $\mathrm{Ad}(\overline{\rho})$ as $\mathbb{F}_p[H]$-modules. Then

$$\mathcal{R}(\overline{\rho}) \cong \mathbb{Z}_p[[T_1, T_2, T_3]],$$

and one can give an explicit description of $\boldsymbol{\rho}$ on (well-chosen) generators for $\tilde{\Pi} = G(L/\mathbb{Q})$.

Proof. (Just a sketch.) From Theorem 5.5, one can see that $\overline{P}$ is generated by

i. an element $\overline{x}$ which is fixed under H,
ii. an element $\overline{y}$ such that $\overline{y}^c = (\overline{y})^{-1}$, where c is a chosen complex conjugation in G,
iii. other elements which generate a prime-to-adjoint $\mathbb{F}_p[H]$-module.

Using theorem 5.3, we see that P is generated by x (fixed under H), y such that $y^c = y^{-1}$, and other generators which are prime-to-adjoint. From this we see that the dimension of the tangent space is three and we can define a deformation to $\mathbb{Z}_p[[T_1, T_2, T_3]]$ by

$$x \mapsto \begin{pmatrix} 1+T_1 & 0 \\ 0 & 1+T_1 \end{pmatrix}$$
$$y \mapsto \begin{pmatrix} (1+T_2T_3)^{1/2} & T_2 \\ T_3 & (1+T_2T_3)^{1/2} \end{pmatrix}$$

and mapping all the other generators (which are prime-to-adjoint) to the identity. This gives the universal deformation. □

While this result covers a case where the deformation problem is unobstructed, Boston's work also includes various results for obstructed cases. See, for example, [**9**].

Flach, Boston, and Ullom have similar results for deformations of residual representations $\overline{\rho}$ which come from the p-division points of an elliptic curve. For example, the following is a result of Flach. Suppose E is an elliptic curve over $\mathbb{Q}$, $p \geq 5$, and assume that E has good reduction at p. Take

$$S = \{\text{primes of bad reduction}\} \cup \{p, \infty\},$$

let $E[p]$ be the points of order p on E, and let

$$\overline{\rho} = \overline{\rho}_{\mathrm{E}} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathbb{F}_p)$$

be the representation given by the action of $G_{\mathbb{Q},S}$ on $E[p]$.

Theorem 5.8 (Flach). *Suppose that*

i. $\overline{\rho} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathbb{F}_p)$ *is surjective,*
ii. for all $r \in S$, $\mathrm{H}^0(\mathbb{Q}_r, E[p] \otimes E[p]) = 0$,
iii. p *does not divide* $\Omega^{-1}L(M, 2)$*, where* $M = Sym^2(E)$ *and* Ω *is a transcendental period.*

Then $\mathcal{R}(\overline{\rho}) \cong \mathbb{Z}_p[[T_1, T_2, T_3]]$.

The point is to show that the deformation problem is unobstructed by exploiting the fact that $\mathrm{Sym}^2(E)$ is closely related to $\mathrm{Ad}(\overline{\rho})$. See Appendix 2 for an expository account of the details and [**50**] for the original publication.

Boston and Ullom have obtained results of this type for the case of an elliptic curve with complex multiplication in [**11**]. Their result also covers some cases where the deformation problem is obstructed.

The most far-reaching results based on this method have been obtained by Böckle. For example, in [**6**], he shows under mild extra hypotheses that if $\overline{\rho}$ is tame and absolutely irreducible then the rigid space $\mathrm{Spf}\,\mathcal{R}(\overline{\rho})^{\mathrm{rig}}$ has the expected Krull dimension. In [**5**] he proves a generalization of a result of Mazur (which we will mention in a later lecture) that implies, in many cases (and without a tameness assumption), that $\mathcal{R}(\overline{\rho})$ has the predicted dimension. Finally (for example, in [**7**]), he has also studied carefully the case in which $\overline{\rho}$ is Borel (therefore reducible) but still satisfies $C(\overline{\rho}) = \mathsf{k}$.

LECTURE 6
Deformations With Prescribed Properties

Suppose we have a residual representation $\overline{\rho}$ which satisfies $C(\overline{\rho}) = \mathsf{k}$. Then we have a universal deformation ring. As we pointed out before, we can think of this ring as defining a universal deformation *space* whose points correspond to actual deformations of $\overline{\rho}$. For example, in the previous lecture we saw cases where the universal deformation ring is $W(\mathsf{k})[[T_1, T_2, T_3]]$; the corresponding space is three-dimensional (over $W(\mathsf{k})$), and the T_i can be thought of as giving "coordinates" for our space: for each triple (m_1, m_2, m_3) in the maximal ideal of some $W(\mathsf{k})$-algebra, we get a representation by mapping T_i to m_i. We would like to understand what these "coordinates" mean in terms of the representations themselves. One strategy for doing this is to consider interesting subspaces (equivalently, quotient rings of the universal deformation ring). The natural way to do this is to consider subspaces of the universal deformation space that correspond to deformations that have certain interesting or desirable properties. This idea was first considered in Mazur's original paper [**97**], where he discusses ordinary deformations and also looks at several other possible conditions.

Even more important is the observation that in many circumstances we do not want to consider all deformations, but rather only those satisfying certain conditions. The best known example of this is when we try to prove modularity of certain deformations. In that situation, it is natural to require our deformations to have those properties that we know modular representations will have. We will discuss modular representations more carefully in the next lecture.

The basic idea for today, then, is this: suppose that the residual representation has a certain property. Then one can ask which deformations retain that property. For well-chosen properties, this allows us to define a representable sub-functor of the general deformation functor, and therefore to obtain a "universal deformation with the given property", which will correspond to a quotient of the universal deformation ring (or, from the geometric point of view, will define a subspace of the full deformation space.)

Deformation Conditions

Let's begin with a general account of "deformation conditions," that is, conditions that give rise to "good" subfunctors of the deformation functor. We follow the discussion in [**101**].

What do we expect of a "deformation condition?" Well, the first thing we need is that the association

$$R \rightsquigarrow \{\text{deformations of } \overline{\rho} \text{ satisfying our condition}\}$$

be a subfunctor of $\mathbf{D}_{\overline{\rho}}$, so that we have a deformation problem to work with. In addition, we would like this functor to be "relatively representable," that is, we would like our subfunctor to be sufficiently well behaved so that it is representable whenever $\mathbf{D}_{\overline{\rho}}$ is.

Recall that we can interpret a representation

$$\rho : \Pi \longrightarrow \mathrm{GL}_n(A)$$

by saying that it gives us a free A-module of rank n with a continuous A-linear action of the profinite group Π. In practice, all of the deformation conditions that have been useful have amounted to specifying properties that this Π-module should have. For technical reasons, it makes sense to specify these properties for representations where A is an artinian coefficient Λ-algebra.

Before we state the definition, let's introduce some terminology. Let A and A_1 be artinian coefficient Λ-algebras. If we are given a representation $\rho : \Pi \longrightarrow \mathrm{GL}_n(A)$ and a homomorphism of artinian coefficient Λ-algebras $\alpha : A \longrightarrow A_1$, then we get a representation $\rho_1 : \Pi \longrightarrow \mathrm{GL}_n(A_1)$ by composing ρ with the homomorphism $\mathrm{GL}_n(A) \longrightarrow \mathrm{GL}_n(A_1)$ induced by α. We will call ρ_1 the *push-forward of ρ by α*, and sometimes denote it by $\alpha_*\rho$. Of course, this works just as well for deformations of a residual representation $\overline{\rho}$; in fact, the push-forward operation is what makes $\mathbf{D}_{\overline{\rho}}$ a functor (we could have denoted the push-forward map by $\mathbf{D}_{\overline{\rho}}(\alpha)$ instead of α_*).

It's perhaps worth remarking that if we interpret a representation to $\mathrm{GL}_n(A)$ as giving a Π-module structure to the free A-module of rank n, then the push-forward operation is just the tensor product: if M is the free A module of rank n with a continuous linear Π-action, then the push-forward is $M \otimes_A A_1$, where the map $\alpha : A \longrightarrow A_1$ gives the A-module structure on A_1.

Now we are ready to define a (good) deformation condition. Informally, we want this to be a condition on deformations that is satisfied by the residual representation $\overline{\rho}$ (otherwise there's no point) and that is functorial (that is, preserved by push-forwards). Finally, we want the resulting functor to be relatively representable, so we require that our condition behave well with respect to fiber products and subrings.

Definition 6.1. Let $\overline{\rho}$ be a residual representation of dimension n. A *deformation condition* on deformations of $\overline{\rho}$ is a property $\mathcal{Q}$ of n-dimensional representations of Π defined over artinian coefficient Λ-algebras (equivalently, of A-modules which are free of rank n over A and have a continuous Π-action) which satisfies the following conditions.

i. The residual representation $\overline{\rho}$ has property $\mathcal{Q}$.

ii. Given a deformation $\rho : \Pi \longrightarrow \mathrm{GL}_n(A)$ of $\overline{\rho}$ and a homomorphism of coefficient Λ-algebras $\alpha : A \longrightarrow A_1$, if ρ has property $\mathcal{Q}$ then so does the push-forward $\alpha_*\rho$.

iii. Let

$$\begin{tikzcd}
& A \times_C B \arrow[dl, "p"'] \arrow[dr, "q"] & \\
A \arrow[dr, "\alpha"'] & & B \arrow[dl, "\beta"] \\
& C &
\end{tikzcd}$$

be a fiber product diagram in $\mathcal{C}_\Lambda^0$, and let

$$\rho : \Pi \longrightarrow \mathrm{GL}_n(A \times_C B)$$

be a deformation of $\overline{\rho}$. Then ρ has property $\mathcal{Q}$ if and only if both $p_*\rho$ and $q_*\rho$ have property $\mathcal{Q}$.

iv. Let $\alpha : A \longrightarrow A_1$ be an injective homomorphism of coefficient Λ-algebras and let $\rho : \Pi \longrightarrow \mathrm{GL}_n(A)$ be a deformation of $\overline{\rho}$. If $\alpha_*\rho$ has property $\mathcal{Q}$ then so does ρ.

It's probably worthwhile to comment a bit on the role of these four conditions. The first is clearly necessary for our subfunctor not to be trivial (and to make sure its value on k is a singleton). The second makes sure that we are going to get a functor. The third is clearly related to the Schlessinger criteria, and its role will be to make sure that the subfunctor is relatively representable. As Mark Dickinson explained to me, the fourth in fact follows from (ii) and (iii)—see the complements to this lecture for a proof.

Given a deformation condition, we can define a subfunctor of the deformation functor $\mathbf{D}_{\overline{\rho}}$:

Definition 6.2. Let $\mathcal{Q}$ be a deformation condition for $\overline{\rho}$. We define a functor

$$\mathbf{D}_{\mathcal{Q}} : \mathcal{C}_\Lambda^0 \rightsquigarrow \underline{\mathrm{Sets}}$$

by setting, for each artinian coefficient Λ-algebra A,

$$\mathbf{D}_{\mathcal{Q}}(A) = \{\text{deformations of } \overline{\rho} \text{ to } A \text{ which have property } \mathcal{Q}\}.$$

We can then extend $\mathbf{D}_{\mathcal{Q}}$ to all of $\mathcal{C}_\Lambda$ by continuity: if R is a coefficient Λ-algebra,

$$\mathbf{D}_{\mathcal{Q}}(R) = \varprojlim_k \mathbf{D}_{\mathcal{Q}}(R/\mathfrak{m}^k).$$

In other words, we are saying that a deformation of $\overline{\rho}$ to a coefficient Λ-algebra R has property $\mathcal{Q}$ if and only if its reductions modulo $\mathfrak{m}^k$ have property $\mathcal{Q}$ for all $k \geq 1$.

Problem 6.1. Check that $\mathbf{D}_{\mathcal{Q}}$ is a subfunctor of $\mathbf{D}_{\overline{\rho}}$. (The main thing to check is that it does the right thing when we have a homomorphism of artinian coefficient Λ-algebras.)

Theorem 6.1. *If $\mathcal{Q}$ is a deformation condition for $\overline{\rho}$, then $\mathbf{D}_{\mathcal{Q}}$ satisfies conditions **H1**, **H2**, and **H3** in Schlessinger's theorem. If $C(\overline{\rho}) = \mathsf{k}$, then $\mathbf{D}_{\mathcal{Q}}$ also satisfies property **H4**, and therefore is representable by a ring $\mathcal{R}_{\mathcal{Q}}$ which is a quotient of the universal deformation ring $\mathcal{R}(\overline{\rho})$.*

Proof. This is pretty much immediate from conditions (ii) and (iii) in the definition of a deformation problem. □

Problem 6.2. Check the details!

Suppose we have a deformation condition $\mathcal{Q}$. Then we can consider the tangent spaces of both $\mathbf{D}_{\mathcal{Q}}$ and $\mathbf{D}_{\overline{\rho}}$:

$$\mathbf{D}_{\mathcal{Q}}(\mathsf{k}[\varepsilon]) \subset \mathbf{D}_{\overline{\rho}}(\mathsf{k}[\varepsilon]).$$

Recall that we have a cohomological interpretation of the larger space:

$$\mathbf{D}_{\overline{\rho}}(\mathsf{k}[\varepsilon]) \cong \mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho})).$$

Definition 6.3. We define $\mathrm{H}^1_{\mathcal{Q}}(\Pi, \mathrm{Ad}(\overline{\rho}))$ to be the subspace of $\mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho}))$ corresponding, under this isomorphism, to the subspace $\mathbf{D}_{\mathcal{Q}}(\mathsf{k}[\varepsilon])$.

It's important to emphasize that this definition, as Mazur says, is really only a "promissory note." What it does is hint at the fact that for most of the interesting deformation conditions $\mathcal{Q}$ it will be possible to describe this cohomology group in an intrinsic way.

Problem 6.3. Use a similar dodge to define the Ext group $\mathrm{Ext}^1_{\mathsf{k}[[\Pi]],\mathcal{Q}}(V_{\overline{\rho}}, V_{\overline{\rho}})$. Can you describe this intrinsically in terms of extensions of Π-modules?

We go on to consider several different possible choices for the deformation condition $\mathcal{Q}$.

Deformations with fixed determinant

Probably the most natural restriction we can put on deformations is to fix their determinant. To see why this is something we might want to do, remember that representations attached to elliptic curves have determinant equal to the cyclotomic character, and that representations attached to modular forms have determinant equal to a character of finite order times a power (related to the weight of the modular form) of the cyclotomic character. Thus, for example, when we are trying to show that an elliptic curve is modular by studying its Galois representation it usually suffices to look only at the deformations that have determinant equal to the cyclotomic character.

In order for it to make sense to say that "all deformations have determinant δ," we need to take δ to be a character with values in $\Lambda^\times$, which we can then view as taking values in any coefficient Λ-algebra via the structural map.

Definition 6.4. Let δ be a continuous homomorphism

$$\delta : \Pi \longrightarrow \Lambda^\times,$$

and for every coefficient Λ-algebra R let δ_R be the composition

$$\delta_R : \Pi \xrightarrow{\delta} \Lambda^\times \longrightarrow R^\times.$$

We say a deformation ρ of $\overline{\rho}$ to R *has determinant* δ if $\det \rho = \delta_R$.

Notice that this abuses the language somewhat, since after all the determinant of a deformation "with determinant δ" is actually not δ! In practice, this does not lead to any trouble, so we won't worry too much about it.

Now suppose $\overline{\rho}$ itself has determinant δ. Let "$\det = \delta$" be shorthand for the property of having determinant δ.

Lemma 6.2. *Suppose $\overline{\rho}$ has determinant δ. Then "$\det = \delta$" is a deformation condition.*

Problem 6.4. Prove the lemma. (This is quite straightforward.)

Suppose, now, that $C(\overline{\rho}) = \mathsf{k}$. Then the lemma implies that there exists a quotient $\mathcal{R}_{\mathrm{det}=\delta}$ of the universal deformation ring $\mathcal{R}$ corresponding to those deformations whose determinant is δ. It's tangent space is not too hard to pin down: let $\mathrm{Ad}^0(\overline{\rho})$ denote the subspace of $\mathrm{Ad}(\overline{\rho})$ consisting of those matrices whose trace is zero. It's clear that $\mathrm{Ad}^0(\overline{\rho})$ is stable under the conjugation action of Π. Then we have

Lemma 6.3. *If $p \nmid n$, the tangent space of the functor $\mathbf{D}_{\mathrm{det}=\delta}$ is given cohomologically by*

$$\mathbf{D}_{\mathrm{det}=\delta}(\mathsf{k}[\varepsilon]) = \mathrm{H}^1(\Pi, \mathrm{Ad}^0(\overline{\rho})) \subset \mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho})).$$

If $p|n$, then $\mathrm{H}^1(\Pi, \mathrm{Ad}^0(\overline{\rho}))$ is in general no longer a subset of $\mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho}))$, but the inclusion $\mathrm{Ad}^0(\overline{\rho}) \hookrightarrow \mathrm{Ad}(\overline{\rho})$ still induces a map from one to the other; in this case

$$\mathbf{D}_{\mathrm{det}=\delta}(\mathsf{k}[\varepsilon]) = \mathrm{Im}\left(\mathrm{H}^1(\Pi, \mathrm{Ad}^0(\overline{\rho})) \longrightarrow \mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho}))\right).$$

Proof. The proof is the same in both cases. Essentially, we just repeat the argument, given in lecture 4, connecting lifts to $\mathsf{k}[\varepsilon]$ with elements in the cohomology, and then note that the diagram

$$\begin{tikzcd}
1+\varepsilon \mathrm{M}_n(\mathsf{k}) \arrow[r] \arrow[d, "\det"] & \mathrm{M}_n(\mathsf{k}) \arrow[d, "\mathrm{Tr}"] \\
1+\varepsilon \mathsf{k} \arrow[r] & \mathsf{k}
\end{tikzcd}$$

is commutative, where the horizontal maps are the isomorphism $1+\varepsilon b \mapsto b$. The requirement of fixed determinant forces the $1+\varepsilon \mathrm{M}_n(\mathsf{k})$ part of the lift to have determinant 1, which translates to trace 0 when we go to $\mathrm{M}_n(\mathsf{k})$. Thus, we get elements in $\mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho}))$ which are represented by cocycles taking values in $\mathrm{Ad}^0(\overline{\rho})$, and hence belong to the image of $\mathrm{H}^1(\Pi, \mathrm{Ad}^0(\overline{\rho}))$ in $\mathrm{H}^1(\Pi, \mathrm{Ad}(\overline{\rho}))$. When $p \nmid n$, this image is just $\mathrm{H}^1(\Pi, \mathrm{Ad}^0(\overline{\rho}))$ itself. □

Problem 6.5. Suppose $\overline{\rho}$ satisfies $C(\overline{\rho}) = \mathsf{k}$ and has determinant δ. Let $\mathcal{R}$ be the universal deformation ring of $\overline{\rho}$, let $\boldsymbol{\rho}$ be the universal deformation, and let $\mathcal{R}_{\mathrm{det}=\delta}$ be the universal "deformation with determinant δ" ring. Let $\delta_{\mathcal{R}}$ be

$$\delta_{\mathcal{R}} : \Pi \xrightarrow{\delta} \Lambda^\times \longrightarrow \mathcal{R}^\times,$$

the composition of δ with the structure homomorphism of $\mathcal{R}$ as a Λ-algebra. Show that $\mathcal{R}_{\mathrm{det}=\delta}$ is the quotient of $\mathcal{R}$ by the closed ideal generated by the elements

$$\delta_{\mathcal{R}}(g) - \det \boldsymbol{\rho}(g),$$

where g runs through a set of topological generators of Π.

Problem 6.6. (From [**6**].) Assume p does not divide n. Suppose $\overline{\rho}$ satisfies $C(\overline{\rho}) = \mathsf{k}$ and has determinant δ. Consider the three universal objects

i. $\mathcal{R}$ is the universal deformation ring and $\boldsymbol{\rho}$ the universal deformation,

ii. $\mathcal{R}_{\mathrm{det}=\delta}$ is the universal ring for deformations of determinant δ and $\boldsymbol{\rho}_\delta$ is the universal deformation of determinant δ,

iii. $\Lambda[[\Gamma]]$ is the universal deformation ring of the trivial character and $\boldsymbol{\epsilon}$ is the universal deformation.

Show that

$$\mathcal{R} = \mathcal{R}_{\det=\delta}\hat{\otimes}_{\Lambda}\Lambda[[\Gamma]]$$

and that

$$\boldsymbol{\rho} = \boldsymbol{\rho}_{\delta} \otimes \boldsymbol{\epsilon}.$$

The result in this last problem is used by Böckle in [**6**] to reduce the dimension conjecture for $\mathcal{R}$ to a question about the dimension of $\mathcal{R}_{\det=\delta}$. In general, it is not too hard to go back and forth between deformations in general (perhaps even subject to other deformation conditions) and deformations (of the same type) with fixed determinant.

Categorical deformation conditions

One interesting class of deformation conditions was introduced by Ramakrishna in [**115**]. The idea is to require that our deformations (at least at the artinian level) define Π-modules that belong to a particularly nice subcategory of the category of Λ-modules of finite length with a continuous Λ-linear action of Π.

To set this up, let $\mathcal{P}$ be a full subcategory of the category of Λ-modules of finite length with a continuous Λ-linear action of Π, and assume $\mathcal{P}$ is closed under passage to sub-objects, quotients, and finite direct sums. Given $\overline{\rho}$, we say that a deformation $\rho : \Pi \longrightarrow \mathrm{GL}_n(R)$ is of type $\mathcal{P}$ if the Π-modules defined by all of its quotients $\rho_k : \Pi \longrightarrow \mathrm{GL}_n(R/\mathfrak{m}^k)$, viewed as Λ-modules with a continuous linear action of Π, are in the category $\mathcal{P}$.

Theorem 6.4 (Ramakrishna). *Suppose $\overline{\rho}$ is of type $\mathcal{P}$. The condition of "being of type $\mathcal{P}$" is a deformation condition.*

Proof. We need to prove that the property of "being of type $\mathcal{P}$" is preserved by push-forwards and that it works well with fiber products. Suppose first that A and A_1 are artinian coefficient rings, that we have a coefficient ring homomorphism $\alpha : A \longrightarrow A_1$, and that ρ is a deformation of $\overline{\rho}$ to A which has property $\mathcal{P}$. Let $M = A^n$ and let $M_1 = A_1^n$, both of which are endowed with continuous linear actions of Π via ρ (and its push-forward to A_1). We think of both M and M_1 as Λ-modules of finite length, and we know that M is in the subcategory $\mathcal{P}$.

Since both A and A_1 are of finite length, there exists an artinian coefficient ring B such that

- $\alpha : A \longrightarrow A_1$ factors through B, that is, there exist coefficient ring homomorphisms $\alpha_1 : A \longrightarrow B$ and $\alpha_2 : B \longrightarrow A_1$ such that $\alpha = \alpha_2 \circ \alpha_1$,
- B is free of finite rank as an A-module, and
- $\alpha_2 : B \longrightarrow A_1$ is surjective.

(For example, we can take B to be a quotient of a power series ring over A by a power of its maximal ideal.) Now the pushforward of M via α_1 is simply a direct sum M^r of copies of M, and therefore is an object of $\mathcal{P}$, and M_1, which is the pushforward of M^r under α_2, is a quotient of M^r. Since $\mathcal{P}$ is closed under finite direct sums and under quotients, it follows that M_1 is an object of $\mathcal{P}$. This proves condition (ii) in the definition of a deformation condition.

Property (iii) is easier. Suppose we have homomorphisms of artinian coefficient rings $A \longrightarrow C$ and $B \longrightarrow C$. Let $R = A \times_C B$, and suppose we have a deformation ρ of $\overline{\rho}$ to R. We already know that if ρ has property $\mathcal{P}$ then so do the push-forwards to A and B. For the converse, let ρ_A and ρ_B be the push-forwards, and suppose both have property $\mathcal{P}$. Notice that R is a subring of $A \oplus B$, and therefore R^n is a submodule of $A^n \oplus B^n$. Since we know both A^n and B^n are in $\mathcal{P}$ and $\mathcal{P}$ is closed under direct sums and sub-objects, it follows that R^n is in $\mathcal{P}$, and we are done. □

The most important example of a property of this type is the one Ramakrishna considered in his original paper and which was later used by Wiles. Suppose $\Pi = G_{\mathbb{Q}_\ell}$ is the absolute Galois group of $\mathbb{Q}_\ell$, and we let $\mathcal{P} = \mathcal{P}_{fl}$ be the category of all $G_{\mathbb{Q}_\ell}$ representations ρ (over artinian rings A) such that the deformation space of ρ is isomorphic to the $G_{\mathbb{Q}_\ell}$-module obtained from the generic fiber of a finite flat group scheme over $\mathrm{Spec}(\mathbb{Z}_\ell)$. The category $\mathcal{P}_{fl}$ is closed under passage to sub-objects, quotients, and direct sums, and therefore defines a deformation condition. We call deformations satisfying this condition "flat deformations."

One can also define the analogous condition for representations of $G_{\mathbb{Q},S}$ if $\ell \in S$. In this case we have a homomorphism $G_{\mathbb{Q}_\ell} \longrightarrow G_{\mathbb{Q},S}$; we say a deformation $\rho : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_n(A)$ is "flat at ℓ" if its composition with this homomorphism defines a representation of $G_{\mathbb{Q}_\ell}$ which is of type $\mathcal{P}_{fl}$.

Problem 6.7. Work out a usable description of the tangent space of the functor $\mathbf{D}_{fl}$ of flat deformations of representations of $G_{\mathbb{Q}_\ell}$. (Hint: what we would like to have is an identification of the tangent space with an Ext^1 group in the category of finite flat group schemes. To make this work, you will need to use Raynaud's result in **[116]** on uniqueness of models; this requires $e < p-1$, and hence you'll want to assume $p > 2$.)

Ordinary deformations

We now restrict ourselves to the case $n = 2$ to talk about the condition of being "ordinary." This actually appears with two different meanings in the literature. We follow Mazur's definition (which I think is a minority opinion); to other authors what we are defining here are "co-ordinary" deformations, i.e., deformations whose dual is ordinary. In many situations, this makes no difference (e.g., because the universal deformation rings of $\overline{\rho}$ and of its contragredient are canonically isomorphic).

Ordinary deformations were first considered by Mazur in connection with Hida's theory of ordinary p-adic modular forms. Let $\mathcal{O}$ be a discrete valuation ring which is finite over $\mathbb{Z}_p$, and suppose we have a Hecke eigenform f of weight k and level N and defined over $\mathcal{O}$ such that $\mathrm{U}_p(f) = \lambda f$ with $\lambda \in \mathcal{O}^\times$ a p-adic unit. Then the attached Galois representation

$$\rho_f : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathcal{O})$$

has the following property. If we set

$$M = \mathcal{O} \times \mathcal{O}$$

endowed with the $G_{\mathbb{Q},S}$-action defined by ρ_f, and if I_p is an inertia subgroup at p, then the submodule M^{I_p} of vectors fixed under I_p is $\mathcal{O}$-free of rank one, and is a direct summand.[1]

[1] Caveat: for this to be true, one must define the representation attached to a modular form f in terms of the geometric Frobenius transformation, which induces $x \mapsto x^{p^{-1}}$ on the residue field. As a result, our "representation attached to a modular form f" is the contragredient of the usual one constructed in étale cohomology. With the arithmetic Frobenius, one has an invariant quotient instead of an invariant subspace.

One can show, in fact, that all of the deformations which Hida's theory gives us have this property. Thus, it makes sense to try to work out which piece of the full deformation space is cut out by this condition.

Definition 6.5. Fix Π and k as above, let R be a ring in $\mathcal{C}$, and choose a closed subgroup $I \subset \Pi$. Let

$$\rho : \Pi \longrightarrow \mathrm{GL}_2(R)$$

be a representation, and let $M = R \times R$ with the Π-module structure determined by ρ. We say ρ is I-ordinary if the sub-R-module $M^I \subset M$ is free of rank 1 over R and a direct summand of M.

Notice that by this definition a representation satisfying $M^I = M$ (we might call it I-unramified) does *not* qualify as being I-ordinary.

Problem 6.8. Suppose $\overline{\rho}$ is I-ordinary and that ρ is a representation lifting $\overline{\rho}$. Check that if ρ is strictly equivalent to σ and ρ is ordinary, then σ is ordinary. In other words, the property of being ordinary is invariant under strict equivalence.

Theorem 6.5. *Suppose $\overline{\rho}$ is I-ordinary. Then the condition of being I-ordinary is a deformation condition for $\overline{\rho}$.*

Proof. This is pretty much straightforward. The first condition we need to check is that if a deformation $\rho : \Pi \longrightarrow \mathrm{GL}_2(A)$ is I-ordinary and $\pi : A \longrightarrow A'$ is a homomorphism of coefficient Λ-algebras, then the deformation $\pi_*\rho : \Pi \longrightarrow \mathrm{GL}_2(A')$ obtained from π is also ordinary. Changing basis if necessary, we can assume that the image of any $x \in I$ under ρ is of the form

$$\rho(x) = \begin{pmatrix} 1 & * \\ 0 & * \end{pmatrix}.$$

But then it's clear that the image of this matrix under π is a matrix of the same form, which shows that $\pi_*\rho$ is I-ordinary.

The second condition asks us to look at a fiber product situation, So suppose we have rings R_0, R_1, and R_2 in $\mathcal{C}_0$, and morphisms

$$\phi_1 : R_1 \longrightarrow R_0 \qquad \text{and} \qquad \phi_2 : R_2 \longrightarrow R_0.$$

Let

$$R_3 = R_1 \times_{R_0} R_2.$$

We need to show that if we have a deformation to R_3 such that the induced deformations to R_1 and R_2 are both I-ordinary, then so is R_3. The basic strategy is similar to what we have done before: we must find homomorphisms in the strict equivalence classes of the deformations to R_1 and R_2 such that the rank-one subspaces fixed by I map to the same subspace of $R_0 \times R_0$, and so on. We leave the details to the reader. □

Problem 6.9. Finish the proof of the theorem.

Given an I-ordinary residual representation

$$\overline{\rho} : \Pi \longrightarrow \mathrm{GL}_2(\mathsf{k}),$$

it now makes sense to consider the subfunctor $\mathbf{D}_I$ of $\mathbf{D}$ such that, for any ring R in $\mathcal{C}_\Lambda$,

$$\mathbf{D}_I(R) = \{I\text{-ordinary deformations of } \overline{\rho} \text{ to } R\}.$$

Since being I-ordinary is a deformation condition, this functor is representable whenever $\mathbf{D}$ is.

Corollary 6.6. *Suppose $\overline{\rho}$ is I-ordinary and that $C(\overline{\rho}) = \mathsf{k}$. Then there exists a universal I-ordinary deformation of $\overline{\rho}$. Specifically, there exists a ring $\mathcal{R}_I = \mathcal{R}_I(\Pi, \mathsf{k}, \overline{\rho}, I)$ and an I-ordinary deformation*

$$\boldsymbol{\rho}_I : \Pi \longrightarrow \mathrm{GL}_2(\mathcal{R}_I)$$

such that any I-ordinary deformation of $\overline{\rho}$ to a ring A is $\mathcal{C}_\Lambda$ is obtained from $\boldsymbol{\rho}_I$ via a unique homomorphism $\mathcal{R}_I \longrightarrow A$.

It is easy to see that one can extend this result to show that if we consider a set of closed subgroups $I_1, I_2, \ldots, I_n$ and the residual representation $\overline{\rho}$ is ordinary for each of these subgroups, then there exists a universal deformation of that type.

Problem 6.10. A representation $\rho : \Pi \longrightarrow \mathrm{GL}_2(A)$ is called I-co-ordinary if its representation space M has a submodule M_1 which is free of rank 1 as an A-module, and is a direct summand of M, and such that M/M_1 is invariant under I. Show that being I-co-ordinary is a deformation condition.

As discussed above, the tangent space to the I-ordinary deformation subfunctor corresponds to a subspace

$$\mathrm{H}^1_I(\Pi, \mathrm{Ad}(\overline{\rho})) \subset \mathrm{H}^1(\Pi, Ad).$$

Problem 6.11. Let $V_{\overline{\rho}}$ be the representation space for $\overline{\rho}$, and let V^I be the subspace fixed by inertia. Assume $\overline{\rho}$ is ordinary, so that V^I is one-dimensional. Let $\mathrm{Ad}_I(\overline{\rho})$ denote the subspace of $\mathrm{Ad}(\overline{\rho})$ consisting of those matrices which correspond to endomorphisms of V which factor through V/V^I. Show that

$$\mathrm{H}^1_I(\Pi, \mathrm{Ad}(\overline{\rho})) = \mathrm{H}^1(\Pi, \mathrm{Ad}_I(\overline{\rho})) \subset \mathrm{H}^1(\Pi, Ad).$$

Deformation conditions for global Galois representations

Now let's go back to the situation of most interest for us, when $\Pi = G_{\mathbb{Q},S}$ is the Galois group of the maximal extension of $\mathbb{Q}$ unramified outside a finite set of primes S. As always, we will assume that S contains both p and the prime at infinity. The first thing to consider is why one would want to impose deformation conditions to begin with, and which conditions they might be.

Suppose, for example, that our residual representation $\overline{\rho}$ is the Galois representation arising from the p-division points of an elliptic curve E, and let's assume $\overline{\rho}$ is absolutely irreducible. Then we can take S to consist of the primes of bad reduction of E, plus p and ∞, and we have at hand at least one lift to characteristic zero, the representation

$$\rho_{E,p} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathbb{Z}_p)$$

arising from the Galois action on the Tate module of E. In this situation, the thing to note is that we know quite a bit about these representations. For example, we know that the determinant of ρ is the p-adic cyclotomic character, and we know that (the prime-to-p part of) the Artin conductor of $\rho_{E,p}$ is equal to (the prime-to-p part of) the conductor of E. So, for each of the primes in S, we have some information on how the representation behaves when restricted to the decomposition group at that prime. If we are trying to make our deformation problem "as tight as possible" (for example, so that $\rho_{E,p}$ is a deformation captured by our problem but at the same time so that it is possible that every representation captured by our problem is modular), we need to impose deformation conditions that reflect the properties of these deformations. These conditions are of two types. First, there is the condition on the determinant. As the discussion above suggests, this is not too serious an issue. Second, there are local conditions at primes $\ell \in S$. These are deformation conditions that are posed in terms of the restriction of ρ to the decomposition groups $G_{\mathbb{Q}_\ell}$ (as we did above for the example of the "finite flat" condition).

Let $\overline{\rho} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_n(\mathsf{k})$ be a residual representation. Formally, a *global Galois deformation problem* $\mathcal{Q}$ is the problem of representing a subfunctor of $\mathbf{D}_{\overline{\rho}}$ defined by giving, for each non-archimedean prime $\ell \in S$, a deformation condition $\mathcal{Q}_\ell$ for the local residual representation $\overline{\rho}|_{G_{\mathbb{Q}_\ell}}$. A *global Galois deformation problem with fixed determinant* is the same, with an added "det $= \delta$" condition.

The following is easy to check:

Lemma 6.7. *A global Galois deformation problem is a deformation condition for representations of $G_{\mathbb{Q},S}$.*

Proof. Easy, given that we know that each of the local "pieces" is a deformation condition. □

As always, we want to understand the tangent space to the subfunctor $\mathbf{D}_{\mathcal{Q}}$ associated to a global Galois deformation problem. The main thing we need to be careful about is that the local conditions $\mathcal{Q}_\ell$ are defined only in terms of the restriction of the representations to $G_{\mathbb{Q}_\ell}$. As before, let $\mathrm{H}^1_{\mathcal{Q}}(G_{\mathbb{Q},S}, \mathrm{Ad}(\overline{\rho}))$ be the subspace of $\mathrm{H}^1(G_{\mathbb{Q},S}, \mathrm{Ad}(\overline{\rho}))$ corresponding to the tangent space of $\mathbf{D}_{\mathcal{Q}}$, and let $\mathrm{H}^1_{\mathcal{Q}_\ell}(G_{\mathbb{Q}_\ell}, \mathrm{Ad}(\overline{\rho}))$ be the subspace of $\mathrm{H}^1(G_{\mathbb{Q}_\ell}, \mathrm{Ad}(\overline{\rho}))$ corresponding to the tangent space of the subfunctor of the local deformation functor defined by the local condition $\mathcal{Q}_\ell$.

Theorem 6.8. *The diagram*

$$\begin{tikzcd}
\mathrm{H}^1_{\mathcal{Q}}(G_{\mathbb{Q},S}, \mathrm{Ad}(\overline{\rho})) \arrow[r] \arrow[d] & \mathrm{H}^1(G_{\mathbb{Q},S}, \mathrm{Ad}(\overline{\rho})) \arrow[d] \\
\bigoplus_{\ell \in S} \mathrm{H}^1_{\mathcal{Q}_\ell}(G_{\mathbb{Q}_\ell}, \mathrm{Ad}(\overline{\rho})) \arrow[r] & \bigoplus_{\ell \in S} \mathrm{H}^1(G_{\mathbb{Q}_\ell}, \mathrm{Ad}(\overline{\rho}))
\end{tikzcd}$$

where the horizontal arrows are inclusions and the vertical arrow on the right is the restriction map on cohomology, is Cartesian, that is, it identifies $\mathrm{H}^1_{\mathcal{Q}}(G_{\mathbb{Q},S}, \mathrm{Ad}(\overline{\rho}))$ with the set of elements of $\mathrm{H}^1_{\mathcal{Q}}(G_{\mathbb{Q},S}, \mathrm{Ad}(\overline{\rho}))$ which, for each $\ell \in S$, map under restriction to the image of $\mathrm{H}^1_{\mathcal{Q}_\ell}(G_{\mathbb{Q}_\ell}, \mathrm{Ad}(\overline{\rho}))$.

Proof. Straightforward from the definitions. □

This shows that the tangent space to the functor $\mathbf{D}_{\mathcal{Q}}$ is a kind of "Selmer group," i.e., a part of the global cohomology group defined by local conditions for each $\ell \in S$.

Representations that are ordinary at p

An interesting application of the ideas in this lecture is when $\overline{\rho}$ is a two-dimensional global Galois representation, so that $\Pi = G_{\mathbb{Q},S}$ for a finite set of primes S, and we choose only one local deformation condition, requiring that our deformations be I-ordinary, when $I = I_p$ is an inertia subgroup at p. We say that such representations are *ordinary at* p.

So suppose $\overline{\rho}$ is ordinary at p, and consider the functor $\mathbf{D}^0$ which is defined by

$$\mathbf{D}^0(R) = \{\text{deformations of } \overline{\rho} \text{ to } R \text{ which are ordinary at } p\}.$$

As before, this is representable, and we call the representing ring $\mathcal{R}^0(\overline{\rho})$ the *universal ordinary deformation ring.* This is a quotient of the full universal deformation ring and parametrizes deformations of $\overline{\rho}$ which are ordinary at p.

There are two reasons to give special attention to the universal ordinary deformation ring. The first is that, in many cases, it follows from Wiles' work that if $\overline{\rho}$ is modular, then any ordinary deformation of $\overline{\rho}$ is also modular (perhaps in an extended sense). Another way of saying this is that we can prove, in many cases, that the ring $\mathcal{R}^0(\overline{\rho})$ can be identified with a certain p-adic Hecke algebra (a localization of the Hida algebra, to be discussed further in the next lecture).

The other reason the universal ordinary deformation ring is interesting is that, in certain cases, the homomorphism $\mathcal{R}(\overline{\rho}) \longrightarrow \mathcal{R}^0(\overline{\rho})$ is well understood. For example, one can prove the following result.

Proposition 6.9 (Mazur, Martin). *Let* $S = \{p, \infty\}$, $\mathsf{k} = \mathbb{F}_p$, *and let*

$$\overline{\rho} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathbb{F}_p)$$

be ordinary (at p*). Let* ω *denote the Teichmüller character* $G_{\mathbb{Q},S} \longrightarrow \mathbb{F}_p^\times$. *Suppose either that*

i. $\det\overline{\rho} \neq 1, \omega, \omega^{-1}, \omega^{\frac{p-1}{2}}$, *or that*
ii. $\overline{\rho}$ *is tamely ramified.*

Then the kernel of the canonical homomorphism $\mathcal{R}(\overline{\rho}) \longrightarrow \mathcal{R}^0(\overline{\rho})$ *is generated by two elements.*

See [**98**] and [**93**]. Böckle has now improved considerably on this; see [**5**]. The reason such results are interesting is that they give us a way from lifting information about $\mathcal{R}^0$ to information about $\mathcal{R}$.

Corollary 6.10. *Under the assumptions of the Proposition, we have*

$$\text{Krull dim}\left(\mathcal{R}^0(\overline{\rho})/p\mathcal{R}^0(\overline{\rho})\right) \geq 1.$$

If in addition we know that $\dim t_{\mathcal{R}^0(\overline{\rho})} \leq 1$, *then we have* $\mathcal{R}^0(\overline{\rho}) \cong \mathbb{Z}_p[[T]]$ *and* $\mathcal{R}(\overline{\rho})$ *is a power series ring in two variables over* $\mathcal{R}^0(\overline{\rho})$.

One might describe this result as saying that the ordinary deformation ring "controls" the full deformation ring. In practical terms, this means that studying ordinary deformations (when they exist, that is, when $\overline{\rho}$ is ordinary) can lead to results about all deformations.

Complements to lecture six

As we pointed out above, condition (iv) in the definition of a "deformation condition" follows from conditions (i) to (iii). The argument we give here is due to Mark Dickinson.

Suppose, then, that we have a deformation condition $\mathcal{Q}$ satisfying conditions (i) to (iii) in the definition. Then (by property (ii)) we have a subfunctor $\mathbf{D}_{\mathcal{Q}}$ of the deformation functor $\mathbf{D}$, and (by property (i)) we know that $\mathbf{D}_{\mathcal{Q}}(\mathsf{k}) = \mathbf{D}(\mathsf{k})$ are both sets with exactly one element. The result we want to prove is essentially the content of the following proposition:

Proposition 6.11 (Dickinson)**.** *Let* $\mathbf{F}$ *be a set-valued functor on* $\mathcal{C}^0$*, and let* $\mathbf{G}$ *be a subfunctor of* $\mathbf{F}$*. Suppose that the following condition holds:*

> ***Property (*):*** *If* $A \longrightarrow C$ *and* $B \longrightarrow C$ *are homomorphisms of artinian coefficient rings, and the push-forwards to* $\mathbf{F}(A)$ *and* $\mathbf{F}(B)$ *of an element* x *in* $\mathbf{F}(A \times_C B)$ *are in* $\mathbf{G}(A)$ *and* $\mathbf{G}(B)$*, respectively, then* x *is in* $\mathbf{G}(A \times_C B)$*.*

Suppose $j : A \longrightarrow B$ *is an inclusion of artinian coefficient rings and* x *is an element of* $\mathbf{F}(A)$ *whose push-forward by* j *lies in* $\mathbf{G}(B)$*. Then* x *is in* $\mathbf{G}(A)$*.*

Proof. The main ingredient in the proof is the notion of an *equalizer* in the category of artinian coefficient rings, Suppose we have two homomorphisms of artinian coefficient rings $f, g : B \longrightarrow R$; we say a homomorphism $j : A \longrightarrow B$ is the equalizer of f and g if two conditions are satisfied:

i. The composite maps are equal: $f \circ j = g \circ j$.

ii. The map j is "universal" in the sense that given any other homomorphism $h : A' \longrightarrow B$ such that $f \circ h = g \circ h$ there exists a homomorphism $h' : A' \longrightarrow A$ such that $h = j \circ h'$.

In other words, any other homomorphism which "equalizes" f and g factors through the equalizer map j. It is easy to see that equalizers are always injective.

The strategy of the proof is, first, to show that the result is true when $j : A \longrightarrow B$ is the equalizer of some pair of homomorphisms, and second, to show that we can deduce the general case from this one.

Step One: *If* $j : A \longrightarrow B$ *is an equalizer in the category of artinian coefficient rings, then the result holds.* Suppose there are homomorphisms $f, g : B \longrightarrow R$ such that j is the equalizer of f and g. Then we have a Cartesian diagram[2]

$$\begin{tikzcd}
A \arrow[r, "j"] \arrow[d, "f\circ j = g\circ j"'] & B \arrow[d, "f \times_{\mathsf{k}} g"] \\
R \arrow[r, "\mathrm{diag}"] & R \times_{\mathsf{k}} R
\end{tikzcd}$$

Now $j_* x \in \mathbf{G}(B)$ and, by functoriality, $(f \circ j)_* x \in \mathbf{G}(R)$, and Property (*) tells us that $x \in \mathbf{G}(A)$.

Step Two: *If the inclusion* j *is not a surjection, then there exists a proper subring* C *of* B *which contains the image of* j *and for which the inclusion* $C \longrightarrow B$ *is an equalizer.* If j is not surjective, then neither is the induced map on cotangent spaces, and so the dual map

$$\mathrm{Hom}(B, \mathsf{k}[\varepsilon]) \longrightarrow \mathrm{Hom}(A, \mathsf{k}[\varepsilon])$$

is not injective. Thus, there exist distinct maps $f, g : B \longrightarrow \mathsf{k}[\varepsilon]$ which agree on A. We take C to be the equalizer of these two maps.

