Hecke operators and Hilbert modular forms

Paul E. Gunnells and Dan Yasaki

University of Massachusetts Amherst, Amherst, MA 01003, USA

Abstract. Let F be a real quadratic field with ring of integers $\mathcal{O}$ and with class number 1. Let Γ be a congruence subgroup of $\mathrm{GL}_2(\mathcal{O})$. We describe a technique to compute the action of the Hecke operators on the cohomology $H^3(\Gamma; \mathbb{C})$. For F real quadratic this cohomology group contains the cuspidal cohomology corresponding to cuspidal Hilbert modular forms of parallel weight 2. Hence this technique gives a way to compute the Hecke action on these Hilbert modular forms.

1 Introduction

1.1

Let $\mathbf{G}$ be a reductive algebraic group defined over $\mathbb{Q}$, and let $\Gamma \subset \mathbf{G}(\mathbb{Q})$ be an arithmetic subgroup. Let $Y = \Gamma \backslash X$ be the locally symmetric attached to $G = \mathbf{G}(\mathbb{R})$ and Γ, where X is the global symmetric space, and let $\mathcal{M}$ be a local system on Y attached to a rational finite-dimensional complex representation of Γ. The cohomology $H^*(Y; \mathcal{M})$ plays an important role in number theory, through its connection with automorphic forms and (mostly conjectural) relationship to representations of the absolute Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ (cf. [3, 7, 16, 26]). This relationship is revealed in part through the action of the *Hecke operators* on the cohomology spaces. Hecke operators are endomorphisms induced from a family of correspondences associated to the pair $(\Gamma, \mathbf{G}(\mathbb{Q}))$; the arithmetic nature of the cohomology is contained in the eigenvalues of these linear maps.

For $\Gamma \subset \mathrm{SL}_n(\mathbb{Z})$, *modular symbols* provide a concrete method to compute the Hecke eigenvalues in $H^\nu(Y; \mathcal{M})$, where $\nu = n(n+1)/2 - 1$ is the top nonvanishing degree [8, 23]. Using modular symbols many people have studied the arithmetic significance of this cohomology group, especially for $n = 2$ and 3 [3,6,7,12,25,26]; these are the only two values of n for which $H^\nu(Y; \mathcal{M})$ can contain cuspidal cohomology classes, in other words cohomology classes coming from cuspidal automorphic forms on $\mathrm{GL}(n)$. Another setting where automorphic cohomology has been profitably studied using modular symbols is that of $\Gamma \subset \mathrm{SL}_2(\mathcal{O})$, where $\mathcal{O}$ is the ring of integers in a complex quadratic field [10,11,13,22]. In this case Y is a three-dimensional hyperbolic orbifold; modular symbols allow investigation of $H^2(Y; \mathcal{M})$, which again contains cuspidal cohomology classes.

1.2

Now let F be a real quadratic field with ring of integers $\mathcal{O}$, and let $\mathbf{G}$ be the $\mathbb{Q}$-group $\mathrm{Res}_{F/\mathbb{Q}}(\mathrm{GL}_2)$. Let $\Gamma \subseteq \mathbf{G}(\mathbb{Q})$ be a congruence subgroup. In this case we

have $X \simeq \mathfrak{H} \times \mathfrak{H} \times \mathbb{R}$, where $\mathfrak{H}$ is the upper halfplane (§2.2). The locally symmetric space Y is topologically a circle bundle over a Hilbert modular surface, possibly with orbifold singularities if Γ has torsion. The cuspidal cohomology of Y is built from cuspidal Hilbert modular forms. Hence an algorithm to compute the Hecke eigenvalues on the cuspidal cohomology gives a topological technique to compute the Hecke eigenvalues of such forms. But in this case there is a big difference from the setting in §1.1: the top degree cohomology occurs in degree $\nu = 4$, but the cuspidal cohomology appears in degrees $2, 3$.[1] Thus modular symbols cannot "see" the cuspidal Hilbert modular forms, and cannot directly be used to compute the Hecke eigenvalues.

