## Overview

We give an introduction to the theory of automorphic forms on the multiplicative group of a quaternion algebra over $\mathbb{Q}$ and over totally real fields $F$ (including Hilbert modular forms). We know traditionally from the time of Gauss and Eisenstein that modular forms on a congruence subgroup $\Gamma$ of $SL_2(\mathbb{Z})$ contain a striking amount of arithmetic information. Here are some examples of applications. This is just to exhibit usefulness of automorphihc forms and only a part of them will be discussed in the class.

An easiest way of constructing a modular form is to make an averaging sum of its factors of automorphy: an Eisenstein series. There is another explicit way of constructing modular forms. As an application of Poisson's summation formula, an infinite series attached to each quadratic form $Q(x)$ on a $\mathbb{Q}$–vector space (of dimension $m$) has been used to construct elliptic modular forms explicitly: a theta series. Since the theta series of $Q$ is defined by

$$\theta(z) = \sum_{x \in \mathbb{Z}^m} \exp(2\pi i Q(x) z)$$

if $Q$ is positive definite, one is able to count the number of integer solutions of $Q(x) = n$ for a given positive integer $n$ by studying the theta series, which is a modular form of weight $\frac{m}{2}$ (see [HMI, Theorem 2.65]). For small $m$, one can prove an exact formula of the number of solutions as an explicitly given function of $n$. This is the case for the sum of squares $Q(x) = \sum_{j=1}^m x_j^2$ for $2 \leq m \leq 8$, because one can explicitely write $\theta$ down as a constant multiple of Eisenstein series and Fourier coefficients of an Eisenstein series can be computed explicitly. The idea of relating theta series and Eisenstein series to find such a formula is classical going back to the days of Gauss and Jacobi and has been developed much by Siegel, Weil and Shimura more recently.

Writing theta series as a linear combination of Eisenstein series, we get examples of the formula for rthe number of expressions of each positive integers as a sum of squares ($2 \leq m \leq 8$). Write $S_m(n)$ for the number of representations of an integer $n > 0$ as sums of $m$ squares. Assuming for simplicity $n$ to be odd square-free (see [Sh] 3.9 for the general cases), we have, for the quadratic residue symbol $\left(\frac{q}{p}\right)$ (primitive with respect to $q$),

- $S_2(n) = 2\left(1 + \left(\frac{-1}{n}\right)\right) \sum_{0<d|n} \left(\frac{-1}{d}\right)$ (Lagrange, Gauss, Jacobi);
- $S_3(n) = 24\pi^{-1}\sqrt{n} L(1, \left(\frac{-n}{\cdot}\right))$ (Gauss, Dirichlet, Shimura);
- $S_4(n) = 8 \sum_{0<d|n} d$ (Jacobi);
- $S_5(n) = 2^7(2\pi)^{-2}(\sqrt{n})^3 b_5(n) L(2, \left(\frac{n}{\cdot}\right))$ (Eisenstein, Smith, Minkowski, Shimura). Here $b_5(n) = 5$ if $n \equiv 3 \mod 4$, and $b_5(n) = 2^{-3} \cdot 3 \cdot 5, 2^{-3} \cdot 5 \cdot 7$ according as $n \equiv 1, 5 \mod 8$;
- $S_6(n) = (\left(\frac{-1}{n}\right) 2^4 - 4) \sum_{0<d|n} \left(\frac{-1}{d}\right) d^2$ (Jacobi);
- $S_7(n) = 2^9(2\pi)^{-3}(\sqrt{n})^3 b_7(n) L(3, \left(\frac{-n}{\cdot}\right))$ (Shimura);
  Here $b_7(n) = 7$ if $n \equiv 1 \mod 4$, and $b_7(n) = 2^{-5} \cdot 3^2 \cdot 5 \cdot 7, 2^{-5} \cdot 7 \cdot 37$ according as $n \equiv 3, 7 \mod 8$.
- $S_8(n) = 16 \sum_{0<d|n} d^3$ (Jacobi, Siegel).