Wrap-up: The result now follows by induction on the length over $W(\mathsf{k})$ of B/A. If j is surjective (the base case), there is nothing to prove. If not, we use step two to factor $j : A \longrightarrow B$ as $A \longrightarrow C \longrightarrow B$ where the second map is an equalizer. By step one, the image of x in $\mathbf{F}(C)$ lies in $\mathbf{G}(C)$. Since the length of

[2] i.e., a fiber product diagram, see page 40.

C/A is smaller than the length of B/A, x is in $\mathbf{G}(A)$ by the induction hypothesis, and we are done. □

LECTURE 7
Modular Deformations

The idea that "naturally occurring" Galois representations should "come from" modular forms has become an important number-theoretical principle, and the success of Wiles, Taylor, Diamond, Conrad, and Breuil in proving that every elliptic curve over $\mathbb{Q}$ is modular strengthens this expectation. Our theory has produced for us a plethora of representations; is there any sense in which they all come from modular forms? Alternatively, can we make sense of the idea that "most" of them do?

First of all, in a deformation theory setting the question makes the most sense when we start from an absolutely irreducible odd two-dimensional Galois representation $\overline{\rho} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathsf{k})$ *which is attached to a modular form.* The Serre Conjecture claims that all such residual representations are indeed attached to a modular form, and we know that if this is true then (with some caveats if $p = 2$ or $p = 3$) we can require this modular form to have "optimal weight and level" in the sense of [**117**]. If we know that the residual representation is modular, then we can ask whether the deformations are modular too.

The first difficulty one faces is making precise what is meant by "modular form." The most obvious thing is to consider the classical modular forms (we'll recall the theory below). It's clear, however, that unless we add some very stringent deformation conditions there is no chance that *all* deformations will be modular, though it makes sense to ask whether the modular deformations are "dense" (in various senses of the word) in the space of all deformations.

The other option is to generalize the notion of "modular." The correct theory here would be some sort of p-adic theory of modular forms, which should be the "p-adic completion" of the usual theory. Such a theory is in fact available, and produces enough "modular deformations" so that there is some chance that these are in fact all the deformations. We will only mention some of the bare facts of this theory, and then explain (in the next lecture) how in at least one case one can prove that "all deformations are (in this extended sense) modular." This still leaves us, of course, with the problem of "locating" the deformations which are modular in the "classical" sense among the others.

An introduction to modular forms can be found in [**13**]; for more details, see the references therein. For the p-adic theory, see the expository account by Matthew Emerton in Appendix 3, which includes a survey of the literature.

Classical modular forms and their representations

We begin from the "classical" theory of modular forms. Let N be an integer relatively prime to p. Given non-negative integers k and ν, we will write $S_k(\Gamma_1(Np^\nu),\mathbb{Z}_p)$ for the space of cuspidal modular forms of weight k on $\Gamma_1(Np^\nu)$ defined over $\mathbb{Z}_p$ (i.e., whose q-expansions have coefficients in $\mathbb{Z}_p$). What this means is the following. The classical space $S_k(\Gamma_1(Np^\nu),\mathbb{C})$ of complex modular forms of weight k on $\Gamma_1(Np^\nu)$ has a basis consisting of modular forms whose q-expansions at infinity are such that all the coefficients are integers. By a "modular form defined over $\mathbb{Z}_p$" we just mean a $\mathbb{Z}_p$-linear combination of this integral basis. Equivalently, if we write $S_k(\Gamma_1(Np^\nu),\mathbb{Z})$ for the subspace of $S_k(\Gamma_1(Np^\nu),\mathbb{C})$ consisting of modular forms whose q-expansion coefficients at infinity are integral, then $S_k(\Gamma_1(Np^\nu),\mathbb{Z})$ is a finite $\mathbb{Z}$-module and it is known that

$$S_k(\Gamma_1(Np^\nu),\mathbb{C}) = S_k(\Gamma_1(Np^\nu),\mathbb{Z})\otimes_{\mathbb{Z}}\mathbb{C},$$

and our definition amounts to defining the $\mathbb{Z}_p$-space analogously:

$$S_k(\Gamma_1(Np^\nu),\mathbb{Z}_p) = S_k(\Gamma_1(Np^\nu),\mathbb{Z})\otimes_{\mathbb{Z}}\mathbb{Z}_p.$$

The choice of $\mathbb{Z}_p$ as the ring over which our forms are defined is made for simplicity; any discrete valuation ring $\mathcal{O}$ finite over $\mathbb{Z}_p$ would work just as well, and the definition would be the same.

This definition has a somewhat arbitrary "feel" to it, of course, not least because it privileges the cusp at infinity over the other cusps (the only reason to do this is not to have to bother with adjoining roots of unity to our base ring). In fact, one can do much better by using a geometric definition. Suppose $N \geq 5$ (and, as we agreed above, not divisible by p), and let $\mathbb{Z}_{(p)}$ be the localization of $\mathbb{Z}$ at p (so that $\mathbb{Z}_{(p)} = \mathbb{Q}\cap\mathbb{Z}_p$). Then there is an algebraic curve $X_1(Np^\nu)$ defined over $\mathbb{Z}_{(p)}$ which parametrizes (generalized) elliptic curves with an appropriately-defined Np^ν-level structure (basically, the choice of a point of order Np^ν, but this can't be taken literally when $\nu > 0$, so one must find a more sophisticated description that will work in our situation). There is a canonical invertible sheaf ω on $X_1(Np^\nu)$ constructed in terms of the universal (generalized) elliptic curve over $X_1(Np^\nu)$, and we can define $S_k(\Gamma_1(Np^\nu),\mathbb{Z}_{(p)})$ as the global sections of ω^k. If we extend scalars from $\mathbb{Z}_{(p)}$ to $\mathbb{C}$ this space of global sections is exactly the classical space $S_k(\Gamma_1(Np^\nu),\mathbb{C})$, and thus it makes sense to define $S_k(\Gamma_1(Np^\nu),\mathbb{Z}_p)$ as the global sections of ω^k over $X_1(Np^\nu)_{/\mathbb{Z}_p}$. This gives *almost* the same space as above; in fact, it is contained in the space above with finite index equal to a power of p. The difference is that the geometric definition automatically forces integrality (suitably defined) at *all* the cusps rather than only at the cusp at infinity. The case of $N < 5$ can then be dealt with (at least for $p \neq 2, 3$) by taking fixed points under the appropriate group. For an introduction to this view of things, start with Appendix 3, then see [**79**] for the case $\nu = 0$; for the case $\nu > 0$ it is harder to give good references,[1] but [**39**] and [**85**] are the natural starting points.

In what follows, one can work with either definition of $S_k(\Gamma_1(Np^\nu),\mathbb{Z}_p)$. In fact, more often than not we will want to work with $S_k(\Gamma_1(Np^\nu),\mathcal{O})$, where $\mathcal{O}$ is a discrete valuation ring which is finite over $\mathbb{Z}_p$. If $\mathcal{O}$ contains enough roots of unity, then it makes sense to require that the q-expansions at all the cusps have

[1] This is particularly true if one wants to define the Hecke operators geometrically, the trickiest one in our case being the U_p operator.

coefficients in $\mathcal{O}$, and we recover the same space as the one given by the geometric definition. It's important to note, however, that it is the geometric definition that really allows us to understand the situation.

The space $S_k(\Gamma_1(Np^\nu), \mathbb{Z}_p)$ is a finite free $\mathbb{Z}_p$-module on which act the Hecke operators T_ℓ, $\ell \nmid Np$, defined in the usual way. One can also define[2] operators U_ℓ for $\ell | N$ and, when $\nu > 0$, an operator U_p which acts on q-expansions by

$$\mathrm{U}_p\left(\sum a_n q^n\right) = \sum a_{np} q^n.$$

We will be particularly interested in *eigenforms*, that is, forms which are simultaneous eigenfunctions all for the Hecke operators.

There is another family of operators acting on our space, the diamond operators[3] $\langle n \rangle$, $n \in (\mathbb{Z}/N\mathbb{Z})^\times$. If a modular form is an eigenfunction for these operators, then we have $\langle n \rangle f = \varepsilon(n) f$ for some character ε which we call the *tame nebentypus character* of f. We extend this action to an action of $\mathbb{Z}_p^\times \times (\mathbb{Z}/N\mathbb{Z})^\times$, which we call the "double diamond" action $f \mapsto \langle x, y \rangle f$. The action of $x \in \mathbb{Z}_p^\times$ on a form f is determined by a combination of its weight and the p-part of its nebentypus character. In particular, if f is of weight k on $\Gamma_1(Np^\nu)$, and has nebentypus character $\varepsilon = \varepsilon_N \varepsilon_p$, where ε_N is a character on $(\mathbb{Z}/N\mathbb{Z})^\times$ and ε_p is a character on $(\mathbb{Z}/p^\nu\mathbb{Z})^\times$, then

$$\langle x, y \rangle f = \varepsilon_N(y) \varepsilon_p(x) x^k f.$$

(For more details about, and a more natural definition of, this action, see [**57**].) We will require eigenforms to also be eigenfunctions for the diamond action (which will usually require them to be defined over an extension $\mathcal{O}$ of $\mathbb{Z}_p$, as we mentioned above).

The reason for "twisting" the diamond action by the factor x^k is that this results in operators that act on the *sum* (over k) of the spaces $S_k(\Gamma_1(Np^\nu), \mathbb{Z}_p)$ in a way that preserves integrality of q-expansions. (See [**81**] for an example of how this can be used to find congruences between modular forms). This will be helpful for the p-adic theory in the next section.

We will say an eigenform is *normalized* if its q-expansion (at infinity) is of the form

$$f(q) = q + a_2 q^2 + a_3 q^3 + \dots.$$

In this case, it is easy to see that $\mathrm{T}_\ell(f) = a_\ell f$ for all $\ell \nmid Np$.

Let $\mathcal{O}$ be the valuation ring in a finite extension K of $\mathbb{Q}_p$. Given a form $f \in S_k(\Gamma_1(Np^\nu), \mathcal{O})$ which is an eigenform for the T_ℓ for all $\ell \nmid Np$ and for the diamond operators, one can construct a Galois representation ρ_f which is attached to f in the following sense. Suppose we have $\mathrm{T}_\ell f = a_\ell f$ and $\langle \ell, \ell \rangle f = \lambda(\ell) f$ for all $\ell \nmid Np$. Let

$$S = \{\text{primes dividing N}\} \cup \{p, \infty\},$$

[2] There seems to be no agreement about whether one should call these operators U_ℓ or simply define T_ℓ differently when $\ell | N$. The advantage of retaining the distinction is that then the action of T_ℓ on q-expansions is always the same as one varies k, N, and ν, and is different from the action of U_ℓ. In the case when $\ell = p$, the distinction between T_p (which acts on our space when $\nu = 0$) and U_p (which acts when $\nu > 0$) does matter for what we want to do.

[3] Notice the notational weirdness here: these are only *some* of the usual diamond operators on $S_k(\Gamma_1(Np^\nu), \mathbb{Z}_p)$. We will add the p-part when we move to the "double diamonds" in a minute.

and let $\Phi_\ell \in G_{\mathbb{Q},S}$ denote a geometric Frobenius transformation at ℓ. Then one can construct a Galois representation

$$\rho_f : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathcal{O})$$

such that $\rho_f \otimes_{\mathcal{O}} K$ is semisimple, and such that f and ρ_f are related by the formulas

$$\det \rho_f(\Phi_\ell) = \frac{1}{\ell}\lambda(\ell) \qquad \text{and} \qquad \operatorname{Tr} \rho_f(\Phi_\ell) = a_\ell$$

for each $\ell \nmid Np$. (Readers familiar with the usual definition should remember that if f is of weight k with nebentypus χ we have $\lambda(\ell) = \chi(\ell)\ell^k$.) It is easy to see (because $\rho_f \otimes_{\mathcal{O}} K$ is semisimple) that ρ_f is completely determined (up to equivalence over K) by f.

How is the representation attached to f obtained? The construction is due to Eichler, Shimura, Deligne, and Serre, and it is quite complicated. When the form f is of weight 2, one finds the representation by considering the Jacobian of the modular curve $X_1(Np^\nu)$. There is an action of the Hecke algebra on the Jacobian, and the eigenform gives a map from the Hecke algebra to R which allows us to "cut out" a piece of the étale cohomology of the Jacobian where the Galois group acts as we want (this is the dual of the usual approach, which finds the representation in the Tate module of the Jacobian). For general weights k, things get a lot more complicated. See [**117**, Appendix 2] for an account of how this works for $k = 2$, and [**144**], [**37**], and [**40**] for the beginnings of the rest of story.

Given such a ρ_f associated to a form f defined over some discrete valuation ring $\mathcal{O}$ which is a finite extension of $\mathbb{Z}_p$, we can reduce this modulo the maximal ideal $\mathfrak{m} \subset \mathcal{O}$. The resulting representation (which is defined over the residue field of $\mathcal{O}$) may not be semisimple, so we take its semisimplification, call it the reduction modulo the maximal ideal of the representation attached to f, and denote it by $\overline{\rho}_f$. The actual reduction may depend on the homomorphism $G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathcal{O})$ rather than just on its equivalence class; the semisimplification, however, is the same for any homomorphism in the equivalence class. (Notice that if $\overline{\rho}_f$ is absolutely irreducible, we do not need to take the semisimplification step, and $\overline{\rho}_f$ is just the reduction of ρ_f. We will always restrict to this situation in what follows.)

From our point of view, we should note that if we have two normalized eigenforms f and g (perhaps of different weights and with different values of ν) which are congruent modulo $\mathfrak{m}$ (in the sense that their q-expansion coefficients are congruent modulo $\mathfrak{m}$) then their Hecke and diamond eigenvalues are the same modulo p, so that the reductions modulo p of their Galois representations are the same also. In other words, if $\bar{f} = \bar{g}$, then the two representations ρ_f and ρ_g are (different) deformations of the same residual representation $\overline{\rho} = \overline{\rho}_f = \overline{\rho}_g$. In fact, we can weaken this somewhat: all we need is for the eigenvalues for T_ℓ with $\ell \nmid Np$ to be the same, and this will happen if

$$a_n(f) \equiv a_n(g) \pmod{\mathfrak{m}}$$

for all n such that $\gcd(n, Np) = 1$, where $a_n(f)$ (resp, $a_n(g)$) denotes the n-th q-expansion coefficient of f (resp, g).

In order to be able to think about whether all (or many) deformations are modular, we need to collect some information about these representations. The properties of residual modular representations are discussed in detail in [**117**], to which we refer the reader. We record here only a few useful facts, especially having to do with ramification. Given a residual Galois representation

$$\overline{\rho} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathsf{k}),$$

one may measure its ramification outside p by its conductor, which we denote by $\mathrm{N}(\overline{\rho})$. Since $\overline{\rho}$ is certainly not ramified at primes that do not divide Np, the conductor is a product of powers of primes dividing N,

$$\mathrm{N}(\overline{\rho}) = \prod_{\ell \in S - \{p\}} \ell^{n(\ell,\overline{\rho})},$$

where the numbers $n(\ell,\overline{\rho})$ are defined as follows: choose a place of $\overline{\mathbb{Q}}$ over ℓ, and let $I = I_\ell$ be the corresponding inertia group; let $\overline{V} = \mathsf{k} \times \mathsf{k}$ with the $G_{\mathbb{Q},S}$-action given by $\overline{\rho}$, and let $\overline{V}_0$ be the subspace of $\overline{V}$ fixed by $\overline{\rho}(I)$; then

$$n(\ell,\overline{\rho}) = 2 - \dim \overline{V}_0 + sw(\overline{\rho}),$$

where $sw(\overline{\rho})$ is the Swan conductor of (the restriction to a decomposition group at ℓ of) $\overline{\rho}$. (For a definition, see, for example, [**137**].) Note that if $\dim \overline{V}_0 = 2$, then $\overline{\rho}$ is unramified at ℓ and $n(\ell,\overline{\rho}) = 0$ (and in particular $sw(\overline{\rho}) = 0$ in this case). We know that $\overline{\rho}$ is tamely ramified at ℓ if and only if $sw(\overline{\rho}) = 0$, and that $\overline{\rho}$ is ℓ-ordinary exactly when $\dim \overline{V}_0 = 1$.

Of course, the conductor of $\overline{\rho}$ can be much smaller than the tame level N of the modular form from which it comes. In fact, that is a major theme of the recent work on Serre's conjecture reported in [**117**].

Now let's consider the situation for a lift of $\overline{\rho}$ to characteristic zero. As before, let K be a finite extension of $\mathbb{Q}_p$, let $\mathcal{O}$ be its valuation ring, and assume the residue field of $\mathcal{O}$ is k. The conductor of a deformation

$$\rho : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathcal{O})$$

of $\overline{\rho}$ (which may or may not be modular) is defined in an analogous way, as

$$\mathrm{N}(\rho) = \prod_{\ell \in S - \{p\}} \ell^{n(\ell,\rho)},$$

where we take $V = \mathcal{O} \times \mathcal{O}$ with the action of $G_{\mathbb{Q},S}$ given by ρ, V_0 is the submodule of invariants under the action of an inertia group at ℓ, and set

$$n(\ell,\rho) = 2 - \mathrm{rank}_{\mathcal{O}} V_0 + sw(\rho) = 2 - \mathrm{rank}_{\mathcal{O}} V_0 + sw(\overline{\rho}).$$

(The fact that the Swan conductors of ρ and of $\overline{\rho}$ are equal is well known; it is so because the wild inertia group at ℓ is a pro-ℓ-group, while, as noted in Lecture 5, the kernel of the reduction map $\mathrm{GL}_2(A) \longrightarrow \mathrm{GL}_2(\mathsf{k})$ is a pro-p-group, so that all the wild ramification will already occur in the $\overline{\rho}$.)

Note that we have eliminated powers of p from the conductor; this is inevitable, since we are not assuming ρ is a part of a compatible family of ℓ-adic representations. (For a modular deformation, we have an ℓ-adic representation for every prime ℓ, and the exponents in the conductor agree. This allows us to define the p-part of the conductor by looking at an ℓ-adic representation for $\ell \neq p$. If we want to do deformation theory, however, we have to stick to a p-adic setting.) That it is the "tame level" that is detectable from the representation is also pleasantly consistent with the fact that only the prime-to-p part of the level is relevant in the context of p-adic modular forms, as we will see below.

What do we know about these representations (and specifically, their conductors)? In characteristic zero, we know

- If f is a newform of level Np^ν, then the (prime-to-p part of the) Artin conductor of the representation ρ_f is equal to N. (This is an important theorem of Carayol.)
- The local representations $\rho_f|_{G_{\mathbb{Q}_\ell}}$, for $\ell|N$, $\ell \neq p$, are well-understood via the "local Hecke correspondence." (See [**15**] or the more accessible [**14**] for the details, which are complicated.)

Modulo p, we know a little more. In particular, the recent work on Serre's conjecture involved a quite detailed understanding of the local representations modulo p. See [**117**] for more details.

For the deformation theory, it is important to compare the situation in characteristic zero with the situation in characteristic p. For example, the following result captures what can happen to the conductor:

Proposition 7.1. *Let $\overline{\rho}$ be a residual Galois representation*

$$\overline{\rho} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathsf{k}),$$

and let ρ be any deformation of $\overline{\rho}$ to characteristic zero. Then, for each $\ell \in S$:

i. if $\overline{\rho}$ is unramified at ℓ, then $n(\ell,\rho) \leq 2$;
ii. if $\overline{\rho}$ is ℓ-ordinary, then $n(\ell,\rho) \leq n(\ell,\overline{\rho}) + 1$;
iii. if $\overline{\rho}$ and ρ are both ℓ-ordinary, then we have $n(\ell,\rho) = n(\ell,\overline{\rho})$;
iv. if $\overline{\rho}$ is ramified at ℓ but not ℓ-ordinary, then $n(\ell,\rho) = n(\ell,\overline{\rho})$.

Proof. This all follows immediately from the fact that

$$n(\ell,\rho) - n(\ell,\overline{\rho}) = \dim \overline{V}_0 - \operatorname{rank} V_0,$$

since $\operatorname{rank} V_0 \leq \dim \overline{V}_0 \leq 2$. □

The local representation $\rho_f|_{G_{\mathbb{Q}_p}}$ is somewhat harder to understand. The main tool for studying it is Fontaine's theory of p-adic $G_{\mathbb{Q}_p}$-modules. Using this language, we can describe what we know about the local representation attached to a modular form by saying that it is either *crystalline* (if $\nu = 0$) or, in general, *potentially semistable*. See [**52**] for definitions, discussion, and a study of the deformation problems attached to such conditions.

We can also use the p-adic Hodge theory of Tate and Sen to study the local representation. We know that the local representation $\rho_f|_{G_{\mathbb{Q}_p}}$ is of Hodge-Tate type, and the Hodge-Tate-Sen weights (as defined in [**129**] and [**130**] ; see also [**97**] and [**100**]) attached to it are $(0, k-1)$. In particular, any representation that comes from a classical modular form must be in the "Sen null subspace" of the deformation space, which consist of those deformations which have one of their Sen weights equal to zero. This is known to be a codimension one analytic subspace of the (rigid-analytic subpsace attached to the) full deformation space. See the discussion in §7 of [**100**] for more details.

One can "move" from the Sen null subspace by twisting deformations by characters of infinite order which reduce to the trivial character modulo p (sometimes called "wild" characters of infinite order). See [**97**] and [**100**] for a more careful discussion of how twisting can be interpreted as the action of a formal group on the deformation space that is "essentially transversal" to the Sen null subspace. (We will get back to this in the next lecture.)

One final bit of information is also quite useful. If the eigenvalue for the p-th Hecke operator (either T_p or U_p, depending on whether $\nu = 0$ or $\nu > 0$) is a p-adic unit, then the local representation $\rho_f|_{G_{\mathbb{Q}_p}}$ is p-ordinary (see [**106**], for example). For residual representations, the converse is true (see [**78**]); the converse for representations in characteristic zero is known in some cases (basically, whenever one knows an "all ordinary deformations are modular" theorem, see below).

So let's summarize what we know. In the full deformation space, which we expect to be three-dimensional by the dimension conjecture (see page 55), the representations that come from (classical) modular forms sit in various smaller subspaces: first, they all sit in the Sen null subspace, which (since we know it is of codimension one) we expect to be two-dimensional. Second, a representation coming from a modular of weight k has determinant equal to (a finite character times) the $(k-1)$-th power of the cyclotomic character, and we can consider the subspace of the deformation space corresponding to deformations with that determinant. This again gives a space of codimension one. If the modular form is one whose p-th Hecke eigenvalue is a p-adic unit, the representation is p-ordinary, and we can consider the subspace corresponding to p-ordinary deformations (by contrast with the others, this is usually one-dimensional). If we're even more ambitious, we might try to pin down the subspace corresponding to deformations that are potentially semi-stable (if there is one). Finally, we can try to get the conductor right by controlling the ramification at the primes $\ell \neq p$.

This raises a number of interesting questions about the deformation space. Which of these conditions are "deformation conditions" in the sense of Lecture 6, and hence define algebraic subspaces of the deformation space? How do these various subspaces intersect?

On the other hand, this discussion also points out that modular deformations are very special, and that they at best fill up only a small part of the deformation space. If we hope to have a theorem saying something like "all deformations are modular," we will either have to restrict the meaning of "deformation" (by introducing deformation conditions, as in Lecture 6) or extend the meaning of "modular."

p-adic modular forms

One way of extending the meaning of a "modular deformation" is to work with "p-adic modular forms." In this section we give a very rough description of what this theory looks like. See Matthew Emerton's survey in Appendix 3 for a much more careful description, and [**134**], [**79**], [**82**], [**57**], [**26**], [**27**], and [**28**] for various accounts of the details.

The theory of p-adic modular forms gives us a very large space $\mathbf{V}(N, \mathbb{Z}_p)$ of "parabolic p-adic modular functions defined over $\mathbb{Z}_p$." We can go from forms over $\mathbb{Z}_p$ to forms over more general p-adically complete and separated rings R simply by defining $\mathbf{V}(N, R) = \mathbf{V}(N, \mathbb{Z}_p)\hat{\otimes}R$. Rather than describe this space and its construction, we note only that for any weight k and any $\nu \geq 0$ there exist inclusions

$$S_k(\Gamma_1(Np^\nu), \mathbb{Z}_p) \hookrightarrow \mathbf{V}(N, \mathbb{Z}_p),$$

and that the union of the images of these inclusions is dense in $\mathbf{V}(N, \mathbb{Z}_p)$ with respect to the p-adic topology derived from the q-expansions. Thus, $\mathbf{V}(N, \mathbb{Z}_p)$

contains every single one of the eigenforms we have considered so far, plus many more that are obtained by some sort of limiting process.[4]

There are naturally-defined Hecke operators T_ℓ (for $\ell \nmid Np$) and diamond operators $\langle x, y\rangle$ on $\mathbf{V}(N, \mathbb{Z}_p)$. The Hecke operators act as expected on q-expansions, and both the Hecke and the diamond operators restrict to the ones we defined above when we apply them to classical modular forms. (The other Hecke operators, and in particular the U_p operator, also extend, and the U_p operator in fact plays an important role in the theory. For now, however, we will stick with a smaller set of operators.)

The easiest way to define the Hecke operators on $\mathbf{V}(N, \mathbb{Z}_p)$ is to use the fact (first proved by Hida) that for any fix ν the union over k of the spaces $S_k(\Gamma_1(Np^\nu), \mathbb{Z}_p)$ is dense in $\mathbf{V}(N, \mathbb{Z}_p)$. Using this, we can define the Hecke algebra $\mathbf{T}$ as the inverse limit over k of the (restricted) Hecke algebras[5] acting on $S_k(\Gamma_1(Np^\nu), \mathbb{Z}_p)$. The action of $\mathbf{T}$ extends to all of $\mathbf{V}(N, \mathbb{Z}_p)$ by continuity. We can check that $\mathbf{T}$ is independent of ν, so that we get an algebra of operators on $\mathbf{V}(N, \mathbb{Z}_p)$, and that the T_ℓ act as expected on q-expansions. Each of the Hecke algebras at finite level has a natural p-adic topology, and we give $\mathbf{T}$ the inverse limit topology. Any p-adic modular function $f \in \mathbf{V}(N, \mathbb{Z}_p)$ determines a continuous map $\mathbf{T} \longrightarrow \mathbb{Z}_p$ by mapping an operator T to $a_1(Tf)$, the first q-expansion coefficient of Tf. This map is a homomorphism if and only if f is a normalized eigenform; in that case, it maps each operator to its eigenvalue. More generally, for any p-adically complete and separated ring R, any normalized eigenform $f \in \mathbf{V}(N, R)$ gives a continuous homomorphism $\mathbf{T} \longrightarrow R$.

(One might suspect that any such homomorphism determines a form f. This would certainly be true if we had included *all* the Hecke operators in $\mathbf{T}$. It is a little more problematic with our setup.)

Now suppose R is a ring in $\mathcal{C}$ and suppose that we have a form $f \in \mathbf{V}(N, R)$ which is an eigenform for T_ℓ (for all $\ell \nmid Np$) and for the diamond operators. Then f determines a homomorphism $\phi_f : \mathbf{T} \longrightarrow R$, and hence, after reduction modulo the maximal ideal, a homomorphism $\mathbf{T} \longrightarrow \mathsf{k}$. Let $\mathfrak{m}$ be the kernel of this homomorphism, and let $\mathcal{R}_m(\bar{f}) = \mathbf{T}_\mathfrak{m}$ be the completion of $\mathbf{T}$ at $\mathfrak{m}$.[6] Given another eigenform $g \in \mathbf{V}(N, R)$, the corresponding homomorphism ϕ_g factors through $\mathcal{R}_m(\bar{f})$ if and only if $\bar{f}$ and $\bar{g}$ have the same Hecke eigenvalues for all T_ℓ with $\ell \nmid Np$ and for all the diamond operators (i.e., if and only if the eigenvalues of f and of g are the same modulo the maximal ideal). In other words, $\mathcal{R}_m(\bar{f})$ is a kind of "universal deformation ring" for the packet of Hecke eigenvalues coming from the residual eigenform $\bar{f}$.

Now let $\overline{\rho}$ be the residual Galois representation attached to $\bar{f}$. If $\overline{\rho}$ is absolutely irreducible, then it has a universal deformation ring $\mathcal{R}(\overline{\rho})$. To get a "universal modular deformation" of $\overline{\rho}$, we need to construct a deformation of $\overline{\rho}$ to the completed Hecke algebra $\mathcal{R}_m(\bar{f})$. By the universal property of $\mathcal{R}(\overline{\rho})$ constructing such a thing

[4] Eigenforms in $\mathbf{V}(N, \mathbb{Z}_p)$ are sometimes limits of classical eigenforms, but more commonly they are limits of classical modular forms which are not eigenforms, but which are "eigenforms modulo p^n" for bigger and bigger n.

[5] That is, the $\mathbb{Z}_p$-submodule of the algebra of endomorphisms of $S_k(\Gamma_1(Np^\nu), \mathbb{Z}_p)$ which is generated by the T_ℓ with $\ell \nmid Np$ and by the diamond operators.

[6] The m in the notation $\mathcal{R}_m(\bar{f})$ stands for "modular." The point is that (in many cases) this ring will turn out to give a "universal modular deformation" of the representation $\overline{\rho}$ corresponding to $\bar{f}$.

is the same as constructing a homomorphism $\mathcal{R}(\overline{\rho}) \longrightarrow \mathcal{R}_m(\bar{f})$. This turns out to be possible.

What we get is the following. Fix the "tame level" N, and let

$$S = \{\text{primes dividing } N\} \cup \{p, \infty\}.$$

Suppose we have an eigenform $\bar{f} \in \mathbf{V}(N, \mathbb{F}_p)$. Since the classical forms are dense in $\mathbf{V}(N, R)$, we know that $\bar{f}$ is equal to the reduction of a classical eigenform, and in fact we can assume that this classical eigenform is of level N. In particular, there is a Galois representation attached to $\bar{f}$; let $\overline{\rho} : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathsf{k})$ be the attached Galois representation.

Theorem 7.2 (Gouvêa, Hida). *There exists a representation*

$$\boldsymbol{\rho}_m : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathcal{R}_m(\bar{f}))$$

such that, if Φ_ℓ *is a geometric Frobenius at* $\ell \nmid Np$,

$$\det \boldsymbol{\rho}_m(\Phi_\ell) = \frac{1}{\ell}\langle \ell, \ell \rangle \qquad \textit{and} \qquad \operatorname{Tr} \boldsymbol{\rho}_m(\Phi_\ell) = \mathrm{T}_\ell.$$

In [**57**], this theorem is proved by constructing a homomorphism from the universal deformation ring of $\overline{\rho}$ to the completed Hecke algebra. In the original version, there were technical assumptions which have since been removed by work of Carayol (see [**16**] and [**101**, §6]). Hida's proof (in [**73**]; see also [**72**]) gets around these technicalities by using the theory of pseudo-representations (also known as pseudo-characters).

A (perhaps surprising) consequence of this theorem is that *there exists a Galois representation attached to a p-adic eigenform f even when that eigenform is not a classical modular form*, provided that the residual representation attached to a classical eigenform that is congruent to f is absolutely irreducible. These representations are considerably more mysterious than the representations attached to classical eigenforms. For example, they are not necessarily of Hodge-Tate type (but see [**86**] for a positive result along these lines).

If $\overline{\rho}$ is absolutely irreducible, so that the theorem applies, we call $\boldsymbol{\rho}_m$ the *universal modular deformation* of $\overline{\rho}$. It parametrizes all the deformations of $\overline{\rho}$ which come from p-adic modular forms (which of course includes the ones which come from classical modular forms). Since the universal modular deformation is itself a deformation of $\overline{\rho}$, we get a homomorphism

$$\mathcal{R}(\overline{\rho}) \longrightarrow \mathcal{R}_m(\bar{f}).$$

The trace of the Frobenius Φ_ℓ in the universal deformation ring must map to T_ℓ in $\mathcal{R}_m(\bar{f})$. Since the Hecke operators topologically generate $\mathcal{R}_m(\bar{f})$, the homomorphism is surjective.

Thus, we have constructed a "universal modular deformation" which cuts out the portion of the deformation space which corresponds to deformations attached to p-adic modular forms. This can be viewed as a sort of Zariski closure of the points in the deformation space corresponding to modular deformations in the classical sense. (We do not expect it to be the closure in the p-adic topology, because p-adic eigenforms need not be limits of classical *eigen*forms.) The question, then, is how big a portion of the deformation space we have obtained. It will be the full deformation space exactly when the homomorphism

$$\mathcal{R}(\overline{\rho}) \longrightarrow \mathcal{R}_m(\bar{f})$$

is an isomorphism. As we will see in the next lecture, in at least one case we know that it is in fact the full deformation space, i.e., we know that every deformation is "pro-modular" (i.e., comes from a p-adic modular form). On the other hand, we will see that unless $N = 1$ we cannot escape imposing some sort of deformation condition before we can get a positive result.

The game we just played with the big p-adic Hecke algebra can also be played with classical Hecke algebras. Thus, suppose we have an eigenform

$$f \in S_k(\Gamma_1(N), \mathbb{Z}_p),$$

and let $\overline{\rho}$ be the residual representation attached to f. (Eigenforms defined over finite extensions of $\mathbb{Z}_p$ work exactly the same way; we work with $\mathbb{Z}_p$ for simplicity.) Let $\mathbf{T}_k(N)$ be the subalgebra of the endomorphisms of $S_k(\Gamma_1(N), \mathbb{Z}_p)$ generated by the Hecke operators T_ℓ for $\ell \nmid Np$ and by the diamond operators. As before, the eigenform f determines a homomorphism $\mathbf{T}_k(N) \longrightarrow \mathbb{Z}_p$; reducing modulo the maximal ideal gives a homomorphism $\mathbf{T}_k(N) \longrightarrow \mathbb{F}_p$ whose kernel is a maximal ideal $\mathfrak{m}$. Let $\mathbf{T}(\bar{f})$ be the completion of $\mathbf{T}_k(N)$ at the maximal ideal $\mathfrak{m}$. Then there is a surjective homomorphism from the universal modular deformation ring $\mathcal{R}_m(\bar{f})$ to $\mathbf{T}(\bar{f})$, and therefore there exists a deformation ρ_f of $\overline{\rho}$ to $\mathbf{T}(\bar{f})$. (Note that this is generally not the same as the Galois representation attached to f. In fact, they are the same only if $\mathbf{T}(\bar{f}) = \mathbb{Z}_p$.) This deformation parametrizes all deformations of $\overline{\rho}$ which come from classical modular forms of weight k and level N, i.e., it plays the same role, for weight k and level N, that the big Hecke algebra $\mathcal{R}_m(\bar{f})$ plays for all p-adic modular forms.

If we could determine the representation-theoretic properties of ρ_f with sufficient precision, we could hope to write down deformation conditions that would restrict our deformation problem to "those deformations that look as if they come from modular forms of weight k and level N." One part of this is not difficult: to fix the weight, we need a determinant condition. Fixing the level, as before, boils down to imposing local deformation conditions at the primes dividing N. The subtle part is finding a deformation condition that will restrict us to classical modular forms. This would presumably be a local condition at p. In Wiles' work, for example, this condition was either that the deformation be p-ordinary or that it be flat at p.

If we successfully find the correct set $\mathcal{Q}$ of deformation conditions, then we get a ring $\mathcal{R}_\mathcal{Q}(\overline{\rho})$ that gives us the universal deformation subject to those conditions. This should give us a surjective homomorphism

$$\mathcal{R}_\mathcal{Q}(\overline{\rho}) \longrightarrow \mathbf{T}(\bar{f}).$$

As before, asking whether this map is in fact an isomorphism amounts to asking whether "all deformations (of this kind) are modular (of weight k and level N)." In comparison to the overall question about whether all deformations are modular, this is both more precise (in particular, a yes answer to the question about the big deformation ring does not imply a yes answer here without some extra work) and perhaps more accessible. One reason for this is that the Hecke algebras $\mathbf{T}_k(N)$ and $\mathbf{T}(\bar{f})$ are relatively well understood, while the big Hecke algebra $\mathcal{R}_m(f)$ is much more mysterious (as are, of course, the universal deformation rings).

This approach to the problem, which works with fixed weight and level, is what appears in the work of Wiles, Taylor, Diamond, Breuil, and Conrad on the Shimura-Taniyama conjecture. For the most part, they work with weight $k = 2$, and consider several different deformation problems attached to $k = 2$ and varying levels. (For the crucial application to the representation attached to an elliptic curve, it's essential *not* to impose the condition that the residual modular form $\bar{f}$ is of minimal level.) The hardest part of this work, as noted above, is to pin down the deformation conditions that will restrict us to classical modular forms. As we will see in the next section, this is easy to do in the ordinary case, because of Hida's "control theorem." In the non-ordinary case, they use (variants of) the flat deformation condition we discussed in Lecture 6 to force the restriction to classical modular forms.

The ordinary case

The case of representations coming from ordinary p-adic modular forms is much better understood than the general case. We say that a p-adic modular function $f \in \mathbf{V}(N, R)$ is *ordinary* if f belongs to the R-submodule of $\mathbf{V}(N, R)$ (topologically) spanned by generalized eigenforms[7] for the U_p operator corresponding to eigenvalues which are p-adic units. As Hida shows, there exists an idempotent e in the endomorphism ring of $\mathbf{V}(N, \mathbb{Z}_p)$ which commutes with the Hecke and diamond operators and "picks out" the ordinary forms, so that $ef = f$ if and only if f is ordinary. On the representation theory side, Mazur and Wiles have shown in [**106**] that if f is ordinary then the associated representation is also I_p-ordinary in the representation-theoretic sense discussed above, that is, the subspace fixed by inertia in the representation space is of rank one and a direct summand. This nice match between a "modular" condition and a representation condition, together with the fact that thanks to Hida we know quite a lot about the Hecke algebra associated to ordinary modular forms, allows us to set things up nicely ... and some of the questions actually have answers.

First, we let $\mathbf{T}^0 = e\mathbf{T}$ and call it the ordinary part of the Hecke algebra (with our definitions, e does not belong to $\mathbf{T}$, so that $\mathbf{T}^0$ is not actually a "part" of $\mathbf{T}$; nevertheless, this wording is instructive). As before, an ordinary eigenform $\bar{f} \in \mathbf{V}(N, \mathsf{k})$ gives a map $\mathbf{T}^0 \longrightarrow \mathsf{k}$ whose kernel is a maximal ideal $\mathfrak{m}$, and we write $\mathcal{R}^0_m(\bar{f})$ for the completion of $\mathbf{T}^0$ at $\mathfrak{m}$. Then Hida proves the following theorem.

Theorem 7.3 (Hida). *If $p \geq 5$, and $\overline{\rho}$ is absolutely irreducible, there exists a representation*

$$\boldsymbol{\rho}_h : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathcal{R}^0_m(\bar{f}))$$

such that, if Φ_ℓ is a geometric Frobenius at $\ell \nmid Np$,

$$\det \boldsymbol{\rho}_h(\Phi_\ell) = \frac{1}{\ell}\langle \ell, \ell\rangle \qquad \textit{and} \qquad \operatorname{Tr} \boldsymbol{\rho}_h(\Phi_\ell) = \mathrm{T}_\ell.$$

We can think of Hida's representation as the "universal modular-ordinary deformation" of $\overline{\rho}$, since it parametrizes all deformations of $\overline{\rho}$ which come from ordinary p-adic modular forms. One of the more important things about $\mathcal{R}^0_m(\bar{f})$ is that one knows what sort of ring it is: Hida showed that it is a finite flat algebra over $\Lambda = \mathbb{Z}_p[[\Gamma]]$. One also knows, by work of Mazur and Wiles, that the representation $\boldsymbol{\rho}_h$ is ordinary in the sense of representation theory.

Suppose $\overline{\rho}$ is absolutely irreducible and attached to an ordinary modular form $\bar{f}$, so that we also know that $\overline{\rho}$ is ordinary in the representation-theoretic sense.

[7] An eigenform f for the U_p operator satisfies $(\mathrm{U}_p - \lambda)f = 0$ for some λ; we say f is a *generalized eigenform* attached to the eigenvalue λ if it satisfies $(\mathrm{U}_p - \lambda)^n f = 0$ for some $n \geq 1$.

Then we have constructed various deformation rings: the universal deformation ring $\mathcal{R}(\overline{\rho})$, the universal modular deformation $\mathcal{R}_m(\bar{f})$, the universal ordinary deformation $\mathcal{R}^0(\overline{\rho})$, and the universal modular-ordinary deformation $\mathcal{R}^0_m(\bar{f})$. These fit together in a diagram:

$$\begin{tikzcd}
\mathcal{R}(\overline{\rho}) \arrow[r] \arrow[d] & \mathcal{R}^0(\overline{\rho}) \arrow[d] \\
\mathcal{R}_m(\bar{f}) \arrow[r] & \mathcal{R}^0_m(\bar{f})
\end{tikzcd}$$

with all of the maps in the diagram surjective. Around 1990, Mazur and I both stated the conjecture that the vertical maps are in fact isomorphisms, that is, that *all deformations of a modular residual representation are (p-adically) modular.* Similarly, we would conjecture that *all ordinary deformations of a residual representation coming from an ordinary modular form themselves come from ordinary p-adic modular forms.* As they stand, both conjectures seem unlikely to be true unless $N = 1$, simply because, as the notation indicates, the top row depends on $\overline{\rho}$ and the bottom row depends on $\bar{f}$. This seems innocuous until we realize that in the deformation theory we simply fixed a set of primes at which we allowed ramification, while in the modular theory we fixed the tame level N. If we take a form f of level N and think of it as of level N^2, say, the top row does not change, while the bottom row does. This requires us to make things a little more precise before we make our conjecture.

Imposing deformation conditions

There are several choices as to how to proceed.

i. We can impose local conditions that force the conductor to remain equal to N. If we assume that our form was chosen with "optimal level," this amounts to making ρ be "as unramified as possible."

ii. In addition to the local conditions, we can impose conditions that force the representations to look like those that come from classical modular forms. For example, in the weight two case, we could follow Wiles and Taylor (see above) and require that the determinant be the cyclotomic character, impose local conditions at $\ell \neq p$, and at p restrict ourselves to representations that are p-ordinary (if $\overline{\rho}$ is) or finite flat (if $\overline{\rho}$ is).

iii. We can also *relax* some of the local conditions, provided we understand where to look for modular forms that produce representations satisfying the relaxed conditions.

Here's a sketch of the first approach.

To be able to formulate the right conjecture about modularity, we must understand the relation between the level N of a modular form and the (tame part of the) conductor of the attached Galois representation. Suppose we have an absolutely irreducible representation $\overline{\rho}$ which comes from a modular form $\bar{f} \in S_k(\Gamma_1(N), \mathsf{k})$ (as we mentioned above, there is no loss of generality in assuming $\bar{f}$ is of level exactly N).

For this whole section,[8] let $p \geq 5$. To measure the ramification of $\overline{\rho}$, we use, as discussed above, its conductor. Since $\overline{\rho}$ is attached to a modular form of level N, it follows from the discussion above that the conductor of $\overline{\rho}$ will be a divisor of N, and in fact Proposition 7.1 gives a quite precise description of how the two can differ. This lets us "control the conductor" in the deformation theory by imposing local deformation conditions at the primes dividing N.

We describe the strategy in the "minimal case," in which we assume that we have chosen our modular form $\bar{f}$ so that the level N is optimal. (That this is possible for $p \geq 5$ is one of the main theorems described in [**117**].) Once we have done that, the conductor of $\overline{\rho}$ will be exactly equal to N. We will look for deformations that "look as if they might correspond to forms of level N" by imposing local ramification conditions at some of the primes $\ell | N$.

As usual, let

$$S = \{\text{primes dividing } N\} \cup \{p, \infty\}.$$

Now suppose that we take a discrete valuation ring R which is a finite extension of $\mathbb{Z}_p$ and has residue field k, and suppose we have a classical eigenform $g \in S_{k'}(\Gamma_1(N), R)$ such that $\bar{g} = \bar{f}$. Then it follows from Carayol's main theorem in [**15**] that the conductor of the corresponding representation ρ_g is exactly N.

On the other hand, suppose ρ_1 is unramified outside S and $\overline{\rho}_1 = \overline{\rho}$. Then the conductor of ρ_1 can indeed be bigger than N. Proposition 7.1 shows that this will happen if and only if there exists a prime $\ell \in S$ such that $\overline{\rho}$ is I_ℓ-ordinary and ρ_1 is not I_ℓ-ordinary[9]. This gives us the clue about how to fix the problem.