1.3

In this article we discuss a technique, based on constructions in [17], that in practice allows one to compute the Hecke action on the cohomology space $H^3(Y; \mathbb{C})$. Moreover it is easy to modify our technique to compute with other local systems; all the geometric complexity occurs for trivial coefficients. Here we must stress the phrase *in practice*, since we cannot prove that our technique will actually work. Nevertheless, the ideas in [17] have been successfully used in practice [4, 5], and the modifications presented here have been extensively tested for $F = \mathbb{Q}(\sqrt{2}), \mathbb{Q}(\sqrt{3})$.

The basic idea is the following. We first identify a finite topological model for $H^3(Y; \mathbb{C})$, the *Voronoĭ reduced* cocycles. This uses a generalization of Voronoĭ's reduction theory for positive definite quadratic forms [1, 20], which constructs a Γ-equivariant tessellation of X (§2.2). The Hecke operators do not act directly on this model, and to accommodate the Hecke translates of reduced cocycles we work with a larger model for the cohomology, the (infinite-dimensional) space $S_1(\Gamma)$ of 1-*sharblies* modulo Γ (§2.3). The space $S_1(\Gamma)$ is part of a homological complex $S_*(\Gamma)$ with Hecke action that naturally computes the cohomology of Y. Any Voronoĭ reduced cocycle in H^3 gives rise to a 1-sharbly cycle, which allows us to identify a finite dimensional subspace $S_1^{\mathrm{red}}(\Gamma) \subset S_1(\Gamma)$.

The main construction is then to take a general 1-sharbly cycle ξ and to modify it by subtracting an appropriate coboundary to obtain a homologous cycle ξ' that is closer to being Voronoĭ reduced (§3). By iterating this process, we eventually obtain a cycle that lies in our finite-dimensional subspace $S_1^{\mathrm{red}}(\Gamma)$. Unfortunately, we are unable to prove that at each step the output cycle ξ' is better than the input cycle ξ, in other words that it is somehow "more reduced." However, in practice this always works.

The passage from ξ to ξ' is based on ideas found in [17], which describes an algorithm to compute the Hecke action on H^5 of congruence subgroups of

---

[1] The reader is probably more familiar with the case of $\mathbf{G}' = \mathrm{Res}_{F/\mathbb{Q}}\mathrm{SL}_2$. In this case the locally symmetric space is a Hilbert modular surface, and the cuspidal Hilbert modular forms contribute to H^2. Our symmetric space is slightly larger since the real rank of $\mathbf{G}$ is larger than that of $\mathbf{G}'$. However, regardless of whether one studies the Hilbert modular surface or our GL_2 symmetric space, the cusp forms contribute to the cohomology in degree one below the top nonvanishing degree.

$\mathrm{SL}_4(\mathbb{Z})$. The common feature that this case has with that of subgroups of $\mathrm{GL}_2(\mathcal{O})$ is that the cuspidal cohomology appears in the degree one less than the highest. This means that from our point of view the two cases are geometrically very similar. There are some complications, however, coming from the presence of non-torsion units in $\mathcal{O}$, complications leading to new phenomena requiring ideas not found in [17]. This is discussed in §4. We conclude the article by exhibiting the reduction of a 1-sharbly to a sum of Voronoĭ reduced 1-sharblies where the base field is $\mathbb{Q}(\sqrt{2})$ (§5).

We remark that there is another case sharing these same geometric features, namely that of subgroups of $\mathrm{GL}_2(\mathcal{O}_K)$, where K is a complex quartic field. We are currently applying the algorithm in joint work with F. Hajir and D. Ramakrishnan for $K = \mathbb{Q}(\zeta_5)$ to compute the cohomology of congruence subgroups of $\mathrm{GL}_2(\mathcal{O}_K)$ and to investigate the connections between automorphic cohomology and elliptic curves over K. Details of these cohomology computations, including some special features of the field K, will appear in [18]; the present paper focuses on the Hilbert modular case.

Finally, we remark that there is a rather different method to compute the Hecke action on Hilbert modular forms using the Jacquet–Langlands correspondence. For details we refer to work of L. Dembélé [14,15]. However, the Jacquet–Langlands technique works only with the complex cohomology of subgroups of $\mathrm{GL}_2(\mathcal{O})$, whereas our method in principle allows one to compute with torsion classes in the cohomology.