When $m > 8$, there is a small but nontrivial contribution of cusp forms; so, we cannot have a precise formula but an asymptotic formula.

The case of $m = 4$ is related quaternionic automorphic forms, and because of this, contribution of cusp forms is quite subtle when $m = 4$, which we study to good extent in this course for norm forms (of four variables) of quaternion algebras. If we start with the quaternion algebra

$$H = \mathbb{Q} + \mathbb{Q}i + \mathbb{Q}j + \mathbb{Q}k \text{ with } \mathbb{H} = H \otimes_{\mathbb{Q}} \mathbb{R}$$

such that $i^2 = j^2 = k^2 = -1$, $ij = -ji = k$, $jk = -kj = i$ and $ki = -ik = j$, the norm form is exactly the sum of four squares: $N(x) = \overline{x}x = x_1^2 + x_2^2 + x_3^2 + x_4^2$ for $x = x_1 + x_2 i + x_3 j + x_4 k \in H$ and $\overline{x} = x_1 - x_2 i - x_3 j - x_4 k$.

To get the formula of $S_2(n)$, a key point is that the ring of Gaussian integers $\mathbb{Z}[\sqrt{-1}]$ is an Euclidean domain (so, a PID) and has four units $\{\pm 1, \pm\sqrt{-1}\}$. For an odd prime $\ell$, as Fermat observed,

$$\begin{gathered}\ell = x_1^2 + x_2^2 \text{ with } x_1, x_2 \in \mathbb{Z} \iff \\ \ell = \alpha\overline{\alpha} \text{ for } \alpha \in \mathbb{Z}[\sqrt{-1}] \iff \\ \left(\tfrac{-1}{\ell}\right) = 1 \ (\Leftrightarrow \ell \equiv 1 \mod 4).\end{gathered}$$

Thus $S_2(\ell) = 4, 0$ according as $\ell \equiv 1 \mod 4$ or not.

As for $S_4(\ell)$, we need to look into the order $R = \mathbb{Z} + \mathbb{Z}i + \mathbb{Z}j + \mathbb{Z}k \subset H$ and study right ideals of $R$. This order is not maximal; that is, there is a maximal subring $O_H$ containing $R$ which is a lattice of the $\mathbb{Q}$-vector space $H$ and maximal among such subrings. The ring $O_H$ is called the Hurwitz order. We have the index $[O_H : R] = 2$ with $\frac{1+i+j+k}{2} \in O_H$ (see [Hz]). Since $O_H$ is a noncommutative Euclidean domain, all right ideals of $O_H$ are principal, and hence a right $R$-ideal $\mathfrak{a}$ is principal if $N(\mathfrak{a}) = [R : \mathfrak{a}]$ is odd: $\mathfrak{a} = \alpha R$ for $\alpha \in R$. Since the quaternion conjugation $x \mapsto \overline{x}$ turns right ideals into left ideals, we find $\overline{\mathfrak{a}}\mathfrak{a} = R\overline{\alpha}\alpha R$, which is a two-sided ideal generated by $N(\alpha) = \overline{\alpha}\alpha \in \mathbb{Z}$. Thus $S_4(\ell)/8 = 1 + \ell$ is the number of such factorizations $\ell = \overline{\alpha}\alpha$, because $R$ has 8 units: $\{\pm 1, \pm i, \pm j, \pm k\}$.

We can think of another quaternion algebra $D = M_2(\mathbb{Q})$. Then a maximal order is given by $M_2(\mathbb{Z})$. The unit group of $M_2(\mathbb{Z})$ is infinite and given by $GL_2(\mathbb{Z}) = SL_2(\mathbb{Z}) \sqcup SL_2(\mathbb{Z})\epsilon$ for $\epsilon = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. Again all right ideals of $M_2(\mathbb{Z})$ is principal. We define the norm form of $M_2(\mathbb{Q})$ to be $N(x) = \det(x)$. We also have an $M_2(\mathbb{Q})$–conjugation given by $\iota : \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mapsto \left(\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right)$. Then $N(x) = x^\iota x$. We see easily that up to right multiplication by units of $M_2(\mathbb{Z})$, we have $1 + \ell$ elements $\alpha$ in $M_2(\mathbb{Z})$ with $N(\alpha) = \ell$:

$$\left\{\left(\begin{smallmatrix} 1 & 0 \\ 0 & \ell \end{smallmatrix}\right) \text{ and } \left(\begin{smallmatrix} \ell & u \\ 0 & 1 \end{smallmatrix}\right) \text{ for } u = 1, \ldots, \ell\right\}. \tag{0.1}$$