Suppose, then, that $\overline{\rho}$ is an absolutely irreducible Galois representation arising from a modular form $\bar{f} \in S_k(\Gamma_1(N), \mathsf{k})$ which is the reduction of a classical modular form $f \in S_k(\Gamma_1(N), R)$ (with R as above), and assume that the conductor of $\overline{\rho}$ is exactly N, i.e., that f is of optimal level. Let

$$S = \{\text{primes dividing } N\} \cup \{p, \infty\},$$

and let

$$S_0 = \{\ell | N \text{ such that } \overline{\rho} \text{ is } I_\ell\text{-ordinary}\}.$$

(Let's note in passing that one can determine which primes are in S_0 in strictly "modular" terms—see [**58**] for the details.) By our assumptions on f, the representation ρ_f is also I_ℓ-ordinary for every $\ell \in S_0$. Let $\mathcal{Q}$ denote the condition that any deformation $\overline{\rho}$ be I_ℓ-ordinary for all $\ell \in S_0$. This is easily checked to be a deformation condition, and hence it defines a deformation ring:

$$\mathcal{R}_N(\overline{\rho}) = \mathcal{R}_{\mathcal{Q}}(\overline{\rho})$$

This is the universal deformation ring for deformations unramified outside S and ordinary at each $\ell \in S_0$, and we have a corresponding universal deformation

$$\boldsymbol{\rho}_N : G_{\mathbb{Q},S} \longrightarrow \mathrm{GL}_2(\mathcal{R}_N(\overline{\rho})).$$

We might call these the *universal level N deformation ring* and the *universal level N deformation*, respectively.

[8] For $p = 2$ or 3 one needs to be careful with the issue of adjusting the level, and the theory of p-adic forms requires a bit more care, so we prefer not to consider them here.

[9] This is true only because of our assumption that $\bar{f}$ has been chosen of minimal level. Without this assumption, it could also be the case that $\overline{\rho}$ is unramified at ℓ while ρ is ramified. When one considers a non-minimal deformation problem, this case must also be taken into account.

If $\overline{\rho}$ is ordinary at p, we can do the analogous thing with the added deformation condition of being I_p-ordinary, and define $\mathcal{R}^0_N(\overline{\rho})$, the universal ($p$-)ordinary level N deformation ring.

With these definitions, we can show that the homomorphism $\mathcal{R}(\overline{\rho}) \longrightarrow \mathcal{R}_m(\bar{f})$ and (if $\overline{\rho}$ is ordinary at p and f is an ordinary modular form) the homomorphism $\mathcal{R}^0(\overline{\rho}) \longrightarrow \mathcal{R}^0_m(\bar{f})$ factor through the level N deformation rings, giving maps

$$\mathcal{R}_N(\overline{\rho}) \longrightarrow \mathcal{R}_m(\bar{f}) \tag{I}$$

and (if $\overline{\rho}$ is ordinary at p and f is an ordinary modular form)

$$\mathcal{R}^0_N(\overline{\rho}) \longrightarrow \mathcal{R}^0_m(\bar{f}) \tag{II}$$

Both of these maps are known to be surjective (see above, or [**58**] for more detail), and it is now reasonable to conjecture that they are in fact isomorphisms.

Conjecture. The maps (I) and (II) above are isomorphisms.

This conjecture is due to Mazur, though it seems that it was first stated in print in [**58**]. What it says is that any deformation of a modular residual representation is p-adically modular, i.e., attached to a p-adic modular form. The work of Wiles, Taylor-Wiles, et. al. is sufficient to establish in many cases that (II) is an isomorphism. As for (I), it seems much harder to get a handle on it, basically because we do not really know very much about the big Hecke algebra. We will sketch later an argument (involving the "infinite fern" construction) that proves that (I) is true in a particularly simple case (basically, when $N = 1$ and the deformation problem is unobstructed).

Suppose we can prove (I). Then we know that every (appropriately ramified) deformation comes from a p-adic modular form of level N. The question of how to locate the deformations attached to classical modular forms now amounts to asking whether we can locate the classical modular forms within the p-adic modular forms. One important result is Hida's "control theorem." Let k be an integer; we say that a p-adic modular form f is of weight k if the left diamond operators act via k-th powers:

$$\langle x, 1\rangle f = x^k f \qquad \text{for all } x \in \mathbb{Z}_p^\times.$$

Theorem 7.4 (Hida). *Let f be an ordinary p-adic modular form of weight $k \geq 2$ and level N. Then f is a classical modular modular form of weight k on $\Gamma_1(N) \cap \Gamma_0(p)$.*

Two remarks are in order. First, if $k \geq 3$ the form f is in fact of level N. (This follows, for example, from the discussion of p-old and p-new forms in the next lecture.) Second, Hida in fact shows a much more general result which captures classical forms of level Np^ν with $\nu > 0$; the main change is that one must consider characters of $\mathbb{Z}_p^\times$ that are more complicated than raising to the k-th power.

For the non-ordinary case, things are much less satisfactory. Coleman has proved a generalized control theorem (we will discuss it in the next lecture), but it applies to p-adic modular forms which have a special property (they are "overconvergent"). The problem is that we do not yet know how to distinguish the representations attached to overconvergent forms from representations that come from non-overconvergent forms.

LECTURE 8
p-adic families and infinite ferns

The goal of this final lecture is to explain how Mazur and I showed, in a very special case, that "all deformations are modular." This involves using Coleman's work on families of modular forms and the theory of p-old and p-new forms to produce an intricate structure inside the deformation ring. The existence of this structure, together with the assumption that the deformation problem is unobstructed, then yields the result.

Here's the basic setup. We'll assume we are given a residual representation

$$\overline{\rho} : G_{\mathbb{Q},\{p,\infty\}} \longrightarrow \mathrm{GL}_2(\mathbb{F}_p)$$

which is absolutely irreducible and comes from some eigenform of weight k and level 1 defined over $\mathbb{Z}_p$. Let $f \in S_k(\Gamma_1(1), \mathbb{Z}_p)$ be the eigenform attached to $\overline{\rho}$.

As before, we can consider the universal deformation ring $\mathcal{R}(\overline{\rho})$ and the universal modular deformation ring $\mathcal{R}_m(\bar{f})$ (which is just a completion of the big p-adic Hecke algebra). Since we are assuming that $N = 1$, we don't have to worry about imposing local deformation conditions. As before, we then have a homomorphism

$$\mathcal{R}(\overline{\rho}) \longrightarrow \mathcal{R}_m(\bar{f}),$$

and we want to prove this is in fact an isomorphism.

The crucial assumption we will make is the following:

Assume that the deformation problem for $\overline{\rho}$ is unobstructed.

In particular, we have $\mathcal{R}(\overline{\rho}) \cong \mathbb{Z}_p[[T_1, T_2, T_3]]$. This allows us to think of the deformation space in a very concrete way: every triple of p-adic integers $(a_1, a_2, a_3) \in p\mathbb{Z}_p \times p\mathbb{Z}_p \times p\mathbb{Z}_p$ defines a homomorphism

$$\mathcal{R}(\overline{\rho}) = \mathbb{Z}_p[[T_1, T_2, T_3]] \longrightarrow \mathbb{Z}_p,$$

and this describes all such homomorphisms. Hence, we can think of the space of deformations of $\overline{\rho}$ to $\mathbb{Z}_p$ as a "cube with side $p\mathbb{Z}_p$," i.e., a kind of affine three-dimensional space. (The same is true over any extension of $\mathbb{Z}_p$ also, of course.)

Our goal is to exploit two very simple ideas (the "slope" of an eigenform and the theory of p-old and p-new forms), together with a powerful theorem of Coleman, to produce a large number of points in our space that are attached to (classical) modular forms. Under a mild technical assumption on the form f, there turn out to be enough such points that one can conclude that the homomorphism

$$\mathcal{R}(\overline{\rho}) \longrightarrow \mathcal{R}_m(\bar{f})$$

must in fact be an isomorphism. Rather than give the full proof, we will set up the ideas that allow us to construct many modular points and use them to construct a very complex object inside the deformation space. The details of how to prove that the existence of this object implies that the homomorphism above must be an isomorphism can be found in [**62**].

The slope of an eigenform

As before, let N be a number not divisible by p. We'll want to consider modular forms on $\Gamma_1(N)$, but for the p-adic theory it's important to work with the U_p operator rather than the more natural T_p. Since the space of forms on $\Gamma_1(N)$ is not stable under U_p, we move to the next largest space that is, i.e., we look at modular forms on the group $\Gamma_1(N) \cap \Gamma_0(p)$. Notice that $\Gamma_1(N) \supset \Gamma_1(N) \cap \Gamma_0(p) \supset \Gamma_1(Np)$.

Suppose, then, that $f \in S_k(\Gamma_1(N) \cap \Gamma_0(p), \mathbb{C}_p)$, where $\mathbb{C}_p$ is the completion of an algebraic closure of $\mathbb{Q}_p$. (We extend the field to $\mathbb{C}_p$ to avoid having to worry about the field of definition of our eigenforms. This way, our definition is as general as possible.) Suppose f is an eigenform under the action of U_p, and that the eigenvalue is λ_p.

Definition 8.1. If $f \in S_k(\Gamma_1(N) \cap \Gamma_0(p), \mathbb{C}_p)$ and $U_p(f) = \lambda_p f$, we define the *slope* of f to be the p-adic valuation of λ_p:

$$\mathrm{slope}(f) = \mathrm{ord}_p(\lambda_p),$$

where the p-adic valuation ord_p is normalized by $\mathrm{ord}_p(p) = 1$.

The reason for the name "slope" is the following. We have an operator U_p acting on a finite-dimensional vector space $S_k(\Gamma_1(N) \cap \Gamma_0(p), \mathbb{C}_p)$, and so we can compute its characteristic polynomial $P_k(t) = \det(1 - tU_p)$. We have $P_k(t) \in \mathbb{Z}_p[t]$, because the U_p operator (like all the other Hecke operators) is in fact rationally defined. We can construct, in the usual way, the *Newton polygon* of this polynomial. (This is the lower convex hull of the points $(i, \mathrm{ord}_p(c_i))$, where c_i is the i-th coefficient of $P_k(t)$; see, for example, [**59**] for more on Newton polygons.) The slopes of the eigenforms in $S_k(\Gamma_1(N) \cap \Gamma_0(p), \mathbb{C}_p)$ are exactly the slopes of the line segments making up the polygon, and the length of (the projection on the x-axis of) the segments gives the number of times each slope occurs among the eigenforms in this space.

We have considered, in the previous lecture, the situation in which $\mathrm{slope}(f) = 0$, i.e., in which the eigenvalue λ_p is a p-adic unit, and we called such eigenforms *ordinary.* As we pointed out, the Galois representation attached to such a form is also (I_p-)ordinary in the representation-theoretic sense, i.e., the representation space contains a one-dimensional direct summand which is fixed under the image of the inertia group at p.

When f has non-zero slope, it is far less clear how the slope may be understood in terms of the representation. In fact, as we'll soon see, there are almost always forms of different slope which produce the same Galois representation.

On the other hand, the slope plays a very important role in the theory of p-adic modular forms, especially in the case of "overconvergent" p-adic modular forms. A first example of this, mentioned above, is Hida's control theorem (Theorem 7.4), which shows that ordinary p-adic modular forms of low weight are automatically classical.

p-old and p-new

To understand a little better the slopes of modular forms in $S_k(\Gamma_1(N)\cap\Gamma_0(p),\mathbb{C}_p)$, we need to introduce the idea of p-old and p-new forms, and then consider what it tells us about slopes.

The starting point is to notice that there are two inclusion maps

$$S_k(\Gamma_1(N),\mathbb{C}_p)\hookrightarrow S_k(\Gamma_1(N)\cap\Gamma_0(p),\mathbb{C}_p).$$

The first is essentially the "forgetful" map: a form that is modular under the action of the larger group $\Gamma_1(N)$ is certainly also modular under the subgroup $\Gamma_1(N)\cap\Gamma_0(p)$. This gives an inclusion

$$i_p : S_k(\Gamma_1(N),\mathbb{C}_p)\hookrightarrow S_k(\Gamma_1(N)\cap\Gamma_0(p),\mathbb{C}_p)$$

which induces the identity map on q-expansions, so that

$$(i_pf)(q)=f(q).$$

The second map is a little bit harder to describe; let's just say that there is an inclusion

$$v_p : S_k(\Gamma_1(N),\mathbb{C}_p)\hookrightarrow S_k(\Gamma_1(N)\cap\Gamma_0(p),\mathbb{C}_p)$$

which acts on q-expansions by replacing q by q^p:

$$(v_pf)(q)=f(q^p).$$

The map v_p is a standard tool in the theory of newforms, which is due to Atkin, Lehner, Miyake, Casselman, and Li; see, for example, [**108**] for more details.

The subspace of $S_k(\Gamma_1(N)\cap\Gamma_0(p),\mathbb{C}_p)$ spanned by the images of i_p and v_p is called the *p-old subspace*, and the eigenforms in this space are called *p-old eigenforms*. There is a natural inner product on $S_k(\Gamma_1(N)\cap\Gamma_0(p),\mathbb{C}_p)$, and using that inner product we can define the *p-new subspace* as the orthogonal complement of the p-old subspace. Eigenforms that belong to the p-new subspace are called *p-new eigenforms*, or sometimes just *p-newforms*. It is known that every eigenform is either in the p-old space or in the p-new space.

Our goal here is to understand what this structure can tell us about the slopes of eigenforms. For the p-new part, the answer turns out to be very simple:

Theorem 8.1. *The slope of a p-new eigenform of weight k on $\Gamma_1(N)\cap\Gamma_0(p)$ is always $(k-2)/2$. More specifically, if ϵ is the nebentypus character of f and a_p is the eigenvalue for U_p, we have $a_p^2=\epsilon(p)p^{k-2}$.*

The proof can be found in the standard accounts of the theory of newforms; for example, see [**91**, Theorem 3]. An important consequence for the application we want to make is the contrapositive: if the slope of f is not $(k-2)/2$, then f must be p-old.

From our point of view, this tells us that all p-newforms have the same slope: $(k-2)/2$. The situation for p-oldforms is very different, and in some ways much more interesting.

To understand what happens, consider an eigenform $f\in S_k(\Gamma_1(N),\mathbb{C}_p)$, and let a_p be the eigenvalue of f under the action of the p-th Hecke operator T_p. As above, f has two images, i_pf and v_pf, in the bigger space $S_k(\Gamma_1(N)\cap\Gamma_0(p),\mathbb{C}_p)$. If $\ell\neq p$, both i_pf and v_pf are also eigenforms under the action of the ℓ-th Hecke operator, and the eigenvalue is the same as the one for f. What we need to analyze, then, is the action of U_p. This is easy to work out. As before, let ϵ be the nebentypus character of f (a Dirichlet character modulo N, therefore). Then we have

$$\begin{aligned} U_p(i_p f) &= a_p i_p f - \epsilon(p)p^{k-1} v_p f \\ U_p(v_p f) &= i_p f \end{aligned}$$

In other words, the two-dimensional subspace spanned by $i_p f$ and $v_p f$ is stable under U_p, which acts as the matrix

$$\begin{pmatrix} a_p & 1 \\ -\epsilon(p)p^{k-1} & 0 \end{pmatrix}.$$

The characteristic polynomial of U_p is then $t^2 - a_p t + \epsilon(p)p^{k-1}$.

Suppose that there are two distinct roots of this polynomial, λ_1 and λ_2. Then we can construct two eigenforms

$$\begin{aligned} f_1 &= i_p f - \lambda_2 v_p f \\ f_2 &= i_p f - \lambda_1 v_p f, \end{aligned}$$

And then we will have $U_p f_1 = \lambda_1 f_1$ and $U_p f_2 = \lambda_2 f_2$. Notice that f_1 and f_2 are still eigenforms for all the other Hecke operators, with the same eigenvalues as f. In this situation, we call f_1 and f_2 *twin eigenforms*.

What can we say about the slopes of f_1 and f_2? Well, we know that $\lambda_1\lambda_2 = \epsilon(p)p^{k-1}$, so we know that

$$\operatorname{slope}(f_1) + \operatorname{slope}(f_2) = k - 1.$$

We can, and do, pick the indices so that $\operatorname{slope}(f_1) \leq \operatorname{slope}(f_2)$. In addition, we know that $\lambda_1 + \lambda_2 = a_p$, from which we can conclude that one of two things happen:

- if $\operatorname{ord}_p(a_p) < (k-1)/2$, then we have $\operatorname{slope}(f_1) = \operatorname{ord}_p(a_p)$ and $\operatorname{slope}(f_2) = k - 1 - \operatorname{ord}_p(a_p)$;
- if $\operatorname{ord}_p(a_p) \geq (k-1)/2$, then we have $\operatorname{slope}(f_1) = \operatorname{slope}(f_2) = (k-1)/2$.

In particular, we have

$$0 \leq \operatorname{slope}(f_1) \leq \operatorname{slope}(f_2) \leq k - 1.$$

All of this depends on assuming that $\lambda_1 \neq \lambda_2$. If there were only one eigenvalue, then U_p would not be diagonalizable on the two-dimensional subspace spanned by $i_p f$ and $v_p f$. Notice that this would imply that $a_p = 2\lambda_1$ and hence $\operatorname{ord}_p(a_p) = (k-1)/2$ (unless $p = 2$, in which case we would have $\operatorname{ord}_p(a_p) = (k+1)/2$). The conjecture is that this cannot happen:

Conjecture (Ulmer)**.** The action of U_p on $S_k(\Gamma_1(N) \cap \Gamma_0(p), \mathbb{C}_p)$ is semisimple. In particular, we always have $\lambda_1 \neq \lambda_2$.

This has been proved for the case of forms of weight $k = 2$ on $\Gamma_0(Np)$ by Coleman and Edixhoven in [**30**]. Under the assumption that the Tate Conjecture is true, they show that in fact it is true for all weights $k \geq 2$. Ulmer has shown a different conditional result in [**153**]: the case $k = 3$ follows from the conjecture of Birch and Swinnerton-Dyer for elliptic curves over function fields.

Experimentally, the conjecture certainly seems to hold. In computations for the case $N = 1$, small primes $p \leq 100$ and small weights $k \leq 100$, we always find that

$$\operatorname{ord}_p(a_p) < \frac{k-1}{2}.$$

In fact, we "almost always" find that a much stronger inequality holds, namely that

$$\operatorname{ord}_p(a_p) < \frac{k-1}{p+1},$$

where by "almost always" we mean that the number of counterexamples seems to be quite small. We will discuss these computations in a forthcoming paper.

Every p-old eigenform must arise in this way from an eigenform in level N. The upshot, then, is that p-oldforms usually (if we assume Ulmer's conjecture, always) come in twin pairs f_1 and f_2 whose slopes add up to $k-1$. These twin forms have the same eigenvalues under all the Hecke operators except for U_p. In particular, since the Galois representation is determined by the eigenvalues of the T_ℓ with $\ell \nmid Np$, the Galois representations attached to f_1 and to f_2 (or, for that matter, to f) are exactly the same. It is this fact, that a single representation can "come from" forms of different slopes, that will fuel the construction of the "infinite fern" in the deformation space.

p-adic families of modular forms

The final ingredient in our witches' brew is Coleman's theorem on p-adic families of modular forms. To state a version of this theorem, suppose that we start with an eigenform $f \in S_{k_0}(\Gamma_1(N) \cap \Gamma_0(p), \mathbb{Z}_p)$ (note that now we are asking for coefficients in $\mathbb{Z}_p$!) and suppose that the slope of f is not equal to $k_0 - 1$. Coleman has proved (see [**26**] and [**28**]) that any such eigenform fits into a one-parameter p-adic analytic family of overconvergent p-adic modular forms with Fourier coefficients in $\mathbb{Z}_p$ which are eigenforms for T_ℓ for all $\ell \neq p$ and for U_p, have constant slope α, are all congruent modulo p, and where the "one parameter" is given by the weight. Furthermore, he has shown in [**26**] that the forms in this family corresponding to weights that are rational integers bigger than $\alpha + 1$ are classical modular forms; this, of course, is Coleman's generalization of Hida's "control theorem" for ordinary modular forms. (For an expository formulation of some of these results of Coleman, see [**100**].)

One can think of this analytic family as a family of q-expansions

$$f_k = q + a_2(k)q^2 + a_3(k)q^3 + \dots,$$

where each of the $a_n(k)$ is an analytic function of k and where specialization to the original weight k_0 gives the form from which we started. Since each f_k is an (p-adic) eigenform, there is an associated representation ρ_k; since the f_k are all congruent modulo p, these ρ_k are all deformations of the residual representation attached to our original form. Hence, Coleman's theorem gives us a "p-adic analytic curve" in the deformation space, consisting of representations all of which are attached to forms of slope α.

Infinite ferns

We now focus on the special case we want to study more closely. Let $N = 1$ and assume $\overline{\rho}$ is an absolutely irreducible representation coming from a (classical) modular form f of weight k and level 1 with coefficients in $\mathbb{Z}_p$. As above, we can think of f as an oldform on $\Gamma_0(p)$. Take $S = \{p, \infty\}$, and $\Lambda = \mathbb{Z}_p$. Let $\mathcal{R}(\overline{\rho})$ be the universal deformation ring and (as before) let $\mathcal{R}_m(\bar{f})$ be the completion of the big p-adic Hecke algebra at the maximal ideal corresponding to f. **We assume**

that the deformation problem is unobstructed, so that in particular we have $\mathcal{R}(\overline{\rho}) \cong \mathbb{Z}_p[[T_1, T_2, T_3]]$. Let X be the universal deformation space, which is just the "cube" $p\mathbb{Z}_p \times p\mathbb{Z}_p \times p\mathbb{Z}_p$, thought of as a p-adic analytic space.[1] Let X_0 be the "Sen null subspace" which we discussed in the previous lecture, i.e., the subspace of X corresponding to representations one of whose Sen weights is zero. As we pointed out above, X_0 is a analytic subspace of X of dimension two.

Let B^1 be the rigid-analytic closed unit ball over $\mathbb{Q}_p$, so that $B^1(\mathbb{Q}_p) = \mathbb{Z}_p$. Here is the version of Coleman's theorem which we will use:

Theorem 8.2 (Coleman). *Let f be an eigenform of level p and weight κ_0 and let $x \in X(\mathbb{Q}_p)$ be the point such that the representation ρ_x is attached to f. Suppose the p-adic valuation of the eigenvalue of* U_p *acting on f is not equal to $\kappa_0 - 1$. Then there exists an open neighborhood $D \subset B^1$ of $\kappa_0 \in \mathbb{Z} \subset \mathbb{Z}_p = B^1(\mathbb{Q}_p)$, and a p-adic analytic mapping*

$$z \colon D \longrightarrow X_0 \subset X$$

of D to the subspace X_0 of the p-adic analytic manifold X, and a p-adic analytic function

$$u \colon D \longrightarrow B^1,$$

such that, for an arithmetic progression of (positive, rational) integers $\mathcal{K} \subset D$ which is topologically dense in D, the image of each $\kappa \in \mathcal{K}$ under the mapping z is a point $z(\kappa)$ whose associated representation is the representation attached to a modular eigenform f_κ of level p, weight κ, and U_p*-eigenvalue equal to $u(\kappa)$. Finally, $f_{\kappa_0} = f$.*

The crucial point is that the family depends not only on the representation we start with, but also on the *slope* of the modular form attached to that representation. As a result, when the Galois representation ρ is attached (in the above sense) to a pair of "twins," it follows that there exist *two distinct* one-parameter families of deformations of ρ. This is what allows us to construct the structures Mazur and I call "infinite ferns" in the deformation space of ρ.

Let $\Gamma \subset \mathbb{Z}_p^\times$ denote the group of 1-units in $\mathbb{Z}_p$, i.e., the multiplicative group of p-adic integers congruent to 1 mod p. Twisting the representation ρ_x corresponding to a point $x \in X(\mathbb{Q}_p)$ by a one-dimensional "wild" character $\psi : G_{\mathbb{Q},\{p\}} \longrightarrow \Gamma$,

$$\rho_x \mapsto \psi \otimes \rho_x,$$

induces a p-adic analytic action of the group of wild characters (i.e., the formal group in one parameter, call it $\Psi = \mathrm{Hom}_{\mathrm{cont}}(G_{\mathbb{Q},\{p\}}, \Gamma)$) on the p-adic manifold X. For a very brief discussion of this action, see §5 of [**100**]. Let us denote the point of X which corresponds to the representation $\psi \otimes \rho_x$ by $\psi \circ x$.

As noted above, X contains the p-adic analytic "surface" X_0. This space is "essentially transversal" to the action of Ψ on X, in the sense that the natural mapping

$$\begin{aligned} \pi \colon X_0 \times \Psi &\longrightarrow X \\ (x_0, \psi) &\mapsto \psi \circ x_0 \end{aligned}$$

[1] Well, really $X = \mathrm{Spf}(\mathcal{R}(\overline{\rho}))^{\mathrm{rig}}$ is the open unit rigid-analytic 3-ball over $\mathbb{Q}_p$. The "cube" is actually $X(\mathbb{Q}_p)$, i.e., it consists of the points in X that are defined over $\mathbb{Q}_p$. All the points in the "infinite fern" we are about to construct *are* in fact defined over $\mathbb{Q}_p$, so the mental image of the "cube" is not too misleading.

has fibers of cardinality ≤ 2 (in fact, π is the restriction to $X_0 \times \Psi$ of a mapping of degree 2 on analytic spaces) and π is unramified off the locus $X_{00} \times \Psi \subset X_0 \times \Psi$, where X_{00} is the analytic subset of X whose $\mathbb{Q}_p$-points x are those for which the Hodge-Tate-Sen weights of the associated representation ρ_x are $\{0, 0\}$ (cf. the main proposition of §8 in [**100**]).

We need one further assumption before we can proceed:

> **We assume the slope of the modular form f is not equal to 0 or to $k - 1$.**

Note that forms of slope 0 always come in twin pairs with forms of slope $k - 1$, so that it makes sense to assume both of those together. We call forms whose slope is 0 or $k-1$ "forms of critical slope;" thus, our assumption is that f is of non-critical slope.[2]

Given that whole setup, we do the following.

- To begin with, we have a modular residual representation whose associated modular form $f = f_0$ is of weight $k = k_0$ and slope α_0 not equal to 0 or $k_0 - 1$.
- Use Coleman's theorem to produce a curve of deformations containing f_0, each attached to a modular form of some weight and slope α_0. (This might be the curve C in the picture.)
- If the initial form is not a newform, then it has a twin, and we can construct a curve corresponding to that twin. It goes through the same initial point, since twin forms give the same representation, but it corresponds to forms of weight $k_0 - 1 - \alpha_0$. (If the first curve is C, this is $C^{(\kappa)}$.)
- "Move" along either of the families to a classical form f_1 of weight k_1 and slope α_0. Make sure $k_1 > \alpha_0 + 1$, $k_1 \neq 2\alpha_0 + 1$, and $k_1 \neq 2\alpha_0 + 2$. Notice that there are infinitely many integers k_1 with these properties. In fact, the set of such integers is dense in $\mathbb{Z}_p$.
- The last inequality means that f_1 is p-old, so it has a twin f_1 of weight k_1 and slope $\alpha_1 = k_1 - 1 - \alpha_0$. Notice that $\alpha_1 \neq 0, k_1 - 1, \alpha_0$. (This is where we need to know that the initial form did not have critical slope.)
- Now repeat the process starting from f_1. In fact, repeat this at all points that satisfy the weight-slope constraints.

This produces an amazingly complex structure, which we call an "infinite fern," made up of infinitely many analytic arcs, all contained in X_0. (See figure 1 and the discussion in §18 of [**100**].) In the diagram, each curve segment corresponds to a modular arc in X_0, which is an embedded p-adic analytic image of a disk in $\mathbb{Z}_p$. For a topologically dense set of the points κ on any given modular arc C there is another arc $C^{(\kappa)}$ crossing C at κ. More pictorially, calling any given modular arc a "spine" and calling the modular arcs crossing it "needles," we have that each "spine" has a topologically dense set of "needles."

In fact, not only do we get an infinite fern growing around our initial point, but also we see that every modular form satisfying our constraints which is congruent to f corresponds to a point in X_0 which has a topological neighborhood containing the image of an infinite fern. This structure "fills up" the subvariety X_0, and we can "thicken it" by considering all possible twists of the whole structure. This

[2] This is actually a minor assumption: if our initial form happens to have slope 0, we can replace it by another form of non-zero slope which produces the same residual representation, at the cost of enlarging the base ring over which we are working, using the trick given on page 111 of [**57**] (but note that this forces us to move from $\mathbb{Z}_p$ to some ramified extension).

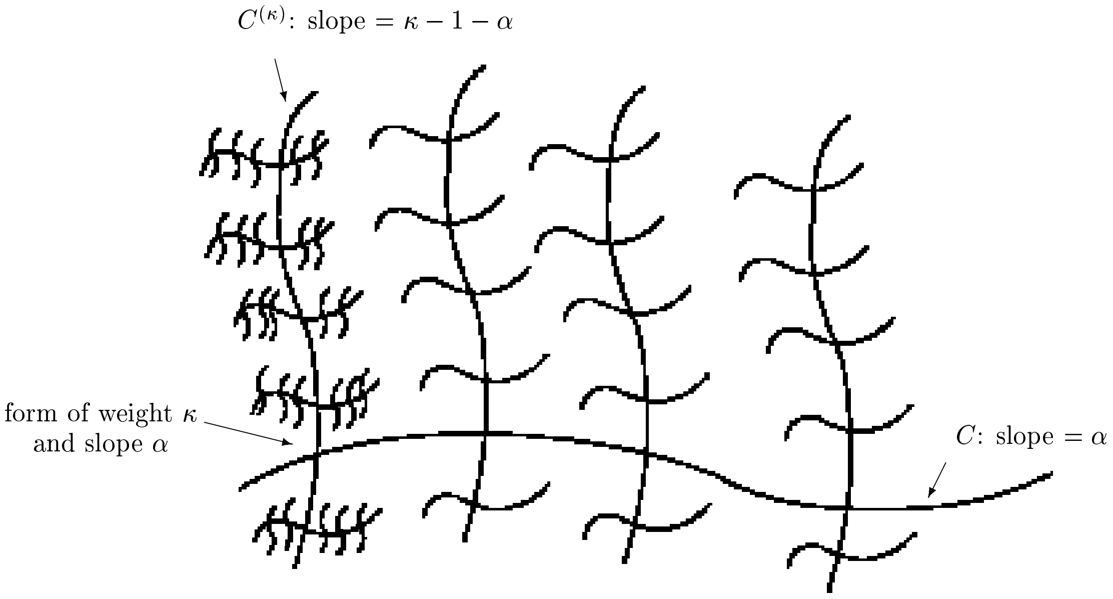

$C^{(\kappa)}$: slope = $\kappa - 1 - \alpha$
form of weight κ and slope α
C: slope = α

Figure 1. An infinite fern

gives a structure that is sufficiently "big" to prove that the modular points must be Zariski-dense in the full deformation space. Using this, we can prove the following result:

Proposition 8.3. *Suppose $\overline{\rho}$ is absolutely irreducible, unramified outside p and infinity, and attached to a (classical) modular form on $\Gamma_0(p)$ of non-critical slope and with Fourier coefficients in $\mathbb{Z}_p$. Suppose also that the deformation problem associated to $\overline{\rho}$ is unobstructed. Then the map $\mathcal{R}(\overline{\rho}) \longrightarrow \mathcal{R}(\bar{f})$ given by the deformation theory is an isomorphism.*

It is possible that a similar approach will work whenever we know the dimension (perhaps even the local dimension?) of the deformation space.

In the case where the residual representation comes from an ordinary modular form, another approach to theorems of this sort seems plausible: one can use Wiles' methods to show that all ordinary deformations are modular, and then use Böckle's results (in [**5**]) on the relation between the ordinary deformation space and the full deformation space to compare the full deformation space to the modular deformation space. Böckle has recently announced a result along these lines.

APPENDIX 1
A criterion for existence of a universal deformation ring by Mark Dickinson

The purpose of this note is to give an elementary proof of the existence of the universal deformation ring, using a representability criterion of Grothendieck. In particular the proof makes no use of Schlessinger's criteria or of noetherian hypotheses. I would like to thank Brian Conrad for suggesting that this be written up and Fernando Gouvêa for allowing me to include it here. I would also like to thank Sam Williams for helpful comments.

Let k be a field and let Λ be a topological ring which is the inverse limit of a system of artinian local rings, each with the discrete topology and with residue field k. (For example let k be finite and Λ the ring of Witt vectors of k.) We define two full subcategories of the category of topological Λ-algebras R. Let $\mathscr{C}_\Lambda^{\mathrm{fl}}$ be the full subcategory whose objects are discrete finite-length local Λ-algebras R for which the structure map induces an isomorphism on residue fields, and define $\mathscr{C}_\Lambda$ to be the full subcategory whose objects are those arising as an inverse limit of objects of $\mathscr{C}_\Lambda^{\mathrm{fl}}$. (The objects of $\mathscr{C}_\Lambda$ are examples of 'pseudo-compact' rings; for basic properties of these rings see section 0 of [**56**].)

Now let G be a profinite group, d a positive integer and $\bar{\rho}\colon G \to \mathrm{GL}_d(k)$ a continuous representation of G. Assume that the only matrices in $\mathrm{M}_d(k)$ which commute with every element of the image of $\bar{\rho}$ are the scalar matrices.

Definition 9.1. A *lifting* of $\bar{\rho}$ (to R) is a pair (R, ρ) consisting of an object R of $\mathscr{C}_\Lambda$ together with a continuous representation $\rho\colon G \to \mathrm{GL}_d(R)$ whose pushforward by the natural reduction map $R \to k$ is conjugate to $\bar{\rho}$. Two liftings (R, ρ) and (R, σ) of $\bar{\rho}$ to the same ring are *conjugate* if the representations ρ and σ are conjugate. A *deformation* of $\bar{\rho}$ to R is a conjugacy class of liftings of $\bar{\rho}$ to R; the notation (R, ρ) will also be used for the deformation represented by a lifting (R, ρ).

Note that, if we assume $C(\bar{\rho}) = k$, the conjugating matrix relating two (conjugate) representations (R, ρ) and (R, σ) lifting $\bar{\rho}$ can be chosen to be congruent to the identity modulo the maximal ideal of R.

If (R, ρ) is a deformation of $\bar{\rho}$ and $\phi\colon R \to S$ is a morphism of $\mathscr{C}_\Lambda$ then the pushforward $(S, \phi_*\rho)$ of (R, ρ) by ϕ is a deformation of $\bar{\rho}$ to S. Thus there is a well-defined functor

$$\mathrm{Def}\colon \mathscr{C}_\Lambda \to \mathbf{Sets}$$

which sends an object R of $\mathscr{C}_\Lambda$ to the set of deformations of $\bar\rho$ to R. Suppose that certain deformations of $\bar\rho$ are designated 'of type $\mathscr{P}$', and that the pushforward of a deformation (R,ρ) of type $\mathscr{P}$ by a morphism $\phi\colon R \to S$ is again of type $\mathscr{P}$. Then one can define a subfunctor

$$\mathrm{Def}_{\mathscr{P}}\colon \mathscr{C}_\Lambda \to \mathbf{Sets}$$

of Def which sends an object R to the set of deformations of $\bar\rho$ to R of type $\mathscr{P}$.

Definition 9.2. We define a *universal deformation* of $\bar\rho$ of type $\mathscr{P}$ to be a deformation $(R^{\mathrm{univ}}_{\mathscr{P}}, \rho^{\mathrm{univ}}_{\mathscr{P}})$ of $\bar\rho$ of type $\mathscr{P}$ such that for any given deformation (R,ρ) of $\bar\rho$ of type $\mathscr{P}$ there is a unique morphism $\phi\colon R^{\mathrm{univ}}_{\mathscr{P}} \to R$ for which the pushforward of $(R^{\mathrm{univ}}_{\mathscr{P}}, \rho^{\mathrm{univ}}_{\mathscr{P}})$ by ϕ is equal to (R,ρ).

By the Yoneda Lemma, to give such a universal deformation is equivalent to giving an object $R^{\mathrm{univ}}_{\mathscr{P}}$ of $\mathscr{C}_\Lambda$ along with an isomorphism

$$\mathrm{Hom}_{\mathscr{C}_\Lambda}(R^{\mathrm{univ}}_{\mathscr{P}}, -) \cong \mathrm{Def}_{\mathscr{P}}$$

of functors; thus a universal deformation of $\bar\rho$ of type $\mathscr{P}$ exists if and only if the functor $\mathrm{Def}_{\mathscr{P}}$ is representable. We call the object $R^{\mathrm{univ}}_{\mathscr{P}}$ a *universal deformation ring* for deformations of $\bar\rho$ of type $\mathscr{P}$. The following theorem tells us when we can expect a universal deformation of type $\mathscr{P}$ to exist.

Theorem 9.1. *The following three conditions are necessary and sufficient for the existence of a universal deformation of $\bar\rho$ of type $\mathscr{P}$:*

- *the trivial deformation $(k,\bar\rho)$ of $\bar\rho$ is of type $\mathscr{P}$,*
- *given a diagram $R \xrightarrow{\phi} T \xleftarrow{\psi} S$ in $\mathscr{C}^{\mathrm{fl}}_\Lambda$, any deformation of $\bar\rho$ to the fiber product $R \times_T S$ whose pushforwards to R and to S are both of type $\mathscr{P}$ is itself of type $\mathscr{P}$, and*
- *if R in $\mathscr{C}_\Lambda$ is a filtered limit of objects $(R_i)_{i\in\mathscr{I}}$ of $\mathscr{C}^{\mathrm{fl}}_\Lambda$ and the pushforward of a deformation (R,ρ) by the natural reduction map $R \to R_i$ is of type $\mathscr{P}$ for each i, then (R,ρ) is of type $\mathscr{P}$.*

Note especially that in the case where *every* deformation of $\bar\rho$ is of type $\mathscr{P}$ the conditions of the theorem are trivially satisfied and so a universal deformation exists.

To prove the theorem, we first give Grothendieck's criterion for a set-valued covariant functor on $\mathscr{C}_\Lambda$ to be representable. Recall that a functor is called left exact if it is compatible with the formation of finite limits and that being left exact is equivalent to taking terminal objects (resp., fiber products) to terminal objects (resp., fiber products).

Proposition 9.2. *A functor $X\colon \mathscr{C}_\Lambda \to$ **Sets** is representable if and only if the restriction of X to $\mathscr{C}^{\mathrm{fl}}_\Lambda$ is left exact and X preserves filtered limits, taken in $\mathscr{C}_\Lambda$, of objects of $\mathscr{C}^{\mathrm{fl}}_\Lambda$.*

Proof. First note that the functor which sends an object R of $\mathscr{C}_\Lambda$ to the system of finite-length discrete quotients of R gives an equivalence of categories between the category $\mathscr{C}_\Lambda$ and the category of pro-objects of $\mathscr{C}^{\mathrm{fl}}_\Lambda$, as defined in section A2 of [**67**]. By the corollary to Proposition 3.1 of section A of [**67**] there is an object R of $\mathscr{C}_\Lambda$ and an isomorphism $\mathrm{Hom}_{\mathscr{C}_\Lambda}(R,-) \cong X$ of functors on $\mathscr{C}^{\mathrm{fl}}_\Lambda$ if and only if the restriction of X to $\mathscr{C}^{\mathrm{fl}}_\Lambda$ is left exact. This isomorphism extends to an isomorphism of functors on $\mathscr{C}_\Lambda$ if and only if X preserves filtered limits of objects of $\mathscr{C}^{\mathrm{fl}}_\Lambda$. □

So in order to prove Theorem 9.1 it is enough to check that the functor $\mathrm{Def}_{\mathscr{P}}$ satisfies the hypotheses of Proposition 9.2. We first prove that the full deformation functor $\mathrm{Def}\colon \mathscr{C}_\Lambda \to \mathbf{Sets}$ satisfies these hypotheses. We begin with the following easy consequence of the fact that the centralizer of the image of $\bar\rho$ is the scalar matrices.

Lemma 9.3. *Let W be a subspace of a finite-dimensional k-vector space V, and B a $d\times d$ matrix with entries in V. If $B\bar\rho(g)-\bar\rho(g)B$ has entries in W for all g in G, then $B = vI + C$ for some element v of V and a matrix C all of whose entries lie in W.*

Proof. Take a basis $\{e_1,\dots,e_r\}$ of W over k and extend it to a basis $\{e_1,\dots,e_s\}$ of V. Write B as $\sum_{i=1}^{s} B_i e_i$ where each B_i is an element of $\mathrm{M}_d(k)$. Then for $r < i \le s$ the matrix B_i commutes with each element of the image of $\bar\rho$ and hence is a scalar matrix. □

Lemma 9.4. *Suppose that R is an object of $\mathscr{C}_\Lambda^{\mathrm{fl}}$ and (R,ρ) is a lifting of $\bar\rho$. Then any element of $\mathrm{M}_d(R)$ which centralizes the image of ρ is a scalar matrix.*

Proof. Since R is a local artinian ring the nth power of the maximal ideal $\mathfrak{m}_R$ of R is trivial for some $n \ge 1$; we prove the result by induction on n. For $n = 1$ the result is immediate, since we assumed that only the scalar matrices in $\mathrm{M}_d(k)$ centralize the image of $\bar\rho$. Suppose that $n > 1$ and that A is an element of $\mathrm{M}_d(R)$ such that $A\rho = \rho A$. By the induction hypothesis the reduction of A modulo $\mathfrak{m}_R^{n-1}$ is a scalar matrix and we can write $A = \lambda I + B$ for some element λ of R and some matrix B with entries in the finite-dimensional k-vector space $\mathfrak{m}_R^{n-1}$. Then $(\lambda I + B)\rho = \rho(\lambda I + B)$ and so $B\rho - \rho B = 0$. Since B has entries killed by $\mathfrak{m}_R$, we can rewrite this equation as $B\bar\rho - \bar\rho B = 0$. So B is a scalar matrix by Lemma 9.3, hence so is $A = \lambda I + B$. □

Lemma 9.5. *The functor*

$$\mathrm{Def}\colon \mathscr{C}_\Lambda \to \mathbf{Sets}$$

which sends an object R of $\mathscr{C}_\Lambda$ to the set of deformations of $\bar\rho$ over R is representable.

Proof. By Proposition 9.2 we need to show that Def preserves filtered limits of objects of $\mathscr{C}_\Lambda^{\mathrm{fl}}$ and that Def restricted to $\mathscr{C}_\Lambda^{\mathrm{fl}}$ is left exact. To show the former requires, not surprisingly, an application of Zorn's Lemma. Let R be the filtered limit of a system $(R_i)_{i\in\mathscr{I}}$ of objects and maps of $\mathscr{C}_\Lambda^{\mathrm{fl}}$ indexed by a set $\mathscr{I}$ and suppose that for each $i \in \mathscr{I}$ we are given a deformation (R_i,ρ_i) of $\bar\rho$ and that for every morphism $i \to j$ of $\mathscr{I}$ the pushforward of (R_i,ρ_i) by the corresponding map $R_i \to R_j$ is equal to (R_j,ρ_j). We must show that there is a unique deformation (R,ρ) whose pushforward by each natural projection map $R \to R_i$ is equal to (R_i,ρ_i). For each object i of $\mathscr{I}$ let S_i be the set of liftings of $\bar\rho$ to R_i which represent (R_i,ρ_i) and which reduce to $\bar\rho$, and consider the system of subsets of S_i consisting of the orbits of subgroups of $\mathrm{GL}_d(R_i)$ of the form $I+\mathrm{M}_d(J)$ for some proper ideal J of R_i. Using these subsets we can apply Theorem 1 of section 7.4 of Chapter 3 of [**12**] to deduce that the inverse limit over i in $\mathscr{I}$ of the sets S_i is non-empty, hence that there is a deformation (R,ρ) as desired. To show that this deformation is unique, suppose that (R,σ) is another such and that for every object i of $\mathscr{I}$ there is a matrix A_i, unique up to scalar multiplication by Lemma 9.4, which conjugates the pushforward to R_i of ρ to the pushforward of σ. If we assume that both ρ and σ

reduce to $\bar{\rho}$ then it follows from the triviality of the centralizer of the image of $\bar{\rho}$ that each A_i reduces to a scalar matrix. Thus we may assume that the top left entry of A_i is equal to 1 for each i in $\mathscr{I}$; then the A_i form a compatible system of matrices and so give a matrix A with entries in R which conjugates ρ to σ.