2   Background

2.1

Let F be a real quadratic field with class number 1. Let $\mathcal{O} \subset F$ denote the ring of integers. Let $\mathbf{G}$ be the $\mathbb{Q}$-group $\mathrm{Res}_{F/\mathbb{Q}}(\mathrm{GL}_2)$ and let $G = \mathbf{G}(\mathbb{R})$ the corresponding group of real points. Let $K \subset G$ be a maximal compact subgroup, and let A_G be the identity component of the maximal $\mathbb{Q}$-split torus in the center of G. Then the symmetric space associated to G is $X = G/KA_G$. Let $\Gamma \subseteq \mathrm{GL}_2(\mathcal{O})$ be a finite index subgroup.

In §2.2 we present an explicit model of X in terms of positive-definite binary quadratic forms over F and construct a $\mathrm{GL}_2(\mathcal{O})$-equivariant tessellation of X following [1, 20]. Section 2.3 recalls the sharbly complex [9, 17, 21].

2.2   Voronoĭ polyhedron

Let ι_1, ι_2 be the two real embeddings of F into $\mathbb{R}$. These maps give an isomorphism $F \otimes_{\mathbb{Q}} \mathbb{R} \simeq \mathbb{R}^2$, and more generally, an isomorphism

$$G \xrightarrow{\sim} \mathrm{GL}_2(\mathbb{R}) \times \mathrm{GL}_2(\mathbb{R}). \tag{1}$$

When the meaning is clear from the context, we use ι_1, ι_2 to denote all such induced maps. In particular, (1) is the map

$$g \longmapsto (\iota_1(g), \iota_2(g)). \tag{2}$$

Under this identification, A_G corresponds to $\{(rI, rI) \mid r > 0\}$, where I is the 2×2 identity matrix.

Let C be the cone of real positive definite binary quadratic forms, viewed as a subset of V, the $\mathbb{R}$-vector space of 2×2 real symmetric matrices. The usual action of $\mathrm{GL}_2(\mathbb{R})$ on C is given by

$$(g \cdot \phi)(v) = \phi({}^t g v), \quad \text{where } g \in \mathrm{GL}_2(\mathbb{R}) \text{ and } \phi \in C. \tag{3}$$

Equivalently, if A_ϕ is the symmetric matrix representing ϕ, then $g \cdot \phi = g A_\phi {}^t g$. In particular a coset $g O(2) \in \mathrm{GL}_2(\mathbb{R})/O(2)$ can be viewed as the positive definite quadratic form associated to the symmetric matrix $g {}^t g$.

Let $\mathcal{C} = C \times C$. Then (2) and (3) define an action of G on $\mathcal{C}$. Specifically, $g \cdot (\phi_1, \phi_2) = (\alpha_1, \alpha_2)$, where α_i is represented by $\iota_i(g) A_{\phi_i} \iota_i({}^t g)$. Let ϕ_0 denote the quadratic form represented by the identity matrix. Then the stabilizer in G of (ϕ_0, ϕ_0) is a maximal compact subgroup K. The group A_G acts on $\mathcal{C}$ by positive real homotheties, and we have

$$X = \mathcal{C}/\mathbb{R}_{>0} = (C \times C)/\mathbb{R}_{>0} \simeq \mathfrak{H} \times \mathfrak{H} \times \mathbb{R},$$

where $\mathfrak{H}$ is the upper halfplane.

Let $\bar{\mathcal{C}}$ denote the closure of $\mathcal{C}$ in $V \times V$. Each vector $w \in \mathbb{R}^2$ gives a rank 1 positive semi-definite form $w {}^t w$ (here w is regarded as a column vector). Combined with ι_1 and ι_2, we get a map $L : k^2 \to \bar{\mathcal{C}}$ given by

$$L(v) = \left(\iota_1(v) \cdot {}^t(\iota_1(v)), \iota_2(v) \cdot {}^t(\iota_2(v)) \right). \tag{4}$$

Let $R(v)$ be the ray $\mathbb{R}_{>0} \cdot L(v) \subset \bar{\mathcal{C}}$. Note that

$$L(cv) = (\iota_1(c)^2 L_1(v), \iota_2(c)^2 L_2(v))$$

so that if $c \in \mathbb{Q}$, then $L(cv) \in R(v)$, and in particular $L(-v) = L(v)$. The set of *rational boundary components* $\mathcal{C}_1$ of $\mathcal{C}$ is the set of rays of the form $R(v)$, $v \in F^2$ [1]. These are the rays in $\bar{\mathcal{C}}$ that correspond to the usual cusps of the Hilbert modular variety.

