3. Hecke algebras and unitary groups

It is finally time for the unitary group G_0 introduced in the first lecture to make its appearance. Recall that our basic object is a cohomological automorphic representation Π of $GL(n,F)$, satisfying several axioms that guarantee its descent to both unitary groups G and G_0. We henceforward assume that Π has cohomology with trivial coefficients. We fix a level subgroup $K_f \subset G_0(\mathbf{A}_f)$ such that $\pi'^{,K_f} \neq 0$. With some additional work we can even assume $\dim \pi'^{,K_f} = 1$ in practice, but this will not be necessary. Then the L-packet π' has non-zero intersection with the space of functions

$$M'(K_f,\mathbb{C}) = C(G_0(\mathbb{Q})\backslash G_0(\mathbf{A})/G_0(\mathbb{R})\cdot K_f,\mathbb{C}) = M'(K_f,\mathbb{Z})\otimes\mathbb{C}$$

where $K_f \subset G_0(\mathbf{A}_f)$ is a compact open subgroup and $M'(K_f,\mathbb{Z})$ is the free $\mathbb{Z}$-module of integer-valued functions. Fixing the level K_f guarantees that the module $M'(K_f,\mathbb{Z})$ is free of finite rank. We will impose the following conditions on K_f:

(1) $K_f = \prod_v K_v$ as v ranges over finite primes of F^+.
(2) At primes that remain inert in F/F^+, K_v is a hyperspecial maximal compact subgroup.
(3) At primes that ramify in F/F^+, we could assume K_v is a "very special" maximal compact subgroup, in the terminology of Labesse. However, this condition can be ignored, because we can always assume after a quadratic base change that F/F^+ is everywhere unramified.
(4) If v splits in F/F^+, and if $v \notin S$, then $K_v \simeq GL(n,\mathcal{O}_v)$.
(5) If $v \in S$, K_v is adapted to the situation. In practice, $S = Q\cup R$ where Q are the Taylor-Wiles primes, at which

$$K_v = \{g \in GL(n,\mathcal{O}_v) \mid g \equiv \begin{pmatrix} *_{n-1,n-1} & *_{n-1,1} \\ 0_{1,n-1} & 1 \end{pmatrix} \pmod{\mathfrak{m}_v}\}.$$

and R are the level-raising primes, at which

$$K_v = \{g \in GL(n,\mathcal{O}_v) \mid g \pmod{\mathfrak{m}_v} \text{ is upper-triangular unipotent }\}$$

We will not really be able to give a precise account of the final condition. In the present version of the article we can always use base change to assume K_v contains an Iwahori subgroup of $GL(n,\mathcal{O}_v)$, but it is unnecessarily restrictive to make this hypothesis.

Let Π and $\rho = \rho_{\Pi,\ell}$ be as in the previous lectures, with L_{Π_∞} the finite-dimensional representation of the compact group $G_{0,\mathbb{R}}$ in which Π_∞ has cohomology.. For now we work on the Π side. Henceforward we make the following simplifying assumption:

$$L(\Pi_\infty) = \mathbb{C}. \tag{3.1}$$

It follows that, if π_0 is the G_0-avatar of Π, then $\pi_{0,\infty}$ is the trivial representation of $G_0(F^+_\infty)$; moreover, the Galois representation ρ_ℓ will ultimately be realized in the middle-dimensional cohomology of an $n-1$-dimensional Shimura variety with $\mathbb{Q}_\ell$-coefficients. This hypothesis is irrelevant to the modularity theorems but it suffices for the applications to the Sato-Tate Conjecture, and it spares us a lot of notation. In particular, the non-trivial Hodge-Tate numbers are all of the form $h^{i,n-1-i} = 1$ with $0 \leq i \leq n-1$.

As is customary we begin by introducing an ℓ-adic integer ring $\mathcal{O}$, with fraction field K and residue field k, a finite extension of $\mathbb{F}_\ell$. Our Hecke algebras and deformation rings will all be $\mathcal{O}$-algebras.

The subspace of the space of $\mathbb{C}$-valued automorphic forms on G_0 generated by automorphic representations π_0 with $\pi_{0,\infty} = \mathbb{C}$ is just the space of automorphic forms on G_0 on which $G_0(\mathbb{R})$ acts trivially, namely

$$S(G_0, \mathbb{C})) = S_{triv}(G_0, \mathbb{C})) := C^\infty(G_0(F^+)\backslash G_0(\mathbf{A})/G_0(\mathbb{R})), \mathbb{C}). \tag{3.2}$$

The space $Sh(G_0) = G_0(F^+)\backslash G_0(\mathbf{A})/G_0(\mathbb{R}))$ is a profinite set, in fact a zero-dimensional Shimura variety, and the notation C^∞ denotes the space of locally constant functions. In (3.2) these functions are taken with values in $\mathbb{C}$, but we could just as well take values in $\mathcal{O}$, or more generally in any $\mathcal{O}$-algebra A:

$$S(G_0, A)) := C^\infty(Sh(G_0), A). \tag{3.3}$$

This can be viewed as the cohomology in degree zero of $Sh(G_0)$, and obviously behaves well with respect to base change: if $A \to B$ is a homomorphism of $\mathcal{O}$-algebras, then $S(G_0, A) \otimes_{\mathcal{O}} B \xrightarrow{\sim} S(G_0, B)$ under the natural map. This is not always true for cohomology in higher degrees of more general Shimura varieties, not to mention the locally symmetric spaces attached to $GL(n)$, and is one of the advantages of working with G_0.

We fix a set T of primes of F^+ which will be the primes at which our π_0 (or Π, or ρ) will be allowed to ramify. We assume

$$T = S(B) \cup S_\ell \cup S_1 \cup R$$

where S_ℓ is the set of divisors of ℓ, S_1 is a *non-empty* set of auxiliary primes (descended from the $\mathfrak{r}$ of the original Taylor-Wiles paper) which allows us to eliminate elliptic fixed points in $Sh(G_0)$, and R is the set of primes at which Taylor studies possible level-raising in [T]. There will also be sets of primes disjoint from T, denoted Q_N, as N varies among positive integers; these are the Taylor-Wiles primes, used in the patching method. We let $T(Q_N) = T \cup Q_N$. These primes have the following properties:

3.4.1 All primes in $T(Q_N)$ split in F/F^+.
3.4.2 If $v \in S_1$ lies above a rational prime p then $[F(\zeta_p) : F] > n$.
3.4.3 If $v \in R$ then $\mathbf{N}v \equiv 1 \pmod{\ell}$.
3.4.4 If $v \in Q_N$ then $\mathbf{N}v \equiv 1 \pmod{\ell^N}$.