To check that Def is left exact on $\mathscr{C}_\Lambda^{\mathrm{fl}}$ it suffices to check that it preserves fiber products and the terminal object. The only deformation of $\bar{\rho}$ to k is $(k, \bar{\rho})$ itself, so $\mathrm{Def}(k)$ is a one point set and Def preserves the terminal object. It remains to show that Def preserves fiber products in $\mathscr{C}_\Lambda^{\mathrm{fl}}$. Suppose that we have a fiber product

$$\begin{tikzcd}
R \times_T S \arrow[r] \arrow[d] & R \arrow[d, "\phi"] \\
S \arrow[r, "\psi"] & T
\end{tikzcd}$$

of objects of $\mathscr{C}_\Lambda^{\mathrm{fl}}$, and suppose that (R, ρ) and (S, σ) are liftings of $\bar{\rho}$ whose pushforwards $(T, \phi_*\rho)$ and $(T, \psi_*\sigma)$ are conjugate. We will show that it is possible to replace (R, ρ) and (S, σ) with conjugate liftings whose pushforwards to T are identical. Then we obtain a lifting $(R \times_T S, \pi)$ of $\bar{\rho}$ whose pushforwards to R and S are conjugate to (R, ρ) and (S, σ) respectively; a similar argument to the one above for filtered limits shows that if this lifting exists then it is unique up to conjugation.

We suppose that the nth power of the maximal ideal $\mathfrak{m}_T$ of T is zero and prove the existence of π as above by induction on n. We may assume that ρ and σ each reduce to $\bar{\rho}$ (rather than just to a conjugate of $\bar{\rho}$), and in the case $n = 1$ there is nothing more to do; now suppose that $n > 1$ and that

$$\phi_*\rho = \psi_*\sigma \quad \text{modulo } \mathfrak{m}_T^{n-1}.$$

By assumption there is a matrix C in $\mathrm{GL}_d(T)$ such that $C\psi_*\sigma C^{-1} = \phi_*\rho$. The reduction of C modulo $\mathfrak{m}_T^{n-1}$ centralizes the image of the reduction of the representation $\psi_*\sigma$ modulo $\mathfrak{m}_T^{n-1}$ so is a scalar matrix by Lemma 9.4. So without loss of generality we may assume that $C = I + L$ for some matrix L with entries in $\mathfrak{m}_T^{n-1}$. Now

$$(I + L)\psi_*\sigma(I - L) = \phi_*\rho$$

and since L is annihilated by $\mathfrak{m}_T$ we can rewrite this as

$$L\bar{\rho} - \bar{\rho}L = \phi_*\rho - \psi_*\sigma.$$

From Lemma 9.3 it follows that $L = \lambda I + \phi(M) - \psi(N)$ for some element λ of $\mathfrak{m}_T^{n-1}$ and for some matrices M and N with entries in R and S respectively; by replacing M and N by $M - P$ and $N + P$ for some P in $\mathrm{M}_d(\Lambda)$ lifting the reduction of M to $\mathrm{M}_d(k)$ we may assume that M and N have entries in the maximal ideals $\mathfrak{m}_R$ and $\mathfrak{m}_S$ of R and S. Then $\phi(M)(\phi(M) - \psi(N)) = 0$ and it follows that $I + L = (1 + \lambda)(I - \phi(M))^{-1}(I - \psi(N))$ and hence that

$$\psi_*((I - N)\sigma(I - N)^{-1}) = \phi_*((I - M)\rho(I - M)^{-1})$$

thus giving conjugates of ρ and σ whose pushforwards to T are identical, as required. □

Proof of Theorem 9.1. The proof of Theorem 9.1 now follows easily: Def is left exact and preserves filtered limits by Lemma 9.5 and Proposition 9.2. Now the first two conditions in the statement of the theorem assert that the subfunctor $\mathrm{Def}_{\mathscr{P}}$ of Def is left exact on $\mathscr{C}_\Lambda^{\mathrm{fl}}$ and the third that $\mathrm{Def}_{\mathscr{P}}$ preserves filtered limits of objects of $\mathscr{C}_\Lambda^{\mathrm{fl}}$. So by Proposition 9.2 again the functor $\mathrm{Def}_{\mathscr{P}}$ is representable (equivalently, a universal deformation of $\bar{\rho}$ of type $\mathscr{P}$ exists) if and only if the conditions of the theorem are satisfied. □

To end we note that Grothendieck's results also yield a criterion for the universal deformation ring to be noetherian. Let $k[\varepsilon]$ denote the object $k[X]/(X^2)$ of $\mathscr{C}_\Lambda^{\mathrm{fl}}$ (where ε corresponds to the image of X). It is a k-vector space object of the category $\mathscr{C}_\Lambda^{\mathrm{fl}}$, with addition $k[\varepsilon] \times_k k[\varepsilon] \to k[\varepsilon]$ defined by sending $(a\varepsilon, b\varepsilon)$ to $(a+b)\varepsilon$ and scalar multiplication by a defined by sending $b\varepsilon$ to $ab\varepsilon$. For any left exact functor $X\colon \mathscr{C}_\Lambda^{\mathrm{fl}} \to \mathbf{Sets}$, these maps provide the set $X(k[\varepsilon])$ with the structure of a k-vector space, and for any natural transformation of left-exact functors $\alpha\colon X \to Y$ the map $\alpha_{k[\varepsilon]}\colon X(k[\varepsilon]) \to Y(k[\varepsilon])$ is k-linear. In particular if the deformation functor $\mathrm{Def}_{\mathscr{P}}$ is representable then $\mathrm{Def}_{\mathscr{P}}(k[\varepsilon])$ is a k-vector space and is isomorphic to $\mathrm{Hom}_{\mathscr{C}_\Lambda}(R_{\mathscr{P}}^{\mathrm{univ}}, k[\varepsilon])$.

Proposition 9.6. *Suppose that Λ is noetherian and that the functor $\mathrm{Def}_{\mathscr{P}}$ satisfies the conditions of Theorem 9.1 and so is representable. Then the universal deformation ring $R_{\mathscr{P}}^{\mathrm{univ}}$ is noetherian if and only if the k-vector space $\mathrm{Def}_{\mathscr{P}}(k[\varepsilon])$ is finite-dimensional; furthermore, if $\mathrm{Def}_{\mathscr{P}}(k[\varepsilon])$ has dimension d then $R_{\mathscr{P}}^{\mathrm{univ}}$ is a quotient of the power-series ring $\Lambda[[X_1, \ldots, X_d]]$.*

Proof. This follows from Proposition 5.1 of section A of [**67**]. □

Mark Dickinson
Department of Mathematics
Harvard University
Cambridge, MA 02138
dickinso@math.harvard.edu

APPENDIX 2
An overview of a theorem of Flach
by Tom Weston

In recent years, the study of the deformation theory of Galois representations has become of central importance in arithmetic algebraic geometry. The most fundamental question in this field is the explicit determination of universal deformation rings associated to given residual representations. It was observed by Mazur in his first paper [**97**, Section 1.6] on the subject that the solution of this problem is immediate if a certain Galois cohomology group associated to the residual representation vanishes.

The goal of this appendix is to provide an overview of a theorem of Flach which yields the vanishing of this cohomology group for many mod l representations coming from rational elliptic curves. We do not seek to give a complete proof; we hope only to make clear the main ideas. In the process we will touch on many facets of arithmetic algebraic geometry, including Tate's duality theorems in Galois cohomology, generalized Selmer groups, Kolyvagin's theory of Euler systems and the geometry of modular curves.

The work we will describe actually has another, more direct, application: it can be used in many cases to prove the Taylor-Wiles isomorphism between a certain universal deformation ring and a certain Hecke algebra. We will not touch on this aspect; for details, see [**99**] or [**158**].

We have tried to keep the prerequisites to a minimum. The main requirement is a good familiarity with Galois cohomology. The algebraic geometry we use is mostly at the level of [**146**, Chapters 1 and 2], with the exception of Appendix B, which is significantly more advanced. Some familiarity with elliptic curves is helpful, although with the exception of Appendix A we will use little more than the Tate module and the existence of the Weil pairing.

I would like to thank Brian Conrad, Matthew Emerton and Karl Rubin for teaching me much of the material presented here. I would also like to thank Fernando Gouvêa for encouraging the writing of this paper. Above all, I would like to thank Barry Mazur for his constant help and insights; I can only hope that his point of view is visible in the mathematics below.

Unobstructed deformation problems

Let $G_{\mathbb{Q},S}$ be the maximal quotient of the absolute Galois group of $\mathbb{Q}$ unramified away from a finite set of places S. Let l be a prime number and let $\bar{\rho} : G_{\mathbb{Q},S} \to \mathrm{GL}_2(\mathbb{F}_l)$ be a Galois representation. Under certain additional hypotheses (for example, if $\bar{\rho}$ is absolutely irreducible) we can associate a *universal deformation ring* $\mathcal{R}(\bar{\rho})$ to such a residual representation; see Lecture 3 in these notes, and [**97**] or [**101**] for details.

In general the determination of the structure of the ring $\mathcal{R}(\bar{\rho})$ is quite difficult. However, in at least one case the determination is easy: let $\mathrm{ad}(\bar{\rho})$ be the $G_{\mathbb{Q},S}$-module of 2×2 matrices over $\mathbb{F}_l$ on which $\gamma \in G_{\mathbb{Q},S}$ acts via conjugation by $\bar{\rho}(\gamma)$. If

$$\mathrm{H}^2(G_{\mathbb{Q},S}, \mathrm{ad}(\bar{\rho})) = 0,$$

then $\mathcal{R}(\bar{\rho})$ is isomorphic to a power series ring in $\dim_{\mathbb{F}_l} \mathrm{H}^1(G_{\mathbb{Q},S}, \mathrm{ad}(\bar{\rho}))$ variables over $\mathbb{Z}_l$; see [**97**, Section 1.6, Proposition 2]. If this is the case, we say that the deformation problem for $\bar{\rho}$ is *unobstructed.*

Our goal in this paper is to explain the main ideas of the proof of the following theorem of Flach (this is [**50**, Theorem 2]).

Theorem 10.1. *Let E be an elliptic curve over $\mathbb{Q}$, let $l \geq 5$ be a prime and let S be the set of places of $\mathbb{Q}$ at which E has bad reduction, together with l and ∞. Let $\rho : G_{\mathbb{Q},S} \to \mathrm{GL}_2(\mathbb{Z}_l)$ be the representation of $G_{\mathbb{Q},S}$ on the l-adic Tate module of E and let $\bar{\rho} : G_{\mathbb{Q},S} \to \mathrm{GL}_2(\mathbb{F}_l)$ be the residual representation. Assume further that:*

- *E has good reduction at l;*
- *ρ is surjective;*
- *For all $p \in S - \{\infty\}$, $E[l] \otimes E[l]$ has no $G_{\mathbb{Q}_p}$-invariants;*
- *l does not divide the rational number $L(\mathrm{Sym}^2 T_l E, 0)/\Omega$, where $\mathrm{Sym}^2 T_l E$ is the symmetric square of the l-adic Tate module of E and Ω is a certain period.*

Then the deformation problem for $\bar{\rho}$ is unobstructed.

In Appendix A we discuss precisely how stringent these hypotheses are; the main result is that for fixed E which does not have complex multiplication, then they are satisfied for a set of primes l of density 1. We will explain the fourth hypothesis in Section 5.

We should note that this theorem uses in a crucial way the fact that E is modular, and thus stating it in the form above relies heavily on the recent proof of the Shimura-Taniyama conjecture.

Galois modules and the calculus of Tate twists

We begin with some formalities on Galois actions and certain commutative algebra operations. Let S be a finite set of places of $\mathbb{Q}$ including the prime l. Let M and N be $\mathbb{Z}_l$-modules with $G_{\mathbb{Q},S}$-actions. We make the tensor product $M \otimes_{\mathbb{Z}_l} N$ a $G_{\mathbb{Q},S}$-module via the diagonal action: $\gamma(m \otimes n) = \gamma m \otimes \gamma n$. We make $\mathrm{Hom}_{\mathbb{Z}_l}(M, N)$ a $G_{\mathbb{Q},S}$-module via the adjoint action: $\gamma f(m) = \gamma \cdot f(\gamma^{-1} m)$ for $f \in \mathrm{Hom}_{\mathbb{Z}_l}(M, N)$. Note that the $G_{\mathbb{Q},S}$-invariants of $\mathrm{Hom}_{\mathbb{Z}_l}(M, N)$ are precisely the $G_{\mathbb{Q},S}$-equivariant homomorphisms $\mathrm{Hom}_{\mathbb{Z}_l[G_{\mathbb{Q},S}]}(M, N)$. Throughout this paper we assume that the base ring for any of these constructions is $\mathbb{Z}_l$; we will usually omit it from the notation.

Now assume that M is free over $\mathbb{Z}_l$ of rank 2. We define the *symmetric square* $\operatorname{Sym}^2 M$ of M to be the submodule of $M \otimes M$ which is invariant under the automorphism of $M \otimes M$ interchanging the two factors. If x, y is a basis for M, then $x \otimes x, x \otimes y + y \otimes x, y \otimes y$ is a basis for $\operatorname{Sym}^2 M$, so that $\operatorname{Sym}^2 M$ is free over $\mathbb{Z}_l$ of rank 3. In fact, if $l \neq 2$, then $\operatorname{Sym}^2 M$ is a direct summand of $M \otimes M$; the complimentary summand is the alternating square $\wedge^2 M$, which has basis $x \otimes y - y \otimes x$:

$$M \otimes M = \wedge^2 M \oplus \operatorname{Sym}^2 M. \tag{1}$$

One checks easily that $\operatorname{Sym}^2 M$ is stable under the action of $G_{\mathbb{Q},S}$, so that it can also be considered as a $G_{\mathbb{Q},S}$-module. We also have an induced action of $G_{\mathbb{Q},S}$ on $\wedge^2 M$, and with these actions the decomposition (1) is a decomposition of $G_{\mathbb{Q},S}$-modules.

The module of endomorphisms $\operatorname{End}(M)$ admits a similar decomposition if $\ell \neq 2$. (As always we let $G_{\mathbb{Q},S}$ act on $\operatorname{End}(M)$ via the adjoint action.) The scalar matrices in $\operatorname{End}(M)$ are a free $\mathbb{Z}_l$-module of rank 1 with trivial Galois action, since conjugation is trivial on scalars. Thus, for $\ell \neq 2$ we have a canonical decomposition

$$\operatorname{End}(M) = \mathbb{Z}_l \oplus \operatorname{End}^0(M)$$

where $\operatorname{End}^0(M)$ denotes the trace zero matrices in $\operatorname{End}(M)$ and the first summand corresponds to the scalar matrices. (We always take $\mathbb{Z}_l$ itself to have trivial $G_{\mathbb{Q},S}$-action.)

Now let E be a rational elliptic curve (i.e., an elliptic curve defined over $\mathbb{Q}$), and let l be an odd prime. Recall that the *l-adic Tate module* of E is the free $\mathbb{Z}_l$-module of rank 2 defined by

$$T_l E = \varprojlim E[l^n].$$

If S is any set of places of $\mathbb{Q}$ including l and the places where E has bad reduction, then $T_l E$ carries a natural action of $G_{\mathbb{Q},S}$. Upon choosing a basis for $T_l E$ we can view this as a representation

$$\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_l).$$

The Galois module which will actually prove most relevant to the proof of Theorem 10.1 is the symmetric square of $T_l E$.

We can perform a similar construction with the l-power roots of unity: this Tate module, $\varprojlim \mu_{l^n}$, is written as $\mathbb{Z}_l(1)$. Thus $\mathbb{Z}_l(1)$ is a free $\mathbb{Z}_l$-module of rank 1 on which $G_{\mathbb{Q},S}$ acts via the ℓ-adic cyclotomic character

$$\varepsilon : G_{\mathbb{Q},S} \to \mathbb{Z}_l^*;$$

here S is any set of places of $\mathbb{Q}$ containing l and ∞. For any $n > 0$, define $\mathbb{Z}_l(n)$ to be the tensor product (over $\mathbb{Z}_l$) of $\mathbb{Z}_l(1)$ with itself n times. Define $\mathbb{Z}_l(-1)$ to be the integral Pontrjagin dual $\operatorname{Hom}(\mathbb{Z}_l(1), \mathbb{Z}_l)$ of $\mathbb{Z}_l(1)$, and define $\mathbb{Z}_l(-n)$ as the tensor product of $\mathbb{Z}_l(-1)$ with itself n times. Thus $\mathbb{Z}_l(n)$ is a free $\mathbb{Z}_l$-module of rank one on which $G_{\mathbb{Q},S}$ acts via ε^n. If M is any $\mathbb{Z}_l$-module with an action of $G_{\mathbb{Q}}$, we define its n^{th} *Tate twist* by

$$M(n) = M \otimes_{\mathbb{Z}_l} \mathbb{Z}_l(n).$$

$M(n)$ is isomorphic to M as a $\mathbb{Z}_l$-module, but they usually have different $G_{\mathbb{Q}}$-actions.

A key property of the Tate module of an elliptic curve is that the Weil pairings

$$E[l^n] \otimes E[l^n] \to \mu_{l^n}$$

compile to yield a perfect, skew-symmetric, Galois equivariant pairing

$$e : T_l E \otimes T_l E \to \mathbb{Z}_l(1).$$

See [**146**, Proposition III.8.3]. Since e is skew-symmetric, this implies that $\wedge^2 T_l E \cong \mathbb{Z}_l(1)$. We record some additional consequences below.

Lemma 10.2. *The Weil pairing induces a Galois equivariant isomorphism*

$$\operatorname{End}(T_l E)(1) \cong T_l E \otimes T_l E.$$

This isomorphism restricts to an isomorphism

$$\operatorname{End}^0(T_l E)(1) \cong \operatorname{Sym}^2 T_l E$$

of direct summands.

Proof. The Weil pairing yields a duality isomorphism

$$T_l E \cong \operatorname{Hom}(T_l E, \mathbb{Z}_l(1)),$$

essentially by the definition of a perfect pairing. Galois equivariance of the Weil pairing implies precisely that this identification respects Galois action, thanks to the definition of the adjoint Galois action. Tensoring with $T_l E$ now yields the first statement of the lemma, since $\operatorname{Hom}(T_l E, T_l E(1))$ is visibly isomorphic to $\operatorname{End}(T_l E)(1)$.

Explicitly, the above isomorphism sends $t \otimes t' \in T_l E \otimes T_l E$ to the function $t' \otimes e(t, \cdot) \in \operatorname{Hom}(T_l E, T_l E \otimes \mathbb{Z}_l(1))$. To check the second statement, we can ignore Galois actions and we simply have to check that symmetric elements of $T_l E \otimes T_l E$ correspond to trace zero matrices in $\operatorname{End}(T_l E)$. This follows immediately from the fact that the Weil pairing is alternating; we leave it as an exercise. □

If M is any finite free $\mathbb{Z}_l$-module with an action of $G_{\mathbb{Q},S}$, we define its *integral Cartier dual* M^* to be the $G_{\mathbb{Q},S}$-module $\operatorname{Hom}(M, \mathbb{Z}_l(1))$. (Often the term *Cartier dual* is used for the module $\operatorname{Hom}(M, \mathbb{Q}_l/\mathbb{Z}_l(1))$.)

Lemma 10.3. *The Weil pairing induces an isomorphism*

$$(T_l E \otimes T_l E)^* \cong T_l E \otimes T_l E(-1).$$

This isomorphism restricts to an isomorphism

$$(\operatorname{Sym}^2 T_l E)^* \cong (\operatorname{Sym}^2 T_l E)(-1).$$

Proof. In general, if A and B are any free $\mathbb{Z}_l$-modules with $G_{\mathbb{Q},S}$-actions, then $(A \otimes B)^* \cong A^* \otimes B^*(-1)$, as one checks easily from the definition. In our case, the Weil pairing shows that $(T_l E)^* \cong T_l E$, and the first statement follows. The second statement is immediate once the first isomorphism has been made explicit; we omit the details. □

First reductions

In this section we will use various global duality theorems of Tate to reduce our calculation of $\mathrm{H}^2(G_{\mathbb{Q},S}, \operatorname{ad}(\bar{\rho}))$ to the vanishing of a certain Shafarevich-Tate group. For the remainder of the paper we fix a rational elliptic curve E and a prime l satisfying the hypotheses of Theorem 10.1. Let S be the corresponding set of places of $\mathbb{Q}$. Note that as Galois modules $\operatorname{ad}(\bar{\rho}) \cong \operatorname{End}(E[l])$; we will use the notation $\operatorname{End}(E[l])$ from now on.

We begin with the following small piece of the Poitou-Tate exact sequence (see [**156**, Section 8] for statements and [**107**, Chapter 1, Section 4] for a proof; here we are using the fact that $l \neq 2$ to eliminate the terms at infinity):

$$\prod_{p \in S - \{\infty\}} \mathrm{H}^0(\mathbb{Q}_p, E[l] \otimes E[l]) \to \mathrm{Hom}\big(\mathrm{H}^2(G_{\mathbb{Q},S}, (E[l] \otimes E[l])^*), \mathbb{Z}/l\mathbb{Z}\big) \to$$
$$\mathrm{H}^1(G_{\mathbb{Q},S}, E[l] \otimes E[l]) \to \prod_{p \in S} \mathrm{H}^1(\mathbb{Q}_p, E[l] \otimes E[l]).$$

Tensoring the first isomorphism of Lemma 10.2 with $\mathbb{Z}/l\mathbb{Z}$ yields an isomorphism

$$E[l] \otimes E[l] \cong \mathrm{End}(E[l])(1).$$

Together with Lemma 10.3, this implies that

$$(E[l] \otimes E[l])^* \cong \mathrm{End}(E[l]).$$

Thus the above exact sequence contains a term which is the dual vector space to $\mathrm{H}^2(G_{\mathbb{Q},S}, \mathrm{End}(E[l]))$; since to prove Theorem 10.1 we need to show that this group vanishes, we see that it will suffice to show that the two groups

$$\prod_{p \in S - \{\infty\}} \mathrm{H}^0(\mathbb{Q}_p, E[l] \otimes E[l])$$

$$Ш^1(G_{\mathbb{Q},S}, E[l] \otimes E[l])) = \ker \left(\mathrm{H}^1(G_{\mathbb{Q},S}, E[l] \otimes E[l]) \to \prod_{p \in S} \mathrm{H}^1(\mathbb{Q}_p, E[l] \otimes E[l]) \right)$$

both vanish.

The vanishing of the first of these groups is the third hypothesis in the statement of Theorem 10.1. For the second, we first write

$$E[l] \otimes E[l] = \wedge^2 E[l] \oplus \mathrm{Sym}^2 E[l] \cong \mu_l \oplus \mathrm{Sym}^2 E[l].$$

(The isomorphism of $\wedge^2 E[l]$ and μ_l comes from the Weil pairing.) One sees immediately from this that there is a corresponding decomposition

$$Ш^1(G_{\mathbb{Q},S}, E[l] \otimes E[l]) \cong Ш^1(G_{\mathbb{Q},S}, \mu_l) \oplus Ш^1(G_{\mathbb{Q},S}, \mathrm{Sym}^2 E[l]).$$

The first term is easily dealt with. We will need the following results, which will also be useful later when dealing with Selmer groups.

Lemma 10.4. *Let A be a $G_{\mathbb{Q}}$-module which is unramified away from a finite set of primes S. Then*

$$\mathrm{H}^1(G_{\mathbb{Q},S}, A) \cong \ker \left(\mathrm{H}^1(\mathbb{Q}, A) \to \prod_{p \notin S} \mathrm{H}^1(I_p, A) \right).$$

Here $I_p \subseteq G_{\mathbb{Q}}$ is the inertia group at p.

Proof. See [**156**, Proposition 6] for a proof. The idea is simply that cohomology classes for $G_{\mathbb{Q},S}$ are automatically unramified away from S and therefore are trivial when restricted to the corresponding inertia groups. □

Lemma 10.5. *Let p be a prime different from l. Then the maximal pro-l quotient of the inertia group I_p is isomorphic to $\mathbb{Z}_l$ as a topological group. If $G_{\mathbb{Q}_p}$ is made to act on I_p by conjugation, then the maximal pro-l quotient of I_p is isomorphic to $\mathbb{Z}_l(1)$ as a $G_{\mathbb{Q}_p}$-module.*

Proof. Recall that $I_p = \mathrm{Gal}(\bar{\mathbb{Q}}_p/\mathbb{Q}_p^{\mathrm{ur}})$. It is shown in [**54**, Section 8, Corollary 3] that the maximal pro-l quotient of this group is $\mathrm{Gal}(\mathbb{Q}_p^{\mathrm{ur}}(p^{1/l^\infty})/\mathbb{Q}_p^{\mathrm{ur}})$. This is seen to be isomorphic to $\mathbb{Z}_l(1)$ using the isomorphisms

$$\mathrm{Gal}(\mathbb{Q}_p^{\mathrm{ur}}(p^{1/l^\infty})/\mathbb{Q}_p^{\mathrm{ur}}) \cong \varprojlim \ \mathrm{Gal}(\mathbb{Q}_p^{\mathrm{ur}}(p^{1/l^n})/\mathbb{Q}_p^{\mathrm{ur}}) \cong \varprojlim \ \mu_{\ell^n} \cong \mathbb{Z}_l(1).$$

We leave the verification that the conjugation action of $G_{\mathbb{Q}_p}$ is cyclotomic as an exercise. □

Lemma 10.6. $\text{Ш}^1(G_{\mathbb{Q},S}, \mu_l) = 0$.

Proof. By definition,

$$\text{Ш}^1(G_{\mathbb{Q},S}, \mu_l) = \ker\left(\mathrm{H}^1(G_{\mathbb{Q},S}, \mu_l) \to \prod_{p\in S} \mathrm{H}^1(\mathbb{Q}_p, \mu_l)\right).$$

Lemma 10.4 shows that we can rewrite this as

$$\text{Ш}^1(G_{\mathbb{Q},S}, \mu_l) = \ker\left(\mathrm{H}^1(\mathbb{Q}, \mu_l) \to \prod_{p\notin S} \mathrm{H}^1(I_p, \mu_l) \times \prod_{p\in S} \mathrm{H}^1(\mathbb{Q}_p, \mu_l)\right).$$

We will compute these groups.

We begin by working in some generality. Let K be any perfect field of characteristic different from l, and consider the exact sequence

$$0 \to \mu_l \to \bar{K}^* \xrightarrow{l} \bar{K}^* \to 0$$

of G_K-modules. Hilbert's theorem 90 (see [**140**, Chapter II.1, Proposition 1]) says that $\mathrm{H}^1(K, \bar{K}^*) = 0$, so the long exact sequence in G_K-cohomology coming from the short exact sequence above yields an isomorphism

$$K^* \otimes_{\mathbb{Z}} \mathbb{Z}/l\mathbb{Z} \cong \mathrm{H}^1(K, \mu_l).$$

This applies in particular to the fields $\mathbb{Q}_p$:

$$\mathbb{Q}_p^* \otimes_{\mathbb{Z}} \mathbb{Z}/l\mathbb{Z} \cong \mathrm{H}^1(\mathbb{Q}_p, \mu_l).$$

For $p \notin S$ we also need to compute $\mathrm{H}^1(I_p, \mu_l)$. Since $p \notin S$, we know that $p \neq l$; thus I_p acts trivially on μ_l. Thus $\mathrm{H}^1(I_p, \mu_l) \cong \mathrm{Hom}(I_p, \mu_l)$. Any such homomorphism must factor through the maximal pro-l quotient of I_p, and now Lemma 10.5 shows that this group is just μ_ℓ. It is easily checked that the restriction map

$$\mathrm{H}^1(\mathbb{Q}_p, \mu_l) \to \mathrm{H}^1(I_p, \mu_l) \cong \mathrm{Hom}(\mu_\ell, \mu_\ell) \cong \mathbb{Z}/l\mathbb{Z}$$

is just the natural map

$$\mathbb{Q}_p^\times \otimes_{\mathbb{Z}} \mathbb{Z}/l\mathbb{Z} \cong (\mathbb{Z} \times \mathbb{Z}_p^\times) \otimes_{\mathbb{Z}} \mathbb{Z}/l\mathbb{Z} \to \mathbb{Z}/l\mathbb{Z}$$

which is trivial on $\mathbb{Z}_p^\times$; that is, it is the p-adic valuation map modulo l.

The group $\text{Ш}^1(G_{\mathbb{Q},S}, \mu_l)$ is therefore the kernel of the map

$$\mathbb{Q}^\times \otimes_{\mathbb{Z}} \mathbb{Z}/l\mathbb{Z} \to \prod_{p\notin S} \mathbb{Z}/l\mathbb{Z} \times \prod_{p\in S-\{\infty\}} \mathbb{Q}_p^\times \otimes_{\mathbb{Z}} \mathbb{Z}/l\mathbb{Z}.$$

(Since $l \neq 2$ the term at ∞ vanishes.) Our calculations above show that this kernel consists only of elements of $\mathbb{Q}^\times \otimes_{\mathbb{Z}} \mathbb{Z}/l\mathbb{Z}$ which have p-adic valuation divisible by l for all primes p. (In fact, at $p \in S$ the conditions are even stronger, but we won't need that.) Unique factorization in $\mathbb{Z}$ implies that any such rational number is an l^{th}-power in $\mathbb{Q}^\times$, and therefore is zero in $\mathbb{Q}^\times \otimes \mathbb{Z}/l\mathbb{Z}$. (Here we also need to use the fact that the units $\mathbb{Z}^\times$ are just ± 1 and disappear on tensoring with $\mathbb{Z}/l\mathbb{Z}$.) Thus $Ш^1(G_{\mathbb{Q},S}, \mu_l) = 0$, as claimed. Note that the fact that $\mathbb{Z}$ has unit rank 0 and class number 1 was essential to this argument. □

We have now reduced the proof of Theorem 10.1 to showing that the Shafarevich-Tate group $Ш^1(G_{\mathbb{Q},S}, \operatorname{Sym}^2 E[l])$ is trivial. We will first show that it embeds into an a priori larger group. Before we can do this, however, we need to define the Selmer groups of our Galois modules.

Selmer groups

Recall that the Selmer group of an elliptic curve over $\mathbb{Q}$ is defined to be the subgroup of $\mathrm{H}^1(\mathbb{Q}, T_l E \otimes \mathbb{Q}_l/\mathbb{Z}_l)$ of cohomology classes which for all p are locally in the image of $E(\mathbb{Q}_p)$ under the Kummer map. (See [**63**].) We will be working with $\operatorname{Sym}^2 T_l E \otimes \mathbb{Q}_l/\mathbb{Z}_l$, but here we no longer have any natural geometric object on which to base our local conditions in the definition of the Selmer group. The key to the general definition of a Selmer group is the fact that the image of the local Kummer map consists precisely of those cohomology classes which are unramified, in a sense which we shall make precise below.

Let us fix some notation for the remainder of the paper: set $T = \operatorname{Sym}^2 T_l E$ (a free $\mathbb{Z}_l$-module of rank 3), $V = T \otimes_{\mathbb{Z}_l} \mathbb{Q}_l$ (a 3 dimensional $\mathbb{Q}_l$-vector space) and $A = V/T = \operatorname{Sym}^2 T_l E \otimes_{\mathbb{Z}_l} \mathbb{Q}_l/\mathbb{Z}_l$ (which is isomorphic as an abelian group to $(\mathbb{Q}_l/\mathbb{Z}_l)^3$). T, V and A are to be regarded as three different incarnations of the same Galois module. We will also need to consider $A^* = \operatorname{Hom}_{\mathbb{Z}_l}(T, \mu_{l^\infty})$, which is also isomorphic as an abelian group to $(\mathbb{Q}_l/\mathbb{Z}_l)^3$. Lastly, for technical reasons we will later need to consider the finite modules $T_n = T/l^n T$ and $A_n^* = A^*[l^n]$ (both of which are isomorphic to $(\mathbb{Z}/l^n\mathbb{Z})^3$ as abelian groups).

To formalize the notion of a Selmer group, we wish to define "unramified subgroups" (the usual term is *finite subgroups*) $\mathrm{H}^1_f(\mathbb{Q}_p, A)$ of $\mathrm{H}^1(\mathbb{Q}_p, A)$ for every prime p. We will then define the *Selmer group* $\mathrm{H}^1_f(\mathbb{Q}, A)$ of A by

$$\begin{aligned}\mathrm{H}^1_f(\mathbb{Q}, A) &= \ker\left(\mathrm{H}^1(\mathbb{Q}, A) \to \prod_p \mathrm{H}^1(\mathbb{Q}_p, A)/\mathrm{H}^1_f(\mathbb{Q}_p, A)\right) \\ &= \{c \in \mathrm{H}^1(\mathbb{Q}, A) \mid \operatorname{res}_p c \in \mathrm{H}^1_f(\mathbb{Q}_p, A) \text{ for all } p\}.\end{aligned}$$

Here res_p is the restriction map from $\mathrm{H}^1(\mathbb{Q}, A)$ to $\mathrm{H}^1(\mathbb{Q}_p, A)$.

The definition of $\mathrm{H}^1_f(\mathbb{Q}_p, A)$ is fairly straightforward for primes p which do not lie in S. Indeed, the most obvious notion of an unramified cocycle is one which becomes trivial when restricted to inertia; that is, it should lie in the kernel of the restriction map

$$\mathrm{H}^1(\mathbb{Q}_p, A) \to \mathrm{H}^1(I_p, A)$$

where I_p is the inertia subgroup of $G_{\mathbb{Q}_p}$. The inflation-restriction sequence (see [**156**, Proposition 2 and the discussion following]) identifies this kernel with

$$\mathrm{H}^1(G_{\mathbb{Q}_p}/I_p, A) = \mathrm{H}^1(\mathbb{F}_p, A).$$

(Here we are using that I_p acts trivially on A to see that A is a $G_{\mathbb{Q}_p}/I_p$-module.) We take this as our definition of the finite subgroup, at least for $p \notin S$:

$$\mathrm{H}^1_f(\mathbb{Q}_p, A) = \mathrm{H}^1(\mathbb{F}_p, A),$$

or more honestly its image in $\mathrm{H}^1(\mathbb{Q}_p, A)$ under inflation.

Note that the inflation-restriction exact sequence now takes the form

$$0 \to \mathrm{H}^1_f(\mathbb{Q}_p, A) \to \mathrm{H}^1(\mathbb{Q}_p, A) \to \mathrm{H}^1(I_p, A)^{G_{\mathbb{F}_p}} \to \mathrm{H}^2(\mathbb{F}_p, A).$$

In fact, $\mathrm{H}^2(\mathbb{F}_p, A)$ vanishes because $G_{\mathbb{F}_p} \cong \hat{\mathbb{Z}}$; see [**140**, Chapter II.3]. We will call $\mathrm{H}^1(I_p, A)^{G_{\mathbb{F}_p}}$ the *singular quotient* of $\mathrm{H}^1(\mathbb{Q}_p, A)$ and write it as $\mathrm{H}^1_s(\mathbb{Q}_p, A)$, so that we have an exact sequence

$$0 \to \mathrm{H}^1_f(\mathbb{Q}_p, A) \to \mathrm{H}^1(\mathbb{Q}_p, A) \to \mathrm{H}^1_s(\mathbb{Q}_p, A) \to 0.$$

In analogy with this definition, for $p \in S$ it might seem most reasonable to define the finite part of $\mathrm{H}^1(\mathbb{Q}_p, A)$ also as the kernel of $\mathrm{H}^1(\mathbb{Q}_p, A) \to \mathrm{H}^1(I_p, A)$, which equals $\mathrm{H}^1(\mathbb{F}_p, A^{I_p})$. However, for reasons which will not become apparent in this paper this definition turns out to be inadequate. It turns out that the correct definition is as follows: first assume $p \neq l$. Let $\pi : V \to A$ be the natural quotient map. We define

$$\mathrm{H}^1_f(\mathbb{Q}_p, V) = \mathrm{H}^1(\mathbb{F}_p, V^{I_p})$$

and

$$\mathrm{H}^1_f(\mathbb{Q}_p, A) = \pi_* \mathrm{H}^1_f(\mathbb{Q}_p, V).$$

Here $\pi_* : \mathrm{H}^1(\mathbb{Q}_p, V) \to \mathrm{H}^1(\mathbb{Q}_p, A)$ is the induced map on cohomology. Despite appearances, $\pi_* \mathrm{H}^1_f(\mathbb{Q}_p, V)$ need not be the same as $\mathrm{H}^1(\mathbb{F}_p, V^{I_p})$, at least for $p \in S$. For later use, let us also set $\mathrm{H}^1_f(\mathbb{Q}_p, V) = \mathrm{H}^1(\mathbb{F}_p, V)$ for $p \notin S$. One checks easily that in this case $\pi_* \mathrm{H}^1_f(\mathbb{Q}_p, V)$ does agree with our previous definition of $\mathrm{H}^1_f(\mathbb{Q}_p, A)$.

The definition of $\mathrm{H}^1_f(\mathbb{Q}_l, A)$ is much more subtle. The generally accepted definition (which recovers the usual definition in the case of the Tate module of an elliptic curve) is

$$\mathrm{H}^1_f(\mathbb{Q}_p, V) = \ker\bigl(\mathrm{H}^1(\mathbb{Q}_p, V) \to \mathrm{H}^1(\mathbb{Q}_p, V \otimes B_{\mathrm{crys}})\bigr)$$

and $\mathrm{H}^1_f(\mathbb{Q}_p, A) = \pi_* \mathrm{H}^1_f(\mathbb{Q}_p, V)$. Here B_{crys} is one of Fontaine's "big rings". We will not concern ourselves very much with this definition in this paper, although in many ways it is one of the most important topics. It is possible to make this definition much more concrete, but even that does not really make this condition any easier to deal with.

In passing, we should note that $\mathrm{H}^1(\mathbb{R}, A) = 0$ (since $l \neq 2$), so we need not concern ourselves with any definitions at infinity. We have now defined $\mathrm{H}^1_f(\mathbb{Q}_p, A)$ for all primes p, and with it the Selmer group $\mathrm{H}^1_f(\mathbb{Q}, A)$. Of course, we can mimic the identical construction with A^* instead of A, and thus we also have a Selmer group $\mathrm{H}^1_f(\mathbb{Q}, A^*)$.

We can also redo the construction for the Galois module T. Note that T is fundamentally quite different from A, in that it is free over $\mathbb{Z}_l$ rather than isomorphic to several copies of $\mathbb{Q}_l/\mathbb{Z}_l$. The definitions are nevertheless quite analogous. Let $i : T \to V$ be the natural inclusion, and for all p define

$$\mathrm{H}^1_f(\mathbb{Q}_p, T) = i_*^{-1} \mathrm{H}^1_f(\mathbb{Q}_p, V).$$

We also define singular quotients $\mathrm{H}^1_s(\mathbb{Q}_p, T) = \mathrm{H}^1(\mathbb{Q}_p, T)/\mathrm{H}^1_f(\mathbb{Q}_p, T)$. In this case, it will turn out that the singular part of the local cohomology is that which we can most easily work with. One can also define a Selmer group for T, although we will have no need to consider it.

We will also need corresponding subgroups for the finite modules T_n and $A^*[l^n]$. For any prime p, we simply take $\mathrm{H}^1_f(\mathbb{Q}_p, A_n^*)$ to be the inverse image of $\mathrm{H}^1_f(\mathbb{Q}_p, A^*)$

under the natural map $\mathrm{H}^1(\mathbb{Q}_p, A_n^*) \to \mathrm{H}^1(\mathbb{Q}_p, A^*)$. Similarly, we define $\mathrm{H}^1_f(\mathbb{Q}_p, T_n)$ to be the image of $\mathrm{H}^1_f(\mathbb{Q}_p, T)$ under the natural map $\mathrm{H}^1(\mathbb{Q}_p, T) \to \mathrm{H}^1(\mathbb{Q}_p, T_n)$. One can now define singular quotients and Selmer groups in the usual way.

It is worth noting the general philosophy: we took the natural definition of unramified cocycles in $\mathrm{H}^1(\mathbb{Q}_p, V)$ (with the exception of the case $p = l$, which was more complicated) and we then let these choices propagate down to all of the related Galois modules.

Returning to the case of $T_l E$, recall that one defines the Shafarevich-Tate group $Ш(E/\mathbb{Q})$ as the cokernel of the Kummer map

$$E(\mathbb{Q}) \otimes_{\mathbb{Z}} \mathbb{Q}_l/\mathbb{Z}_l \to \mathrm{H}^1_f(\mathbb{Q}, T_l E).$$

As before we have no obvious analogue of $E(\mathbb{Q})$ in our situation. The work of Bloch and Kato suggests that the correct analogue is the following: let

$$\mathrm{H}^1_f(\mathbb{Q}, V) = \ker\left(\mathrm{H}^1(\mathbb{Q}, V) \to \prod_p \mathrm{H}^1(\mathbb{Q}_p, V)/\mathrm{H}^1_f(\mathbb{Q}_p, V)\right)$$

be the Selmer group for V, and define the "rational points of A" to be

$$A(\mathbb{Q}) = \pi_* \mathrm{H}^1_f(\mathbb{Q}, V) \subseteq \mathrm{H}^1(\mathbb{Q}, A).$$

Note that it follows immediately from the definition of the $\mathrm{H}^1_f(\mathbb{Q}_p, A)$ that $A(\mathbb{Q})$ actually lies in $\mathrm{H}^1_f(\mathbb{Q}, A)$, although it could conceivably be smaller. We define the Shafarevich-Tate group $Ш(A/\mathbb{Q})$ to be the quotient $\mathrm{H}^1_f(\mathbb{Q}, A)/A(\mathbb{Q})$, so that there is an exact sequence

$$0 \to A(\mathbb{Q}) \to \mathrm{H}^1_f(\mathbb{Q}, A) \to Ш(A/\mathbb{Q}) \to 0.$$

$Ш(A/\mathbb{Q})$ is to be thought of as elements of the Selmer group which appear over A but not over V. Note also that despite the similar notation, $Ш(A/\mathbb{Q})$ is not the same as any of the Shafarevich-Tate groups we considered earlier.

Again, we can make analogous definitions for $A^*(\mathbb{Q})$, yielding an exact sequence

$$0 \to A^*(\mathbb{Q}) \to \mathrm{H}^1_f(\mathbb{Q}, A^*) \to Ш(A^*/\mathbb{Q}) \to 0.$$

We are now in a position to finish our reductions of the previous section.

Lemma 10.7. *$Ш^1(G_{\mathbb{Q},S}, \mathrm{Sym}^2 E[l])$ injects into $\mathrm{H}^1_f(\mathbb{Q}, A)$.*

Proof. As an abelian group A is isomorphic to $(\mathbb{Q}_l/\mathbb{Z}_l)^3$, so multiplication by l is surjective on A. Furthermore, the kernel of multiplication by l on A,

$$\mathrm{Sym}^2 T_l E \otimes \frac{1}{l}\mathbb{Z}_l/\mathbb{Z}_l,$$

naturally identifies with $\mathrm{Sym}^2 E[l]$, so there is an exact sequence

$$0 \to \mathrm{Sym}^2 E[l] \to A \xrightarrow{l} A \to 0. \tag{2}$$

The fact that $E[l] \otimes E[l]$ is assumed to have no $G_{\mathbb{Q}_p}$-invariants for any $p \in S - \{\infty\}$ insures that the direct summand $\mathrm{Sym}^2 E[l]$ has no $G_{\mathbb{Q},S}$-invariants; indeed, knowing that it had no invariants at any one place would suffice. It follows easily from this that A itself has no $G_{\mathbb{Q},S}$-invariants, as if there were any, then they could be realized in $\mathrm{Sym}^2 E[l]$ by multiplication by an appropriate power of l. Thus the long exact sequence in $G_{\mathbb{Q},S}$-cohomology associated to (2) yields an injection

$$\mathrm{H}^1(G_{\mathbb{Q},S}, \mathrm{Sym}^2 E[l]) \hookrightarrow \mathrm{H}^1(G_{\mathbb{Q},S}, A).$$

Furthermore, under this injection $\text{Ш}^1(G_{\mathbb{Q},S}, \operatorname{Sym}^2 E[l])$ maps into $\text{Ш}^1(G_{\mathbb{Q},S}, A)$. Indeed, there is a commutative diagram

(3)
$$\begin{tikzcd}
\mathrm{H}^1(G_{\mathbb{Q},S}, \operatorname{Sym}^2 E[l]) \arrow[r] \arrow[d] & \mathrm{H}^1(G_{\mathbb{Q},S}, A) \arrow[d] \\
\prod_{p \in S} \mathrm{H}^1(\mathbb{Q}_p, \operatorname{Sym}^2 E[l]) \arrow[r] & \prod_{p \in S} \mathrm{H}^1(\mathbb{Q}_p, A)
\end{tikzcd}$$

This shows that any element of $\text{Ш}^1(G_{\mathbb{Q},S}, \operatorname{Sym}^2 E[l])$, which is by definition trivial in each $\mathrm{H}^1(\mathbb{Q}_p, \operatorname{Sym}^2 E[l])$, is automatically trivial in each $\mathrm{H}^1(\mathbb{Q}_p, A)$, and thus lies in $\text{Ш}^1(G_{\mathbb{Q},S}, A)$. In other words, the induced map on the kernels of the vertical maps in (3) is the desired injection

$$\text{Ш}^1(G_{\mathbb{Q},S}, \operatorname{Sym}^2 E[l]) \hookrightarrow \text{Ш}^1(G_{\mathbb{Q},S}, A).$$

To prove the lemma it will therefore suffice to show that $\text{Ш}^1(G_{\mathbb{Q},S}, A)$ injects into $\mathrm{H}^1_f(\mathbb{Q}, A)$.