Thus we conclude that $S_4(\ell)/8$ gives the number of solutions $\det(\alpha) = \ell$ in $M_2(\mathbb{Z})$ up to units. This is the simplest example of intricate relations between different quaternion algebras, which we study in details in this chapter (as an introduction to the theory of quaternionic automorphic forms). Elliptic modular forms and Hilbert modular forms are particular cases of such quaternionic automorphic forms coming from $M_2(\mathbb{Q})$ and $M_2(F)$ for a totally real field $F$.

To motivate our study of quaternionic automorphic forms, let us continue to give examples of classical theorems whose proof relies essentially on elliptic modular forms and quaternion algebras. If one wants to solve a degree five non-soluble rational equation, what we need is a few elliptic functions in addition to classical operation of taking radicals (a result of F. Klein), and the solution is given in terms of the coordinate of a 5–torsion point on a rational elliptic curve (without complex multiplication; see [D]).

If one wants to find explicit generators (behaving nicely under Galois action) of an abelian extension of the rational number field $\mathbb{Q}$, we only need the exponential function $z \mapsto \mathbf{e}(z) = \exp(2\pi i z)$, which uniformizes the multiplicative group $\mathbb{G}_m$ ($\mathbf{e} : \mathbb{C} \to \mathbb{G}_m(\mathbb{C}) = \mathbb{C}^\times$ is the universal covering). The generators are roots of unity $\{\mathbf{e}(\frac{1}{N})\}_{0 \neq N \in \mathbb{Z}}$ (a theorem of Kronecker-Weber and Hilbert; see [ICF] Chapter 14).

If one wants to generalize this to abelian extensions of an imaginary quadratic field $K$, one need to consider (all) torsion points of an elliptic curve $E$ with complex multiplication by $K$. Thus the desired generator is given again by an elliptic function. This is the famous "Kronecker's dream of his youth" (Kronecker's Jugendtraum) and the origin of Hilbert's twelfth problem (see [Hl]).

Since modular functions $f : \mathfrak{H} \to \mathbb{C}$ (that is, modular forms of weight 0) on a congruence subgroup $\Gamma$ of $SL_2(\mathbb{Z})$ can be considered as classifying functions of "all" elliptic curves with some extra structures (for example, a point on the curve of a given order $N$), because over $\mathbb{C}$, any elliptic curve $E$ can be uniformized as $E(\mathbb{C}) = \mathbb{C}/\mathbb{Z}z + \mathbb{Z}$ for a point $z \in \mathfrak{H} = \{z \in \mathbb{C} | i(\overline{z} - z) > 0\}$. Thus all information we get as above can be formulated more naturally using elliptic modular forms and functions. Among elliptic modular forms, those forms $f$ which are eigenforms of all Hecke operators $T(n)$ are particularly important. As was shown by Hecke and Shimura, the eigenvalues $a_n$ of $T(n)$: $f|T(n) = a_n f$ generate a number field $\mathbb{Q}(f)$ (that is a finite extension of $\mathbb{Q}$ called a *Hecke field*). When $\mathbb{Q}(f) = \mathbb{Q}$, we call $f$ a *rational Hecke eigenform*.