Let $\tilde{T}$ denote a set of liftings of T to primes of F, so that $\tilde{T} \coprod \tilde{T}^c$ is the set of all primes of F above T; if $v \in T$ let $\tilde{v}$ be the corresponding element of $\tilde{T}$. For any Q_N we define $\tilde{T}(Q_N)$ in the same way.

For split primes v we identify $G_0(F_v^+)$ with $GL(n, F_w)$ for some w dividing v (we choose $\tilde{v}$ for $v \in T$). Now choose an open compact subgroup U of $G_0(\mathbf{A}_f)$, $U = \prod_v U_v$, where v runs over finite primes of F^+, such that

3.5.1 If $v \notin T$, or if $v \in S_\ell$, then U_v is a hyperspecial maximal compact subgroup of $G_0(F_v^+)$.

3.5.2 If $v \in R$ then U_v is an Iwahori subgroup.
3.5.3 If $v \in S_1$ then U_v is the principal congruence subgroup of level v:

$$U_v = \{g \in GL(n, \mathcal{O}_{F,\tilde{v}}) \mid g \equiv 1 \pmod{\mathfrak{m}_{\tilde{v}}}\}$$

where $\mathfrak{m}_{\tilde{v}}$ is the maximal ideal of $\mathcal{O}_{F,\tilde{v}}$.
3.5.4 If $v \in Q_N$ then

$$U_v = U_{1,v} := \{g \in GL(n, \mathcal{O}_{F,\tilde{v}}) \mid g \equiv \begin{pmatrix} g_{n-1} & *_{n-1} \\ 0_{n-1} & 1 \end{pmatrix} \pmod{\mathfrak{m}_{\tilde{v}}}\}$$

where $g_{n-1} \in GL(n-1, \mathcal{O}_{F,\tilde{v}})$ and $*_{n-1}$ (resp. 0_{n-1}) is an arbitrary column matrix of height $n-1$ (resp. the zero row matrix of width $n-1$).

We write $\mathcal{O}_v = \mathcal{O}_{F,\tilde{v}}$ for simplicity, and let $k(v)$ denote its residue field. One likewise defines $U_{0,v} \supset U_{1,v}$ by weaking the condition in (3.5.4) so that the lower right-hand entry is an arbitrary element of $k(v)^\times$. For $v \in S(B)$, (3.5.2) implies that U_v is a maximal compact subgroup, the multiplicative group of a maximal order of B_v; for $v \in R$ U_v can be identified with integral matrices whose reduction modulo $\tilde{v}$ is upper-triangular, which we denote I_v (Iw_v in [CHT,T]). Let q_v be the order of the residue field $k(v)$, a power of the prime p_v. Let $I(1)_v \subset I_v$ be the p_v-Sylow subgroup, the matrices whose reduction modulo $\tilde{v}$ is upper-triangular unipotent; mapping to the diagonal entries thus identifies

$$I_v/I(1)_v \xrightarrow{\sim} (k(v)^\times)^n \tag{3.6}$$

A character of $I_v/I(1)_v$ is denoted $\chi_v = (\chi_{1,v}, \ldots, \chi_{n,v})$ where each $\chi_{i,v}$ is a character of $k(v)^\times$. A character of $U_{0,v}/U_{1,v} \xrightarrow{\sim} k(v)^\times$ is denoted ψ_v^0.

Let χ_v be as above, for $v \in R$, and define

$$S_{\{\chi_v\}}(U, A) = \{f \in S(G_0, A) \mid f(gu) = \prod_{v \in R} \chi_v^{-1}(u_v) f(g)\} \tag{3.7}$$

for all $g \in G_0(\mathbf{A}_f)$ and $u = \prod u_v \in U$. This is the module on which our Hecke algebras act. Suppose $A = \mathbb{C}$ (don't worry about its $\mathcal{O}$-algebra structure); then $S_{\{\chi_v\}}(U, \mathbb{C})$ is the space of vectors in the space of automorphic forms on

$$G_0(F^+)\backslash G_0(\mathbf{A})/G_0(\mathbb{R}) \cdot \prod_{v \notin R} U_v$$

on which $\prod_{v \in R} U_v$ acts by the indicated character. In particular, the only automorphic representations π_0 that contribute to $S_{\{\chi_v\}}(U, \mathbb{C})$ are those with non-trivial fixed vectors under $\prod_{v \notin R} U_v \times \prod_{v \in R} I(1)_v$ Our choice of U_v for $v \in S(B)$ implies that any π_0 has a base change Π to $GL(n, F)$ for which Π_v is an abelian twist of the Steinberg representation. In order to allow more discrete series local factors at $v \in S(B)$ (as required by condition (3)) we would need to allow representations of U_v of dimension > 1 and consider vector-valued forms with values in these representations, tensored over the places in $S(B)$. This is the point of view of [CHT] and [T]. For simplicity we prefer not to work with vector-valued forms in these notes.

However, the reader is advised that certain steps in the proof of the Sato-Tate conjecture require the use of such vector-valued forms.

For a place v of F^+ we let Γ_v denote a decomposition group at v. Here is how the conditions on primes in $T(Q_N)$ translate into conditions on the Galois representation $\rho = \rho_{\Pi_\ell}$, which we assume takes values in $GL(n, \mathcal{O})$[1] We write $\bar{\rho}$ for the reduction of ρ modulo the maximal ideal of $\mathcal{O}$.

3.8.1(a) If $v \notin T$, then $\rho \mid_{\Gamma_v}$ is unramified;

3.8.1(b) If $v \in S_\ell$, then $\rho \mid_{\Gamma_v}$ is crystalline.

3.8.2 If $v \in R$ then $\rho \mid_{\Gamma_v}$ may be more ramified than $\bar{\rho}$, and that is the issue resolved in [T].