It follows from Lemma 10.4 that there is an injection

$$\mathrm{H}^1(G_{\mathbb{Q},S}, A) \hookrightarrow \mathrm{H}^1(\mathbb{Q}, A)$$

and that the composite maps

(4)
$$\mathrm{H}^1(G_{\mathbb{Q},S}, A) \to \mathrm{H}^1(\mathbb{Q}, A) \to \mathrm{H}^1(I_p, A) \cong \mathrm{H}^1(\mathbb{Q}_p, A)/\mathrm{H}^1_f(\mathbb{Q}_p, A)$$

are zero for all $p \notin S$. We must show that the image of $\text{Ш}^1(G_{\mathbb{Q},S}, A)$ in $\mathrm{H}^1(\mathbb{Q}, A)$ lies in $\mathrm{H}^1_f(\mathbb{Q}, A)$. By (4), this image is automatically locally unramified for all $p \notin S$. Furthermore, an argument using a diagram analogous to (3) above shows that the maps

$$\text{Ш}^1(G_{\mathbb{Q},S}, A) \to \mathrm{H}^1(\mathbb{Q}, A) \to \mathrm{H}^1(\mathbb{Q}_p, A)$$

are zero for $p \in S$. Thus the image of $\text{Ш}^1(G_{\mathbb{Q},S}, A)$ trivially lands in $\mathrm{H}^1_f(\mathbb{Q}_p, A)$ for such p. Thus now $\text{Ш}^1(G_{\mathbb{Q},S}, A)$ maps to $\mathrm{H}^1_f(\mathbb{Q}_p, A)$ for every prime p, so its image in $\mathrm{H}^1(\mathbb{Q}, A)$ lies in $\mathrm{H}^1_f(\mathbb{Q}, A)$. Thus $\text{Ш}^1(G_{\mathbb{Q},S}, A)$ injects into $\mathrm{H}^1_f(\mathbb{Q}, A)$, which proves the lemma. □

At this point, we have reduced the proof of Theorem 10.1 to the vanishing of the Selmer group $\mathrm{H}^1_f(\mathbb{Q}, A)$.

The L-function of $\operatorname{Sym}^2 T_l E$

Before we complete our final reformulation of Theorem 10.1, we should give the long promised explanation of the term $L(\operatorname{Sym}^2 T_l E, 0)/\Omega$. Recall that the L-function of the Tate module of an elliptic curve (also known as the L-function of the elliptic curve) is defined using the characteristic polynomials of Frobenius elements acting on $T_l E$. We use an analogous method to define $L(\operatorname{Sym}^2 T_l E, s)$. Specifically, the action of $G_{\mathbb{Q}_p}$ on T is unramified for every $p \notin S$, so it makes sense to talk about the action of an arithmetic Frobenius element Fr_p on T. Let $P_p(t)$ be the characteristic polynomial of Fr_p^{-1} acting on T:

$$P_p(t) = \det(1 - (\mathrm{Fr}_p |_T) t).$$

For $p \in S - \{l\}$, Fr_p is only well-defined acting on the inertia invariants T^{I_p}, so we define

$$P_p(t) = \det(1 - (\mathrm{Fr}_p |_{T^{I_p}}) t|).$$

These are all initially polynomials with coefficients in $\mathbb{Z}_l$, but it turns out that $P_p(t)$ actually has coefficients in $\mathbb{Z}$ and the polynomial $P_p(t)$ does not depend on the distinguished prime l, so long as $l \neq p$. (Again, this is all completely analogous to the $T_l E$ case.)

This suggests that to define the factor $P_l(t)$ itself, we should not work directly with T, but rather switch to $\mathrm{Sym}^2 T_p E$ for some $p \neq l$. $\mathrm{Sym}^2 T_p E$ is unramified at l (since E is assumed to have good reduction at l), so Fr_l is well-defined here and we define

$$P_l(t) = \det(1 - \mathrm{Fr}_l \,|_{\mathrm{Sym}^2 T_p E} t).$$

As before this is independent of the choice of auxiliary $p \neq l$.

We now define

$$L(T, s) = \prod_p P_p(p^{-s})^{-1}.$$

It is shown in [**19**] that $L(T, s)$ is an entire function of s.

Since E is modular, one can use the work of Shimura [**145**] to compute special values of this L-function. Let N be the conductor of E and fix a modular parameterization

$$\phi : X_0(N) \to E$$

of E. We assume that ϕ is minimal in the sense that $\deg \phi$ is as small as possible for our fixed E and N. Let $f(z)$ be the newform corresponding to ϕ; this means that for all p not dividing N, the p^{th} Fourier coefficient of $f(z)$ equals $p + 1 - \#E(\mathbb{F}_p)$. Let ω be the Néron differential on E. (If E is given in the form $y^2 = x^3 + ax + b$, ω is just $dx/2y$.) Since ϕ is defined over $\mathbb{Q}$, one can show that the pullback of ω under ϕ must be a rational multiple of the differential $2\pi i f(z)dz$ on $X_0(N)$. We define the *Manin constant* $c \in \mathbb{Q}^\times$ by the equality

$$\phi^* \omega = c 2\pi i f(z) dz.$$

Work of Mazur shows that c is divisible only by 2 and primes of bad reduction for E; see [**95**, Corollary 4.1].

We also use ω to define the period Ω by

$$\Omega = \pi i \int_{E(\mathbb{C})} \omega \wedge \bar{\omega}.$$

Shimura's formula is

$$\frac{L(T, 0)}{\Omega} = \frac{\deg \phi}{N c^2} \prod_{p \in S'} P_p(1). \tag{5}$$

Here S' is the subset of S of places where E has potentially good reduction. Note in particular that $L(T, 0)/\Omega$ is rational and non-zero.

We now state the theorem which we will prove in the remainder of this paper and explain how it implies Theorem 10.1.

Theorem 10.8. *Let E be a rational elliptic curve and let $\phi : X_0(N) \to E$ be a modular parameterization. Let l be a prime such that*

- *E has good reduction at l;*
- *$l \geq 5$;*
- *The Tate module representation $\rho : G_{\mathbb{Q},S} \to \mathrm{GL}_2(\mathbb{Z}_l)$ is surjective.*

Then $\deg \phi \cdot \mathrm{H}^1_f(\mathbb{Q}, A^*) = 0$.

To see that this implies Theorem 10.1, recall that we had already reduced the proof to showing that $\mathrm{H}^1_f(\mathbb{Q}, A)$ vanishes. We had also assumed that l does not divide $L(T,0)/\Omega$. We first must show that l does not divide $\deg\phi$ either. To see this, by (5) we need to show that the rational number

$$\frac{1}{Nc^2}\prod_{p\in S'} P_p(1)$$

has no factors of l in the denominator. But N and c^2 are divisible only by 2 and primes of bad reduction for E, and each $P_p(1)$ is an integer, so this is clear. Thus l does not divide $\deg\phi$.

Now, by Theorem 10.8, $\mathrm{H}^1_f(\mathbb{Q}, A^*)$ is annihilated by the l-adic unit $\deg\phi$. But $\mathrm{H}^1_f(\mathbb{Q}, A^*)$ is an l-power torsion group since A^* is, and it follows that it must be 0. This in turn implies that both $A^*(\mathbb{Q})$ and $\text{Ш}(A^*/\mathbb{Q})$ vanish, as they are a subgroup and a quotient of $\mathrm{H}^1_f(\mathbb{Q}, A^*)$, respectively. Flach has shown (see [**49**]) that the vanishing of $A^*(\mathbb{Q})$ implies that of $A(\mathbb{Q})$. Furthermore, he constructs (generalizing ideas of Cassels and Tate; see [**48**]) a perfect pairing

$$\text{Ш}(A/\mathbb{Q}) \otimes \text{Ш}(A^*/\mathbb{Q}) \to \mathbb{Q}_l/\mathbb{Z}_l;$$

thus the vanishing of $\text{Ш}(A^*/\mathbb{Q})$ implies that of $\text{Ш}(A/\mathbb{Q})$. Since both $A(\mathbb{Q})$ and $\text{Ш}(A/\mathbb{Q})$ vanish, this implies that the Selmer group $\mathrm{H}^1_f(\mathbb{Q}, A)$ vanishes, which completes the proof of Theorem 10.1.

Kolyvagin's theory of Euler systems

Until the mid-eighties the problem of bounding Selmer groups was nearly hopeless; there were no methods which worked in any generality. This changed dramatically with the work of Thaine and Rubin and finally Kolyvagin's theory of Euler systems. Since we only seek to annihilate $\mathrm{H}^1_f(\mathbb{Q}, A^*)$, rather than actually bound its order, we will need only the rudiments of the theory. There are several good sources for more extensive treatments: see [**155**, Chapter 15] for a nice introduction, [**66**] and [**122**] for applications to the arithmetic of elliptic curves, and [**124**] for a general theory. In fact, all of these sources deal with a slightly different type of Euler system than we will use. We will have more to say about this later.

We only sketch the main ideas. For a proof of the result we will need, see [**50**, Proposition 1.1] or [**157**].

Fix a power l^n of l and set $T_n = T/l^nT$, $A^*_n = A^*[l^n]$. We must work with these finite modules for technical reasons; passing from them to the full modules will be easy. The basic idea is the following: recall that since $A^*_n = \mathrm{Hom}(T_n, \mu_{l^n})$ there is a Tate local duality

$$\mathrm{H}^1(\mathbb{Q}_p, T_n) \otimes \mathrm{H}^1(\mathbb{Q}_p, A^*_n) \to \mathbb{Q}_l/\mathbb{Z}_l;$$

see [**156**, Theorem 1]. One can show easily that $\mathrm{H}^1_f(\mathbb{Q}_p, T_n)$ and $\mathrm{H}^1_f(\mathbb{Q}_p, A^*_n)$ are exact annihilators of each other (see [**157**, Lectures 5 and 6]), so that restricting the right-hand factor to $\mathrm{H}^1_f(\mathbb{Q}_p, A^*_n)$ gives a perfect pairing

$$\mathrm{H}^1_s(\mathbb{Q}_p, T_n) \otimes \mathrm{H}^1_f(\mathbb{Q}_p, A^*_n) \to \mathbb{Q}_l/\mathbb{Z}_l. \tag{6}$$

These local pairings sum to a global pairing

$$\left(\bigoplus_p \mathrm{H}^1_s(\mathbb{Q}_p, T_n)\right) \otimes \mathrm{H}^1_f(\mathbb{Q}, A^*_n) \to \mathbb{Q}_l/\mathbb{Z}_l; \tag{7}$$

the pairing of an element $(c_p) \in \oplus \mathrm{H}^1_s(\mathbb{Q}_p, T_n)$ and $d \in \mathrm{H}^1_f(\mathbb{Q}, A_n^*)$ is simply the sum of the local pairings (6) of c_p and $\mathrm{res}_p\, d$; since $c_p = 0$ for almost all p, this is well-defined. This pairing is not perfect, but it does have the following key property, which is a consequence of global class field theory: the image of $\mathrm{H}^1(\mathbb{Q}, T_n)$ under the natural map

$$\mathrm{H}^1(\mathbb{Q}, T_n) \to \prod_p \mathrm{H}^1(\mathbb{Q}_p, T_n) \to \prod_p \mathrm{H}^1_s(\mathbb{Q}_p, T_n)$$

actually lands in $\oplus_p \mathrm{H}^1_s(\mathbb{Q}_p, T_n)$ (see [**157**, Lecture 7, Section 2.1]; this is one place where it is critical that we dropped to a finite quotient of T) and it is orthogonal to all of $\mathrm{H}^1_f(\mathbb{Q}, A_n^*)$ under the global pairing (7); see [**157**, Lecture 8].

The significance of this to our problem is the following: suppose that for lots of primes r we can exhibit elements $c_r \in \mathrm{H}^1(\mathbb{Q}, T_n)$ with the property that they restrict to 0 in $\mathrm{H}^1_s(\mathbb{Q}_p, T_n)$ for all p *except* for r itself, where they restrict to something non-zero. Under the global pairing, the image of c_r in $\oplus \mathrm{H}^1_s(\mathbb{Q}_p, T_n)$ is orthogonal to all of $\mathrm{H}^1_f(\mathbb{Q}, A_n^*)$. But the definition of the global pairing together with the fact that c_r restricts to 0 in $\mathrm{H}^1_s(\mathbb{Q}_p, T_n)$ away from r now shows that $\mathrm{res}_r\, c_r \in \mathrm{H}^1_s(\mathbb{Q}_r, T_n)$ is orthogonal to the image of $\mathrm{H}^1_f(\mathbb{Q}, A_n^*)$ in $\mathrm{H}^1_f(\mathbb{Q}_r, A_n^*)$ under the Tate local pairing at r. If $\mathrm{res}_r\, c_r$ generates a submodule of $\mathrm{H}^1_s(\mathbb{Q}_r, T_n)$ of small index, then this orthogonality and the fact that the Tate local pairing is perfect will force $\mathrm{H}^1_f(\mathbb{Q}, A_n^*)$ to map into a small subgroup of $\mathrm{H}^1_f(\mathbb{Q}_r, A_n^*)$. Since we can do this for lots of r, we obtain conditions on the local behavior of $\mathrm{H}^1_f(\mathbb{Q}, A_n^*)$ at many primes r. Hopefully if we could do this for enough primes r we could somehow show that the local conditions are so stringent that the group $\mathrm{H}^1_f(\mathbb{Q}, A_n^*)$ itself must be small.

Before we state all of this somewhat more formally, we prove the following fundamental lemma. We will call a prime *p good* if it is not in S and if a Frobenius element at p acts on $E[l]$ as complex conjugation. This is equivalent to Fr_p being a complex conjugation element on the splitting field $\mathbb{Q}(E[l])$ of $E[l]$. This field is a finite extension of $\mathbb{Q}$ since $E[l]$ is finite, so by the Chebotarev density theorem there are infinitely many good primes.

Lemma 10.9. *Assume $p \notin S$. Then*

$$\mathrm{H}^1_s(\mathbb{Q}_p, T) \cong T(-1)^{G_{\mathbb{F}_p}}.$$

If p is good, then this group is a free $\mathbb{Z}_l$-module of rank 1. In particular, each $\mathrm{H}^1_s(\mathbb{Q}_p, T_n)$ is a free $\mathbb{Z}/l^n\mathbb{Z}$-module of rank 1.

Proof. For the first isomorphism, recall that

$$\mathrm{H}^1_s(\mathbb{Q}_p, T) \cong \mathrm{H}^0(\mathbb{F}_p, \mathrm{H}^1(I_p, T)).$$

Since I_p acts trivially on T, $\mathrm{H}^1(I_p, T)$ is nothing more than $\mathrm{Hom}(I_p, T)$. Now, T is a pro-l group, so only the pro-l part of I_p can map to it non-trivially. By Lemma 10.5, this quotient is isomorphic to $\mathbb{Z}_l(1)$ as a $G_{\mathbb{F}_p}$-module. We conclude that

$$\mathrm{H}^0(\mathbb{F}_p, \mathrm{H}^1(I_p, T)) \cong \mathrm{H}^0(\mathbb{F}_p, \mathrm{Hom}(\mathbb{Z}_l(1), T)).$$

But $\mathrm{Hom}(\mathbb{Z}_l(1), T)$ is canonically isomorphic to $\mathrm{Hom}(\mathbb{Z}_l, T(-1))$, which in turn is just $T(-1)$. This proves the first statement.

Now assume that p is good. This means that Fr_p acts on $E[l]$ as complex conjugation. In particular, it is a non-scalar involution, which one easily shows implies that it acts diagonally on $E[l]$ with eigenvalues 1 and -1. A Nakayama's lemma argument together with Hensel's lemma and a dimension count allows one to conclude that there is a basis x, y of $T_l E$ over $\mathbb{Z}_l$ with respect to Fr_p acts diagonally; that is, $\mathrm{Fr}_p(x) = ux$ and $\mathrm{Fr}_p(y) = vy$, and we must have

$$u \equiv -v \equiv 1 \pmod{l}.$$

Note that uv is the determinant of Fr_p acting on $T_l E$, which is just $\varepsilon(\mathrm{Fr}_p) = p$ since $T_l E$ has cyclotomic determinant by the Weil pairing. In particular, $p \equiv -1 \pmod{l}$.

A basis for $T = \mathrm{Sym}^2 T_l E$ is given by $x \otimes x, x \otimes y + y \otimes x, y \otimes y$. Fr_p acts on the first by multiplication by u^2; on the second by multiplication by $uv = p$; and on the third by multiplication by v^2. Note that $u^2 \equiv v^2 \equiv 1 \pmod{l}$, which implies that neither u^2 nor v^2 equals p.

Now consider $T(-1)$. Fr_p acts on $\mathbb{Z}_l(1)$ by multiplication by $\varepsilon(\mathrm{Fr}_p) = p$, so it acts on $\mathbb{Z}_l(-1)$ by multiplication by p^{-1}. Thus Fr_p acts on our basis of $T(-1)$ by multiplication by $u^2 p^{-1}$, 1 and $v^2 p^{-1}$ respectively. As we saw above, the first and last terms are different from 1. It follows that only the rank one subspace of multiples of $x \otimes y + y \otimes x$ is invariant under the $G_{\mathbb{F}_p}$-action, so $\mathrm{H}^1_s(\mathbb{Q}_p, T) = T(-1)^{G_{\mathbb{F}_p}}$ is free of rank one over $\mathbb{Z}_l$, as claimed.

The result for T_n follows exactly the same argument. □

We are now in a position to give a precise definition of the sort of set of cohomology classes we seek: let η be an integer. We define a *Flach system of depth η for T_n* to be a collection of cohomology classes $c_r \in \mathrm{H}^1(\mathbb{Q}, T_n)$, one for each good prime r, such that:

- $\mathrm{res}_p\, c_r$ lies in $\mathrm{H}^1_f(\mathbb{Q}_p, T_n)$ for $p \neq r$;
- $\mathbb{Z}_l \cdot \mathrm{res}_r\, c_r$ contains $\eta \mathrm{H}^1_s(\mathbb{Q}_r, T_n)$.

The second condition is equivalent to the quotient $\mathrm{H}^1_s(\mathbb{Q}_r, T_n)/\mathbb{Z}_l \cdot \mathrm{res}_r\, c_r$ being annihilated by η. By Lemma 10.9, $\mathrm{H}^1_s(\mathbb{Q}_r, T_n)$ is a free $\mathbb{Z}/l^n\mathbb{Z}$-module of rank 1, so this condition is reasonable. Note that to check both of the conditions in the definition above, we simply need a good understanding of the singular restriction maps

$$\mathrm{H}^1(\mathbb{Q}, T_n) \to \mathrm{H}^1(\mathbb{Q}_p, T_n) \to \mathrm{H}^1_s(\mathbb{Q}_p, T_n)$$

for all primes p.

Of course, it is trivial and not very useful to write down a Flach system of depth l^n for T_n; to make this a useful notion, we will want η to be independent of n.

Recall that we have assumed that the Tate module representation $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_l)$ is surjective. This implies immediately that $E[l]$ is an absolutely irreducible $G_{\mathbb{Q}}$-representation, which in turn one can check implies that $A^*[l] \cong (\mathrm{Sym}^2 E[l])^*$ is absolutely irreducible. Even though we are assuming all of these hypotheses, we will include the relevant ones in the hypotheses of each result below.

Given what we have said so far, the following lemma is fairly straightforward.

Lemma 10.10. *Assume that T_n admits a Flach system of depth η. Then for every $d \in \mathrm{H}^1_f(\mathbb{Q}, A^*_n)$ and every good prime r, $\mathrm{res}_r\, d$ lies in $\mathrm{H}^1_f(\mathbb{Q}_p, A^*_n)[\eta]$.*

Proof. See [**157**, Lecture 15, Section 1.2]. □

More difficult is the next result, which goes from this local annihilation result to a global annihilation result. Let K be the fixed field of the kernel of $G_{\mathbb{Q},S}$

acting on $E[l^n]$; it is a finite extension of $\mathbb{Q}$ since $E[l^n]^*$ is finite. Note that this field also lies in the kernel of the $G_{\mathbb{Q},S}$ action on $A^*[l^n]$ (since its Galois action is entirely derived from $E[l^n]$ and the cyclotomic character) and that there is a natural inflation injection

$$\mathrm{H}^1(K/\mathbb{Q}, A_n^*) \hookrightarrow \mathrm{H}^1(\mathbb{Q}, A_n^*).$$

Lemma 10.11. *Assume that $l \neq 2$ and that $A^*[l]$ is absolutely irreducible as a $G_{\mathbb{Q},S}$-module. Let $d \in \mathrm{H}^1(\mathbb{Q}, A_n^*)$ be such that $\operatorname{res}_r d = 0$ for every good prime r. Then d lies in the image of $\mathrm{H}^1(K/\mathbb{Q}, A_n^*)$.*

Proof. See [**157**, Lecture 15, Section 1.3]. □

Lemma 10.10 and Lemma 10.11 combine to show that

$$\eta \mathrm{H}^1_f(\mathbb{Q}, A_n^*) \subseteq \mathrm{H}^1(K/\mathbb{Q}, A_n^*).$$

Since the l^n-torsion representation $\rho_n : G_{\mathbb{Q},S} \to \mathrm{GL}_2(\mathbb{Z}/l^n\mathbb{Z})$ is surjective, we will have $\mathrm{Gal}(K/\mathbb{Q}) \cong \mathrm{GL}_2(\mathbb{Z}/l^n\mathbb{Z})$. Furthermore, Lemma 10.2 and Lemma 10.3 show that $A_n^* \cong \mathrm{ad}^0(\rho_n)$. The next result, which is purely a statement about group cohomology, now finishes our proof, at least at the level of l^n-torsion.

Lemma 10.12. *Let $\mathrm{GL}_2(\mathbb{Z}/l^n\mathbb{Z})$ act on $\mathrm{End}^0(\mathbb{Z}/l^n\mathbb{Z})$ via the adjoint action. Then*

$$\mathrm{H}^1(\mathrm{GL}_2(\mathbb{Z}/l^n\mathbb{Z}), \mathrm{End}^0(\mathbb{Z}/l^n\mathbb{Z})) = 0.$$

Proof. See [**34**, Lemma 2.48] and [**50**]. □

Combining all of this, we have the following theorem.

Theorem 10.13. *Let E be a rational elliptic curve and let l be a prime. Let $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_l)$ be the associated Tate module representation. Assume that ρ is surjective. Further assume that for every good prime r there is a class $c_r \in \mathrm{H}^1(\mathbb{Q}, T)$ such that*

- *$\operatorname{res}_p c_r$ lies in $\mathrm{H}^1_f(\mathbb{Q}_p, T)$ for $r \neq p$;*
- *$\mathbb{Z}_l \cdot \operatorname{res}_r c_r$ contains $\eta \mathrm{H}^1_s(\mathbb{Q}_r, T_n)$.*

Then η annihilates $\mathrm{H}^1_f(\mathbb{Q}, A^*)$.

Proof. The given Flach system for T induces one for each T_n. The results to this point have thus shown that $\eta \mathrm{H}^1_f(\mathbb{Q}, A_n^*) = 0$ for each n. But any class $d \in \mathrm{H}^1_f(\mathbb{Q}, A^*)$ must be annihilated by some power of l, so it lies in the image of some $\mathrm{H}^1_f(\mathbb{Q}, A_n^*)$. (Note that $\mathrm{H}^1_f(\mathbb{Q}, A_n^*)$ maps into $\mathrm{H}^1_f(\mathbb{Q}, A^*)$ since we defined the finite subgroups for A_n^* using those for A^*.) Thus $\eta d = 0$, which completes the proof. □

The proof of Theorem 10.13 is purely a Galois cohomology argument, and therefore there is no actual need to assume that the representation ρ comes from an elliptic curve. For example, in [**51**] Galois representations coming from more general modular forms are considered.

The machinery we have given is sufficient for annihilation and finiteness results. To actually obtain a bound on the order of $\mathrm{H}^1_f(\mathbb{Q}, A^*)$, one has to exhibit classes not only for prime levels (like the c_r) but also for composite levels. Kolyvagin's derivative construction is then used to turn these classes into better and better annihilators. We should note, however, that with most Euler systems which have been studied the classes c_n are defined over larger and larger fields, depending on n. In our case, the classes would all be defined over $\mathbb{Q}$. Such an Euler system is often called a *geometric Euler system*, and there is not yet a general theory of such objects. For one example, see [**123**]. In fact, no one has succeeded in extending the Flach system above to a full geometric Euler system; this was the "gap" in the original proof of semistable Shimura-Taniyama by Wiles, which was eventually filled in by Taylor-Wiles using different methods.

The Flach map

We continue to let E be an elliptic curve over $\mathbb{Q}$ and $\phi : X_0(N) \to E$ a modular parameterization. It remains to construct a Flach system for T of depth $\deg\phi$. This construction lies at the heart of Flach's proof. These classes will come from certain well-chosen geometric objects on the surface $E \times E$, although in order to actually exhibit them we will need to work on the surface $X_0(N) \times X_0(N)$, which has a much richer intrinsic geometry. These objects are then transformed into classes in $\mathrm{H}^1(\mathbb{Q}, T)$ via a certain Chern class map. The key to Flach's construction is that it is possible to read off local properties of these classes in $\mathrm{H}^1_s(\mathbb{Q}_p, T)$ from corresponding local properties of the associated geometric objects. That is, we will begin with a map $\sigma : \mathcal{C}(E \times E) \to \mathrm{H}^1(\mathbb{Q}, T)$, where $\mathcal{C}(E \times E)$ will be defined in a moment purely geometrically. We cannot describe the image of σ directly (we can't even really describe $\mathrm{H}^1(\mathbb{Q}, T)$ effectively), but there is (for p not lying in S) a commutative diagram

$$\begin{tikzcd}
\mathcal{C}(E \times E) \arrow[rr] \arrow[d, "\sigma"'] & & \cdot \arrow[d] \\
\mathrm{H}^1(\mathbb{Q}, T) \arrow[r] & \mathrm{H}^1(\mathbb{Q}_p, T) \arrow[r] & \mathrm{H}^1_s(\mathbb{Q}_p, T)
\end{tikzcd} \tag{8}$$

to be filled in later. Since all we care about for the production of our Flach system is the restriction of classes to $\mathrm{H}^1_s(\mathbb{Q}_p, T)$, to check that classes c_r really form a Flach system we will be able to bypass the complicated $\mathrm{H}^1(\mathbb{Q}, T)$ entirely and work instead with much more concrete geometric objects.

We begin by defining $\mathcal{C}(E \times E)$, which will involve working with curves lying in the surface $E \times E$. Let C be any projective geometrically integral algebraic curve over $\mathbb{Q}$; we do *not* assume that C is non-singular. We will be interested in rational functions on C which have trivial Weil divisor. (Recall that the *Weil divisor* of the function f on C is the formal sum of the points at which it has zeros minus the formal sum of the points at which it has poles, all counted with multiplicity. Often Weil divisors are only defined for *nonsingular* curves, but it is possible to define them more generally. The simplest approach is to define Weil divisors on singular curves by considering Weil divisors on their normalizations and then identifying points which become identified on the singular model. We will see an example of this in a moment.) If C is nonsingular, then it is a standard fact that the only such functions are constant. However, if C is singular, it is possible to exhibit non-constant rational functions with trivial Weil divisor.

For an example, consider a curve C with a nodal singularity $P \in C(\mathbb{Q})$. Let C' be its normalization, with P_1 and P_2 the points lying above P. Let f be a rational function on C' with divisor $nP_1 - nP_2$ for some n. (Such a function may or may not exist for a general C'; it will certainly exist if C' has genus 0, for example.) C' and C are birational, so f can also be interpreted as a rational function on C, and both P_1 and P_2 map to P. Thus f has trivial Weil divisor on C, even though it is a non-constant rational function.

Now consider the non-singular projective surface $E \times E$. We define $\mathcal{C}(E \times E)$ as follows: elements are pairs (C, f) of (possibly singular) curves C contained in $E \times E$ together with a rational function f on C with trivial Weil divisor. We also require that both C and f are defined over $\mathbb{Q}$. Flach defines a map

$$\sigma : \mathcal{C}(E \times E) \to \mathrm{H}^1(\mathbb{Q}, T)$$

which is what we will use to generate our Flach system. For the definition of σ, which involves étale cohomology and algebraic K-theory, see Appendix B.

Let us try briefly to explain the underlying philosophy by analogy with algebraic topology. Begin with the genus one complex curve $E(\mathbb{C})$, which we can regard as $\mathbb{C}/\Lambda$ for an appropriate lattice Λ. One way to obtain Λ is as the integral homology group $\mathrm{H}_1(E(\mathbb{C}), \mathbb{Z})$ (just thing about the standard homology generators on a torus), which is a lattice of maximal rank in $\mathrm{H}_1(E(\mathbb{C}), \mathbb{R}) \cong \mathbb{C}$. From this point of view the l^n-torsion on E is $\frac{1}{l^n}\Lambda$, so that the l-adic Tate module of E is $\Lambda \otimes_{\mathbb{Z}} \mathbb{Z}_l$; thus we can regard the l-adic Tate module of E as $\mathrm{H}_1(E(\mathbb{C}), \mathbb{Z}_l)$. Of course, we have lost any trace of Galois actions, but let us not concern ourselves with this at the moment.

Now consider the complex surface $E(\mathbb{C}) \times E(\mathbb{C})$ and its second homology group $\mathrm{H}_2(E(\mathbb{C}) \times E(\mathbb{C}), \mathbb{Z})$. The Künneth theorem shows that this group surjects onto

$$\mathrm{H}_1(E(\mathbb{C}), \mathbb{Z}) \otimes_{\mathbb{Z}} \mathrm{H}_1(E(\mathbb{C}), \mathbb{Z}).$$

Tensoring this with $\mathbb{Z}_l$ yields

$$\mathrm{H}_1(E(\mathbb{C}), \mathbb{Z}_l) \otimes_{\mathbb{Z}_l} \mathrm{H}_1(E(\mathbb{C}), \mathbb{Z}_l)$$

which by the above discussion contains $\mathrm{Sym}^2 T_l E = T$ as a direct summand. Combining all of this we see that we have a map

$$\mathrm{H}_2(E(\mathbb{C}) \times E(\mathbb{C}), \mathbb{Z}_l) \to T.$$

By Poincaré duality, we can also regard this as a map

$$\mathrm{H}^2(E(\mathbb{C}) \times E(\mathbb{C}), \mathbb{Z}_l)^\vee \to T \tag{9}$$

where the $\vee$ denotes the Poincaré dual. We will return to this map later.

Now consider a pair (C, f). The curve C has real dimension 2, and therefore determines in a natural way an element of the homology group

$$\mathrm{H}_2(E(\mathbb{C}) \times E(\mathbb{C}), \mathbb{Z}).$$

If f has non-trivial divisor on C, then we could also use this divisor (which has real dimension 0) to determine an element of

$$\mathrm{H}_0(E(\mathbb{C}) \times E(\mathbb{C}), \mathbb{Z}).$$

In our case, however, f has trivial divisor. In this situation, the pair (C, f) does not determine anything of dimension 0, but still somehow contains more information than just C itself. This extra information has the effect of cutting down the relevant dimension by 1, so that (C, f) determines an element of

$$\mathrm{H}_1(E(\mathbb{C}) \times E(\mathbb{C}), \mathbb{Z}).$$

(This is where the algebraic K-theory comes in; K-theory is very good at keeping track of dimensions like this which may not make all that much sense purely geometrically.) Applying Poincaré duality we can regard this as an element of $\mathrm{H}^3(E(\mathbb{C}) \times E(\mathbb{C}), \mathbb{Z})^\vee$. So far, then, we have a map

$$\mathcal{C}(E \times E) \to \mathrm{H}^3(E(\mathbb{C}) \times E(\mathbb{C}), \mathbb{Z})^\vee. \tag{10}$$

(At this point we should confess that this map turns out to just be the zero map. It will nevertheless serve our motivational purposes.)

Étale cohomology is the algebraic analogue of singular cohomology, and the first miracle of étale cohomology is that the preceding construction can be carried out over $\bar{\mathbb{Q}}$ rather than $\mathbb{C}$, so long as we always use l-adic coefficients. Thus the pair (C, f) should give rise to an element of the dual $\mathrm{H}^3_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l)^\vee$ of the étale cohomology group $\mathrm{H}^3_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l)$; this last cohomology group is isomorphic to $\mathrm{H}^3(E(\mathbb{C}) \times E(\mathbb{C}), \mathbb{Z}_l)$ as an abelian group, but has the advantage of having a Galois action.

In fact, since (C, f) is defined over $\mathbb{Q}$, this element should be Galois invariant, yielding a map

$$\mathcal{C}(E \times E) \to \left(\mathrm{H}^3_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l)^\vee\right)^{G_{\mathbb{Q}}}$$

analogous to (10). Unfortunately, $\mathrm{H}^3_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l)^\vee$ can be shown to have no non-zero Galois invariants, so at the moment all of this work has produced 0.

The second miracle of étale cohomology is that we can carry our construction out over $\mathbb{Q}$, rather than $\bar{\mathbb{Q}}$. Thus (C, f) yields an element of $\mathrm{H}^3_{\text{ét}}(E \times E, \mathbb{Z}_l)^\vee$. This group is no longer isomorphic to any singular cohomology group, but rather is a complicated combination of various $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$-cohomology groups of étale cohomology groups of $E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}$. It admits a natural map via a spectral sequence to

$$\mathrm{H}^0\left(\mathbb{Q}, \mathrm{H}^3_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l)^\vee\right).$$

However, as we said above, this group vanishes, and from this one shows that the spectral sequence yields a map

$$\mathrm{H}^3_{\text{ét}}(E \times E, \mathbb{Z}_l)^\vee \to \mathrm{H}^1\left(\mathbb{Q}, \mathrm{H}^2_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l)^\vee\right).$$

Thus we finally have a map

$$\mathcal{C}(E \times E) \to \mathrm{H}^1\left(\mathbb{Q}, \mathrm{H}^2_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l)^\vee\right).$$

Combining this with the étale analogue of (9), we finally obtain our desired Flach map

$$\sigma : \mathcal{C}(E \times E) \to \mathrm{H}^1(\mathbb{Q}, T).$$

We now discuss the local behavior of σ. Let p be a prime not lying in S. We will define a map

$$d_p : \mathcal{C}(E \times E) \to \mathrm{Div}(E_{\mathbb{F}_p} \times E_{\mathbb{F}_p})$$

where $\mathrm{Div}(E_{\mathbb{F}_p} \times E_{\mathbb{F}_p})$ is the group of Weil divisors (defined over $\mathbb{F}_p$) on the non-singular surface $E_{\mathbb{F}_p} \times E_{\mathbb{F}_p}$. (Recall that a Weil divisor on a surface is a formal sum of curves on the surface.) d_p is the map which will go on the top of (8) above. To define $d_p(C, f)$, first consider the reduction of C modulo p. (Technically, by the reduction of C modulo p we mean the base change to $\mathbb{F}_p$ of the scheme-theoretic closure of C in a model of E over $\mathbb{Z}_p$. However, one loses nothing by simply regarding this as considering the equations defining C modulo p.) This may well have several geometric components $C_1, \ldots, C_n$, even if in characteristic 0 it did not. We claim that if the function f has a zero or pole at any point of a component C_i, then it has a zero or pole of the same order along the entire component C_i. (Actually, poles and zeros can combine at the points where the C_i intersect, but this doesn't matter much.) The idea is that if f had a zero or pole at an isolated point of C_i, then we could lift this to a zero or pole of f over $\mathbb{Q}$, which is not possible by the definition of $\mathcal{C}(E \times E)$. Given this, for any component C_i we can let m_i be the order of the zero or pole of f on C_i: of course, we could have $m_i = 0$. We define

$$d_p(C, f) = \sum m_i C_i.$$

The last thing we will need to connect the geometry to the behavior of cohomology classes is a map from $\mathrm{Div}(E_{\mathbb{F}_p} \times E_{\mathbb{F}_p})$ to $\mathrm{H}^1_s(\mathbb{Q}_p, T)$. Recall that by Lemma 10.9 we have $\mathrm{H}^1_s(\mathbb{Q}_p, T) \cong T(-1)^{G_{\mathbb{F}_p}}$. This in turn is isomorphic to $\mathrm{End}^0_{G_{\mathbb{F}_p}}(T_l E)$, by Lemma 10.2. That is, the singular quotient at p corresponds precisely to trace zero $G_{\mathbb{F}_p}$-equivariant maps of the l-adic Tate module of E. The map we seek is a standard one in algebraic geometry, called a *cycle class map*:

$$s : \mathrm{Div}(E_{\mathbb{F}_p} \times E_{\mathbb{F}_p}) \to \mathrm{End}_{G_{\mathbb{F}_p}}(T_l E) \to \mathrm{End}^0_{G_{\mathbb{F}_p}}(T_l E)$$

(the second map is simply projection onto a direct summand). We will not give a general description of the map $\mathrm{Div}(E_{\mathbb{F}_p} \times E_{\mathbb{F}_p}) \to \mathrm{End}_{G_{\mathbb{F}_p}}(T_l E)$, except in the special case we will need. (In fact, if $\mathrm{End}_{G_{\mathbb{F}_p}}(T_l E)$ is given an appropriate cohomological interpretation, then the cycle class map is really just an algebraic version of part of the algebraic topology construction discussed above.) Let $g : E_{\mathbb{F}_p} \to E_{\mathbb{F}_p}$ be some map. Let Γ_g be the graph of g, by which we mean the image of the product map

$$\mathrm{id} \times g : E_{\mathbb{F}_p} \to E_{\mathbb{F}_p} \times E_{\mathbb{F}_p}.$$

Then Γ_g has codimension 1 in $E_{\mathbb{F}_p} \times E_{\mathbb{F}_p}$, and thus is an element of $\mathrm{Div}(E_{\mathbb{F}_p} \times E_{\mathbb{F}_p})$. The image of Γ_g under s is nothing other than the endomorphism of $T_l E$ induced by the map g (or more honestly its projection onto the trace zero direct summand). This endomorphism is $G_{\mathbb{F}_p}$-equivariant since g is defined over $\mathbb{F}_p$.

We are finally in a position to state the fundamental local description of the map σ: for every prime p not in S, there is a commutative diagram

(11)
$$\begin{tikzcd}
\mathcal{C}(E \times E) \arrow[rrr, "d_p"] \arrow[d, "\sigma"'] & & & \mathrm{Div}(E_{\mathbb{F}_p} \times E_{\mathbb{F}_p}) \arrow[d, "s"] \\
\mathrm{H}^1(\mathbb{Q}, T) \arrow[r] & \mathrm{H}^1(\mathbb{Q}_p, T) \arrow[r] & \mathrm{H}^1_s(\mathbb{Q}_p, T) \arrow[r, "\cong"] & \mathrm{End}^0_{G_{\mathbb{F}_p}}(T_l E)
\end{tikzcd}$$

"All" we have to do to generate our Flach system, then, is to exhibit appropriate elements of $\mathcal{C}(E \times E)$ and compute their image in $\mathrm{End}^0_{G_{\mathbb{F}_p}}(T_l E)$ via the clockwise maps. Unfortunately, in general this would be extremely difficult. That is, given an arbitrary surface S, it is a very difficult problem in algebraic geometry to write down many particularly useful curves on S. In order to do this in our case we will have to take full advantage of the fact that E is modular.

The geometry of modular curves

In this section we review the facts we will need about the geometry of modular curves. For a more thorough treatment and references to the standard sources, see [**41**, Part II].

Let N be a positive integer and let $\Gamma_0(N)$ be the usual congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$. Recall that orbits for the $\Gamma_0(N)$-action on the upper half plane $\mathfrak{H}$

correspond to isomorphism classes of pairs of complex elliptic curves E and cyclic subgroups of $E(\mathbb{C})$ of order N. Furthermore, the orbit space $\Gamma_0(N)\backslash\mathfrak{H}$ can be given the structure of a non-compact Riemann surface. We will write $Y_0(N)^{\mathrm{an}}$ for this Riemann surface. (The "an" is for analytic.) We can also compactify $Y_0(N)^{\mathrm{an}}$ by adding a finite number of points called the *cusps*; we write $X_0(N)^{\mathrm{an}}$ for the resulting compact Riemann surface.

It is a classical fact of algebraic geometry that every compact Riemann surface can be realized as a nonsingular projective complex algebraic curve; that is, there is an algebraic curve $X_0(N)_{\mathbb{C}}$ over the complex numbers such that the $\mathbb{C}$-valued points of $X_0(N)_{\mathbb{C}}$ recover $X_0(N)^{\mathrm{an}}$. A fundamental fact in the arithmetic theory of modular curves is that this curve can actually be canonically defined over the rational numbers $\mathbb{Q}$. That is, the polynomial equations which define $X_0(N)_{\mathbb{C}}$ can be canonically chosen in such a way that all of the coefficients are rational. We will write $X_0(N)_{\mathbb{Q}}$ for this nonsingular projective rational algebraic curve. The subscheme of cusps of $X_0(N)_{\mathbb{Q}}$ is canonically defined over $\mathbb{Q}$, so $Y_0(N)^{\mathrm{an}}$ can also be realized as the complex points of a nonsingular algebraic curve $Y_0(N)_{\mathbb{Q}}$; of course, $Y_0(N)_{\mathbb{Q}}$ is only quasi-projective.

Now that we have a model for our modular curve over $\mathbb{Q}$, we can ask how the equations for $X_0(N)_{\mathbb{Q}}$ reduce modulo primes p. The fundamental result is that $X_0(N)_{\mathbb{Q}}$ reduces to a nonsingular projective algebraic curve over $\mathbb{F}_p$ for every prime p which does not divide N. (Perhaps the most compact way to say this is that $X_0(N)_{\mathbb{Q}}$ is the generic fiber of a canonical smooth proper $\mathbb{Z}[1/N]$-scheme. This description allows one to work with $X_0(N)_{\mathbb{Q}}$ and all of its reductions simultaneously, which is often convenient; nevertheless, we will content ourselves below with working one prime at a time.) From now on we will just write $X_0(N)$ when it does not matter what field (of characteristic 0 or p not dividing N) we are working over; the behavior of the modular curves over the various fields is virtually identical.

If p is a prime which divides N, then $X_0(N)$ will pick up singularities over $\mathbb{F}_p$, but at least in the case that p divides N exactly once it is possible to very explicitly describe $X_0(N)_{\mathbb{F}_p}$: it has two irreducible components, each isomorphic to $X_0(N/p)_{\mathbb{F}_p}$ (which is nonsingular by what we said above), and they intersect transversally at a finite number of points.

The other question one might ask about our models $Y_0(N)$ is whether or not they still classify pairs of elliptic curves and cyclic subgroups of order N. Let us say that a pair of an elliptic curve E and a cyclic subgroup C of order N is *defined over a field* K (of characteristic 0 or p not dividing N) if E is defined over K and if C is mapped to itself under the action of every element of the absolute Galois group of K. (Note that we do not require that C is fixed pointwise by the Galois group, which is equivalent to C actually lying in $E(K)$.) We could hope that the K-rational points of $Y_0(N)$ correspond to the pairs (E, C) as above which are defined over K. If this were the case, we would call these modular curves *fine moduli spaces* for classifying such pairs. Unfortunately, this is simply not true. This is well-known in the case $N = 1$: $Y_0(1)_{\mathbb{Q}}$ is isomorphic to the affine line $\mathbb{A}^1_{\mathbb{Q}}$ via the j-invariant, but the j-invariant is not enough to determine the isomorphism class of elliptic curves over fields which are not algebraically closed. The problem in the general case is similar, although somewhat less severe.

However, these modular curves are at least *coarse moduli spaces*. We will not give the technical definition of this, except to say that it means that the modular curves are as close to fine moduli spaces for classifying pairs as it is possible for them to be. In particular, every pair of an elliptic curve E and a cyclic subgroup C of order N, defined over a field K, does give rise to a point of $Y_0(N)(K)$.

In passing, we should note that it is also possible to give a modular interpretation to the cusps of $X_0(N)$ in terms of *generalized elliptic curves.* We will not give a description of this here; it is, however, extremely useful for computations involving the cusps.

We can use our moduli interpretations to define various maps between modular curves. (Actually, the description we gave above is not enough to actually make rigorous the upcoming definitions. However, we assure the reader that these constructions can be made entirely rigorous with a more thorough understanding of the "moduli interpretation" of $X_0(N)$.) Let K be a field as above, and let r be a prime not dividing N. A pair of an elliptic curve E and a cyclic subgroup C of order Nr, defined over K, gives rise in a natural way to a corresponding pair with respect to N: take the same elliptic curve E and take the unique cyclic subgroup of C of order N; call it C_N. We define

$$j_r : X_0(Nr) \to X_0(N)$$

to be the corresponding map; that is, it sends the point corresponding to the pair (E, C) to the point corresponding to the pair (E, C_N). (As we said above, this requires more work to be a rigorous definition, but it is possible to give it a better interpretation.) The fact that we can make this definition on the level of points for any field K (of the appropriate characteristics) insures that j_r can actually be defined over any of the fields $\mathbb{Q}$ or $\mathbb{F}_p$ for p not dividing N. Note that there is a slight subtlety (which we will ignore) for the field $\mathbb{F}_r$, as there we have not said anything about the moduli interpretation of $X_0(Nr)$.