One of the spectacular achievements in the recent history of Number theory is the proof of the Shimura-Taniyama conjecture by Wiles and Taylor et al (see [BCDT] and [HMI, 1.3.4]). This could be (rather in an over-simplified way) formulated as follows. Starting from a rational Hecke eigenform $f$ of weight 2 on the congruence subgroup $\Gamma_0(N)$ of $SL_2(\mathbb{Z})$, Eichler (for $N = 11$) and Shimura (in general) in the 1950s created a rational elliptic curve $E_{f/\mathbb{Q}}$ whose $L$–function $L(s, E_f)$ is identical to $L(s, f)$ (so $L(s, E_f)$ has analytic continuation to whole $s$–plane, proving the conjecture of Hasse-Weil for this particular $E_f$; see [GME] Section 4.2). If we use the classical definition of $L$–functions of elliptic curve, this could be formulated as $1 + \ell - a_\ell = |E(\mathbb{F}_\ell)|$ as long as

(U1) the equation of the curve modulo $\ell$ gives an elliptic curve over the finite field $\mathbb{F}_\ell$ (that is, $E$ has good reduction modulo $\ell$).

If we use a slightly more modern formulation, $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ acts naturally and continuously on the étale cohomology group $H^1(E_{f/\overline{\mathbb{Q}}}, \mathbb{Z}_p) \cong {\mathbb{Z}_p}^2$ (for any prime $p$), and the Galois action is characterized so that $\mathrm{Tr}(Frob_\ell) = a_\ell$ for almost all primes $p \neq \ell$ (independently of $p$ different from $\ell$), where $Frob_\ell$ is the (geometric) Frobenius element of $\ell$ in the Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. Thus the Galois action on $H^1(E_{f/\overline{\mathbb{Q}}}, \mathbb{Z}_p)$ gives a family of Galois

representations $\{\rho_p\}_p$ indexed by primes $p$ with independent trace $\mathrm{Tr}(\rho_\ell(Frob_\ell)) = a_\ell \in \mathbb{Z}$ as long as

(U2) the image of the inertia group at $\ell$ under $\rho_p$ is trivial ($\rho_p$ is called *unramified* at $\ell$ in this case).

The condition (U2) is actually a consequence of (U1) (a result of Hasse–Deuring and Shimura, e.g., [ACM] Chapter III) and (U2) implies (U1) (a later result of Serre-Tate, [SeT]).

The conjecture then states that any rational elliptic curve $E$ is isogenous over $\mathbb{Q}$ to $E_f$ for a suitably chosen rational Hecke eigenform $f$. An isogeny is a morphism of group schemes: $E \to E_f$ which is surjective (so, having finite kernel because of $\dim E = \dim E_f = 1$). The $L$–function is an isogeny invariant.

In the spirit of Shimura and Langlands, we may generalize this modularity problem to general compatible families $\{\rho_{\mathfrak{p}}\}_{\mathfrak{p}}$ of Galois representations. Here $\mathfrak{p}$ runs over all prime ideals of a number field $E$, and $\rho_{\mathfrak{p}} : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL_2(E_{\mathfrak{p}})$ with $\mathrm{Tr}(\rho_{\mathfrak{p}}(Frob_\ell)) \in E$ is independent of $\mathfrak{p}$ as long as $\rho_{\mathfrak{p}}$ is unramified at $\ell$. Such a family can be created for any given Hecke eigenform $f$ so that $\mathrm{Tr}(\rho(Frob_\ell)) = a_\ell$ (so $E = \mathbb{Q}(f)$): This is a result of Shimura when the weight $k$ is equal to 2, of Deligne if $k > 2$ (although Shimura also obtained a slightly weaker form of Deligne's result (of 1969) for more general automorphic forms: see [68c] in [CPS]) and of Deligne-Serre for $k = 1$ (some of them will be described in the lecture notes). Thus if $\det(\rho_{\mathfrak{p}})/\mathcal{N}^{k-1}$ is of finite order for the $p$–adic cyclotomic character $\mathcal{N}$, one expects to have a Hecke eigenform $f$ of weight $k$ giving rise to the compatible family $\{\rho_{\mathfrak{p}}\}_{\mathfrak{p}}$. This generalized form of the conjecture is also known in many cases of weight $k \geq 2$ as was summarized in [HMI, §1.3.4], and also, some cases of $k = 1$ has been successfully attacked by Langlands and R. Taylor et al (see [BCG], [BDST] and [T03]).