3.8.3 If $v \in S_1$ then ρ is unramified at v and $\bar{\rho}$ has no deformation to a representation ramified at v.

3.8.4 If $v \in Q_N$ then ρ is unramified at v but $\bar{\rho}$ has certain deformations to representations ramified at v, and the point of the Taylor-Wiles method, as generalized in [CHT] and [T], is to use these additional deformations to bound the size of the ring of all deformations in terms of the Hecke algebra; see §3.13 for details.

The assertions (3.8.1) and (3.8.2) can be justified on the basis of the information presented up to now. This is not true of (3.8.3) and (3.8.4). The need to choose sets S_1 and Q_N with these properties requires us to impose additional hypotheses on $Im(\bar{\rho})$. That such choices are possible then follows from an argument using Chebotarev density, as in the original article of Taylor-Wiles. This will be explained below.

Now let $A = \mathcal{O}$. Our Hecke algebra is a finite free $\mathcal{O}$-algebra, given with an explicit infinite family of generators. Each *split* prime $v \notin T(Q_N)$ contributes n generators. One can also include generators at non-split primes outside T, but these are unnecessary, basically because the split primes of F have Dirichlet density 1 (this is not true of the primes of F^+ that split in F!).

Let w be a prime of F split over F^+, v its restriction to F^+, so that v factors as $w \cdot w^c$. Then $G_0(F_v^+) \xrightarrow{\sim} GL(n, F_w)$, and $\pi_{0,v} \xrightarrow{\sim} \Pi_w$, as we saw above. The Hecke polynomial attached to the unramified representation Π_w of $GL(n, F_w)$ was originally determined by Shimura and is presented in his red book on modular forms. The coefficients of this polynomial define the n Hecke operators at v; replacing the Hecke operators by their eigenvalues on the 1-dimensional $U_v \simeq GL(n, \mathcal{O}_w)$-fixed subspace of $\pi_{0,v}$, and the variable by q_v^{-s}, yields the inverse of the local Euler factor $L(s, \Pi_w)$.

Explicitly, let ϖ_w be a uniformizer at w. The Hecke operators are double coset operators

$$T_v^{(j)} = U_v(diag(\varpi_w I_j, I_{n-j}))U_v \subset G_0(F_v^+), j = 1, \ldots, n, \tag{3.9}$$

These operators act in the usual way on $S_{\{\chi_v\}}(U, A)$ for any A, as does $(T_v^{(n)})^{-1}$, which is just translation by an element of the center. Explicitly, if $f \in S_{\{\chi_v\}}(U, \mathbb{Z})$,

[1]This is true when ρ is viewed as a representation of $Gal(\overline{\mathbb{Q}}/F)$; ρ extends to a representation of $Gal(\overline{\mathbb{Q}}/F^+)$ with values in the L-group of G_0), as discussed above.

we have

$$T_v^{(j)} f(h) = \int_{G_{0,v}} f(hg')T_v^{(j)}(g')dg'$$

where dg' is the $\mathbb{Q}$-valued bi-invariant Haar measure on $G_{0,v}$ that gives volume 1 to K_v. Since the integrand is right-invariant under K_v, this is a sum of values of a $\mathbb{Z}$-valued function with $\mathbb{Z}$ coefficients.

We assume $\mathcal{O}$ chosen sufficiently large to include all eigenvalues of all Hecke operators introduced thus far acting on the finite-dimensional $\mathbb{Q}$-vector space $S_{\{\chi_v\}}(U, \mathbb{Q})$. We denote

$$\mathbb{T}^T_{\{\chi_v\}}(U)$$

the $\mathcal{O}$-subalgebra of $End(S_{\{\chi_v\}}(U, \mathcal{O}))$ generated by the $T_v^{(j)}$, $j = 1, \ldots, n$, together with $(T_v^{(n)})^{-1}$, for all split unramified $v \notin T$. For $v \in Q_N$ we need to add additional (non-spherical) Hecke operators V_v defined in the appendix, (A.3), for the reasons explained there (see also the discussion above Theorem 3.15.4). This is a finite free commutative $\mathcal{O}$-subalgebra of $End(S_{\{\chi_v\}}(U, \mathcal{O}))$. The commutativity is a consequence of the commutativity of the local Hecke operators $T_v^{(j)}$ for each v, and this is part of the theory of the Satake transform. Since $\mathbb{T}^T_{\{\chi_v\}}(U)$ is finite over $\mathcal{O}$, it is a semi-local ring, and we let $\mathfrak{m}_i \subset \mathbb{T}^T_{\{\chi_v\}}(U)$ denote its maximal ideals, as i runs over a finite index set.

Remark. Let G be a unitary group of signature $(n-1, 1)$ at one prime, definite at other primes. The point of this construction is that $\mathbb{T}^T_{\{\chi_v\}}(U)$ is the same product of local Hecke algebras (over primes $v \notin S$ that split in F/F^+) that acts on the cohomological automorphic forms on $GL(n, F)$ and on the middle-dimensional cohomology $H^{n-1}(Sh(G), \mathbb{Q}_\ell)$, where the avatar of Π is the L-packet we've denoted $\{\pi\}$.

The unramified local Langlands correspondence considered in §1 is normalized so that

$$P_w(X) = X^n + \sum_{j=1}^{n} (-1)^j q_w^{j(j-1)/2} T_v^{(j)} X^{n-j},$$

with each $T_v^{(j)}$ specialized to its eigenvalue for Π_w, is the characteristic polynomial of $\rho_{\Pi,\ell}(Frob_w)$ for any ℓ, where $Frob_w$ is geometric Frobenius.

We will be using

Proposition 3.10. *The algebra* $\mathbb{T}^T_{\{\chi_v\}}(U)$ *is reduced.*

This follows easily from the fact that $S_{\{\chi_v\}}(U, \mathbb{C})$ is a semisimple $T(K_f)$-module, which in turn comes from the fact that $T(K_f)$ is hermitian with respect to the L_2-inner product (Petersson norm). Recall that G_0 is anisotropic, so all automorphic forms on G_0 are square integrable.