There is a second way to obtain a map between these modular curves: let C_r be the unique subgroup of C of order r, and now send the pair (E, C) to the pair $(E/C_r, C/C_r)$. This gives rise to a second map

$$j'_r : X_0(Nr) \to X_0(N).$$

We define the r^{th} *Hecke correspondence* T_r on $X_0(N)$ to be the image of the product map

$$j_r \times j'_r : X_0(Nr) \to X_0(N) \times X_0(N).$$

It can be shown that T_r is a singular curve which is birational to $X_0(Nr)$. Furthermore, it is possible to give a very precise description of T_r in characteristic r. Recall that a curve (or more generally any scheme) over $\mathbb{F}_r$ has a Frobenius endomorphism induced by the r-power map on the function field. Define $\Gamma \subseteq X_0(N)_{\mathbb{F}_r} \times X_0(N)_{\mathbb{F}_r}$ as the graph of the Frobenius map $\mathrm{Fr} : X_0(N)_{\mathbb{F}_r} \to X_0(N)_{\mathbb{F}_r}$; that is, Γ is the image of the product map

$$\mathrm{id} \times \mathrm{Fr} : X_0(N)_{\mathbb{F}_r} \to X_0(N)_{\mathbb{F}_r} \times X_0(N)_{\mathbb{F}_r}.$$

We define the *Verschiebung* $\Gamma' \subseteq X_0(N)_{\mathbb{F}_r} \times X_0(N)_{\mathbb{F}_r}$ to be the image of the transpose map

$$\mathrm{Fr} \times \mathrm{id} : X_0(N)_{\mathbb{F}_r} \to X_0(N)_{\mathbb{F}_r} \times X_0(N)_{\mathbb{F}_r}.$$

Note that $T_{r,\mathbb{F}_r}$, Γ and Γ' are all of codimension one in the surface $X_0(N)_{\mathbb{F}_r} \times X_0(N)_{\mathbb{F}_r}$, and thus are divisors. The *Eichler-Shimura relation* states that there is an equality

$$T_{r,\mathbb{F}_r} = \Gamma + \Gamma'.$$

of divisors on $X_0(N)_{\mathbb{F}_r} \times X_0(N)_{\mathbb{F}_r}$. In fact, each of Γ and Γ' is the isomorphic image of one of the irreducible components of $X_0(Nr)_{\mathbb{F}_r}$ (both isomorphic to $X_0(N)_{\mathbb{F}_r}$) which we discussed above. This relation will be the key to our computations below.

Some modular units

Modular curves will be of use to us since the surface $X_0(N) \times_{\mathbb{Q}} X_0(N)$ has all of the fairly explicit divisors T_r. Our basic plan at this point is to find an appropriate rational function f_r on $T_{r,\mathbb{Q}}$ (for each r not dividing N) such that $(T_{r,\mathbb{Q}}, f_r) \in \mathcal{C}(X_0(N)_{\mathbb{Q}} \times X_0(N)_{\mathbb{Q}})$. We will then map the pair $(T_{r,\mathbb{Q}}, f_r)$ via $\phi \times \phi$ into $\mathcal{C}(E_{\mathbb{Q}} \times E_{\mathbb{Q}})$, and then by σ into $\mathrm{H}^1(\mathbb{Q}, T)$. If we can also arrange for f_r to have trivial divisor away from characteristic r, then our geometric description of the local behavior of σ will show that f_r maps to 0 in each $\mathrm{H}^1_s(\mathbb{Q}_p, T)$ for p not dividing rlN (recall that our geometric description of the Flach map broke at the primes in S) and we will be well on our way to constructing the desired Flach system.

Since $T_{r,\mathbb{Q}}$ is birational to $X_0(Nr)_{\mathbb{Q}}$, to exhibit rational functions on $T_{r,\mathbb{Q}}$ we can work instead on $X_0(Nr)_{\mathbb{Q}}$. Recall that rational functions on $X_0(Nr)_{\mathbb{Q}}$ are modular functions of level Nr and weight 0, with coefficients in $\mathbb{Q}$. We will define such a function as the ratio of two modular forms of the same weight.

We want f_r to have trivial divisor over $\mathbb{Q}$, so we should start with modular forms with especially simple divisors. Perhaps the best known is $\Delta(z)$, the unique cusp form of level 1 and weight 12. Δ is initially defined on $X_0(1)_{\mathbb{Q}}$, and has a simple zero at the unique cusp ∞ and no other zeros or poles. (Δ is a pluri-canonical form, not a function, so it can have more zeros than poles.) Pulling back Δ via the natural map $\pi : X_0(N)_{\mathbb{Q}} \to X_0(1)_{\mathbb{Q}}$ yields a form $\pi^*\Delta$ on $X_0(N)_{\mathbb{Q}}$ (this is really nothing more than reinterpreting Δ as having level N) which will have zeros of various orders at the cusps and no other zeros or poles:

$$\mathrm{div}_{X_0(N)_{\mathbb{Q}}} \Delta = \sum_{\text{cusps } c_i} n_i c_i$$

for some integers n_i.

We will pull back Δ to $X_0(Nr)_{\mathbb{Q}}$ via the two maps j_r and j'_r. In order to understand the divisors of these forms, we need to know how the cusps behave under these maps. The basic fact is that the preimage of a cusp c_i of $X_0(N)_{\mathbb{Q}}$ under j_r consists of two cusps $c_{i,1}$ and $c_{i,2}$ of $X_0(Nr)_{\mathbb{Q}}$; j_r is unramified at $c_{i,1}$ and ramified of degree r at $c_{i,2}$. Under j'_r we have the opposite behavior: $c_{i,1}$ and $c_{i,2}$ are again the only two points in the preimage of c_i, but now $c_{i,2}$ is unramified and $c_{i,1}$ is ramified of degree r. Combining all of this, we find that

$$\mathrm{div}_{X_0(Nr)_{\mathbb{Q}}} j_r^*\Delta = \sum n_i c_{i,1} + r n_i c_{i,2}$$
$$\mathrm{div}_{X_0(Nr)_{\mathbb{Q}}} j_r'^*\Delta = \sum r n_i c_{i,1} + n_i c_{i,2}.$$

Both of these forms have weight 12, since Δ does, so their ratio is a rational function f_r on $X_0(Nr)_{\mathbb{Q}}$ with divisor

$$\mathrm{div}_{X_0(Nr)_{\mathbb{Q}}} f_r = \sum (1-r) n_i (c_{i,1} - c_{i,2}).$$

We now think of f_r as a rational function on the singular curve $T_{r,\mathbb{Q}}$, which is birational to $X_0(Nr)_{\mathbb{Q}}$. As we said above, both $c_{i,1}$ and $c_{i,2}$ map to c_i under j_r and

j_r'. Since $T_{r,\mathbb{Q}}$ is the image of $X_0(Nr)_{\mathbb{Q}}$ under the map $j_r \times j_r'$, the divisor of f_r on $T_{r,\mathbb{Q}}$ is

$$\operatorname{div}_{T_{r,\mathbb{Q}}} f_r = \sum (1-r) n_i \big((j_r(c_{i,1}), j_r'(c_{i,1})) - (j_r(c_{i,2}), j_r'(c_{i,2})) \big) = 0.$$

Thus $(T_{r,\mathbb{Q}}, f_r) \in \mathcal{C}(X_0(N)_{\mathbb{Q}} \times X_0(N)_{\mathbb{Q}})$, as desired.

We can now define the Flach classes $c_r \in \mathrm{H}^1(\mathbb{Q}, T)$: we first map $(T_{r,\mathbb{Q}}, f_r)$ to $\mathcal{C}(E_{\mathbb{Q}} \times E_{\mathbb{Q}})$ via $(\phi \times \phi)_*$. That is, let $T'_{r,\mathbb{Q}}$ be the image of $T_{r,\mathbb{Q}}$ under $\phi \times \phi$ and let f_r' be the rational function on $T_{r,\mathbb{Q}}$ induced by f_r. (f_r' is really the norm of f_r in the finite extension of function fields $k(X_0(N)_{\mathbb{Q}})/k(E_{\mathbb{Q}})$ induced by ϕ.) One easily checks that

$$\operatorname{div}_{T'_{r,\mathbb{Q}}} f_r' = (\phi \times \phi)_* \operatorname{div}_{T_{r,\mathbb{Q}}} f_r = 0,$$

so $(T'_{r,\mathbb{Q}}, f_r') \in \mathcal{C}(E_{\mathbb{Q}} \times E_{\mathbb{Q}})$. We now map this pair to $\mathrm{H}^1(\mathbb{Q}, T)$ via the Flach map σ. This is the class c_r. Note that we defined these classes for all r not dividing N, even though we only need them for good r. (In fact, one can even define classes for r dividing N with some care.)

Local behavior of the c_r

To complete the construction of our Flach system, we need to analyze the local behavior of the classes c_r in $\mathrm{H}^1_s(\mathbb{Q}_p, T)$ for all p. Specifically, we need to show that they map to 0 for all $p \neq r$ and we need to compute them explicitly for $p = r$. We do this using our description in terms of divisors and cycle classes. We distinguish several cases; for simplicity, we assume that r is good, although this is not critical to these computations.

p does not divide Nlr

This is the easiest case. $d_p(T'_{r,\mathbb{Q}}, f_r')$ is just the divisor of f_r' on $T'_{r,\mathbb{F}_p}$, and the analysis of the preceding section of the divisor of f_r over $\mathbb{Q}$ goes through in exactly the same way over $\mathbb{F}_p$. Thus

$$\operatorname{div}_{T'_{r,\mathbb{F}_p}} f_r' = (\phi \times \phi)_* \operatorname{div}_{T_{r,\mathbb{F}_p}} f_r = 0,$$

so $d_p(T'_{r,\mathbb{Q}}, f_r') = 0$. Commutativity of the diagram (11) now shows that c_r maps to 0 in $\mathrm{H}^1_s(\mathbb{Q}_p, T)$, since it is the image of the pair $(T'_{r,\mathbb{Q}}, f_r')$ which already maps to 0 in $\mathrm{Div}(E_{\mathbb{F}_p} \times E_{\mathbb{F}_p})$. In particular, there is no need to know anything at all about the map s in this case.

p divides N

This is the case of bad reduction of E and the local diagram we used in the first case does not hold in this setting. Flach uses two different arguments to handle this case. If E has potentially multiplicative reduction at p, then one can give a very explicit description of the $G_{\mathbb{Q}_p}$-action on V, and one can compute that $\mathrm{H}^1_f(\mathbb{Q}_p, V) = \mathrm{H}^1(\mathbb{Q}_p, V)$. It follows that $\mathrm{H}^1_s(\mathbb{Q}_p, T) = 0$, so that there is no local condition to check! If E has potentially good reduction at p, Flach mimics the argument above in the case of good reduction, using the Néron model of E; see [**50**, pp. 324–325] and [**51**, Section 5.5.2 and Section 6].

$p = l$

In this case we again do not have the local diagram to fall back on. Flach uses results of Faltings to conclude that $\mathrm{res}_l\, c_r$ lies in $\mathrm{H}^1_f(\mathbb{Q}_l, T)$; see [**50**, pp. 322-324].

$p = r$

This is the key computation. Recall that the Eichler-Shimura relation says that $T_{r,\mathbb{F}_r}$ can be written as a sum $\Gamma + \Gamma'$ of the graph of Frobenius and the Verschiebung. We will work with each piece separately.

We begin with Γ:

$$\begin{tikzcd}
 & & X_0(N)_{\mathbb{F}_r} \\
X_0(Nr)_{\mathbb{F}_r} \arrow[r, dashed] \arrow[urr, "j_r"] \arrow[drr, "j_r'"'] & \Gamma \arrow[r] \arrow[ur, "\mathrm{id}"'] \arrow[dr, "\mathrm{Fr}"] & X_0(N)_{\mathbb{F}_r} \times X_0(N)_{\mathbb{F}_r} \arrow[u] \arrow[d] \\
 & & X_0(N)_{\mathbb{F}_r}
\end{tikzcd}$$

(Only one of the irreducible components of $X_0(Nr)_{\mathbb{F}_r}$ maps to Γ, which is why we have used a dotted line there.) The function on Γ corresponding to $j_r^*\Delta$ is just the pull back of Δ under the identity map; thus $\mathrm{div}_\Gamma\, j_r^*\Delta$ will just be the usual linear combination of points of Γ corresponding to cusps. The function on Γ corresponding to $j_r'^*\Delta$ is the pull back of Δ under Fr. Fr is purely inseparable, and purely inseparable maps are trivial on differentials; see [**146**, Chapter 2, Proposition 4.2]. Thus $\mathrm{Fr}^*\,\Delta = j_r'^*\Delta$ will pick up a zero on Γ in addition to the usual cuspidal divisor. One can check that this zero has order 6 (essentially because Δ has weight $12 = 2 \cdot 6$). As always, the cuspidal parts of the divisor cancel, and we conclude that

$$\mathrm{div}_\Gamma\, f_r = \mathrm{div}_\Gamma\, j_r^*\Delta - \mathrm{div}_\Gamma\, j_r'\Delta = -6\Gamma.$$

The computation for Γ' is virtually identical, except that id and Fr are interchanged. Thus

$$\mathrm{div}_{\Gamma'}\, f_r = 6\Gamma'.$$

We conclude that

$$\mathrm{div}_{T_{r,\mathbb{F}_r}}\, f_r = 6(\Gamma' - \Gamma).$$

The next step is to push our whole construction forward to $E \times E$. If we let Γ_E and Γ_E' be the image of

$$\mathrm{id} \times \mathrm{Fr} : E_{\mathbb{F}_r} \to E_{\mathbb{F}_r} \times E_{\mathbb{F}_r}$$

$$\mathrm{Fr} \times \mathrm{id} : E_{\mathbb{F}_r} \to E_{\mathbb{F}_r} \times E_{\mathbb{F}_r}$$

respectively, then it is clear that $\phi \times \phi$ maps Γ onto Γ_E and Γ' onto Γ_E'. Since each point of Γ_E and Γ_E' is the image of $\deg\phi$ points of Γ and Γ', we have the equalities

$$(\phi \times \phi)_*\Gamma = (\deg\phi)\Gamma_E$$
$$(\phi \times \phi)_*\Gamma' = (\deg\phi)\Gamma_E'$$

as divisors. We conclude finally that

$$d_p(\phi \times \phi)_*(T_{r,\mathbb{Q}}, f_r) = 6(\deg\phi)(\Gamma_E' - \Gamma_E) \in \mathrm{Div}(E_{\mathbb{F}_r} \times E_{\mathbb{F}_r}).$$

We now need to compute the image of this under the cycle class map s. Our description of s shows that Γ_E, being the graph of Frobenius at r maps to precisely the endomorphism of $T_l E$ given by Fr_r. (This is well defined since $T_l E$ is unramified at r.) r is assumed to be good, so the proof of Lemma 10.9 shows that we can choose a basis x, y for $T_l E$ with respect to which Fr_r has matrix

$$\begin{pmatrix} u & 0 \\ 0 & v \end{pmatrix}$$

with $u \equiv -v \equiv 1 \pmod{l}$ and $uv = r$. This matrix is the image of Γ_E in $\mathrm{End}_{G_{\mathbb{F}_p}}(T_l E)$.

To make the corresponding calculation for Γ'_E we will need to reinterpret it as a graph. Since Fr has degree r, there is a map $V : E_{\mathbb{F}_r} \to E_{\mathbb{F}_r}$ with the property that $V \circ \mathrm{Fr} = \mathrm{Fr} \circ V$ is the multiplication by r map on $E_{\mathbb{F}_r}$; see [**146**, Chapter 3, Section 6]. Γ'_E is the image of the map

$$\mathrm{Fr} \times \mathrm{id} : E_{\mathbb{F}_r} \to E_{\mathbb{F}_r} \times E_{\mathbb{F}_r}.$$

If we precompose with the map $V : E_{\mathbb{F}_r} \to E_{\mathbb{F}_r}$, which is a surjective map of degree r, the literal image will not change, but each point will pick up a multiplicity of r. Thus the image of the map $\mathrm{Fr} \circ V \times V = r \times V$ is $r\Gamma'_E$. We claim that we can cancel the two r's, which leaves us with the fact that Γ'_E is the graph of V. The easiest way to do this is to pretend for the moment that multiplication by r has an inverse r^{-1} on E. (Of course, this is absurd, but it is somewhat less absurd when one does the entire computation in the range $\mathrm{End}(T_l E)$, where r is invertible.) Then an argument similar to the one above for precomposing with V shows that the image of $r \times V$ is r^2 times the image of $\mathrm{id} \times V r^{-1}$. This means that $s(r\Gamma'_E) = rs(\Gamma'_E)$ is the same as $r^2 V r^{-1}$, where now V is regarded as an endomorphism of $T_l E$. In other words, $s(\Gamma'_E)$ is just the endomorphism induced by V.

Since $V \circ \mathrm{Fr} = r$, this implies that the cycle class of Γ'_E has matrix

$$r \begin{pmatrix} u & 0 \\ 0 & v \end{pmatrix}^{-1} = \begin{pmatrix} v & 0 \\ 0 & u \end{pmatrix}.$$

We conclude that $6(\deg \phi)(\Gamma'_E - \Gamma_E)$ maps to

$$6(\deg \phi) \begin{pmatrix} (v-u) & 0 \\ 0 & (u-v) \end{pmatrix}$$

in $\mathrm{End}_{G_{\mathbb{F}_r}}(T_l E)$, and even in $\mathrm{End}^0_{G_{\mathbb{F}_r}}(T_l E)$ since this matrix already has trace 0. This, then, is the image of c_r in $\mathrm{H}^1_s(\mathbb{Q}_r, T)$.

Recall that $\mathrm{H}^1_s(\mathbb{Q}_r, T) \cong \mathrm{End}^0_{G_{\mathbb{F}_r}}(T_l E)$ is a free $\mathbb{Z}_l$-module of rank 1. One easily checks that the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

is a generator of this module. Combined with our computation above, we find that $6(\deg \phi)(v-u)$ annihilates

$$\mathrm{H}^1_s(\mathbb{Q}_r, T)/\mathbb{Z}_l \cdot \mathrm{res}_r\, c_r.$$

But 6 is an l-adic unit, and $v - u \equiv -2 \pmod{l}$, so $v - u$ is as well. We conclude that $\deg \phi$ annihilates this module, and thus that the c_r form a Flach system of depth $\deg \phi$ for T. This concludes the proof of Theorem 10.13, and with it the proof of Theorem 10.1.

Appendix A: On the local Galois invariants of $E[l]\otimes E[l]$

The purpose of this appendix is to prove the following result.

Theorem 10.14. *Let E be an elliptic curve over $\mathbb{Q}$ without complex multiplication. Let $\phi : X_0(N) \to E$ be a modular parameterization of E and let S_0 be the set of places of $\mathbb{Q}$ at which E has bad reduction. Then the set of primes such that*

- *E has good reduction at l;*
- *The l-adic representation $\rho : G_{\mathbb{Q},S_0\cup\{l\}} \to \mathrm{GL}_2(\mathbb{Z}_l)$ is surjective;*
- *For all $p\in S_0$, $E[l]\otimes E[l]$ has no $G_{\mathbb{Q}_p}$-invariants;*
- *$E[l]\otimes E[l]$ has no $G_{\mathbb{Q}_l}$-invariants;*
- *l does not divide the degree of ϕ;*

has density 1 *in the set of all primes.*

Of course, the first and fifth conditions are obviously satisfied for almost all l. That the second condition is satisfied for almost all l is a result of Serre; see [**132**]. The new content is in the third and fourth conditions. We will show that the third condition is also satisfied for almost all l, and that the fourth condition is satisfied for a set of primes of density 1.

Recall that by the Weil pairing we can write

$$E[l]\otimes E[l] \cong \mu_l \oplus \mathrm{Sym}^2 E[l].$$

The analysis of $\mathbb{Q}_p$-rational points on the first of these factors is immediate from the fact that (for $p\neq 2$) $\mathbb{Q}_p$ contains precisely the $(p-1)$-th roots of unity: it has $\mathbb{Q}_p$-rational points if and only if $p\equiv 1 \pmod{l}$.

We begin the analysis of $\mathrm{Sym}^2 E[l]$ with a modification of the argument of [**20**, Lemma 2.3(i)].

Lemma 10.15. *Let E be an elliptic curve over $\mathbb{Q}_p$ and let l be any prime. Then* $\mathrm{H}^0(\mathbb{Q}_p,\mathrm{Sym}^2 E[l])\neq 0$ *if and only if $E(K)$ has non-trivial l-torsion for some quadratic extension K of $\mathbb{Q}_p$.*

Proof. We first set some notation. Let $\varepsilon : G_{\mathbb{Q}_p} \to \mathbb{Z}_l^*$ be the cyclotomic character; its image has finite index in $\mathbb{Z}_l^*$. Let $\rho : G_{\mathbb{Q}_p} \to \mathrm{GL}(E[l])$ and $\varphi : G_{\mathbb{Q}_p} \to \mathrm{GL}(\mathrm{Sym}^2 E[l])$ be the Galois representations associated to E. By the Weil pairing we have $\det\rho=\varepsilon$.

If x is a K-rational l-torsion point for some quadratic extension K of $\mathbb{Q}_p$, then one checks immediately that $x\otimes x\in E[l]\otimes E[l]$ is $G_{\mathbb{Q}_p}$-invariant, which proves one direction of the lemma.

Suppose, then, that there exists $t\in\mathrm{Sym}^2 E[l]$ such that $\varphi(\tau)t=t$ for all $\tau\in G_{\mathbb{Q}_p}$. Thus 1 is an eigenvalue of $\varphi(\tau)$ for every $\tau\in G_{\mathbb{Q}_p}$.

Now choose $\sigma_0\in G_{\mathbb{Q}_p}$ such that $\varepsilon(\sigma_0)$ is not a root of unity; this is certainly possible since the image of ε has finite index. Let λ and μ be the eigenvalues of $\rho(\sigma_0)$. Then the eigenvalues of $\varphi(\sigma_0)$ are λ^2, $\lambda\mu=\varepsilon(\sigma_0)$ and μ^2. Since one of these is 1 and $\varepsilon(\sigma_0)$ is not a root of unity, we can assume without loss of generality that $\lambda^2=1$.

Set $\sigma=\sigma_0^2$. The eigenvalues of $\rho(\sigma)$ are $\lambda^2=1$ and μ^2. We have $\mu^2\neq 1$ (since $\lambda^2\mu^2=\varepsilon(\sigma_0)^2$ is not a root of unity), so we can choose a basis x,y for $E[l]$ of eigenvectors for $\rho(\sigma)$, with eigenvalues 1 and $\varepsilon(\sigma)$ respectively. $\varepsilon(\sigma)^2$ is still not 1, from which one easily computes (using the basis $x\otimes x$, $x\otimes y+y\otimes x$, $y\otimes y$ of $\mathrm{Sym}^2 E[l]$) that t is a scalar multiple of $x\otimes x$. It follows easily that x is rational over some quadratic extension of $\mathbb{Q}_p$. □

We now state the general analysis of torsion in elliptic curves over local fields, coming from an analysis of the formal group and the component group of the Néron model.

Proposition 10.16. *Let E be an elliptic curve over a finite extension K of $\mathbb{Q}_p$ and assume $p \neq l$. Let k be the residue field of K.*

- *If E has good reduction over K, then $E(K)$ has non-trivial l-torsion only if l divides $\#E(k)$.*
- *If E has non-split multiplicative reduction over K, then $E(K)$ has non-trivial l-torsion only if $p \equiv 1 \pmod{l}$ or $l \leq 3$.*
- *If E has split multiplicative reduction over K, then $E(K)$ has non-trivial l-torsion only if $p \equiv 1 \pmod{l}$ or $l \leq 11$.*
- *If E has additive reduction over K, then $E(K)$ has non-trivial l-torsion only if $l \leq 3$.*

Proof. By [**146**, Proposition VII.2.1] there is an exact sequence

$$0 \to E_1(\mathfrak{m}_K) \to E_0(K) \to \tilde{E}_{ns}(k) \to 0$$

where E_1 is the formal group of E, $\mathfrak{m}_K$ is the maximal ideal of the ring of integers of K, $E_0(K)$ is the set of points of $E(K)$ with non-singular reduction and $\tilde{E}_{ns}(k)$ are the non-singular points of the reduction. By [**146**, Proposition IV.3.2(b)], $E_1(K)$ has no non-trivial l-torsion, so any l-torsion in $E(K)$ must appear in $E(K)/E_0(K)$ or $\tilde{E}_{ns}(k)$. The proposition now follows from the determination of $\tilde{E}_{ns}(k)$ in the various cases (see [**146**, Proposition VII.5.1]) and the analysis of the component group of the Néron model of E (see [**146**, Theorem VII.6.1] and use that the minimal discriminant has valuation at most 11). □

An entirely similar argument yields the following result for the case $p = l$.

Proposition 10.17. *Let E be an elliptic curve over a quadratic extension K of $\mathbb{Q}_l$ and assume $l \geq 5$. Let k be the residue field of K.*

- *If E has good reduction over K, then $E(K)$ has non-trivial l-torsion only if l divides $\#E(k)$.*
- *If E has non-split multiplicative reduction over K, then $E(K)$ has no non-trivial l-torsion.*
- *If E has split multiplicative reduction over K, then $E(K)$ has non-trivial l-torsion only if $l \leq 11$.*

Proof. The only difference with Proposition 10.16 is the possibility of torsion in $E_1(K)$, but this is ruled out by [**146**, Theorem IV.6.1] and the fact that the valuation of l in K is at most 2. We can make no statement about the case of additive reduction since then $\tilde{E}_{ns}(k)$ always has l-torsion. □

The last ingredient of the proof of Theorem 10.14 is some additional analysis of l-torsion in the case of good reduction in characteristic l. Note that if $K/\mathbb{Q}_l$ is a quadratic extension, then the residue field k is either $\mathbb{F}_l$ or $\mathbb{F}_{l^2}$.

Consider first the case that $k = \mathbb{F}_l$. Then by the Riemann hypothesis for elliptic curves over finite fields (see [**146**, Theorem V.1.1]) we know that

$$-2\sqrt{l} \leq \#E(\mathbb{F}_l) - l - 1 \leq 2\sqrt{l}.$$

It follows easily that for $l \geq 7$ the only way to have l divide $\#E(\mathbb{F}_l)$ is to have $\#E(\mathbb{F}_l) = l$.

Now consider the case $k = \mathbb{F}_{l^2}$. This time the Riemann hypothesis shows that the only way to have l divide $\#E(\mathbb{F}_{l^2})$ is to have

$$\#E(\mathbb{F}_{l^2}) \in \{l^2 - l, l^2, l^2 + l, l^2 + 2l\}.$$

Let α, β be the eigenvalues of Frobenius at l acting on the p-adic Tate module of E for some $p \neq l$; we have $\alpha\beta = l$. Then by [**146**, Section V.2],

$$\#E(\mathbb{F}_l) = 1 + l - \alpha - \beta$$
$$\#E(\mathbb{F}_{l^2}) = 1 + l^2 - \alpha^2 - \beta^2.$$

Since $\alpha\beta = l$, we have

$$\alpha^2 + 2\alpha\beta + \beta^2 = \alpha^2 + 2l + \beta^2,$$

and we conclude that

$$|l + 1 - \#E(\mathbb{F}_l)| = \sqrt{1 + 2l + l^2 - \#E(\mathbb{F}_{l^2})}.$$

This equation has several consequences. First, suppose that $\#E(\mathbb{F}_{l^2}) = l^2 - l$. Then $3l + 1$ must be a perfect square, say n^2. Thus $3l = n^2 - 1$, which implies that $l = 5$. Similarly, the case $\#E(\mathbb{F}_{l^2}) = l^2$ can not occur, and if $\#E(\mathbb{F}_{l^2}) = l^2 + l$, then $l = 3$. If $\#E(\mathbb{F}_{l^2}) = l^2 + 2l$, then we find that

$$\#E(\mathbb{F}_l) \in \{l, l + 2\}.$$

We now state and prove a more precise version of the unresolved part of Theorem 10.14.

Theorem 10.18. *Let E be an elliptic curve over $\mathbb{Q}$ and let S_0 be the set of places of $\mathbb{Q}$ at which E has bad reduction. Let l be a prime such that*

- *$l \geq 13$;*
- *l does not divide $p - 1$ for any $p \in S_0$;*
- *l does not divide $\#E(\mathbb{F}_p)$ or $\#E(\mathbb{F}_{p^2})$ for any $p \in S_0$;*
- *E has good reduction at l;*
- *$\#E(\mathbb{F}_l)$ is not l or $l + 2$.*

Then $\mathrm{H}^0(\mathbb{Q}_p, E[l] \otimes E[l]) = 0$ for all $p \in S_0 \cup \{l\}$. In particular, the set of such l has density 1 in the set of all primes.

Proof. The second condition insures that μ_l has no $\mathbb{Q}_p$-rational points for any $p \in S_0$. To show that $\mathrm{Sym}^2 E[l]$ has no $\mathbb{Q}_p$-rational points for $p \in S_0$, we must (by Lemma 10.15) show that $E(K)$ has no non-trivial l-torsion for any quadratic extension of $\mathbb{Q}_r$. This possibility is ruled out by the first three conditions and Proposition 10.16. Note that we do need to consider the case of good reduction here, as even though E has bad reduction over $\mathbb{Q}_p$, it may attain good reduction over K.

To show that $\mathrm{H}^0(\mathbb{Q}_l, E[l] \otimes E[l]) = 0$, note first that μ_l has no $\mathbb{Q}_l$-rational points, so we must only consider $\mathrm{Sym}^2 E[l]$. By Proposition 10.17 and the first and fourth hypotheses, it suffices to show that l does not divide $\#E(\mathbb{F}_l)$ or $\#E(\mathbb{F}_{l^2})$, and this follows from the preceding discussion and the fifth hypothesis.

It remains to show that the set of such l has density 1. It is clear that the first four conditions eliminate only finitely many primes l. It is shown as a very special case of [**136**, Theorem 20] that the fifth condition is satisfied for a set of primes of density 1. This completes the proof. □

Appendix B: The definition of the Flach map

In this section we give the formal definition of the Flach map. For conceptual clarity we will work in a more general setting. Let X be a nonsingular projective variety of dimension n, defined over $\mathbb{Q}$.

Let X^p denote the set of irreducible subschemes of X of codimension p. Quillen has constructed a spectral sequence from the filtration by codimension of support:

$$E_1^{pq} = \underset{x\in X^p}{\oplus} K_{-p-q}k(x) \Rightarrow K_{-p-q}(X);$$

here $k(x)$ is the function field of the scheme x and the $K_ik(x)$ are Quillen's K-groups. There is an analogous spectral sequence in étale cohomology:

$$(E_1^{pq})'(\mathcal{F}) = \underset{x\in X^p}{\oplus} \mathrm{H}^{q-p}_{\text{ét}}(\operatorname{Spec} k(x), \mathcal{F}(-p)) \Rightarrow \mathrm{H}^{p+q}(X, \mathcal{F})$$

where $\mathcal{F}$ is some Tate twist of the constant étale sheaf $\mathbb{Z}_l$. For any integer i, these spectral sequence are connected by Chern class maps

$$E_r^{pq} \to (E_r^{p,q+2i})'(\mathbb{Z}_l(i))$$

constructed by Gillet.

Now fix an integer m between 0 and n and assume

- $\mathrm{H}^{2m+1}_{\text{ét}}(X_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(m+1))$ has no $G_{\mathbb{Q}}$-invariants.

This is implied by the Weil conjectures if this cohomology group is torsion free, as is the case when X is a curve or a product of curves. We define the Flach map

$$\sigma_m : E_2^{m,-m-1} \to \mathrm{H}^1(\mathbb{Q}, \mathrm{H}^{2m}_{\text{ét}}(X_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(m+1)))$$

as the composition of three maps. The first is the Chern class map above with $p = m$, $q = -m-1$ and $i = m+1$:

$$E_2^{m,-m-1} \to (E_2^{m,m+1})'(\mathbb{Z}_l(m+1)).$$

The second is an edge map in the étale cohomology spectral sequence above:

$$(E_2^{m,m+1})'(\mathbb{Z}_l(m+1)) \to \mathrm{H}^{2m+1}_{\text{ét}}(X, \mathbb{Z}_l(m+1)).$$

(To see that there really is an edge map from this term, one uses the fact that terms of this spectral sequence below the diagonal always vanish, as is clear from the expression above.) This last group appears in the Hochschild-Serre spectral sequence

$$\mathrm{H}^p(\mathbb{Q}, \mathrm{H}^q_{\text{ét}}(X_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(m+1))) \Rightarrow \mathrm{H}^{p+q}_{\text{ét}}(X, \mathbb{Z}_l(m+1)).$$

Our assumption above that $\mathrm{H}^0(\mathbb{Q}, \mathrm{H}^{2m+1}_{\text{ét}}(X_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(m+1))) = 0$ insures that there is an edge map

$$\mathrm{H}^{2m+1}_{\text{ét}}(X, \mathbb{Z}_l(m+1)) \to \mathrm{H}^1(\mathbb{Q}, \mathrm{H}^{2m}(X_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(m+1)))$$

and gives the last map in the definition of σ_m.

The map considered in the text is a slight variant of this. Take $X = E \times E$ and $m = 1$, so that we have a map

$$\sigma_1 : E_2^{1,-2} \to \mathrm{H}^1(\mathbb{Q}, \mathrm{H}^2_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(2))).$$

We now show how to manipulate these terms to obtain the map

$$\sigma : \mathcal{C}(E \times E) \to \mathrm{H}^1(\mathbb{Q}, \mathrm{Sym}^2 T_l E)$$

of the text. Working from the expression above for the Quillen spectral sequence, we see that $E_2^{1,-2}$ is the cohomology of a sequence

$$K_2 k(E \times E) \to \bigoplus_{x \in (E \times E)^1} k(x)^* \to \bigoplus_{y \in (E \times E)^2} \mathbb{Z}.$$

Quillen computes that the second map is just the divisor map sending a term $f \in k(x)^*$ to $\oplus_{y \in x} m_y$, where m_y is the order of f at y. The kernel of this map is precisely the group $\mathcal{C}(E \times E)$; σ_1 is defined on the quotient of this group by the image of $K_2 k(E \times E)$, so we can also regard it as defined on $\mathcal{C}(E \times E)$ itself. This takes care of the domain.

Next, the Kunneth theorem implies that $\mathrm{H}^2_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(2))$ is torsion free and that there is a projection

$$\mathrm{H}^2_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(2)) \to \mathrm{H}^1_{\text{ét}}(E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(1)) \otimes_{\mathbb{Z}_l} \mathrm{H}^1_{\text{ét}}(E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(1)).$$

The Kummer sequence naturally identifies $\mathrm{H}^1_{\text{ét}}(E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(1))$ with the l-adic Tate module $T_l E$, so projecting onto the symmetric direct summand yields a map

$$\mathrm{H}^2_{\text{ét}}(E_{\bar{\mathbb{Q}}} \times E_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(2)) \to \mathrm{Sym}^2 T_l E$$

which is easily used to finish the definition of the map σ.

Returning to the general case, let us now investigate the local behavior of σ_m. Let p be a prime different from l at which X has good reduction (meaning that $X_{\mathbb{Q}_p}$ is the generic fiber of a proper smooth $\operatorname{Spec} \mathbb{Z}_p$-scheme $\mathfrak{X}$) and make the additional assumption:

- $\mathrm{H}^{2m+1}_{\text{ét}}(X_{\bar{\mathbb{Q}}_p}, \mathbb{Z}_l(m+1))$ has no $G_{\mathbb{Q}_p}$-invariants;

Let $T = \mathrm{H}^{2m}(X_{\bar{\mathbb{Q}}}, \mathbb{Z}_l(m+1))$. Then there is a commutative diagram

$$\begin{tikzcd}
E_2^{m,-m-1} \arrow[r, "\mathrm{div}_{\mathbb{F}_p}"] \arrow[d, "\sigma_m"] & A^m \mathfrak{X}_{\mathbb{F}_p} \arrow[d, "s"] \\
\mathrm{H}^1(\mathbb{Q}, T) \arrow[d] & \mathrm{H}^{2m}_{\text{ét}}(\mathfrak{X}_{\bar{\mathbb{F}}_p}, \mathbb{Z}_l(m))^{G_{\mathbb{F}_p}} \arrow[d, "\cong"] \\
\mathrm{H}^1(\mathbb{Q}_p, T) \arrow[r] & \mathrm{H}^1_s(\mathbb{Q}_p, T)
\end{tikzcd}$$

Here $A^m \mathfrak{X}_{\mathbb{F}_p}$ is the codimension m Chow group of $\mathfrak{X}_{\mathbb{F}_p}$, which is just the analogue of the Picard group in higher codimension; $\mathrm{div}_{\mathbb{F}_p}$ sends a pair (x, f) of a cycle and a rational function to its divisor in characteristic p; s is the usual cycle class map in étale cohomology; and the bottom right isomorphism is the natural analogue of the isomorphism of Lemma 10.9, using the smooth base change theorem to identify $\mathrm{H}^{2m}_{\text{ét}}(\mathfrak{X}_{\bar{\mathbb{F}}_p}, \mathbb{Z}_l(m))$ with $T(-1)$. The diagram of the text follows immediately from this one.

Flach defines the map σ (in the case $X = E \times E$, $m = 1$) using a related method in [**50**]. There he proves the commutativity of the above local diagram (in a slightly different form) through explicit computations. In [**51**] he gives the construction of σ (this time in the case $X = X_0(N) \times X_0(N)$, $m = 1$) we gave above and writes down the local diagram, although his proof of commutativity is somewhat incomplete and does not immediately generalize. The general case, which relies heavily on purity conjectures of Grothendieck (which have been proven in the relevant cases by Raskind and Thomason), is the subject of [**158**, Chapters 6 and 7]. The construction of maps similar to σ also appear in the work of Kato; see [**4**] and [**127**]. Mazur offers an alternative construction of the Flach map in [**99**], without any explicit dependence on K-theory. There he also studies some algebraic properties of the map which are not immediately apparent and which permit some Euler system type conclusions even without the existence of an Euler system.

Tom Weston
Department of Mathematics
Harvard University
Cambridge, MA 02138
weston@math.harvard.edu

APPENDIX 3

An introduction to the p-adic geometry of modular curves by Matthew Emerton

The aim of this appendix is to give some feeling for the ideas behind the p-adic theory of modular curves and modular forms (at a more down to earth level than that of [**57**] or [**79**], which are the standard references on this topic). It provides a written approximation to an informal lecture given on this topic by the author at the 1999 PCMI meeting. It was in the informal spirit of the lecture to make many assertions and allusions without providing details. That spirit may also permeate this written version, but some attempt has been made to give relevant references to the literature. In fact, although our coverage of the bibliography is (of course and inevitably) incomplete, we do mention many of the major papers in the field, and the final section is devoted to annotating briefly some of those references.

As one might guess, the background and sophistication that is assumed of the reader varies widely from section to section. A large part of the presentation is devoted to treating the case $p = 2$, where many phenomena can be discovered and investigated via explicit calculation; in these parts of the presentation we assume comparatively little background. On the other hand, in order to understand the theory behind the calculations, and the generalization to arbitrary primes p, more sophisticated ideas (such as the consideration of elliptic curves and modular forms defined over rings other than fields, as well as the techniques of formal and rigid analytic geometry) cannot be avoided. Nevertheless, we have attempted to make the discussion as clear and intuitive as possible; we leave it to the reader and the references to fortify the intuition with correct mathematics, while the author will accept responsibility for any lack of clarity.

Acknowledgments: I would like to thank David Ben-Zvi, Robert Coleman, Brian Conrad and Mike Roth for their helpful comments on earlier drafts of the present work, as well as Fernando Gouvêa for all his assistance in the TEXnical preparation of this appendix.

The curve $X_0(2)$

All the essential ideas in the p-adic theory can already be seen in the case $p = 2$, and this case has the merit that one can easily perform explicit computations. We begin by describing some of these, working first over $\mathbb{C}$.

We let $\mathcal{H}^*$ denote the *extended upper half-plane*, that is, the union of the usual complex upper half-plane with the set $\mathbb{Q} \cup \{\infty\}$. The group $\mathrm{SL}_2(\mathbb{Z})$ acts naturally on $\mathcal{H}^*$ by linear fractional transformations; the quotient is the *modular curve of level one*, which (a little idiosyncratically, but for reasons of consistency) we will denote by $X_0(1)$. It is well known that $\mathrm{SL}_2(\mathbb{Z})$ acts transitively on $\mathbb{Q} \cup \{\infty\}$, and we denote the corresponding point of $X_0(1)$ simply by ∞, and refer to it as the *cusp* of $X_0(1)$ (although of course it is not a singular point; the usage of this term is purely for historical reasons). Let us also write $Y_0(1) = X_0(1) \setminus \{\infty\}$. It is well known that the points of $Y_0(1)$ are in a one-to-one correspondence with the isomorphism classes of elliptic curves over $\mathbb{C}$; the orbit of a point τ in the upper half-plane corresponds to the elliptic curve $E_\tau := \mathbb{C}/\Lambda_\tau$, where Λ_τ is the lattice $2\pi i\tau\mathbb{Z} + 2\pi i\mathbb{Z}$. (The $2\pi i$ factors are included just for the purposes of having a good normalization when it comes to comparing the analytic and algebraic theory of elliptic curves.)

From the algebraic theory of elliptic curves, we know that elliptic curves over any field are also classified up to isomorphism by their j-invariant. If we write $j(\tau) := j(E_\tau)$, then the function $j(\tau)$ is an analytic function on the open subset $Y_0(1)$, which induces an analytic isomorphism $Y_0(1) \xrightarrow{\sim} \mathbb{A}^1$. (This is essentially tautological if one has a sufficient understanding of $X_0(1)$ as a moduli space; alternatively, see [**133**] for a down-to-earth account of j as an analytic function on the upper half-plane and $Y_0(1)$). The function $j(\tau)$ has a simple pole at ∞, and thus this isomorphism extends to an isomorphism

$$X_0(1) \xrightarrow{\sim} \mathbb{P}^1,$$

and by virtue of this isomorphism we will also regard j as a coordinate on $X_0(1)$, and refer to $X_0(1)$ as the j-line.

The curve $X_0(2)$ is obtained by taking the quotient of $\mathcal{H}^*$ by the congruence subgroup $\Gamma_0(2)$ of $\mathrm{SL}_2(\mathbb{Z})$. There are two orbits of $\Gamma_0(2)$ on $\mathbb{Q} \cup \{\infty\}$, the orbit of the point 0 and that of the point ∞. We refer to these points as the *cusps* of $X_0(2)$, and we label them simply by 0 and ∞. We also write $Y_0(2) = X_0(2) \setminus \{0, \infty\}$.

Of course the points $Y_0(2)$ also have an interpretation in terms of moduli of elliptic curves. Namely, given a point τ in the upper half-plane, we can form the curves E_τ and $E_{2\tau}$, and observe that there is an isogeny between them:

$$E_\tau \longrightarrow E_{2\tau} \tag{I}$$

via the map $z \bmod \Lambda_\tau \mapsto 2z \bmod \Lambda_{2\tau}$. The kernel of this isogeny has order two, and is generated by the point $\pi i \bmod \Lambda_\tau$. One easily checks that if we modify τ by an element of $\Gamma_0(2)$ then the isomorphism class of this two-isogeny remains unchanged, and that in fact the points of $Y_0(2)$ are in bijection with the isomorphism classes of two-isogenies of elliptic curves over $\mathbb{C}$.

We were able to explicitly describe $X_0(1)$ as an algebraic curve: it is the j-line. Is there a similar description of $X_0(2)$? Here is a first attempt: since the orbit of a point τ under $\Gamma_0(2)$ determines the two-isogeny (I) up to isomorphism, we certainly know its source and target up to isomorphism, and so to the orbit $\tau \bmod \Gamma_0(2)$ in $Y_0(2)$ we can associate the two complex numbers $j(\tau)$ and $j(2\tau)$, which do classify the isomorphism class of E_τ and $E_{2\tau}$. Thus we get a map

$$Y_0(2) \longrightarrow Y_0(1) \times Y_0(1) \xrightarrow{\sim} \mathbb{A}^2. \tag{II}$$

Since the source of this map is one-dimensional over $\mathbb{C}$, its image in $Y_0(1) \times Y_0(1)$ must be a curve, whose equation will express a (somewhat complicated, as it turns out) relation between $j(\tau)$ and $j(2\tau)$. Such relations are classically called *modular equations*. From our point of view this modular equation has two disadvantages: firstly, it is a little hard to compute explicitly; secondly, the map (II) is not an isomorphism onto its image – it turns out that the image has singularities. What this means in modular terms is that for certain special choices of τ, there is one or more two-isogeny between E_τ and $E_{2\tau}$ which is not isomorphic to the isogeny (I).