We can extend such a principle even to mod $p$ representations. As Serre did in [Se], one would then conjecture any mod $p$ 2-dimensional odd Galois representation $\overline{\rho}$ is a reduction modulo $p$ of a modular Galois representation associated to an elliptic Hecke eigenform of specific weight and level. Taylor found in [T02] a Hilbert modular Hecke eigenform associated to $\overline{\rho}$ restricted to $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$ for an unspecified totally real field $F$. Finally Khare–Wintenberger settled the conjecture for $\overline{\rho}$ in [KW], which uses essentially the results in [T02].

In a reverse direction, we can study deformation of a mod $p$ Galois representation $\overline{\rho}$, creating a "big" Galois representation $\boldsymbol{\rho}$ into $GL_2(\mathbb{T})$ for a big $p$-profinite algebra $\mathbb{T}$ so that, for specific prime ideals $P$ densely populated in $\mathrm{Spf}(\mathbb{T})$, $\rho_P := \boldsymbol{\rho} \mod P$ gives rise to the modular Galois representation as above whose reduction modulo the maximal ideal containing $P$ is isomorphic to $\overline{\rho}$. Examples of the "big" Galois representations were first constructed in [H86] for elliptic modular forms and were later generalized to Hilbert modular forms in [W] and [H89] after the earlier works on modular Galois representations described above. This construction essentially depends on the study of quaternionic automorphic forms. The abstract frame work of Galois deformation theory was given by Mazur in [M], and the principle proposed by Mazur is that the "big" modular Galois representation is universal among all specific deformations.

So it appears to be sufficient only to study elliptic modular forms and automorphic forms on the split $GL(2)$. This is not the case for a general base field $F$. We can consider an arbitrary base field $F$ and a compatible family $\rho = \{\rho_{\mathfrak{p}}\}_{\mathfrak{p}}$ of representations of $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$. We can formulate the conjecture that there should exist a Hecke eigenform $f : GL_2(F)\backslash GL_2(F_{\mathbb{A}}) \to \mathbb{C}$ giving rise to the family, because we can naturally associate with each elliptic Hecke eigenform an adelic Hecke eigenform on $GL_2(\mathbb{A})$ (as we will see later). This direction of the conjecture has been also proven when $F$ is totally real, by K. Fujiwara [HMI, §3.2.4], Skinner-Wiles [SW00] and [SW01] and Kisin [K] under different sets of assumptions ([HMI, Chapter 3]). The "direction" is to find a modular form on $GL_2(F_{\mathbb{A}})$ out of a given (arithmetic family of) Galois representation.

However, there are cases where we have no known way to create Galois representation directly out of a Hecke eigenform on the split $GL(2)$, without relying on some tricks moving to automorphic forms on some other algebraic groups (e.g., [T04]). If $F$ is not totally real, the modular variety $GL_2(F)\backslash GL_2(F_{\mathbb{A}})$ is just a Riemannian manifold (not an algebraic variety); so, there is no way to have subtle arithmetic on the manifold to create Galois representations. As was noticed in the 1960s by Shimura, even if $F$ is totally real, the Hilbert modular variety does not yield desired two-dimensional Galois representations (as can be checked in the real quadratic cases; see [BL] for general totally real fields). *Creating Galois representation (or even creating an elliptic curve from a given Hilbert modular rational Hecke eigenform of weight 2) could be more difficult than finding modular forms out of arithmetic Galois representations or elliptic curves.*

A known systematic way of creating an arithmetic object (see, for example, [67b] in [CPS] and [H81]) out of an automorphic form is to study Shimura curves and varieties obtained from quaternion algebras over a totally real field $F$ whose automorphic manifold is an algebraic variety defined canonically over $F$. The cases where we get algebraic curves in this way are proven to be most useful. There is another possibility of using quaternion algebras over a totally real field producing Shimura surfaces (e.g., [B]), although the above question is still open in general. The utility of such quaternion algebras was first noticed and studied by Shimura. They are not only useful in creating out of quaternionic Hecke eigenforms elliptic curves defined over $F$ (in the rational weight 2 case: [H81]) and Galois representations (cf. [68c] in [CPS], [C86a], [C86b]) but also in solving (cyclotomic and anticyclotomic) Hilbert's twelfth problem for CM fields ([67b] in [CPS] I), using quaternionic automorphic functions.