Proposition 3.11. *Write $\mathbb{T}(U) = \mathbb{T}^T_{\{\chi_v\}}(U)$. (a) There is a representation*

$$\rho_{U,\{\chi_v\}} : \Gamma_F \to GL(n, \mathbb{T}(U)) \otimes \mathbb{Q}_\ell$$

with the property that, for any irreducible representation τ of G_0 as above, admitting base change to a cuspidal cohomological automorphic representation $B(\tau)$ of $GL(n, F)$, the projection

$$\rho_\tau = p_\tau \circ \rho_{U,\{\chi_v\}} : \Gamma_F \to GL(n, \overline{\mathbb{Q}}_\ell)$$

is an l-adic representation of geometric type, unramified outside $S \cup S_\ell$), with

$$L^{S \cup S_\ell)}(s, \rho_\tau) = L^{S \cup S_\ell)}(s, B(\tau)).$$

(b) Let $E(\Pi)$ be the set of τ as in (a) such that (loosely speaking) p_τ factors through $\mathbb{T}_\Pi \cap \mathbb{T}(U)$ ($\mathbb{T}_\Pi$ is the localization at the maximal ideal $\mathfrak{m} = \mathfrak{m}_\Pi$, denoted $\mathbb{T}^T_{\{\chi_v\}}(U)_\mathfrak{m}$ in the next section). The projection

$$\rho_\Pi = \oplus_{\tau \in E(\Pi)} \rho_\tau : G_F \to GL(n, \oplus_{\tau \in E(\Pi)} \overline{\mathbb{Q}}_\ell)$$

is conjugate to a homomorphism $\rho_{\mathbb{T}_\Pi}$ with values in the subgroup $GL(n, \mathbb{T}_\Pi) \subset GL(n, \oplus_{\tau \in E(\Pi)} \overline{\mathbb{Q}}_\ell)$. The reduction of ρ_Π modulo the maximal ideal of $\mathbb{T}_\Pi$ is equivalent to $\bar{\rho}_\Pi$.

(c) The homomorphism $\rho_{\mathbb{T}_\Pi}$ extends to a homomorphism

$$r_{\mathbb{T}_\Pi} : G_{F^+} \to \mathcal{G}_n(T)$$

whose reduction modulo the maximal ideal is equivalent to $\bar{r}_\Pi$.

Assertion (a) is another way of expressing the theorem of Kottwitz and Clozel described in the second lecture. Assertion (b) is a consequence of a very useful theorem of Carayol, and depends on the hypothesis that $\bar{r}_\Pi$ be absolutely irreducible. The last claim in (b) is the assertion that all the Galois representations associated to $\tau \in E(\Pi)$ have the same reduction modulo ℓ. This is not difficult to show – it suffices to calculate traces of Frobenius modulo ℓ, and the point is that these are determined by the eigenvalues of the Hecke operators modulo ℓ, and completion at $\mathfrak{m}_\Pi$ simply picks out all the τ whose Hecke eigenvalues are congruent modulo ℓ. But it deserves to be emphasized, since this is precisely the basis of the theory of congruences of modular forms, as developed by Serre, Mazur, Katz, and Ribet in the 1970s, and extended in various directions by Hida, Wiles, Taylor, and now many others. Assertion (c) makes use of the Petersson pairing.

3.12. Surjectivity of the map $R \to \mathbb{T}$.. Now recall the automorphic representation Π of $GL(n)$. We associate a maximal ideal $\mathfrak{m} = \mathfrak{m}_\Pi \subset \mathbb{T}^T_{\{\chi_v\}}(U)$ to Π. First we descend Π to π_0 as in §2. The Hecke algebra acts on the U-invariants in π_0 by a character λ_Π and we let $\mathfrak{m} \subset \mathbb{T}^T_{\{\chi_v\}}(U)$ be the maximal ideal containing $ker\ \lambda_\Pi$, to which we add the elements $V_v - \alpha_v$ for $v \in Q_N$ with the chosen eigenvalue α_v (or any lifting of $\bar{\alpha}_v$ to $\mathcal{O}$). The localization $S_{\{\chi_v\}}(U, \mathcal{O})$ at $\mathfrak{m}$ consists roughly of those forms congruent to (the U-invariants of) of π_0, and form a module for $\mathbb{T}^T_{\{\chi_v\}}(U)_\mathfrak{m}$. We have already introduced the deformation ring $R^{univ} = R_{\bar{\rho},\mathcal{S}}$ of the residual representation $\bar{\rho}$, with conditions at S corresponding to 3.8.1-4 above. Universality, together with Proposition 3.11, implies there is a map R^{univ} to $\mathbb{T}^T_{\{\chi_v\}}(U)_\mathfrak{m} \otimes \overline{\mathbb{Q}}$. We will assume

Hypothesis 3.12.1. *The ideal* $\mathfrak{m}$ *is not Eisenstein; i.e.* $\bar{\rho}$ *is absolutely irreducible.*

This hypothesis, together with the theorem of Carayol, implies that this factors through a map

$$\phi_U : R^{univ} \to \mathbb{T}^T_{\{\chi_v\}}(U)_{\mathfrak{m}} \tag{3.12.2}$$

The generators of the Hecke algebra at a (good split) prime v appear as the images under ϕ_U of the coefficients of the characteristic polynomials of $Frob_v$. The proof that the image of ϕ_U also contains the operators V_v, for $v \in Q_N$, which depends naturally on the decomposition of $\rho_\Pi(Frob_v)$ discussed below (cf. (3.13.7), is a bit more subtle and is not explained here. If we could show that (3.12.2) is an isomorphism, then the reciprocity conjecture of (1.6) would follow for any lifting of $\bar{\rho}$ satisfying the conditions used to define $R_{\bar{\rho}}$. In fact, it is not known in general that (3.12.2) is an isomorphism, and specifically it is not known in the cases relevant to the Sato-Tate conjecture. But it is known that the map on irreducible components in characteristic zero is a bijection, and this is sufficient.

We write $R_{\bar{r}_\Pi}$ (or $R_{\bar{\rho}_\Pi}$) for R^{univ}, $\mathbb{T}_\Pi$ (or $\mathbb{T}_{\bar{\Pi}}$) for $\mathbb{T}^T_{\{\chi_v\}}(U)_{\mathfrak{m}}$.

This is the place to remark that the set of v that split in F/F^+, with a finite subset removed, suffice to determine any ℓ-adic representation of Γ_F, because their extensions to F have Dirichlet density 1. Practically every step of the argument makes reference to Chebotarev's density theorem, which in the present setting allows us to determine an ℓ-adic representation up to equivalence by the traces of Frobenius at primes v belonging to a set of Dirichlet density 1. This also suffices for the surjectivity assertion in the following Corollary:

Corollary. *There is a surjective homomorphism in* $\hat{\mathcal{C}}_{\mathcal{O}}$*:*

$$\phi_\emptyset : R_{\bar{r}_\Pi,\emptyset} \to \mathbb{T}_\Pi.$$

The $\emptyset$ has been appended to the subscript in anticipation of the introduction of Taylor-Wiles primes. To prove this surjection is an isomorphism, one applies the method of Taylor-Wiles, as simplified by Diamond and Fujiwara (see notes on patching).