There is an alternative approach to describing $X_0(2)$ as an algebraic curve, which we now present. We first proceed algebraically: thus we assume given a two isogeny $\psi : E_1 \longrightarrow E_2$ between two elliptic curves over $\mathbb{C}$. Let ω_2 be a regular differential on E_2, and let $\omega_1 := \psi^*\omega_2$ be the pulled-back regular differential on E_1.

Recall the algebraic definition of modular forms (of level one): a modular form f of weight k is a "rule" which to any pair consisting of an elliptic curve E and a non-zero regular differential ω on E attaches a number $f(E,\omega)$ depending only on the isomorphism class of the pair (E,ω), such that for any non-zero scalar λ, $f(E,\lambda\omega) = \lambda^{-k} f(E,\omega)$, and which "behaves well in families". (We won't make precise the meaning of this statement here; see [**38**] or [**79**].)

There is a canonical modular form of weight 12, the discriminant Δ (see [**38**]). Returning to our two isogeny $\psi : E_1 \longrightarrow E_2$, we define

$$j_2(\psi) = 2^{12}\frac{\Delta(E_1,\omega_1)}{\Delta(E_2,\omega_2)}.$$

Note that if we multiply ω_2 by a non-zero scalar λ, both the numerator and denominator of $j_2(\psi)$ are scaled by the same amount λ^{-12}, and so $j_2(\psi)$ remains invariant. Thus it really does depend only on the isogeny ψ, and not on the auxiliary choice of ω_2.

Now let us interpret this in the analytic picture: on $E_\tau = \mathbb{C}/\Lambda_\tau$ there is a canonical differential ω_τ obtained by reducing the differential dz on $\mathbb{C}$ modulo Λ. (This reduction is possible because the differential dz is invariant under translation by elements of the lattice Λ_τ.) The function

$$\Delta(\tau) := \Delta(E_\tau, \omega_\tau)$$

is a modular form (in the classical analytic sense explained in [**133**] for example) of weight twelve and level one – in fact, it is the unique normalized *cuspform* of this weight and level. Its q-expansion is given by the famous formula

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

(Recall that in the context of modular forms, q denotes the exponential $q = e^{2\pi i\tau}$. The product formula for the q-expansion of Δ is proved in [**133**], for example.) If we let ψ_τ denote the isogeny (I), then we see that $\psi_\tau^*\omega_{2\tau} = 2\omega_\tau$, and so

$$\text{(III)}\qquad j_2(\psi_\tau) = 2^{12}\frac{\Delta(E_\tau, 2\omega_\tau)}{\Delta(E_{2\tau},\omega_{2\tau})} = \frac{\Delta(E_\tau,\omega_\tau)}{\Delta(E_{2\tau},\omega_{2\tau})}.$$

We define the function j_2 on $Y_0(2)$ via the formula

$$j_2(\tau) = j_2(\psi_\tau) \overset{\text{(III)}}{=} \frac{\Delta(\tau)}{\Delta(2\tau)} = \frac{q\prod_{n=1}^{\infty}(1-q^n)^{24}}{q^2\prod_{n=1}^{\infty}(1-q^{2n})^{24}} = q^{-1}\prod_{n=1}^{\infty}(1+q^n)^{-24}.$$

This is an analytic function $Y_0(2) \longrightarrow \mathbb{A}^1$, and from its q-expansion (and the easily verified fact that $q = e^{2\pi i\tau}$ is a uniformizer in the neighbourhood of the point

∞ of $X_0(2)$) we see that is has a simple pole at the point ∞ of $X_0(2)$. Thus j_2 *necessarily* extends to an *isomorphism* $X_0(2) \xrightarrow{\sim} \mathbb{P}^1$. Because of this, we will also refer to $X_0(2)$ as the j_2-line.

Note that we conclude incidentally that any two-isogeny ψ is determined up to isomorphism by the invariant $j_2(\psi)$. One could also establish this fact (with more difficulty) by pure algebra, rather than resorting to the consideration of modular curves, as we did above. However, one motivation for introducing the modular curves (or moduli spaces in general) is to simplify the proof of results such as this.

Now that we have our algebraic description of $X_0(2)$ as the j_2-line, it will be interesting to return to and reinterpret slightly the map (II). The first coordinate of this map extends to a map $X_0(2) \longrightarrow X_0(1)$, which is the natural map arising from the inclusion of $\Gamma_0(2)$ in $\mathrm{SL}_2(\mathbb{Z})$. We denote this map by B_1. The second coordinate of (II) also extends to a map $X_0(2) \longrightarrow X_0(1)$, which we denote by B_2. There is also a natural automorphism of $X_0(2)$, given by the construction of dual isogenies: if $\psi : E_1 \longrightarrow E_2$ is a two-isogeny, it has a dual two-isogeny $\check{\psi} : E_2 \longrightarrow E_1$. If ψ is the isogeny ψ_τ of (I) then $\check{\psi}_\tau$ is the isogeny $E_{2\tau} \longrightarrow E_\tau$ arising from the inclusion of lattices $\Lambda_{2\tau} \subset \Lambda_\tau$. This is isomorphic to the two-isogeny $\psi_{-1/2\tau}$. Thus we see that constructing dual isogenies yields an automorphism of $Y_0(2)$ given by the formula

$$\tau \bmod \Gamma_0(2) \mapsto \frac{-1}{2\tau} \bmod \Gamma_0(2).$$

This extends to an automorphism of $X_0(2)$ which interchanges 0 and ∞. We denote this automorphism by w_2; it is an *involution* (has order two), and is often referred to as the Atkin-Lehner involution. The map B_2 is easily described via w_2: since w_2 interchanges the source and target of a two-isogeny, we see that $B_2 = B_1 \circ w_2$.

How do we describe w_2 in terms of the coordinate j_2 on $X_0(2)$? This is an easy computation, using the modularity properties of Δ; one could make it either algebraically or analytically, but the latter is probably simpler:

$$j_2(-1/2\tau) = \frac{\Delta(-1/2\tau)}{\Delta(-1/\tau)} = \frac{2^{12}\tau^{12}\Delta(2\tau)}{\tau^{12}\Delta(\tau)} = \frac{2^{12}}{j_2(\tau)}.$$

Note that from this formula we also find the value of j_2 at the cusp 0: $j_2(0) = j_2(w_2(\infty)) = 2^{12}/j_2(\infty) = 0$, since j_2 has a pole at ∞.

Here is another question: how do we describe the function B_1 in terms of the coordinates j and j_2?

Lemma 11.1. *The map* $B_1 : X_0(2) \longrightarrow X_0(1)$ *is described by the equation*

$$\text{(IV)} \qquad j \circ B_1 = (j_2 + 256)^3/j_2^2.$$

It is ramified at the point $512 = j_2(i)$ *over the point* $1728 = j(i)$ *with degree two, at the point* $-256 = j_2((-1+\sqrt{-3})/2)$ *over the point* $0 = j((-1+\sqrt{-3})/2)$ *with degree three, and at the point* $0 = j_2(0)$ *over the point* $\infty = j(\infty)$ *with degree two.*

Proof. The map from the upper half-plane to $Y_0(1)$ is ramified over two points: the point $j = 1728$, corresponding to the $\mathrm{SL}_2(\mathbb{Z})$ orbit of $z = i$, with ramification degree two, and the point $j = 0$, corresponding to the $\mathrm{SL}_2(\mathbb{Z})$ orbit of $z = (-1+\sqrt{-3})/2$, with ramification degree three.

The map from the upper half-plane to $Y_0(2)$ is ramified over one point, corresponding to the $\Gamma_0(2)$ orbit of $z = (1+i)/2$, with ramification degree two.

Now $\Gamma_0(2)$ has index three in $\mathrm{SL}_2(\mathbb{Z})$, and so the map B_1 has degree three, and from the preceding two paragraphs we see that the resulting degree three map $Y_0(2) \longrightarrow Y_0(1)$ is ramified at the $\Gamma_0(2)$ orbit of $z = i$, with ramification degree two, and is totally ramified over the point $j = 0$.

Of the two cusps ∞ and 0 on $X_0(2)$, the cusp ∞ is unramified over the cusp ∞ on $X_0(1)$ with respect to B_1 (we saw this implicitly above, when we remarked that q is a uniformizer in a neighbourhood of ∞ on $X_0(2)$, since it is also a uniformizer in a neighbourhood of ∞ on $X_0(1)$), and hence the cusp 0 must be ramified of degree two over the cusp ∞ on $X_0(1)$.

Now pull back j via B_1 to a function on $X_0(2)$. Since we know the zeroes and poles of j on $X_0(1)$, as well as the ramification structure of B_1 above each of these points, we see that $j \circ B_1$ has a pole of order one at ∞, a pole of order two at 0, a zero of order three at $j_2((-1+\sqrt{-3})/2)$, and no other zeroes or poles. Thus $j \circ B_1$ is a scalar multiple of $(j_2 - a)^3/j_2^2$, where $a = j_2((-1+\sqrt{-3})/2)$. One computes that

$$(j_2 - a)^3/j_2^2 = 1/q - (24 + 3a) + \cdots ,$$

while

$$j = 1/q + 744 + \cdots$$

(see [**133**]). Comparing these expressions we find that $a = -256$ and hence that

$$j \circ B_1 = (j_2 + 256)^3/j_2^2.$$

From this equation we see that $-256 = j_2((-1+\sqrt{-3})/2)$ is the point lying over $0 = j((-1+\sqrt{-3})/2)$, and also that

$$j \circ B_1 - 1728 = (j_2 + 64)(j_2 - 512)^2/j_2^2,$$

and thus, from the above description of the ramification of B_1, that $j_2(i) = 512$ and $j_2((1+i)/2) = -64$. We have now verified the asserted formula for the map B_1 and also the claims about its ramification. □

The equation $B_2 = B_1 \circ w_2$ allows us to compute that

$$\text{(V)} \quad j \circ B_2 = ((j_2 + 256)^3/j_2^2) \circ w_2 = (2^{12}/j_2 + 256)^3/(2^{12}/j_2)^2 = (j_2 + 16)^3/j_2.$$

With sufficient enthusiasm, one can eliminate j_2 from the two equations (IV) and (V) and thus find the equation relating $j \circ B_1$ and $j \circ B_2$; this will be the modular equation which describes the image of the map (II). Classically, one of the roles of higher level modular curves such as $X_0(2)$ (or more precisely, the modular functions such as j_2 that are defined on them) was to simplify the shape of the modular equations. In modern terms, we can see this as being related to the simplification that comes from replacing a singular curve by its normalization.

Canonical subgroups of elliptic curves over $\mathbb{C}_2$

The Lefschetz principle assures us that algebraic geometry is the same when studied over any algebraically closed field of characteristic zero. In this section we will take our ground field to be the completion of the algebraic closure of the field of 2-adic numbers $\mathbb{Q}_2$. This field (which is algebraically closed as well as complete) is denoted $\mathbb{C}_2$; it provides a natural location for conducting 2-adic analysis, just as $\mathbb{C}$ provides a natural location for performing classical analysis. In particular, $\mathbb{C}_2$ is equipped with a non-archimedean absolute value, which we denote by $|\ |$, and which we normalize by the condition that $|2| = 1/2$ (although the normalization won't be important).

By the Lefschetz principle, all the algebro-geometric observations about moduli of elliptic curves over $\mathbb{C}$ which we made in the preceding section apply equally well to elliptic curves over $\mathbb{C}_2$. Thus we see that there is j-line whose non-cuspidal points classify isomorphism classes of elliptic curves over $\mathbb{C}_2$, as well as a j_2-line, whose points (other than ∞ and 0) classify isomorphism classes of two-isogenies between elliptic curves over $\mathbb{C}_2$. There is an involution w_2 of the j_2-line given by passing to the dual isogeny, there are two maps B_1 and B_2 mapping the j_2-line to the j-line, and all the formulas for w_2, B_1 and B_2 that we proved in the preceding section continue to hold true.

Let us consider the formula (IV) for B_1. We can rewrite this in the form

$$\frac{256}{j \circ B_1} = \frac{256/j_2}{(1+256/j_2)^2}.$$

Now if $|j_2| > |256|$, then we may expand the right-hand side of this equation in a power series, and so obtain the equation

$$\frac{256}{j \circ B_1} = \sum_{n=1}^{\infty} (-1)^{n-1} n \left(\frac{256}{j_2}\right)^n.$$

Since the leading coefficient of this series is 1 (and so in particular is non-zero) we see by the implicit function theorem (which in this non-archimedean context is just a formal manipulation of power-series) that B_1 establishes an *isomorphism* between the region $|j_2| > |256|$ on the j_2-line and the region $|j| > |256|$ on the j-line. Each of these regions is a disk centred at the point ∞ on the appropriate line; let us denote them by D_2 and D_1 respectively. Thus we see that B_1 induces a 2-adic analytic isomorphism from D_2 to D_1, and so has a 2-adic analytic inverse $B_1^{-1} : D_1 \longrightarrow D_2$. So what?

Well, in modular terms, this means that if E is an elliptic curve over $\mathbb{C}_2$ whose j-invariant satisfies the inequality $|j(E)| > |256|$, then there is a naturally determined two-isogeny whose source is E. Equivalently, thinking of a two-isogeny as being determined by its kernel, we see that there is a naturally determined subgroup of E of order two. Since E has three distinct subgroups of order two, it seems pretty remarkable that there is any way to distinguish one of them as being naturally determined! (Caveat: since the point $B_1^{-1}(j(E))$ in D_2 only determines a two-isogeny up to isomorphism, the subgroup of order two is determined only up to the application of automorphisms of E. Now if E has no automorphism besides ± 1, then since these both fix any subgroup of E, we see that the isomorphism class of the isogeny does determine a well-defined subgroup of order two of E. What if E has extra automorphisms? The point $j = 0$ (whose corresponding elliptic curve has six automorphisms) is not in the disk D_1, so it remains to consider the point $j = 1728$ (whose corresponding elliptic curve has four automorphisms). Now if $j(E) = 1728$, let α denote an automorphism of E of order four. Then $\alpha - 1$ has degree two as an endomorphism of E, and so α fixes exactly one subgroup of order two of E, and interchanges the other two. Since $B_1^{-1}(1728)$ lies in D_2, we see by lemma 1.4 and its proof that $B_1^{-1}(1728) = -64$, and since this is *not* a ramification point of B_1, the kernel of the corresponding two-isogeny must be *fixed* by α. Thus we see that the caveat is no caveat at all: even taking into account possible extra automorphisms of E, we see that the map B_1^{-1} *does* determine an order two subgroup of every E with $j(E)$ lying in D_1.) This subgroup is called the *canonical subgroup* of E.

Now we consider the composite $\phi := B_2 \circ B_1^{-1} : D_1 \longrightarrow X_0(1)$, which is called the *Deligne-Tate map.* The map B_1^{-1} is described by a formula of the form $256/B_1^{-1}(j) = \sum_{n=1}^{\infty} a_n(256/j)^n$, for certain integers a_n such that $a_1 = 1$. The map B_2 is described by (V), which can be rewritten in the form

$$\frac{256}{j} \circ B_2 = \frac{(16/j_2)^2}{(1+16/j_2)^3}.$$

Thus we find that

$$\frac{256}{\phi(j)} = \sum_{n=2}^{\infty} b_n \left(\frac{16}{j}\right)^2,$$

for some integers b_n such that $b_1 = 1$. Actually, although ϕ is defined on all of D_1, this power-series formula only converges on the subdisk D_1' defined by the inequality $|j_2| > |16|$; it shows that the restriction of ϕ to D_1' is a degree two map whose image is D_1 and which satisfies the formula

$$|\phi(j)| = |j|^2 \tag{12}$$

for any point j of D_1'.

What is the modular interpretation of the Deligne-Tate map? Well, if we start with an elliptic curve E such that $j(E)$ lies in D_1, then $B_1^{-1}(j)$ associates to E its canonical subgroup C. Then B_2 associates to the pair (E, C) the *target* of the isogeny whose kernel is C, which is just E/C. Thus in modular terms ϕ associates to any elliptic curve E with $j(E)$ in D_1 its quotient by its canonical subgroup. Now if $j = j(E)$ in fact lies in D_1', we saw that $\phi(j)$ lies in D_1, and so E/C also has a canonical subgroup, which pulls back to a subgroup C' of E of order four.

Lemma 11.2. *In the above notation, C' is a cyclic subgroup of E*

Proof. Either C' is a cyclic subgroup of E of order four, or it is the full two-torsion subgroup of E. In the latter case, we would see that E/C' is isomorphic to E, and thus that $E \longrightarrow E/C$ and $E/C \longrightarrow E/C'$ are dual isogenies.

On the other hand, both of these isogenies have j_2 lying in D_2 (since they are in the image of B_1^{-1}), and the first even satisfies $|j_2| > |16|$, because $|j(E)| > |16|$, and B_1^{-1} preserves absolute values. If $|j_2| > |16|$, then $|2^{12}/j_2| < |256|$, and so $2^{12}/j_2$ does not lie in D_2. Thus they *cannot* be dual isogenies, and we have proved the lemma. □

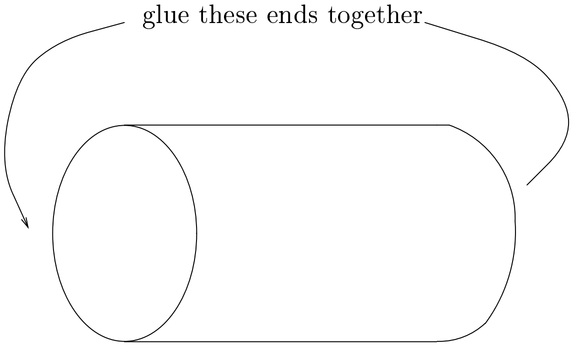
More generally, if we let $D_1^{(n)}$ denote the disk in the j-line determined by the inequality $|j| > |2^{2^{3-n}}|$, then we see that $\phi^n : D_1^{(n)} \longrightarrow D_1$, and that as a consequence if E is an elliptic curve over $\mathbb{C}_2$ for which $j(E)$ lies in $D_1^{(n)}$, then E has a canonical subgroup of order 2^{n+1}, which is cyclic (by the easily proved analogue of lemma 2.2).

The parameter q

This section and the next are a digression on some points of theory which we will need to understand the preceding calculations better. The subject of our first digression is an alternative approach to describing elliptic curves, using the parameter q rather than τ. This leads to a more theoretical interpretation of the

q-expansion of a modular form, and to a modular interpretation of the cusp ∞ on the j-line.

Let us begin by considering the affine j-line $Y_0(1)$ over the complex numbers. If τ is a point in $\mathcal{H}$ its image in $Y_0(1)$ corresponds to the elliptic curve $E_\tau = \mathbb{C}/(2\pi i\tau\mathbb{Z} + 2\pi i\mathbb{Z})$. We can take this quotient in two steps, by noting that the exponential function induces an isomorphism $\mathbb{C}/2\pi i\mathbb{Z} \xrightarrow{\sim} \mathbb{C}^\times$ (where $\mathbb{C}^\times$ denotes the multiplicative group of non-zero complex numbers.) Under the exponential, the lattice $2\pi i\tau\mathbb{Z} + 2\pi i\mathbb{Z}$ has image equal to the cyclic subgroup $q^\mathbb{Z}$ of $\mathbb{C}^\times$, where $q = e^{2\pi i\tau}$. Thus we can describe E_τ as the quotient $\mathbb{C}^\times/q^\mathbb{Z}$. Thinking of $\mathbb{C}^\times$ as a cylinder of infinite length (which it is topologically), this is just the familiar description of a torus as being obtained by gluing together the ends of a finite length cylinder (where in our case, the finite length cylinder will be a fundamental domain for the action of $q^\mathbb{Z}$ on $\mathbb{C}^\times$), as in the following picture:

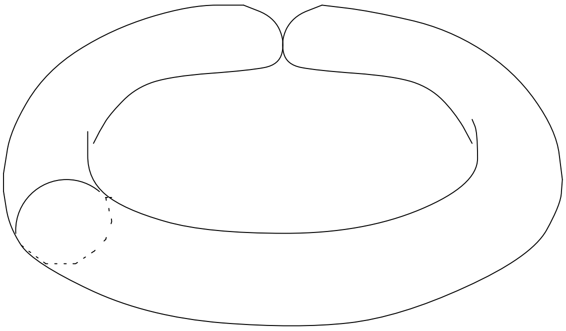


fundamental domain for multiplication by q on $\mathbb{C}^\times$

We will use this description of elliptic curves to obtain a geometric description of the q-expansions of modular forms. We begin with a discussion of differentials. Let z denote the coordinate on $\mathbb{C}$ and t the coordinate on $\mathbb{C}^\times$, so that $t = e^z$. Then the differential dz on $\mathbb{C}$ (which is invariant under the action of $2\pi i\mathbb{Z}$ by translation) descends to the differential dt/t on $\mathbb{C}^\times$, and since this differential is invariant under the action of $q^\mathbb{Z}$ by multiplication, it descends to a differential on $\mathbb{C}^\times/q^\mathbb{Z}$.

Let D^* denote the punctured open unit disk, consisting of those complex numbers q such that $1 > |q| > 0$. As q ranges over all elements of D^*, the elliptic curves $\mathbb{C}^\times/q^\mathbb{Z}$ form a family of elliptic curves lying over D^*, and the differential dt/t on $\mathbb{C}^\times$ descends to a relative differential on this family. Denote this family of elliptic curves by E_{Tate}, and this relative differential by ω_{Tate}. If f is any modular form of some weight k, then we may evaluate f on each member of the family, and thus obtain a function on D^*. More succinctly, we evaluate f on the elliptic curve E_{Tate} over the base D^* to obtain an element $f(E_{\mathrm{Tate}}, \omega)$ of the ring of functions on D^* (necessarily holomorphic – this is a consequence of the condition that f behave well in families). But such a function is just a convergent power-series in q, and this is precisely the q-expansion of the modular form f. Thus the existence of q-expansions of modular forms is related to the existence of the family of elliptic curve E_{Tate} over D^*, equipped with its canonical relative differential ω_{Tate}.

We now turn to an analysis of the cusp ∞ in $X_0(1)$. As τ tends towards the point ∞ in the boundary of $\mathcal{H}^*$, the value of q tends to zero. Thus in the preceding description of E_τ, the circumference of the finite cylinder is remaining fixed, while its length is tending towards zero. What happens when τ reach the point ∞? One way to describe what happens is to remember that the circumferential circle of this cylinder, and the circle obtained by gluing the endpoints of a longitudinal interval, represent two independent generators of the fundamental group of E_τ. What we just observed is that as τ goes to ∞, one of these loops stays a fixed length, while the length of the other tends to zero. This makes it reasonable to think of the point $\tau = \infty$ as corresponding to the curve obtained by completely shrinking one of these loops to a point, as in the following picture:



The indicated generator of the fundamental group is pinched off to zero

In short, the point ∞ on $X_0(1)$ corresponds to a rational curve with one node. In fact one can show that there is a canonical way to extend the family of elliptic curves E_{Tate} over the punctured disk D^* to a family of curves over the (unpunctured) open unit disk D whose fibre at $q = 0$ is the above rational nodal curve.

If one is working over a field K different from $\mathbb{C}$, the above analytic construction of the family E_{Tate} is not available. However, one can construct in an analogous way an elliptic curve E_{Tate} over the field $K((q))$ (which one thinks of as the ring of functions on the *formal punctured disk*), equipped with a canonical nowhere-zero differential ω_{Tate}. Furthermore, the j-invariant of E_{Tate} is an element of $K((q))$ given precisely by the usual formula for j as a function of q [**38, 79**]. Letting $\mathbb{G}_m$ denote the multiplicative group over $K((q))$, one can even describe E_{Tate} as a quotient $\mathbb{G}_m/q^{\mathbb{Z}}$, in a suitable geometric sense, and (if t denotes the coordinate on $\mathbb{G}_m$) ω_{Tate} is obtained by descending the differential dt/t on $\mathbb{G}_m$. Finally, the elliptic curve E_{Tate} extends to a curve over the ring $K[[q]]$, whose fibre over $q = 0$ is a nodal curve. (See [**39**] for the rather technical details of the construction of E_{Tate} over $K[[q]]$ and even its construction over $\mathbb{Z}[[q]]$. With regard to this, note that it is quite important that the Tate curve over $K[[q]]$ for any field K is obtained from a given elliptic curve defined over $\mathbb{Z}[[q]]$, and also that, when $K = \mathbb{C}$, the Tate curve over the formal power-series ring $\mathbb{C}[[q]]$ is 'the same' as the analytic Tate curve over the open unit disk D described in the preceding paragraph; that is, the power-series in $\mathbb{Z}[[q]]$ which appear in the Weierstrass equation for the Tate curve converge on the open unit disc D, and so describe an analytic family of elliptic curves over D, which is the analytic Tate curve described above. These two facts allow one to use analytic reasoning over $\mathbb{C}$ to draw conclusions about the Tate curve which can then be applied in other situations, for example in the p-adic analytic context that we will consider below.)

One can unify the treatment of elliptic curves and singular curves of the type that correspond to the cusp $j = \infty$ by introducing the notion of *generalized elliptic curve* [**39**]. Since such a uniform treatment is possible, at various points in the following text the conceptual distinction between elliptic curves and rational nodal curves will become a little blurred.

If f is any modular form over K then one *defines* the q-expansion of f to be the element of $K((q))$ obtained by evaluating f on the pair $(E_{\mathrm{Tate}}, \omega_{\mathrm{Tate}})$ [**38, 79**]. The q-expansion principle asserts that if f and g are two modular forms of the same weight having the same q-expansion, then they are equal. This is essentially a corollary of the fact that any function on a Zariski open subset of $Y_0(1)$ is determined by its resriction to the formal punctured disk around the cusp $j = \infty$. (More algebraically, and perhaps more precisely, any localization of $K[j]$ injects into the fraction field $K(j)$, and thus into $K((q))$; the point is that $K[j]$ is an integral domain, for any K, or scheme-theoretically speaking, that $Y_0(1)$ is geometrically irreducible and reduced.) (See [**79**] for the details.) The usual requirement of the q-expansion of a modular form, that it lie in $k[[q]]$ rather than just $k((q))$, in the light of this section corresponds to the fact that modular forms extend from functions on elliptic curves to functions on generalized elliptic curves (and so can be evaluated on the fibre of E_{Tate} over the point $q = 0$, for example) [**39**].

The Hasse invariant

In this section we recall briefly the theory of the Hasse invariant of elliptic curves in positive characteristic. We will need to treat elliptic curves defined over rings other than fields. Thus we let p be a prime and R a ring of characteristic p.

Let E be an elliptic curve over R. Recall that there is canonically associated to E another elliptic curve $E^{(p)}$ and a canonical isogeny $\pi : E \longrightarrow E^{(p)}$ of degree p, the *Frobenius*. Furthermore, if ω is a nowhere-zero regular differential on E, then it induces canonically a nowhere-zero regular differential $\omega^{(p)}$ on $E^{(p)}$.

The easiest way to explain these constructions is to note that given E and ω we can find (at least locally over R) a Weierstrass equation

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

for E such that $\omega = dx/(2y + a_1x + a_3)$. Then $E^{(p)}$ is given by the Weierstrass equation

$$y^2 + a_1^pxy + a_3^py = x^3 + a_2^px^2 + a_4^px + a_6^p,$$

the differential $\omega^{(p)}$ is equal to $dx/(2y + a_1^px + a_3^p)$, and the isogeny π is the map $(x, y) \mapsto (x^p, y^p)$.

Let us suppose that E and ω are given as above, and let $\check{\pi}$ be the dual isogeny to π. Then we can construct *two* differentials on $E^{(p)}$: the nowhere-zero regular differential $\omega^{(p)}$, and also the pulled-back differential $(\check{\pi})^*\omega$. Define $A(E, \omega)$ to be that element of R such that

$$(\check{\pi})^*\omega = A(E, \omega)\omega^{(p)}.$$

If we multiply ω by a unit λ of R, then we see that $(\lambda\omega)^{(p)} = \lambda^p\omega^{(p)}$, and consequently $A(E, \lambda\omega) = \lambda^{1-p}A(E, \omega)$. Thus A is a modular form of weight $p - 1$ in characteristic p.

Now suppose that R is a field of characteristic p. In this case, although the value of $A(E,\omega)$ depends on the choice of both E *and* ω, the property of $A(E,\omega)$ being zero or non-zero depends only on E, independent of the choice of ω. If it is non-zero, we say that E is *ordinary*; if it is zero, we say that E is *supersingular*. It is relatively easy to determine the supersingular j-invariants modulo any prime p. There are only finitely many isomorphisms classes of such curves; in fact their j-invariants all lie in the finite field $\mathbb{F}_{p^2}$. Furthermore, there is a precise formula for their number, as well as a precise formula for a polynomial of which their j-invariants are the solutions (see [**74**]).

The remainder of this section is devoted to calculating the q-expansion of the Hasse invariant. Recall from the preceding section that the Tate curve E_{Tate} in characteristic p is the quotient (in an appropriate sense) of the multiplicative group $\mathbb{G}_m$ over $\mathbb{F}_p((q))$ by the cyclic subgroup $q^{\mathbb{Z}}$. Since the multiplicative group is already defined over $\mathbb{F}_p$, we see that in computing $E_{\text{Tate}}^{(p)}$ we just have to raise q to the p^{th} power, and so $E_{\text{Tate}}^{(p)}$ is isomorphic to the quotient of $\mathbb{G}_m$ by the cyclic subgroup $q^{p\mathbb{Z}}$, the Frobenius isogeny is given by $t \bmod q^{\mathbb{Z}} \mapsto t \bmod q^{p\mathbb{Z}}$, and its dual $\check{\pi}$ is just the natural map $\mathbb{G}_m/q^{p\mathbb{Z}} \longrightarrow \mathbb{G}_m/q^{\mathbb{Z}}$. Recall that ω_{Tate} equals dt/t, so that $\omega_{\text{Tate}}^{(p)}$ is again dt/t, and also $(\check{\pi})^*\omega_{\text{Tate}} = dt/t$. We conclude from this that $A(E_{\text{Tate}}, dt/t) = 1$. In other words, the q-expansion of A is just the constant 1!

Note that this would be quite impossible in characteristic zero: the only modular forms over $\mathbb{C}$ with constant q-expansions are those of weight zero. On the other hand, it is the fact that the weight $p-1$ modular form A and the weight zero modular form 1 have the same q-expansion which gives rise to the possibility of congruences of modular forms of different weights, and the beautiful theory of p-adic families of modular forms. (See the discussion and references in the guide to the literature given below.)

Return to $X_0(2)$

In this section we try to shed some theoretical light on the calculations of section 2, to pave the way for the discussion in the case of a general prime p.

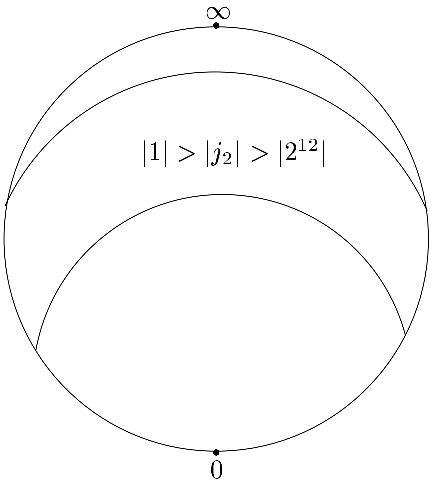
Let $\mathcal{O}$ denote the ring of integers in $\mathbb{C}_2$, that is, the ring of elements r for which $|r| \leq |1|$; geometrically, $\mathcal{O}$ is the closed unit disk of $\mathbb{C}_2$. The open unit disk of elements r such that $|r| < |1|$ is the maximal ideal $\mathfrak{m}$ of the valuation ring $\mathcal{O}$, and the quotient field k of $\mathcal{O}$ by $\mathfrak{m}$ is an algebraic closure of the field $\mathbb{F}_2$, which we denote by k.

Reducing modulo $\mathfrak{m}$ extends to a map from the projective j-line over $\mathbb{C}_2$ to the projective j-line over k, which we call the *specialization map*. This map has the following modular interpretation: if E is an elliptic curve such that $j(E)$ lies in $\mathcal{O}$ then E has a model with coefficients in $\mathcal{O}$ with good reduction modulo $\mathfrak{m}$, and the reduction of $j(E)$ modulo $\mathfrak{m}$ is just the j-invariant of the reduction of this model module $\mathfrak{m}$. On the other hand, if $j(E) \in \mathbb{C}_2 \setminus \mathcal{O}$ then any model of E over $\mathcal{O}$ has singular reduction modulo $\mathfrak{m}$, and the reduction of $j(E)$ modulo $\mathfrak{m}$ is the point ∞, which as we saw in section 3 corresponds to isomorphism class of a generalized elliptic curve which is singular. Finally, the point ∞ on the j-line corresponds to the singular generalized elliptic curve over $\mathbb{C}_2$, any model over $\mathcal{O}$ of which certainly has singular reduction modulo $\mathfrak{m}$, and the point ∞ certainly reduces to the point ∞.

Let us (as is customary) extend the phrase *ordinary reduction* to include either good ordinary reduction or bad reduction. (One should really require bad *multiplicative* reduction. However, $\mathbb{C}_2$ is algebraically closed, and so any E with bad reduction has a model with multiplicative reduction.) Since $j = 0$ is the unique supersingular j-invariant modulo 2, we see that the closed disk $|j| \geq |1|$ is the set of j-invariants in $X_0(1)$ having ordinary reduction. We denote this disk by $X_0(1)^h$ (h is for Hasse). The complement of $X_0(1)^h$ in $X_0(1)$ is the open disk $|j| < |1|$, and consists of the set of j-invariants having good supersingular reduction. We call this the *supersingular disk* in $X_0(1)$.

An examination of equations (IV) and (V) shows that the preimage of $X_0(1)^h$ under both B_1 and B_2 is equal to the union of the two disks $|j_2| \geq 1$ and $|j_2| \leq |2^{12}|$. We denote the former disk by $X_0(2)^h_\infty$ (because it contains the cusp ∞) and the latter by $X_0(2)^h_0$ (because it contains the cusp 0); their union we denote simply by $X_0(2)^h$. Note that $X_0(2)^h$ is preserved by w_2, and that w_2 interchanges $X_0(2)^h_\infty$ and $X_0(2)^h_0$. The fact that $X_0(2)^h$ is preserved by w_2 (which is equivalent to the fact that $X_0(1)^h$ has the same preimage under either B_1 or B_2) reflects the fact that the property of being ordinary or supersingular is an invariant of an isogeny class of elliptic curves in characteristic two.

The complement of $X_0(2)^h$ in $X_0(2)$ is the open annulus $|1| > |j_2| > |2^{12}|$. Its points correspond to those two-isogenies whose source (or equivalently target) has good supersingular reduction. We refer to it as the *supersingular annulus* in $X_0(2)$. Here is the picture:



The j_2-line, drawn as a Riemann sphere, with the supersingular annulus marked

The disk $X_0(1)^h$ is contained in the disk D_1 on which B_1^{-1} is defined, and we see that B_1^{-1} restricts to yield an isomorphism $X_0(1)^h \xrightarrow{\sim} X_0(2)^h_\infty$; in particular, every elliptic curve with ordinary reduction has a canonical subgroup.

We can characterize the disk $X_0(1)^h$ as the set of j-invariants whose reduction modulo $\mathfrak{m}$ has non-zero Hasse invariant. The rest of this section is devoted to finding an analogous description of the larger disk D_1.

Let E_4 denote the weight four Eisenstein series on $\mathrm{SL}_2(\mathbb{Z})$, whose q-expansion is

$$E_4 = 1 + 240 \sum_{n=1}^{\infty} \sigma_3(n) q^n,$$

where $\sigma_3(n) = \sum_{d|n} d^3$. Recall that this modular form is actually well-defined in all characteristics (see [**38**], where it is denoted c_4) and that the j-invariant satisfies (indeed is defined by) the equation $j = E_4^3/\Delta$.

If E is an elliptic curve over $\mathbb{C}_2$ with good reduction, and ω is a differential on E with non-zero reduction modulo $\mathfrak{m}$, then for any modular form f over $\mathbb{C}_2$ we may compute $f(E,\omega)$, and the absolute value $|f(E,\omega)|$ is independent of the choice of ω (because any two such ω differ only by multiplication by a unit of $\mathcal{O}$). Thus we are entitled to write simply $|f(E)|$ in place of $|f(E,\omega)|$. For example, $\Delta(E,\omega)$ must be a unit in $\mathcal{O}$ (since it reduces to the discriminant of the reduction of E, which is non-zero), and so $|\Delta(E)| = |1|$, yielding the formula

$$|j(E)| = \frac{|E_4^3(E)|}{|\Delta(E)|} = |E_4^3(E)|.$$

Hence we see that $|j(E)| > |256|$ if and only if $|E_4(E)| > |2^{8/3}|$.

We now wish to relate E_4 to the Hasse invariant. The Hasse invariant A is a modular form of weight one defined modulo 2. Thus A^4 is a modular form of weight one defined modulo 8, whose q-expansion is the constant 1. (The point being that if x is a number well-defined modulo 2, than x^4 is well-defined modulo 8.) On the other hand, E_4 reduces to a modular form modulo 8 whose q-expansion is also equal to the constant 1 (because 8 divides 240). Thus by the q-expansion principle, we see that $E_4 \equiv A^4 \bmod 8$. Thus $|E_4(E)| > |2^{8/3}|$ if and only if $|A(E)|^4 > |2^{8/3}|$ (because $|2^{8/3}| > |8|$, and so the second equality can be checked after reducing modulo 8, where A^4 makes sense) if and only if $|A(E)| > |2^{2/3}|$ (because $|2^{2/3}| > |2|$, and so the second inequality can be checked after reducing modulo 2, where A makes sense). Here $|A(E)|$ is denoting the valuation of the Hasse invariant of E reduced over $\mathcal{O}/2$ (which is non-Noetherian highly non-reduced local ring!), *not* E reduced over $k = \mathcal{O}/\mathfrak{m}$.

Thus we see that the annulus $|1| \geq |j| > |256|$ consists precisely of those j-invariants for which the corresponding elliptic curve E satisfies $|A(E)| > |2^{2/3}|$. We may think of these as the j-invariants of elliptic curves with "not too supersingular" reduction modulo two (in the words of [**79**]; recall that the annulus $|j| = |1|$ corresponds precisely to the elliptic curves having good ordinary reduction), and so the disk D_1 of elliptic curves consists of those elliptic curves having not too supersingular reduction (in the preceding sense) along with the elliptic curves having bad reduction.

The theory for arbitrary primes p

In this section we explain the generalization to an arbitrary prime p of the results presented above in the case $p = 2$. Let $\mathbb{C}_p$ denote the completion of the algebraic closure of $\mathbb{Q}_p$, let $\mathcal{O}$ denote its ring of integers, and let $\mathfrak{m}$ denote the maximal ideal of $\mathcal{O}$.

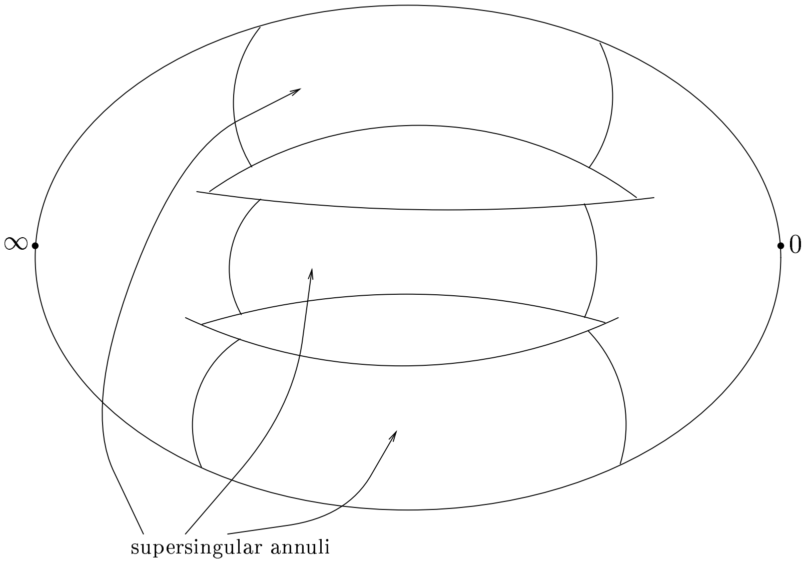
We let $Y_0(p)$ denote the modular curve whose points parameterize p-isogenies between elliptic curves, and $X_0(p)$ its completion. Over $\mathbb{C}$ one can construct $X_0(p)$ as the quotient $\mathcal{H}^*/\Gamma_0(p)$, and so see that it has two cusps, 0 and ∞. As p increases the genus of $X_0(p)$ also increases, and in particular $X_0(p)$ is not a rational curve unless $p = 13$ or $p < 11$. Thus one cannot in general describe $X_0(p)$ by a single parameter, as we described $X_0(2)$ by the parameter j_2.

Nevertheless, $Y_0(p)$ is an affine curve, and so it has an affine ring, which is generated by parameters which classify p-isogenies up to isomorphism (and which can be constructed explicitly via modular forms). By the Lefschetz principle, we see that we may equally regard $Y_0(p)$ as a curve over $\mathbb{C}_p$, since the moduli problem of classifying p-isogenies will depend on the same invariants whether solved over $\mathbb{C}$ or over $\mathbb{C}_p$. Then $X_0(p)$ will be the completion of $Y_0(p)$, a smooth projective curve over $\mathbb{C}_p$.

The process of passing from an isogeny to its dual yields an involution of $X_0(p)$, which we denote by w_p. Passing to the j-invariant of the source of an isogeny is a morphism $B_1 : X_0(p) \longrightarrow X_0(1)$. We let B_2 denote the composition $B_1 \circ w_p$.

We let $X_0(1)^h$ denote the set of j-invariants in $X_0(p)$ corresponding to elliptic curves with ordinary reduction (that is, either bad reduction or good ordinary reduction). The complement of $X_0(1)^h$ in $X_0(1)$ is a disjoint union of disks: each disk is a congruence class modulo $\mathfrak{m}$ of j-invariants which are congruent to a particular supersingular j-invariant in characteristic p. We refer to these disks as the supersingular disks in $X_0(1)$.

We let $X_0(p)^h$ denote the preimage of $X_0(1)^h$ under B_1^{-1}. Just as in the case of $p = 2$, and for the same reason (that being ordinary or supersingular is an isogeny class invariant in characteristic p), $X_0(p)^h$ is invariant under w_p. Furthermore, it is the union of two connected components which are interchanged by w_p, which we label $X_0(p)^h_\infty$ (the component containing ∞) and $X_0(p)^h_0$ (the component containing 0), and the restriction of B_1 to $X_0(p)^h_\infty$ is an isomorphism onto $X_0(1)^h$. (In order to interpret *connected component* in a more than intuitive sense, one must use the language of rigid analytic geometry.) The complement of $X_0(p)^h_\infty$ and $X_0(p)^h_0$ in $X_0(p)$ is a disjoint union of annuli, each the preimage under B_1 of one of the supersingular disks in $X_0(1)$. We refer to these as the *supersingular annuli* in $X_0(p)$. There is the following picture that goes with this discussion, in which one thinks of $X_0(p)$ as being obtained by gluing together two copies of the "sphere with holes" $X_0(1)^h$ (in the guise of $X_0(p)^h_\infty$ and $X_0(p)^h_0$) via the supersingular annuli:



$X_0(p)$ drawn as a union of two "spheres with holes" glued along the supersingular annuli

To prove all the facts just stated, one must study the modular curve $X_0(p)$ as a *scheme* over $\mathbb{Z}_p$, as in [**39, 85**]. The assertions then follow in a fairly standard fashion from the known structure of $X_0(p)$ over $\mathbb{Z}_p$; see the discussion and references in [**25**].

For now, note that since $B_1 : X_0(p)^h_\infty \to X_0(1)^h$ is an isomorphism, it has an inverse $B_1^{-1} : X_0(1)^h \to X_0(p)^h$, and thus any elliptic curve with ordinary reduction is equipped with a canonical subgroup of order p. Just as in the case of $p = 2$, there is a larger set D_1 containing $X_0(1)^h$ to which the map B_1^{-1} extends, and which we now describe.

Let A denote the Hasse invariant in characteristic p. If E is any elliptic curve with good reduction, we may consider its reduction modulo p (which is an elliptic curve over the ring $\mathcal{O}/p$), and the inequality $|A(E)| > |p^{p/(p+1)}|$ makes sense, since $|p^{p/(p+1)}| > |p|$. Define D_1 to the the subset of $X_0(1)$ consisting of the set of points in $X_0(1)^h$ together with those j-invariants of elliptic curves having good reduction whose Hasse invariant satisfies this inequality. Then just as in the case of $p = 2$, the map B_1^1 extends to a map $B_1^{-1} : D_1 \to X_0(p)$ which is an isomorphism between D_1 and its image D_2. If $j(E)$ lies in D_1 then the map B_1^{-1} endows E with a canonical subgroup of order p.