If we start with a quaternionic Hecke eigen automorphic form $f_D$ on a quaternion algebra $D_{/F}$, we have the associated family $\rho$ of Galois representations by the results of Shimura [68c] in [CPS] and Carayol [C86b]. Then in the cases where the modularity problem is solved, we find a Hilbert modular form $f$ having the same eigenvalue as $f_D$. This suggests a natural question if *the Hecke eigenvalues of each quaternionic automorphic form would be realized by a Hilbert modular form.* In other words, as Langlands pointed out, the non-abelian reciprocity law in a rough form depends only on the $\overline{\mathbb{Q}}$–points of the starting algebraic group defined over $F$ (not on its $F$–form; see [MFG] 1.2.1). A genesis of this question can be found in a problem Eichler studied in the 1950s (Eichler's basis problem, which came out in his thought, presumably, without definite knowledge of the non-abelian reciprocity law).

As Gauss and Jacobi knew, positive definite quadratic forms $Q(x)$ of four variable with coefficients in $\mathbb{Q}$ give rise to modular forms of weight 2: $\theta(z) = \sum_{n\in\mathbb{Z}^4} \mathbf{e}(Q(n)z)$. Eichler studied the norm form of an ideal $\mathfrak{a}$ of a definite quaternion algebra $D$ over $\mathbb{Q}$ and asked which subspace of elliptic modular forms can be spanned by such theta series $\theta(\mathfrak{a})$, and more generally, he asked himself to find a natural basis of the space. His result in a special case is as follows: Suppose that $D_\ell = D \otimes_{\mathbb{Q}} \mathbb{Q}_\ell \cong M_2(\mathbb{Q}_\ell)$ for all but one prime, say $p$. Take a maximal order $O_D$ of $D$ with $O_D \otimes_{\mathbb{Z}} \mathbb{Z}_\ell = M_2(\mathbb{Z}_\ell)$ for $\ell \neq p$. In this case, the automorphic variety $D^\times\backslash D_{\mathbb{A}}^\times/\widehat{O}_D^\times D_\infty^\times$ for $D_\infty = D \otimes_{\mathbb{Q}} \mathbb{R}$ and $D_{\mathbb{A}} = D \otimes_{\mathbb{Q}} \mathbb{A}$ is zero-dimensional; so, it is a set in bijection to the $O_D$–right ideal classes: $\{$right $O_D$–ideals $\mathfrak{a}\}$ modulo left multiplication by $D^\times$. For a right $O_D$-ideal $\mathfrak{a}$, the conjugate $\mathfrak{a}O_D\mathfrak{a}^{-1}$ is another maximal order of $D$. Define $e_\mathfrak{a}$ by the order of the unit group $(\mathfrak{a}O_D\mathfrak{a}^{-1})^\times$. Take a Hecke eigenform $f : D^\times\backslash D_{\mathbb{A}}^\times/\widehat{O}_D^\times \to \mathbb{C}$ with eigenvalue $a_\ell$ for $T(\ell)$, and form $\theta_\mathfrak{b}(f) = \sum_\mathfrak{a} e_\mathfrak{a}^{-1} f(\mathfrak{a})\theta(\mathfrak{a}\mathfrak{b}^{-1})$. Then we can find a basis of $S_2(\Gamma_0(p))$ in the set $\{\theta_\mathfrak{b}(f)\}_{f,\mathfrak{b}}$ and $\theta_\mathfrak{b}(f)|T(\ell) = a_\ell\theta_\mathfrak{b}(f)$, as expected. Here $\mathfrak{b}$ runs over right $O_D$-ideals up to left equivalence. A Langlands' version of the basis problem (*Jacquet–Langlands correspondence*) will be studied in the lecture (see [HMI, §2.6] for an original version of Eichler).