3.13. The Taylor-Wiles primes Q_N.

As indicated in 3.8.4, the primes $v \in Q_N$ are chosen so that ρ is unramified at v. It is assumed more pertinently that

$$\bar{\rho} \mid \Gamma_v = \bar{\alpha}_v \oplus \bar{s}_v \tag{3.13.1}$$

where $\bar{\alpha}_v$ is an unramified character that **does not occur as a subquotient** of $\bar{s}_v$. Since $\mathbf{N}_v \equiv 1 \pmod{\ell}$, this means in particular that

3.13.2. *There are no non-trivial Γ_v-extensions between $\bar{\alpha}_v$ and $\bar{s}_v$.*

The proof is by calculating Galois cohomology. Choose a prime $\tilde{v}$ of F dividing v. We consider
(3.13.3)
$H^1(\Gamma_{\tilde{v}}, ad\bar{\rho}) = H^1(\Gamma_{\tilde{v}}, k) \oplus H^1(\Gamma_{\tilde{v}}, \bar{\alpha}_v \otimes \bar{(s)}_v^\vee) \oplus H^1(\Gamma_{\tilde{v}}, \bar{\alpha}_v^{-1} \otimes \bar{s}_v) \oplus H^1(\Gamma_{\tilde{v}}, ad\bar{s}_v)$.

The first factor corresponds to the trivial representation $\alpha_v \otimes \alpha_v^{-1}$. This is calculated by the inflation restriction sequence. Since $\bar{\rho}$ is unramified at v, we have the short exact sequence

$$0 \to H^1(\Gamma_{\tilde{v}}/I_{\tilde{v}}, ad\bar{\rho}) \to H^1(\Gamma_{\tilde{v}}, ad\bar{\rho}) \to H^1(I_{\tilde{v}}, ad\bar{\rho})^{\Gamma_{\tilde{v}}/I_{\tilde{v}}} \to 0$$

where the final 0 is $H^2(\Gamma_{\tilde{v}}/I_{\tilde{v}}, ad\bar{\rho})$ which vanishes because $Gal(\bar{k}(\tilde{v})/k(\tilde{v}))$ is procyclic. We rewrite this

$$0 \to H^1_{unr}(\Gamma_{\tilde{v}}, M) \to H^1(\Gamma_{\tilde{v}}, M) \to Hom_{\Gamma_{\tilde{v}}/I_{\tilde{v}}}(I_{\tilde{v}}/I_{\tilde{v}}^{\ell}, M) \to 0 \tag{3.13.4}$$

where M is either $ad\bar{\rho}$ or any of the four summands that appears in the right hand side of (3.13.3). Now the condition $\mathbf{N}_v \equiv 1 \pmod{\ell}$ is exactly the condition that $I_{\tilde{v}}/I_{\tilde{v}}^{\ell}$ is the trivial $< Frob_v >$-module $\mathbb{F}_\ell$. Thus (3.13.2) implies that

$$Hom_{\Gamma_{\tilde{v}}/I_{\tilde{v}}}(I_{\tilde{v}}/I_{\tilde{v}}^{\ell}, \bar{\alpha}_v \otimes \bar{s}_v^\vee) = Hom_{Frob_v}(\mathbb{F}_\ell, \bar{\alpha}_v \otimes \bar{s}_v^\vee) = 0 \tag{3.13.5}$$

and likewise with $\bar{\alpha}_v^{-1} \otimes \bar{s}_v$. On the other hand, the usual short exact sequence

$$0 \to H^0(\Gamma_{\tilde{v}}, M) \to M \xrightarrow{Frob_v - 1} M \to H^1(\Gamma_{\tilde{v}}/I_{\tilde{v}}, M) \to 0,$$

together with (3.13.2), shows that $H^1_{unr}(\Gamma_{\tilde{v}}, M) = 0$ when M is one of the two middle terms in (3.13.3). Thus (3.13.3) can be rewritten

$$H^1(\Gamma_{\tilde{v}}, ad\bar{\rho}) = H^1(\Gamma_{\tilde{v}}, k) \oplus H^1(\Gamma_{\tilde{v}}, ad\bar{s}_v). \tag{3.13.6}$$

For $v \in Q_N$, we set
$$L_v = H^1(\Gamma_{\tilde{v}}, k) \oplus H^1_{unr}(\Gamma_{\tilde{v}}, ad\bar{s}_v).$$

One verifies easily that this choice of L_v corresponds to the following condition on a deformation ρ of $\bar{\rho}$:

$$\rho\,|_{\Gamma_v} \xrightarrow{\sim} \alpha_v \oplus s_v \tag{3.13.7}$$

where s_v is unramified (but α_v is arbitrary). One also needs to verify that (3.13.7) defines a deformation condition, in other words that this property is relatively representable; this is routine.

The short exact sequence (3.13.4) applied to $M = k$ shows that

$$\dim L_v = 2 + \dim H^1_{unr}(\Gamma_{\tilde{v}}, ad\bar{s}_v) = 1 + \dim H^0(\Gamma_{\tilde{v}}, ad\bar{\rho}),$$

hence $\chi_{v,S} = 1$ for all $v \in Q_N$. Let $S(Q_N) = S \cup Q_N$, where S is the set of ramified primes, including all primes in S_ℓ. In our simplified setup, $S = S_\ell$ or, in the setting of framed deformations, $S = S_\ell \cup R$. Let $\mathcal{S}(Q_N)$ denote the corresponding deformation problem. We return to the Riemann-Roch formula. Recall that $ad\bar{\rho}$ is isomorphic to its linear dual, so that $ad\bar{\rho}^*$ is the Tate twist $ad\bar{\rho}(1)$. Letting $\epsilon_\infty = n\sum_{v\in S_\infty} \frac{1+c_v}{2}$ (the error at archimedean primes, which will be proved to vanish) we find
(3.13.8)

$$\begin{aligned}
&h^1_{\mathcal{S}(Q_N)}(\Gamma_{F^+}, ad\bar{\rho}) \\
&= h^1_{\mathcal{S}(Q_N)^*}(\Gamma_{F^+}, ad\bar{\rho}(1)) + h^0(\Gamma_{F^+}, ad\bar{\rho}) - h^0(\Gamma_{F^+}, ad\bar{\rho}(1)) + |Q_N| - \epsilon_\infty \\
&= h^1_{\mathcal{S}(Q_N)^*}(\Gamma_{F^+}, ad\bar{\rho}(1)) - h^0(\Gamma_{F^+}, ad\bar{\rho}(1)) + |Q_N| - \epsilon_\infty.
\end{aligned}$$

Here the first h^0 term vanishes because $\bar{\rho}$ is absolutely irreducible and complex conjugation acts as -1 on the scalars, as we have already seen.