Let $D_1^{(n)}$ denote the disk which is the union of $X_0(1)^h$ and those j-invariants of good reduction whose corresponding elliptic curve satisfies $|A(E)| > |p^{1/p^{n-1}(p+1)}|$. Just as in the case $p = 2$, the Deligne-Tate map $\phi := B_2 \circ B_1^{-1}$ is a degree p map from $D_1^{(n)}$ onto $D_1^{(n-1)}$, and in particular if $j(E)$ lies in $D_1^{(n)}$ then E has a canonical cyclic subgroup of order p^n.

When written in terms of the uniformizing parameter q at ∞ the Deligne-Tate map has the form $q \mapsto q^p$. We will give a rather lengthy explanation of this fact, which will lead us toward an explanation of the existence of the Deligne-Tate map itself.

The Deligne-Tate map corresponds in modular terms to taking the quotient of an elliptic curve E by its canonical subgroup. Let us first try to understand why it is that the Deligne-Tate map is defined on the formal neighbourhood of the cusp ∞ on $X_0(1)$. In other words, why does E_{Tate} have a canonical subgroup? Or again, why is there a map B_1^{-1} inverse to B_1 defined on the formal neighbourhood of ∞?

This form of the question we can answer: the map B_1 is a unramified at the point ∞ on $X_0(p)$, and so induces an isomorphism of the formal neighbourhood of ∞ in $X_0(p)$ with the formal neighbourhood of ∞ in $X_0(1)$. The map B_1^{-1} is simply the inverse to this isomorphism. The existence of B_1^{-1} means that the Tate curve $E_{\mathrm{Tate}} = \mathbb{G}_m/q^{\mathbb{Z}}$ is equipped with a canonical subgroup of order p. To compute this subgroup, we will look at the analytic picture over $\mathbb{C}$. (This is valid, since the curves $X_0(1)$ and $X_0(p)$, the projection B_1, the formal section B_1^{-1} of B_1 on the formal neighbourhood of $j = \infty$, and the Tate curve $\mathbb{G}_m/q^{\mathbb{Z}}$, are all defined over $\mathbb{Q}$, and thus so is the resulting subgroup of order p in $\mathbb{G}_m/q^{\mathbb{Z}}$. To determine this subgroup, it suffices to compute over the overfield $\mathbb{C}$ of $\mathbb{Q}$, where, as observed above, one can use analytic methods; the information we get is then perfectly applicable to the overfield $\mathbb{C}_p$ of $\mathbb{Q}$ that we are actually interested in.)

If τ is a point in the upper half-plane then the image of τ in $X_0(p)$ is the isogeny $E_\tau \longrightarrow E_{p\tau}$ given by $z \bmod \Lambda_\tau \mapsto pz \bmod \Lambda_{p\tau}$. The kernel of this isogeny is the group $\dfrac{2\pi i}{p}\mathbb{Z}/\mathbb{Z}$. Exponentiating this description of the isogeny, we see that it can also be described as the map $\mathbb{C}^\times/q^{\mathbb{Z}} \longrightarrow \mathbb{C}^\times/q^{p\mathbb{Z}}$ given by $t \mapsto t^p$, whose kernel is the image of the group μ_p of p^{th} roots of unity in $\mathbb{G}_m/q^{\mathbb{Z}}$.

We conclude that the canonical subgroup of $E_{\mathrm{Tate}} = \mathbb{G}_m/q^{\mathbb{Z}}$ is the image of μ_p in $\mathbb{G}_m/q^{\mathbb{Z}}$. The quotient of $\mathbb{G}_m/q^{\mathbb{Z}}$ by this group is isomorphic to $\mathbb{G}_m/q^{p\mathbb{Z}}$, and thus is a Tate elliptic curve with parameter q^p. Hence in terms of the parameter q around the cusp ∞, the Deligne-Tate map is given by the formula $q \mapsto q^p$.

The above argument actually provides the key insight as to the existence of canonical subgroups in general. Let us first explain this for elliptic curves over $\mathbb{C}_p$ having bad (multiplicative) reduction. First note that such curves are parameterized by the disk $|j| > |1|$ in $X_0(1)$. The usual power-series relating j and q (that is, $q = 1/j + 744/j^2 + \cdots$, which has integer coefficients) shows that q provides an analytic isomorphism of the disk $|j| > |1|$ in $X_0(1)$ with the open unit disk in $\mathbb{C}_p$. Since the power-series defining the Tate curve have coefficients lying in $\mathbb{Z}$, we may specialize them at any point q in this open unit disk to obtain an elliptic curve $\mathbb{C}_p^\times/q^{\mathbb{Z}}$, whose j-invariant will be that value of j corresponding to our given choice of q. Thus the elliptic curves over $\mathbb{C}_p$ having bad reduction are obtained as specializations of the Tate curve at points of the open unit q-disk (this is Tate's theory of uniformization of elliptic curves with multiplicative reduction). We may evaluate the formal section B_1^{-1} at any such point, and since this evaluation commutes with specialization, our preceding calculation with the Tate curve shows that the canonical subgroup of the curve $\mathbb{C}_p^\times/q^{\mathbb{Z}}$ will simply be the image of the subgroup μ_p of $\mathbb{C}_p^\times$.

If E is an elliptic curve over $\mathbb{C}_p$ with good ordinary reduction, then E does not admit a description in terms of the Tate curve, but the situation is almost as good. Let E denote a model for with good reduction modulo $\mathfrak{m}$. Then one finds that the formal group of E (thought of as a formal group over $\mathcal{O}$) is isomorphic to the formal multiplicative group $\hat{\mathbb{G}}_m$. (This is a consequence of the fact that the reduction modulo $\mathfrak{m}$ of each of these formal groups is of height one, so that they are isomorphic over k, together with a Newton's method/Hensel's lemma argument, which allows one to lift this isomorphism to an isomorphism over $\mathcal{O}$.) Now the p-torsion in $\hat{\mathbb{G}}_m$ is just the group μ_p, and so this isomorphism yields a copy of μ_p as a subgroup of E. This is the canonical subgroup of E.

To describe the canonical subgroup of a non-ordinary elliptic curve E whose j-invariant lies in D_1 is more difficult. One examines the formal group of E (whose reduction modulo $\mathfrak{m}$ is now of height two) and uses the bound on the Hasse invariant of E (which one knows by virtue of the assumption that $j(E)$ lies in D_1) to construct a canonically determined subgroup of this formal group of order p. The details can be found in [**79**].

p-adic modular forms

Recall again the algebraic definition of a modular form of weight k and of level one defined over the field $\mathbb{C}_p$: it is a rule that attaches to a pair (E, ω) consisting of an elliptic curve and a non-zero regular differential defined over $\mathbb{C}_p$ a number $f(E, \omega)$ which satisfies the weight k transformation rule and which behaves well in families, including those that include generalized elliptic curves among their fibres. From now on we will refer to such forms as *classical* modular forms.

A p-adic modular form of weight k is such a rule which is defined only on those pairs (E, ω) for which E has ordinary reduction (that is, as we said above, has either bad or good ordinary reduction modulo $\mathfrak{m}$). Thus any classical modular form also gives rise to a p-adic modular form, but there are also many p-adic modular forms which are not classical modular forms. For example, if $p = 2$ and E has ordinary reduction, then $E_4(E, \omega) \neq 0$, and so $E_4(E, \omega)^{-1}$ is well-defined. Thus E_4^{-1} is a 2-adic modular form of weight -4. Another way to construct p-adic modular forms for any prime p is to notice that ordinary elliptic curves have canonical subgroups of order p^n for all n, and so any classical modular form on $\Gamma_0(p^n)$ gives rise to a p-adic modular form of level one.

The space of all p-adic modular forms of weight k (for some fixed p and k) is a p-adic Banach space, and just as in the classical theory this space has Hecke operators acting on it. Unfortunately, the action of these operators is rather hard to control, and one does not get, and cannot expect to get, a very good spectral theory. Thus there is not a very good theory of p-adic Hecke eigenforms.

To deal with this, one looks at a more refined type of p-adic modular form. An *overconvergent* p-adic modular form is a rule of the usual type which is defined on those pairs (E, ω) for which $j(E)$ lies in $D_1^{(n)}$ for some n. (The n is fixed for any particular overconvergent form, but may vary from form to form.) All our examples of p-adic modular forms given above are actually overconvergent modular forms: for example, we saw that E_4 is never zero on the disc D_1 in $X_0(2)$, and so E_4^{-1} is defined on that disc. Also, any classical modular form on $\Gamma_0(p^n)$ for some n yields an overconvergent modular form defined on $D_1^{(n)}$, since the elliptic curves with j-invariant in this region are precisely those which have a canonical subgroup of order p^n. Here is an example of a p-adic modular form which is *not* overconvergent: the "weight two Eisenstein series of level one" is the q-expansion

$E_2 = 1 + 24\sum_{n=1}^{\infty} \sigma(n)q^n$, where $\sigma(n) = \sum_{d|n} d$. This is not a classical modular form at all (it doesn't transform correctly under the substitution $\tau \mapsto -1/\tau$), but *is* the q-expansion of a p-adic modular form for every p (see [**79**], in which it is denoted by P), which is *not* overconvergent for any prime p [**31**].

One can again define Hecke operators acting on overconvergent modular forms, and one finds that something special happens. If one works with the Hecke operator U_p (which on q-expansions is defined by $U_p(\sum_n a_n q^n) = \sum_n a_{np} q^n)$) rather than T_p, then there is a good theory of Hecke eigenforms, because the operator U_p is a *completely continuous* (or, in alternative language, *compact*) operator on the space of overconvergent modular forms, and so has a good spectral theory. More precisely, for any non-zero $\lambda \in \mathbb{C}_p$, there is a finite-dimensional space of overconvergent modular forms on which U_p has eigenvalue λ. This space will be preserved by the Hecke operators T_ℓ (as ℓ ranges over all primes different from p), and to construct Hecke eigenforms with U_p-eigenvalue λ it suffices to diagonalize these commuting operators on this finite-dimensional vector space.

The reason that U_p is completely continuous is the following: recall that the Deligne-Tate map is given in terms of the parameter q by the formula $\phi(q) = q^p$. Thus one sees that the operator U_p can be described geometrically via the *trace* of the Deligne-Tate map. More precisely, it is $1/p$ times this trace. It is simplest to explain this for U_p acting on modular forms of weight zero, which is to say modular functions. Then an overconvergent modular function is simply an analytic function f on one of the regions $D_1^{(n)}$. We claim that

$$U_p(f)(j) = \frac{1}{p} \sum_{\phi(j')=j} f(j')$$

for any point $j \in D_1^{(n)}$. To see this, it suffices to verify the formula for those j lying in the residue disc about ∞, where we can compute in terms of q; it will follow in general by analytic continuation. The claim then follows from the above formula for the action of U_p on q-expansions, together with the following simple piece of algebra:

$$\frac{1}{p} \sum_{q'^p=q} \sum_n a_n q'^n = \sum_n a_{np} q^n.$$

(See [**79**] for a precise form of this argument.) Why does this formula imply that U_p is completely continuous? Well, note that if $j \in D_1^{(n)}$ and $\phi(j') = j$, then $j' \in D_1^{(n+1)}$. Thus computing $U_p(f)$ involves restricting f from the region $D_1^{(n)}$ to the proper subregion $D_1^{(n+1)}$, and such restriction operators are always completely continuous. (This is Montel's theorem from complex analysis being applied in the (simpler) p-adic setting.)

The preceding discussion shows that the fact that the Deligne-Tate map is defined on the regions $D_1^{(n)}$ which extend some way into the supersingular annuli is fundamental to obtaining a good spectral theory for overconvergent p-adic modular forms. This explains the importance of the Deligne-Tate map in the theory of p-adic modular forms, and hence why we have devoted our efforts in this appendix to describing the ideas behind the construction of this map.

Guide to the literature

This final section contains a brief review of the literature on the p-adic theory of modular curves and modular forms, which I hope will be helpful to someone trying to enter the field. I have tried to present as accurate an account of the development and current state of the field as I can, within the limits of my own understanding of these matters. I apologize in advance for any omissions or oversights that I may inadvertently have made.

The p-adic aspects of the theory of modular curves that we have discussed in this appendix seem to make their first appearance in the literature in [**43**], in which Dwork constructs and studies "by hand" the Deligne-Tate map on the region D_1. He states that the existence of this map was conjectured by Tate (based on a calculation for $p = 2$; I don't know if it was a similar calculation to that of section 2 above) and first proved in general by Deligne. In subsequent articles [**44, 45**] Dwork studied the spectral theory of the completely continuous operator U_p on the space of overconvergent modular forms of level zero. One of the points emphasized in [**45**] in particular is the importance of restricting ones attention to overconvergent forms if one hopes to obtain a reasonable spectral theory.

In an independent line of research, Swinnerton-Dyer [**150**] began the systematization of the theory of congruences of q-expansions of modular forms (a subject which seems to have begun with the work of Ramanujan), by studying the ring of modular forms modulo p. This work is also reported on in [**131**]. In [**134**] Serre extends these results to develop a theory of congruences of modular forms modulo arbitrary powers of p, and introduces the notion of p-adic modular forms.

In [**79**] Katz gives a systematic presentation of the p-adic geometry of the modular curves, including the construction of canonical subgroups of "not too supersingular elliptic curves" (which construction he attributes to Lubin) and the consequent construction of B_1^{-1} and the Deligne-Tate map, as well as of the theory of congruences of modular forms modulo p. The results of this article, and the later article [**81**] (which extends the techniques of [**79**] to define *generalized p-adic modular functions*, a notion that includes as a special case the p-adic modular forms of [**79**] and [**134**]), subsumed most of those of Dwork, Serre and Swinnerton-Dyer mentioned above, and also generalized them to modular forms of arbitrary level.

However, although [**79**] in some sense unified the various existing p-adic theories, the applications of these theories were in two different directions. On the one hand the papers [**80, 82, 83, 134**] used the theory of congruences of modular forms to construct p-adic analytic families of modular forms (essentially families of Eisenstein series) and hence to construct p-adic L-functions (which appeared as the special values of these Eisenstein series, either at the cusp ∞ or at certain special j-values corresponding to complex-multiplication elliptic curves). On the other hand, the papers [**150, 151**] were concerned with using the same theory of congruences not to construct families of modular forms, but rather to understand the image of the Galois representations attached to Hecke eigenforms (as constructed in [**37**]). Thus their main concern was not congruences between arbitrary modular forms, but the possible congruences that could arise between Hecke eigenforms. This was something that was not dealt with in the theory of [**81**], and this may go some way to explaining the remark on this paper made in the introduction of [**151**].

Studying properties of Galois representations via congruences of modular forms turned out to be a very fruitful idea. One key direction of research was the investigation of congruences between cuspidal Hecke eigenforms and Eisenstein series. This is the main theme of the seminal paper [**94**], and formed the basis for a successful attack on the so-called *main conjecture of Iwasawa theory* over the field $\mathbb{Q}$, beginning in [**118**], continuing in [**159**], and culminating with the proof of the main conjecture in [**105**]. The investigation of possible congruences among the cuspidal eigenforms themselves proved to be an equally important topic, a discussion of which would unfortunately take us too far afield, and which is provided by Ribet's article in this volume.

In Hida's papers [**71, 70**] the properly p-adic aspect of the theory of congruences made its resurgence. In these pivotal papers Hida connected the study of congruences of eigenforms and the associated Galois representations with the study of p-adic families of modular forms by constructing *p-adic analytic families of p-adic Hecke eigenforms* and attaching to them p-adic analytic families of Galois representations. (See also [**106**] for a discussion of Hida's construction and a more refined analysis of these Galois representations.) The main technical constraint on Hida's results is that they only construct families of *ordinary* eigenforms; that is, eigenforms whose U_p eigenvalues are p-adic units.

The influence of Hida's theory was enormous. It allowed a simplification of the proof of the main conjecture of Iwasawa theory, by rephrasing it as the question of analyzing the intersection locus of the Eisenstein family with the cuspidal part of Hida's family. By extending Hida's theory to the context of Hilbert modular forms, Wiles was able to prove the main conjecture for arbitrary totally real fields [**160**]; this harks back to Serre's study of p-adic L-functions as the constant term of families of Eisenstein series. Other developments included the construction of families of p-adic L-functions attached to Hida families of cuspforms [**64, 88, 96**]. Greenberg and Stevens [**64**] used these L-functions to prove the weight two case of the conjecture of Mazur, Tate and Teitelbaum [**104**]. (See also [**65**], which presents the main ideas of their argument in a simplified setting.) of p-adic modular forms, Hida's work also motivated Mazur to develop his theory of deformations of Galois representations [**97**] (for more on which, see the main body of this article!), which proved decisive for the further development of the theory of p-adic modular forms.

In [**57**] Gouvêa used Mazur's theory in order to associate a p-adic Galois representation to each p-adic modular form, whether ordinary or not. This work marked the beginning of line of research aimed at constructing p-adic analytic families of p-adic Hecke eigenforms and Galois representations, analogous to those constructed by Hida, but in the non-ordinary situation. That such families should exist was the principal conjecture of [**61**]. This work also raised a number of questions related to overconvergent p-adic modular forms, refocusing attention on an aspect of the p-adic theory that had languished since the appearance of papers of Dwork cited above. The p-adic analytic viewpoint favoured by Dwork reemerged in a series of papers by Coleman [**25, 26, 27, 28**]. The first of these papers relates to the conjecture of [**104**], while the second and third show that an overconvergent p-adic eigenform whose U_p-eigenvalue is not too divisible by p (in a sense depending on the weight of the form in question) is necessarily classical (extending the analogous result in the ordinary case, due to Hida [**70**], and hence answering one of the questions of [**57**]).

The current course of research in the theory of p-adic modular forms has been set by the papers [**28**] and [**23**]. In the first of these papers, Coleman analyzes the variation of the spectral theory of the U_p operator on overconvergent p-adic modular forms as a function of the weight and hence constructs p-adic analytic families of Hecke eigenforms, in particular proving in a qualitative form the conjectures of [**61**]. In [**23**] this construction is globalized to construct for each prime p a p-adic rigid analytic curve called the *eigencurve.* This is, roughly speaking, the universal parameter space for p-adic analytic families of overconvergent Hecke eigenforms of finite slope (that is, whose U_p-eigenvalue is non-zero; the slope of a Hecke eigenform is the p-adic valuation of its U_p-eigenvalue). The deformation theory of Galois representations plays a key role in the construction of the eigencurve, and the universal family of finite-slope eigenforms comes equipped with a p-adic analytic family of p-adic Galois representations.

With the construction of the eigencurve, it appears that the various directions of research that instigated the development of the theory of p-adic modular forms have achieved synthesis. The overconvergent forms are understood as being the necessary "glue" that binds classical Hecke eigenforms together into analytic families. Furthermore, the existence of positive slope families will certainly have many number-theoretic applications. One such example is [**148**], in which Stevens develops a theory of p-adic L-functions in the context of Coleman's families, and uses them [**149**] to prove the higher-weight case of the conjecture of [**104**] (in the more precise form stated in [**25**]).

Yet the construction of the eigencurve raises as many questions as it answers. For example, while the structure of Hida's families of ordinary forms is very precisely understood, the structure of the higher-slope parts of the eigencurve is essentially a mystery (the references [**24, 47, 147**] provide some information on a small part of this structure for certain small values of p). Relatedly, the quantitative aspects of the conjectures of [**61**] remain unresolved; all that is known are some very special cases dealt with in [**24, 47, 147**].

Another question is that of the existence of p-adic analytic families of eigenforms of infinite slope. Such forms, whose U_p-eigenvalue vanishes, aren't accessible via the standard spectral theory techniques available for the study of the completely continuous operator U_p. One natural question that arises in this context is: are there infinite slope forms that can be written as the limit of finite slope forms? Such forms would correspond to "punctures" in the eigencurve, which would be filled in by the limiting infinite slope form. Coleman [**29**] has shown that forms obtained by twisting finite slope forms with the Teichmüller character are of this type. In the same reference, Coleman reports on a computation of Stein suggesting that some non-twist infinite slope forms might also be of this type. As to whether infinite slope forms move in p-adic families, some computations for forms on $\Gamma_0(4)$ [**46**] suggest that the answer could be yes, but essentially nothing is known in this direction.

One of the most intriguing and general questions raised by the construction of the eigencurve is that of the intrinsic Galois-theoretic interpretation of the Galois representations that it parameterizes (these are just the Galois representations attached to overconvergent p-adic Hecke eigenforms, and this problem was already posed in [**57**]). In [**23**] it is shown that the eigencurve is the rigid analytic Zariski closure (in an appropriate ambient space, which is essentially a deformation space of Galois representations) of those points corresponding to classical modular forms.

The Galois representations associated to such forms are conjecturally characterized as being those which are *potentially semistable* (see [**53**] and the references contained therein for a discussion of this conjecture and the notion of potential semistability). However no analogous intrinsic characterization of the non-classical points on the eigencurve is known, even on the conjectural level (see [**86, 87**] for some investigations in this direction). Finding such a description of the eigencurve, which would not depend on the "extrinsic" construction via the theory of overconvergent p-adic modular forms, is related to the problem of generalizing the construction of the eigencurve to other settings. For a discussion of what such settings might be, and the possible relations to Iwasawa theory and a general theory of p-adic L-functions, see [**102**].

Matthew Emerton
Department of Mathematics
University of Michigan
Ann Arbor, MI 48109
emerton@math.lsa.umich.edu

BIBLIOGRAPHY

1. V. Arnold, M. Atiyah, P. Lax, and B. Mazur (eds.), *Mathematics: Frontiers and perspectives*, American Mathematical Society, 2000.
2. W. Barth and H. Lange (eds.), *Arithmetic of complex manifolds (Erlangen, 1988)*, Lecture Notes in Mathematics, no. 1399, Springer-Verlag, 1989.
3. B. Birch and W. Kuijk (eds.), *Modular functions of one variable IV*, Lecture Notes in Mathematics, vol. 476, Springer-Verlag, Berlin, Heidelberg, New York, 1975.
4. S. Bloch and K. Kato, *L-functions and Tamagawa numbers of motives*, In Cartier et al. [**17**], pp. 333–400.
5. G. Böckle, *Demuškin groups with group actions and applications to deformations of Galois representations*, Preprint, 1998.
6. ———, *The generic fiber of the universal deformation space associated to a tame Galois representation*, Manuscripta Math. **96** (1998), 231–246.
7. G. Böckle and A. Mézard, *The prime-to-adjoint principle and unobstructed Galois deformations in the Borel case*, J. Number Theory **78** (1999), 167–203.
8. N. Boston, *Deformation theory of Galois representations*, Ph.D. thesis, Harvard University, 1987.
9. ———, *Explicit deformation of Galois representations*, Invent. Math. **103** (1991), 181–196.
10. N. Boston and B. Mazur, *Explicit universal deformations of Galois representations*, Advanced Studies in Pure Mathematics **17** (1989), 1–21.
11. N. Boston and Stephen V. Ullom, *Representations related to CM elliptic curves*, Math. Proc. Cambridge Philos. Soc. **113** (1993), 71–85.
12. N. Bourbaki, *Éléments de Mathématique: Théorie des ensembles*, Hermann, Paris, 1968.
13. J. Buhler, *Elliptic curves and modular forms*, in this volume.
14. H. Carayol, *Formes automorphes et représentations galoisiennes*, Seminar on Number Theory, 1981/1982, Univ. Bordeaux I, Talence, 1982, pp. Exp. No. 31, 20.
15. H. Carayol, *Sur les représentations ℓ-adiques associées aux formes modulaires de Hilbert*, Ann. Sci. École Norm. Sup. (4) **19** (1986), 409–468.
16. ———, *Formes modulaires et représentations Galoisiennes à valeurs dans un anneau local complet*, In Mazur and Stevens [**103**], pp. 213–237.

17. P. Cartier, L. Illusie, N. M. Katz, G. Laumon, Y. I. Manin, and K. A. Ribet (eds.), *The Grothendieck Festschrift I*, Birkhauser, 1990.
18. J. W. S. Cassels and J. Fröhlich (eds.), *Algebraic number theory (Brighton, 1965)*, Academic press, 1967.
19. J. Coates and C.-G. Schmidt, *Iwasawa theory for the symmetric square of an elliptic curve*, J. Reine Angew. Math. **375/376** (1987), 104–156.
20. J. Coates and A. Sydenham, *On the symmetric square of a modular elliptic curve*, In Coates and Yau [**22**], pp. 152–171.
21. J. Coates and M. J. Taylor (eds.), *L-functions and arithmetic*, London Mathematical Society Lecture Notes, vol. 153, Cambridge University Press, Cambridge, 1991.
22. J. Coates and S.-T. Yau (eds.), *Elliptic curves, modular forms and Fermat's last theorem (Hong Kong, 1993)*, Cambridge, MA, International Press, 1997.
23. R. Coleman and B. Mazur, *The eigencurve*, (A. J. Scholl and R. L. Taylor, eds.), London Mathematical Society Lecture Note Series, vol. 254, Cambridge University Press, 1998, pp. 1–113.
24. R. Coleman, G. Stevens, and J. Teitelbaum, *Numerical experiments on families of p-adic modular forms*, preprint.
25. R. F. Coleman, *A p-adic Shimura isomorphism and p-adic periods of modular forms*, In Mazur and Stevens [**103**], pp. 21–51.
26. ———, *Classical and overconvergent modular forms*, Invent. Math. **124** (1996), 215–241.
27. ———, *Classical and overconvergent modular forms of higher level*, J. Théor. Nombres Bordeaux **9** (1997), no. 2, 395–403.
28. ———, *p-adic Banach spaces and families of modular forms*, Invent. Math. **127** (1997), 417–479.
29. ———, *Conjectures and rumours of conjectures*, notes to a talk posted on http://www.math.berkeley.edu/ coleman/Sems/Sp00/coleman.html, 2000.
30. R. F. Coleman and B. Edixhoven, *On the semi-simplicity of the U_p-operator on modular forms*, Math. Ann. **310** (1998), 119–127.
31. R. F. Coleman, F. Q. Gouvêa, and N. Jochnowitz, *E_2, Θ, and overconvergence*, Internat. Math. Res. Notices (1995), no. 1, 23–41.
32. B. Conrad, *The flat deformation functor*, In Cornell et al. [**33**], Papers from the Instructional Conference on Number Theory and Arithmetic Geometry held at Boston University, Boston, MA, August 9–18, 1995, pp. 373–420.
33. G. Cornell, J. Silverman, and G. Stevens (eds.), *Modular forms and Fermat's last theorem*, Springer-Verlag, Berlin, Heidelberg,New York, 1997, Papers from the Instructional Conference on Number Theory and Arithmetic Geometry held at Boston University, Boston, MA, August 9–18, 1995.
34. H. Darmon, F. Diamond, and R. Taylor, *Fermat's last theorem*, In Coates and Yau [**22**].
35. A. J. de Jong, *Crystalline Dieudonné module theory via formal and rigid geometry*, Inst. Hautes Études Sci. Publ. Math. (1995), no. 82, 5–96.
36. B de Smit and H. W. Lenstra, *Explicit construction of universal deformation rings*, In Cornell et al. [**33**], Papers from the Instructional Conference on Number Theory and Arithmetic Geometry held at Boston University, Boston, MA, August 9–18, 1995, pp. 313–326.
37. P. Deligne, *Formes modulaires et représentations ℓ-adiques, Exposé 355*, Séminaire Bourbaki 1968/1969, Springer-Verlag, 1971, pp. 139–172.

38. ______, *Courbes elliptiques: formulaire d'après j. tate*, In Birch and Kuijk [**3**], pp. 53–73.
39. P. Deligne and M. Rapoport, *Les schémas de modules de courbes elliptiques*, (Berlin, Heidelberg, New York) (P. Deligne and W. Kuijk, eds.), Lecture Notes in Mathematics, vol. 349, Springer-Verlag, 1973.
40. P. Deligne and J.-P. Serre, *Formes modulaires de poids* 1, Ann. Sci. École Norm. Sup. (4) **7** (1974), 507–530, In J.-P. Serre, *Œuvres*, volume III.
41. F. Diamond and J. Im, *Modular forms and modular curves*, In Murty [**110**], Papers from the seminar held at the Fields Institute for Research in Mathematical Sciences, Toronto, Ontario, 1993–1994, pp. 39–133.
42. C. Doran and S. Wong, *Deformations of Galois representations and modular forms*, Based on lectures of Barry Mazur at Harvard University, Fall 1993.
43. B. Dwork, *p-adic cycles*, Inst. Hautes Études Sci. Publ. Math. **37** (1969), 27–115.
44. ______, *On Hecke polynomials*, Invent. Math. **12** (1971), 249–256.
45. ______, *The* U_p *operator of Atkin on modular functions of level 2 with growth conditions*, In Kuijk and Serre [**89**].
46. M. Emerton, Unpublished computations, 1998.
47. ______, 2*-adic modular forms of minimal slope*, Ph.D. thesis, Harvard University, 1998.
48. M. Flach, *A generalisation of the Cassels-Tate pairing*, J. Reine Angew. Math. **412** (1990), 113–127.
49. ______, *Selmer groups for the symmetric square of an elliptic curve*, Ph.D. thesis, University of Cambridge, 1990.
50. ______, *A finiteness theorem for the symmetric square of an elliptic curve*, Invent. Math. **109** (1992), 307–327.
51. ______, *Annihilation of Selmer groups for the adjoint representation of a modular form*, In Murty [**110**], Papers from the seminar held at the Fields Institute for Research in Mathematical Sciences, Toronto, Ontario, 1993–1994, pp. 249–265.
52. J.-M. Fontaine and B. Mazur, *Geometric galois representations*, In Coates and Yau [**22**], pp. 41–78.
53. ______, *Geometric Galois representations*, In Coates and Yau [**22**], pp. 41–78.
54. J. Fröhlich, *Local fields*, In Cassels and Fröhlich [**18**], pp. 1–41.
55. W. Fulton and J. Harris, *Representation theory: a first course*, Springer-Verlag, 1991.
56. P. Gabriel, *Groupes formels*, Schémas en Groupes (Sém. Géometrie Algébrique, Inst. Hautes Études Sci., 1963/64), vol. Fasc. 2b, Exposé 7b, Inst. Hautes Études Sci., Paris, 1965, pp. 66–152+3.
57. F. Q. Gouvêa, *Arithmetic of p-adic modular forms*, Lecture Notes in Mathematics, vol. 1304, Springer-Verlag, Berlin, Heidelberg, New York, 1988.
58. ______, *Deforming Galois representations: controlling the conductor*, J. Number Theory **34** (1990), 95–113.
59. ______, *p-adic numbers: an introduction*, Springer-Verlag, Berlin, Heidelberg, New York, 1993.
60. ______, *Deforming galois representations: a survey*, In Murty [**110**], Papers from the seminar held at the Fields Institute for Research in Mathematical Sciences, Toronto, Ontario, 1993–1994, pp. 179–207.

61. F. Q. Gouvêa and B. Mazur, *Families of modular eigenforms*, Math. Comp. **58** (1992), 793–806.
62. ———, *On the density of modular representations*, Computational perspectives on number theory (Chicago, IL, 1995), American Mathematical Society, Providence, RI, 1998, pp. 127–142.
63. R. Greenberg, *Iwasawa theory for elliptic curves*, in this volume.
64. R. Greenberg and G. H. Stevens, *p-adic L-functions and p-adic periods of modular forms*, Invent. Math. **111** (1993), 407–447.
65. ———, *On the conjecture of Mazur, Tate and Teitelbaum*, In Mazur and Stevens [**103**], pp. 183–211.
66. B. H. Gross, *Kolyvagin's work on modular elliptic curves*, In Coates and Taylor [**21**], pp. 235–256.
67. Alexander Grothendieck, *Technique de descente et théorèmes d'existence en géométrie algébrique II. Le théorème d'existence en théorie formelle des modules*, Séminaire Bourbaki, Vol. 5 (Paris), Soc. Math. France, 1995, Exp. No. 195, pp. 369–390.
68. K. Haberland, *Galois cohomology of algebraic number fields*, VEB Deutscher Verlag, 1978.
69. Robin Hartshorne (ed.), *Algebraic geometry: Arcata 1974*, American Mathematical Society, 1975.
70. H. Hida, *Galois representations into* $\mathrm{GL}_2(\mathbb{Z}_p[[X]])$ *attached to ordinary cusp forms*, Invent. Math. **85** (1986), 545–613.
71. ———, *Iwasawa modules attached to congruences of cusp forms*, Ann. Sci. École Norm. Sup. (4) **19** (1986), 231–273.
72. ———, *Theory of p-adic Hecke algebras and Galois representations*, Sugaku Expositions **2** (1989), 75–102.
73. ———, *Nearly ordinary Hecke algebras and Galois representations of several variables*, JAMI Innaugural Conference Proceedings, supplement to Amer. J. Math (1990), 115–134.
74. J.-I. Igusa, *Class number of a definite quaternion with prime discriminant*, Proc. Nat. Acad. Sci. USA **44** (1958), 312–314.
75. Y. Ihara, K. A. Ribet, and J.-P. Serre (eds.), *Galois groups over* $\mathbb{Q}$ *(Berkeley 1990)*, Mathematical Sciences Research Institute Publications, no. 16, Springer-Verlag, 1990.
76. U. Jannsen, *Über Galoisgruppen lokaler Körper*, Invent. Math. **70** (1982/83), 53–69.
77. U. Jannsen and K. Wingberg, *Die Struktur der absoluten Galoisgruppe* $\mathfrak{p}$*-adisher Zahlkörper*, Invent. Math. **70** (1982/83), 71–98.
78. K. Joshi and C. Khare, *On ordinary forms and ordinary Galois representations*, Preprint, 1995.
79. N. M. Katz, *p-adic properties of modular schemes and modular forms*, In Kuijk and Serre [**89**].
80. ———, *p-adic L-functions via moduli of elliptic curves*, In Hartshorne [**69**].
81. ———, *Higher congruences between modular forms*, Ann. of Math. (2) **101** (1975), 332–367.
82. ———, *p-adic interpolation of real analytic Eisenstein series*, Ann. of Math. (2) **104** (1976), 459–571.
83. ———, *The Eisenstein measure and p-adic interpolation*, Am. Journ. Math. **99** (1977), 238–311.

84. N. M. Katz and S. Lang, *Finiteness theorems in geometric class field theory*, L'Enseignement Mathématique **XVIII** (1981), 285–319.
85. N. M. Katz and B. Mazur, *Arithmetic moduli of elliptic curves*, Annals of Mathematics Studies, vol. 108, Princeton University Press, Princeton, New Jersey, 1985.
86. M. Kisin, *Periods for p-adic modular forms*, Preprint, 1999.
87. ______, *p-adic modular forms and the Fontaine-Mazur conjecture*, Preprint, 2000.
88. K. Kitigawa, *On standard p-adic l-functions of families of elliptic cusp forms*, In Mazur and Stevens [**103**], pp. 81–110.
89. W. Kuijk and J.-P. Serre (eds.), *Modular functions of one variable III*, Lecture Notes in Mathematics, vol. 350, Springer-Verlag, Berlin, Heidelberg, New York, 1973.
90. S. Lang, *Algebra*, third ed., Addison-Wesley, 1993.
91. W. Li, *Newforms and functional equations*, Math. Ann. **212** (1975), 285–315.
92. Saunders MacLane, *Categories for the working mathematician*, Springer-Verlag, 1971.
93. P. A. Martin, *Deformações de representações Galoisianas ordinárias e de representações não ramificadas*, Ph.D. thesis, Universidade de São Paulo, 1991.
94. B. Mazur, *Modular curves and the Eisenstein ideal*, Inst. Hautes Études Sci. Publ. Math. **47** (1977), 33–186.
95. ______, *Rational isogenies of prime degree*, Invent. Math. **44** (1978), 129–162.
96. ______, *Two-variable p-adic L-functions*, unpublished manuscript, 1985.
97. ______, *Deforming Galois representations*, In Ihara et al. [**75**], pp. 385–437.
98. ______, *Two-dimensional p-adic Galois representations unramified away from p*, Compositio Math. **74** (1990), 115–133.
99. ______, *Galois deformations and Hecke curves*, Harvard University course notes, 1994.
100. ______, *An "infinite fern" in the universal deformation space of galois representations*, Collect. Math. **48** (1997), 155–193, Journées Arithmétiques (Barcelona, 1995).
101. ______, *An introduction to the deformation theory of Galois representations*, In Cornell et al. [**33**], Papers from the Instructional Conference on Number Theory and Arithmetic Geometry held at Boston University, Boston, MA, August 9–18, 1995, pp. 243–311.
102. ______, *The theme of p-adic variation*, In Arnold et al. [**1**], pp. 433–459.
103. B. Mazur and G. Stevens (eds.), *p-adic monodromy and the birch-swinnerton-dyer conjecture*, Contemporary Mathematics, vol. 165, American Mathematical Society, 1994.
104. B. Mazur, J. Tate, and J. Teitelbaum, *On p-adic analogues of the conjectures of Birch and Swinnerton-Dyer*, Invent. Math. **84** (1986), 1–48.
105. B. Mazur and A. Wiles, *Class fields of abelian extensions of* $\mathbb{Q}$, Invent. Math. **76** (1984), 179–330.
106. ______, *On p-adic analytic families of Galois representations*, Compositio Math. **59** (1986), 231–264.
107. J. S. Milne, *Arithmetic duality theorems*, Academic Press, 1986.
108. T. Miyake, *Modular forms*, Springer-Verlag, 1989.
109. P. Morandi, *Field and Galois theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1996.

110. V. K. Murty (ed.), *Seminar on Fermat's Last Theorem*, Providence, RI, American Mathematical Society, 1995, Papers from the seminar held at the Fields Institute for Research in Mathematical Sciences, Toronto, Ontario, 1993–1994.
111. J. Neukirch, A. Schmidt, and K. Wingberg, *Cohomology of number fields*, Springer-Verlag, Berlin, Heidelberg, New York, 2000.
112. J. Neukrich, *Class field theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1986.
113. L. Nyssen, *Pseudo-représentations*, Math. Ann. **306** (1996), 257–283.
114. Joseph Oesterlé, *Travaux de Wiles (et Taylor, ...). II*, Astérisque (1996), no. 237, Exp. No. 804, 5, 333–355, Séminaire Bourbaki, Vol. 1994/95.
115. R. Ramakrishna, *On a variation of Mazur's deformation functor*, Compositio Math. **87** (1993), 269–286.
116. M. Raynaud, *Schémas en groupes de type* $(p,p,\dots,p)$, Bull. Soc. Math. France **102** (1974), 241–280.
117. K. A. Ribet, *Lectures on modular forms*, in this volume.
118. ______, *A modular construction of unramified p-extensions of* $\mathbb{Q}(\mu_p)$, Invent. Math. **34** (1976), 151–162.
119. ______, *Galois representations and modular forms*, Bull. Amer. Math. Soc. (N.S.) **32** (1995), 375–402.
120. D. Robinson, *A course in the theory of groups*, Springer-Verlag, 1982.
121. R. Rouquier, *Caractérisation des caractères et pseudo-caractères*, J. Algebra **180** (1996), 571–586.
122. K. Rubin, *The work of Kolyvagin on the arithmetic of elliptic curves*, In Barth and Lange [**2**], pp. 128–136.
123. ______, *Kolyvagin's system of Gauss sums*, In van der Geer et al. [**154**], pp. 309–324.
124. ______, *Euler systems*, Princeton University Press, Princeton, New Jersey, 1999.
125. M. Schlessinger, *Functors of Artin rings*, Trans. A. M. S. **130** (1968), 208–222.
126. L. Schneps (ed.), *The Grothendieck theory of dessins d'enfants*, London Mathematical Society Lecture Note Series, vol. 200, Cambridge, Cambridge University Press, 1994.
127. A. Scholl, *An introduction to Kato's Euler system*, In Scholl and Taylor [**128**], pp. 379–460.
128. A. Scholl and R. Taylor (eds.), *Galois representations in arithmetic algebraic geometry (Durham, 1996)*, Cambridge, Cambridge Universtiy Press, 1998.
129. S. Sen, *Continuous cohomology and p-adic Galois representations*, Invent. Math. **62** (1980/81), 89–116.
130. ______, *An infinite-dimensional Hodge-Tate theory*, Bull. Soc. Math. France **121** (1993), 13–34.
131. J.-P. Serre, *Congurences et formes modulaires (d'après H. P. F. Swinnerton-Dyer)*, Seminaire Bourbaki (1971/72), no. 416.
132. ______, *Propriétés galoisiennes des points d'ordre fini des courbes elliptiques*, Invent. Math. **15** (1972), 259–331.
133. ______, *A course in arithmetic*, Springer-Verlag, Berlin, Heidelberg, New York, 1973.
134. ______, *Formes modulaires et fonctions zêta p-adiques*, In Kuijk and Serre [**89**], In J.-P. Serre, *Œuvres*, vol III.
135. ______, *Local fields*, Springer-Verlag, Berlin, Heidelberg, New York, 1974.

136. ———, *Quelques applications du théorème de densité de Chebotarev*, Inst. Hautes Études Sci. Publ. Math. **54** (1981), 323–401.

137. ———, *Sur les représentations modulaires de degré* 2 *de* Gal($\overline{\mathbf{Q}}/\mathbf{Q}$), Duke Math. J. **54** (1987), 179–230.

138. ———, *Topics in Galois theory*, Jones and Bartlett Publishers, Boston, MA, 1992, Lecture notes prepared by Henri Damon [Henri Darmon], With a foreword by Darmon and the author.

139. ———, *Cohomologie Galoisienne*, fifth ed., Springer-Verlag, Berlin, 1994.

140. ———, *Galois cohomology*, Springer-Verlag, Berlin, 1997, Translated from the French by Patrick Ion and revised by the author.

141. J.-P. Serre and D. B. Zagier (eds.), *Modular functions of one variable V*, Lecture Notes in Mathematics, vol. 601, Springer-Verlag, Berlin, Heidelberg, New York, 1977.

142. I. R. Shafarevich, *Algebraic number fields*, Proceedings of the International Congress of Mathematicians, Stockholm 1962 (Djursholm), Inst. Mittag-Leffler, 1963, Translated version reprinted in I. R. Shafarevich, *Collected Mathematical Papers* (Springer-Verlag, 1989), pp. 283–294, pp. 163–176.

143. S. S. Shatz, *Profinite groups, arithmetic and geometry*, Annals of Mathematics Studies, vol. 67, Princeton University Press, Princeton, NJ, 1972.

144. G. Shimura, *Introduction to the arithmetic theory of automorphic forms*, Princeton University Press, 1971.

145. ———, *The special values of zeta functions associated with cusp forms*, Comm. Pure Appl. Math. **6** (1976), 783–804.

146. J. H. Silverman, *The arithmetic of elliptic curves*, Springer-Verlag, Berlin, Heidelberg, New York, 1986.

147. L. Smithline, *Exploring slopes of p-adic modular forms*, Ph.D. thesis, University of California at Berkeley, 2000.

148. G. Stevens, *Rigid analytic modular symbols*, unpublished manuscript, 1994.

149. ———, *Coleman's $\mathcal{L}$-invariant and families of modular forms*, unpublished manuscript, 1996.

150. H. P. F. Swinnerton-Dyer, *On ℓ-adic representaions and congruences for coefficients of modular forms*, In Kuijk and Serre [**89**].

151. ———, *On ℓ-adic representations and congruences for coefficients of modular forms*, In Serre and Zagier [**141**], pp. 63–90.

152. J. Tate, *Galois cohomology*, in this volume.

153. D. L. Ulmer, *A construction of local points on elliptic curves over modular curves*, Internat. Math. Res. Notices (1995), 349–363.

154. G. van der Geer, F. Oort, and J. Steenbrink (eds.), *Arithmetic algebraic geometry (Texel, 1989)*, Birkhauser, 1991.

155. L. C. Washington, *Introduction to cyclotomic fields*, Springer-Verlag, Berlin, Heidelberg, New York, 1982.

156. ———, *Galois cohomology*, In Cornell et al. [**33**], Papers from the Instructional Conference on Number Theory and Arithmetic Geometry held at Boston University, Boston, MA, August 9–18, 1995, pp. 101–120.

157. T. Weston, *Euler systems and arithmetic geometry*, Notes from a course given by Barry Mazur at Harvard University, available at `http://www.math.harvard.edu/weston/mazur.html`, 1998.

158. ———, *On Selmer groups of geometric Galois representations*, Ph.D. thesis, Harvard University, 2000.

159. A. Wiles, *Modular curves and the class group of* $\mathbb{Q}(\zeta_p)$, Invent. Math. **58** (1980), 1–35.

160. ———, *The Iwasawa conjecture for totally real fields*, Ann. of Math. **131** (1990), 493–540.

161. K. Wingberg, *Der Eindeutigkeitssatz für Demuškinformationen*, Invent. Math. **70** (1982/83), 99–113.