Let $\Lambda \subset V \times V$ be the lattice

$$\Lambda = \left\{ (\iota_1(A), \iota_2(A)) \,\middle|\, A = \begin{bmatrix} a & c \\ c & b \end{bmatrix}, \quad a, b, c \in \mathcal{O} \right\}.$$

Then $\mathrm{GL}_2(\mathcal{O})$ preserves Λ.

Definition 1. *The* Voronoĭ polyhedron Π *is the closed convex hull in $\bar{\mathcal{C}}$ of the points $\mathcal{C}_1 \cap \Lambda \smallsetminus \{0\}$.*

Since F has class number 1, one can show that any vertex of Π has the form $L(v)$ for $v \in \mathcal{O}^2$. We say that $v \in \mathcal{O}^2$ is *primitive* if $L(v)$ is a vertex of Π. Note that v is primitive only if $L(v)$ is primitive in the usual sense as a lattice point in Λ.

By construction $\mathrm{GL}_2(\mathcal{O})$ acts on Π. By taking the cones on the faces of Π, one obtains a Γ-*admissible decomposition* of $\mathcal{C}$ for $\Gamma = \mathrm{GL}_2(\mathcal{O})$ [1]. Essentially this means that the cones form a fan in $\bar{\mathcal{C}}$ and that there are finitely many cones modulo the action of $\mathrm{GL}_2(\mathcal{O})$. Since the action of $\mathrm{GL}_2(\mathcal{O})$ commutes with the homotheties, this decomposition descends to a $\mathrm{GL}_2(\mathcal{O})$-equivariant tessellation of X.[2]

We call this decomposition the *Voronoĭ decomposition*. We call the cones defined by the faces of Π *Voronoĭ cones*, and we refer to the cones corresponding to the facets of Π as *top cones*. The sets $\sigma \cap \mathcal{C}$, as σ ranges over all top cones, cover $\mathcal{C}$. Given a point $\phi \in \mathcal{C}$, there is a finite algorithm that computes which Voronoĭ cone contains ϕ [19].

For some explicit examples of the Voronoĭ decomposition over real quadratic fields, we refer to [24] (see also §5).

2.3   The sharbly complex

Let S_k, $k \geq 0$, be the Γ-module A_k/C_k, where A_k is the set of formal $\mathbb{Z}$-linear sums of symbols $[v] = [v_1, \cdots, v_{k+2}]$, where each v_i is in F^2, and C_k is the submodule generated by

1. $[v_{\sigma(1)}, \cdots, v_{\sigma(k+2)}] - \mathrm{sgn}(\sigma)[v_1, \cdots, v_{k+2}]$,
2. $[v, v_2, \cdots, v_{k+2}] - [w, v_2, \cdots v_{k+2}]$ if $R(v) = R(w)$, and
3. $[v]$, if v is *degenerate*, i.e., if $v_1, \cdots, v_{k+2}$ are contained in a hyperplane.

We define a boundary map $\partial \colon S_{k+1} \to S_k$ by

$$\partial[v_1, \cdots, v_{k+2}] = \sum_{i=1}^{k+2} (-1)^i [v_1, \cdots, \hat{v}_i, \cdots, v_{k+2}]. \tag{5}$$

This makes S_* into a homological complex, called the *sharbly complex* [2].

The basis elements $\mathbf{u} = [v_1, \cdots, v_{k+2}]$ are called *k-sharblies*. Notice that in our class number 1 setting, using the relations in C_k one can always find a representative for $\mathbf{u}$ with the v_i primitive. In particular, one can always arrange that each $L(v_i)$ is a vertex of Π. When such a representative is chosen, the v_i are unique up to multiplication by ± 1. In this case the v_i—or by abuse of notation the $L(v_i)$—are called the *spanning vectors* for $\mathbf{u}$.