3.14. The Taylor-Wiles primes, part II.. We also want to impose the following conditions on $|Q_N|$.

1. (3.14.1) If $v \in Q_N$ then $\mathbf{N}v \equiv 1 \pmod{\ell^N}$. (this was (3.4.4)
2. (3.14.2) $\bar{\rho} \mid \Gamma_v$ satisfies conditions (3.13.1) and (3.13.2).
3. (3.14.3) There is a number r such that $|Q_N| = r$ for all N.
4. (3.14.4) $h^1_{\mathcal{S}(Q_N)^*}(\Gamma_{F^+}, ad\bar{\rho}(1)) = 0$ for all N.

The existence of Q_N is proved by using the Chebotarev density theorem and calculations of global Galois cohomology similar to but more elaborate than those in (3.13). These are contained in a separate set of notes. The number r in (3.14.3) is just the dimension of the obstruction space $h^1_{\mathcal{S}^*}(\Gamma_{F^+}, ad\bar{\rho}(1))$, and one constructs the set Q_N by adding elements satisfying (3.14.1) and (3.14.2), one at a time, so that in the short exact sequence

$$0 \to H^1_{\mathcal{S}(Q_N)^*}(\Gamma_{F^+}, ad\bar{\rho}(1)) \to H^1_{\mathcal{S}^*}(\Gamma_{F^+}, ad\bar{\rho}(1)) \to \oplus_{v\in Q_N} \xrightarrow{\beta} H^1_{unr}(\Gamma_{\tilde{v}}, ad\bar{\rho}(1))/L_v^\perp$$

the last arrow β is injective. That Q_N satisfying all four conditions exist depends on the hypothesis, described in the notes on the Chebotarev density argument, that $im(\bar{\rho})$ is "big." This hypothesis is always satisfied in the applications (though it has to be checked), and it also implies the vanishing of the remaining global term $h^0(\Gamma_{F^+}, ad\bar{\rho}(1))$ in (3.13.7).

Proposition 3.14.5. *Suppose $im(\bar{\rho})$ is "big." Then for each $N > 1$, one can find a set Q_N satisfying (3.14.1)-(3.14.4) and such that the following dimension formula holds:*

$$h^1_{\mathcal{S}(Q_N)}(\Gamma_{F^+}, ad\bar{\rho}) \ = \ |Q_N| - \epsilon_\infty \ = \ r - \epsilon_\infty.$$

For the general deformation problem, one needs to calculate the relative h^1 as the dimension of the tangent space of the framed deformation problem modulo the dimension of the tangent space of R^{loc}. More on this later.

3.15 Taylor-Wiles patching in the minimal case.

To apply the above calculations to obtain the situation described in the notes on patching, we replace n by r and m by N. The modules H_m are what we have called $S_{\{\chi_v\}}(U(Q_N),\mathcal{O})_{\mathfrak{m}}$ above, where $U(Q_N)$ is the open compact subgroup satisfying condition (3.5.4) at primes in Q_N. We write H_{Q_N} for H_m Ignore the χ_v for the time being, since the set R of non-minimal primes of (3.5.2) is here assumed empty. The ring R_m of Diamond's Corollary 1.6 is our ring $R_{\bar{\rho},\mathcal{S}(Q_N)}$ which we write more simply $R_{\bar{r}_\Pi,Q_N}$. The ring T_m is just the image of $R_{\bar{r}_\Pi,Q_N}$ in $End(H_{Q_N})$, and this is just $\mathbb{T}^T_{\{\chi_v\}}(U(Q_N))_{\mathfrak{m}}$, which we denote $\mathbb{T}_{\bar{r}_\Pi,Q_N}$ for consistency. Indeed, the only way R^{univ} acts on modular forms is through its (surjective) homomorphism to the corresponding Hecke algebra.

We have not yet constructed the maps involving A and B. Recall that A and B are power series ring in r variables, denoted S_i and X_i, respectively. It follows from Proposition 3.14.5 that each $R_{\bar{r}_\Pi,Q_N}$ is generated over $\mathcal{O}$ by $r - n\sum_{v\in S_\infty}\frac{1+c_v}{2}$. elements, hence is a quotient of a power series ring in r variables. One can therefore construct the maps $B \rightarrow R_{\bar{r}_\Pi,Q_N}$ ad hoc.

The maps $A \rightarrow R_m$ are more intrinsic. For each $v \in Q_N$ let Δ_v be the quotient of order ℓ^N of $k(v)^\times$. By (3.14.1) there is such a quotient. Let

$$U_{0,v} = U_{1,v} := \{g \in GL(n,\mathcal{O}_{F,\tilde{v}}) \mid g \equiv \begin{pmatrix} g_{n-1} & *_{n-1} \\ 0_{n-1} & * \end{pmatrix} \pmod{\mathfrak{m}_{\tilde{v}}}\}$$

by analogy to (3.5.4), so that $U_{0,v}/U_{1,v} \xrightarrow{\sim} k(v)^\times$. Let $U^+_{1,v} \subset U_{0,v}$ be the subgroup containing $U_{1,v}$ such that

$$U_{0,v}/U^+_{1,v} \xrightarrow{\sim} \Delta_v \xrightarrow{\sim} \mathbb{Z}/\ell^N\mathbb{Z}$$

We modify our modular forms H_{Q_N} and consider only the submodule of $\prod_{v\in Q_N} U^+_{1,v}$-fixed vectors, but we do not change notation. Let $\Delta_{Q_N} = \prod_{v\in Q_N}\Delta_v$. This group acts on H_{Q_N}, and we have the important

(not quite true) Principle 3.15.1. *For any N, the module H_{Q_N} is free over $\mathcal{O}[\Delta_{Q_N}]$.*

This principle is almost true because the $S_{\{\chi_v\}}(U(Q_N),\mathcal{O})$ are spaces of functions on finite sets on which the group Δ_{Q_N} acts almost freely, and the localization that produces H_{Q_N} does not affect the condition of being free over the group algebra. Since we don't know that Δ_{Q_N} acts freely, we follow Taylor and Wiles and introduce an additional prime of potential ramification, denoted S_1 above. Adding S_1 to the level makes the action of Δ_{Q_N} free, and S_1 is chosen so that no constituent of the localization at $\mathfrak{m}$ is actually ramified at S_1, so the Riemann-Roch calculation is unchanged. The existence of an appropriate S_1 is another condition guaranteed by the hypothesis that the image of $\bar{\rho}$ is "big." I will not dwell on this point.

On the other hand, for $v \in Q_N$, consider the action of inertia $I_{\tilde{v}}$ on the universal deformation r^{univ} of type $\mathcal{S}(Q_N)$ of $\bar{r}_\Pi$. We can restrict our attention to the homomorphism $\rho^{univ}: \Gamma_F \rightarrow GL(n, R_{\bar{r}_\Pi,Q_N}$. Then in an appropriate basis, $\rho^{univ} \mid_{I_{\tilde{v}}}$ can be written as the sum of a trivial $n-1$-dimensional representation (lifting $\bar{s}_v$) and a one-dimensional character $\xi_v : I_{\tilde{v}} \to R^{\times}_{\bar{r}_{\Pi},Q_N}$ on the lifting of the α_v-eigenspace. The character ξ_v is well-defined and independent of the choice of basis, and is tame, hence factors through the tame inertia group $k(\tilde{v})^{\times}$. Moreover, we have

Principle 3.15.2. *The character ξ_v factors through the quotient Δ_v of $k(\tilde{v})^{\times}$, and the action of Δ_v on H_{Q_N} induced by the composition of ξ_v with the homomorphism $R_{\bar{r}_{\Pi},Q_N} \to End(H_{Q_N})$ is the natural group-theoretic action described above.*

Both parts of this principle follow from the compatibility of the local and global Langlands correspondences for the representation ρ_{Π}, proved in my book with Taylor.

Let $A_N = A/J_N$. Choose a generator $\delta_v \in \Delta_v$ for each $v \in Q_N$. The variables S_i in $A = \mathcal{O}[S_1, \ldots S_r]$ are indexed by the elements $v \in Q_N$ for some ordering of the latter – say we write $i = i(v), i = 1, \ldots, r$ – and we identify $A_N = \mathcal{O}[\Delta_N]$ by identifying δ_v with the image of $1 + S_i(v)$ in A_N. In this way, there is a natural map

$$A \to A_N \to R^{\times}_{\bar{r}_{\Pi},Q_N}$$

where the second arrow is the product of the ξ_v of 3.15.2. In this way H_N becomes an A-module for each N, and Diamond's condition (d) is satisfied:

3.15.3. *$Ann_A(H_N) = J_N$ and H_N is a free A_N-module for each N..*

To simplify the notation further, we write R_N and $\mathbb{T}_N$ instead of $R_{\bar{r}_{\Pi},Q_N}$ and $\mathbb{T}_{\bar{r}_{\Pi},Q_N}$ We have already seen Diamond's condition (a) (surjectivity of the maps $R_N \to \mathbb{T}_N$). Condition (b) is not quite true as stated. We have chosen ad hoc maps $B \to R_N$ and we can lift the maps $A \to R_N$ to maps $c_N : A \to B$ in such a way that the map $B \to R_N$ factors through $B_N = B/c_N(J_N)$. In (b) we can replace R_N by B_N, as Diamond did, and then (b) remains true.

Condition (c) is a subtle point. It is not hard to see that $H_N/J_0H_N \xrightarrow{\sim} H_0$ which is the localization at $\mathfrak{m}$ of the automorphic forms invariant under the group $U_0(Q_N)$, which are fixed by $\prod_{v \in Q_N} U_{0,v}$, in the above notation. But condition (c) requires an identification of H_0 with $H_{\Pi,\emptyset}$. There are two independent points, one global, one local, discussed in the appendix.

The global point – see Lemma A.2 of the appendix – is that the condition at $\mathfrak{m}$ guarantees that any representation Π' of type $\mathcal{S}(Q_N)$, with $\bar{\rho}_{\Pi'} \xrightarrow{\sim} \bar{\rho}_{\Pi}$, and with $(\Pi')^{U_0(Q_N)} \neq 0$, is necessarily unramified at Q_N. This is the group-theoretic equivalent of the Galois-theoretic condition (3.13.6) that says that any deformation of $\bar{\rho}_{\Pi}$ of type $\mathcal{S}(Q_N)$ necessarily breaks up as a sum of the unramified $n-1$-dimensional piece and the potentially ramified one-dimensional piece. This heuristic argument can be made rigorous by considering the classification of admissible representations of $GL(n, F_{\tilde{v}})$ with $U_{0,v}$-fixed vectors.

The second point is that H_0 is naturally a space of U_{0,Q_N}-invariant automorphic forms in the space of automorphic forms unramified at Q_N. For each v, the space of $U_{0,v}$-invariant forms in $\Pi_{\tilde{v}}$ is of dimension n, and one needs to pick out a submodule of rank one over $R_\emptyset$ and construct an isomorphism with the module of $\prod_v GL(n, \mathcal{O}_{\tilde{v}})$-invariant forms. It is for this reason that we need the additional operators V_v for $v \in Q_N$ and to include $V_v - \alpha_v$ in the ideal $\mathfrak{m}$. This can be done by means of Hensel's lemma, but the construction depends on an analysis of the reduction modulo $\mathfrak{m}_{\mathcal{O}}$ of principal series representations of $GL(n, F_{\tilde{v}})$ when $Nv - 1$ is divisible by ℓ. This was considered by Vignéras and the results are described in part in the notes entitled modularprincipalseries.pdf.

Admitting this last step, we have completed the verification of Diamond's conditions (or Fujiwara's equivalent conditions). We may therefore conclude as in Taylor-Wiles:

Theorem 3.15.4. *The map*

$$\phi_\emptyset : R_{\bar{r}_\Pi,\emptyset} \to \mathbb{T}_\Pi$$

is an isomorphism of complete intersections, and H_Π is a free module over $\mathbb{T}_\Pi$. In particular, any deformation of $\bar{r}_\Pi$ of minimal type $\mathcal{S}$ is of the form $r_{\Pi'}$ for some automorphic representation Π' of $GL(n)$ of cohomological unitary type.

Finally, the error term $\epsilon_\infty = n \sum_{v \in S_\infty} \frac{1+c_v}{2}$ necessarily vanishes. In other words, $c_v = -1$ for all $v \in S_\infty$.

The generalization to the non-minimal case, where the set R is not necessarily empty, follows broadly the same lines, but all calculations are relative to an algebra R^{loc}. The details will be added to the notes at a later time.

Appendix

Lemma A.1. *Let π be a principal series representation $I(\underline{\psi})$ of $GL(n,K)$ induced by an n-tuple $\underline{\psi} = (\psi_1, \ldots, \psi_n)$ of characters of $K^\times$. Suppose ψ_i is unramified for $i > 1$ and ψ_1 is at most tamely ramified. Let ψ_1^0 be the restriction to $\mathcal{O}_K^\times$, or equivalently to $k(v)^\times$. Then $\dim Hom_{U_0}(\psi_1^0, \pi) = n$ if ψ_1^0 is trivial, $= 1$ otherwise.*

The above Lemma is independent of the coefficients; it remains valid for (smooth) principal series representations with coefficients in $\overline{\mathbb{F}}_\ell$. Since $Nv \equiv 1 \pmod{\ell}$, the representation theory over $\overline{\mathbb{F}}_\ell$ is particularly simple: Vignéras proved in [V] that every $\pmod{\ell}$ principal series representation is completely reducible, and the factors are the same as the factors of the corresponding module over the Hecke algebra, which are easy to identify. In particular, for appropriate $\underline{\psi}$, the $\pmod{\ell}$ Steinberg is a direct summand of $I(\underline{\psi})$. These facts are used in what follows; see also the notes modularprincipalseries.pdf.

The Taylor-Wiles method involves patching spaces of modular forms of level $U = U(Q_N)$, localized at a maximal ideal $\mathfrak{m}$ of $\mathbb{T}^T_{\chi_v}(Q_N)$, for a set of Q_N with $N \to \infty$. The starting point is forms of level $U(\emptyset)$, which are unramified at Q_N. For each

N, we need to compare $S_{\chi_v}(U, \mathcal{O})_{\mathfrak{m}}$ with $S_{\chi_v}(U(Q_N), \mathcal{O})_{\mathfrak{m}}$ as $\mathbb{T}$-modules. Write $U_0(Q_N) \supset U(Q_N)$ the level subgroup with $U_{1,v}$ replaced by $U_{0,v}$ for all $v \in Q_N$.

There are actually two comparisons made. The first is between $S_{\chi_v}(U(Q_N), \mathcal{O})_{\mathfrak{m}}$ and $S_{\chi_v}(U_0(Q_N), \mathcal{O})_{\mathfrak{m}}$. The former contains global π with π_v tamely ramified at $v \in Q_N$ – and with only one degree of freedom for the ramification – whereas the latter contains only those global π with $\pi_v^{U_{0,v}} \neq 0$.

Lemma A.2. *With Q_N as in 3.13, $S_{\chi_v}(U_0(Q_N), \mathcal{O})_{\mathfrak{m}}$ consists only of π with π_v unramified at all $v \in Q_N$.*

One actually uses something stronger: that each π_v is an unramified principal series that is residually irreducible as far as the character $\bar{\alpha}_v$ is concerned (cf. (3.13.1)). In other words, any reducibility comes from reducibility of the unramified principal series of $GL(n-1)$ corresponding to the summand $\bar{s}_v$. I will not attempt to make this more precise. Assume for simplicity that $\pi_v \pmod{\ell}$ is an irreducible unramified principal series. Then we have seen that $\dim(\bar{\pi}_v^{U_{0,v}}) = n$, whereas the tamely ramified constituents of $S_{\chi_v}(U(Q_N), \mathcal{O})_{\mathfrak{m}}$ have only a one-dimensional $U_{0,v}$-invariant subspace. We use a Hecke operator for $U_{0,v}$, or for $U_{1,v}$, to cut out a 1-dimensional subspace of $\dim(\bar{\pi}_v^{U_{0,v}}) = n$. Namely, there is a Hecke operator

$$V_v = U_{?,v}(diag(I_{n-1}, \varpi_v))U_{?,v}, \ ? = 0, 1 \tag{A.3}$$

that acts on the $U_{0,v}$-fixed subspace and decomposes it as a sum of generalized eigenspaces with eigenvalues equal to the eigenvalues of $\bar{\rho}(Frob_v)$. Our assumption that $\pi_v \pmod{\ell}$ is irreducible is basically equivalent (in the classical limit) to the hypothesis that the V_v-eigenvalues are multiplicity free ($\psi_{i,v} \neq \psi_{j,v}$ if $i \neq j$), which is stronger than the hypothesis of (3.13.1).

Let H_{1,Q_N} denote the $\mathcal{O}$-submodule of $S_{\chi_v}(U(Q_N), \mathcal{O})_{\mathfrak{m}}$ on which V_v acts as $\psi_{1,v}(\varpi_v)$, and define $H_{0,Q_N} \subset S_{\{\chi_v\}}(U_0(Q_N), \mathcal{O})_{\mathfrak{m}}$ likewise. One has to be careful in making sense of this: $\psi_{1,v}$ varies with the different automorphic representations π contributing to $S_{\{\chi_v\}}(U(Q_N), \mathcal{O})_{\mathfrak{m}}$, but they are all congruent modulo $\mathfrak{m}$, by construction. This comes down to showing that the characteristic polynomial of V_v over $\mathbb{T}$ has a linear factor. One sees similarly that H_1 and H_0 are direct factors of the appropriate S_{χ_v}.

[V] M.-F. Vignéras, Induced R-representations of p-adic reductive groups, *Selecta Math.*, **4** (1998), 549-623.