Definition 2. *A sharbly is* Voronoĭ reduced *if its spanning vectors are a subset of the vertices of a Voronoĭ cone.*

The geometric meaning of this notion is the following. Each sharbly $\mathbf{u}$ with spanning vectors v_i determines a closed cone $\sigma(\mathbf{u})$ in $\bar{\mathcal{C}}$, by taking the cone generated by the points $L(v_i)$. Then $\mathbf{u}$ is reduced if and only if $\sigma(\mathbf{u})$ is contained in some Voronoĭ cone. It is clear that there are finitely many Voronoĭ reduced sharblies modulo Γ.

Using determinants, we can define a notion of size for 0-sharblies:

---

[2] If one applies this construction to $F = \mathbb{Q}$, one obtains the Farey tessellation of $\mathfrak{H}$, with tiles given by the $\mathrm{SL}_2(\mathbb{Z})$-orbit of the ideal geodesic triangle with vertices at $0, 1, \infty$.

Definition 3. *Given a 0-sharbly* **v**, *the size* Size(**v**) *of* **v** *is given by the absolute value of the norm determinant of the* 2×2 *matrix formed by spanning vectors for* **v**.

By construction Size takes values in $\mathbb{Z}_{>0}$. We remark that the size of a 0-sharbly **v** is related to whether or not **v** is Voronoï reduced, but that in general there exist Voronoï reduced 0-sharblies with size > 1.

The boundary map (5) commutes with the action of Γ, and we let $S_*(\Gamma)$ be the homological complex of coinvariants. Note that $S_*(\Gamma)$ is infinitely generated as a $\mathbb{Z}\Gamma$-module. One can show

$$H_k((S_* \otimes \mathbb{C})(\Gamma)) \xrightarrow{\sim} H^{4-k}(\Gamma; \mathbb{C}) \qquad (6)$$

(cf. [2]), with a similar result holding for cohomology with nontrivial coefficients. Moreover, there is a natural action of the Hecke operators on $S_*(\Gamma)$ (cf. [17]). Thus to compute with $H^3(\Gamma; \mathbb{C})$, which will realize cuspidal Hilbert modular forms over F of weight $(2, 2)$, we work with 1-sharbly cycles. We note that the Voronoï reduced sharblies form a subcomplex of $S_*(\Gamma)$ *finitely generated* subcomplex that also computes the cohomology of Γ as in (6). This is our finite model for the cohomology of Γ.

3 The reduction algorithm

3.1 The strategy

The general idea behind our algorithm is simple. To compute the action of a Hecke operator on the space of 1-sharbly cycles, it suffices to describe to an algorithm that writes a general 1-sharbly cycle as a sum of Voronoï reduced 1-sharblies. Now any basis 1-sharbly **u** contains three sub-0-sharblies (the *edges* of **u**), and the Voronoï reduced 1-sharblies tend to have edges of small size. Thus our first goal is to systematically replace all the 1-sharblies in a cycle with edges of large size with 1-sharblies having smaller size edges. This uses a variation of the classical modular symbol algorithm, although no continued fractions are involved. Eventually we produce a sum of 1-sharblies with all edges Voronoï reduced. However, having all three edges Voronoï reduced is (unfortunately) not a sufficient condition for a 1-sharbly to be Voronoï reduced.[3] Thus a different approach must be taken for such 1-sharblies to finally make the cycle Voronoï reduced. This is discussed further in §4.

3.2 Lifts

We begin by describing one technique to encode a 1-sharbly cycle using some mild extra data, namely that of a choice of *lifts* for its edges:

---

[3] This is quite different from what happens with classical modular symbols, and reflects the infinite units in $\mathcal{O}$.

Definition 4 ([17]). *A 2×2 matrix M with coefficients of F with columns A_1, A_2 is said to be a* lift *of a 0-sharbly $[u, v]$ if $\{R(A_1), R(A_2)\} = \{R(u), R(v)\}$.*

The idea behind the use of lifts is the following. Suppose a linear combination of 1-sharblies $\xi = \sum a(\mathbf{u})\mathbf{u} \in S_1$ becomes a cycle in $S_1(\Gamma)$. Then its boundary must vanish modulo Γ. In the following algorithm, we attempt to pass from ξ to a "more reduced" sharbly ξ' by modifying the edges of each $\mathbf{u}$ in the support of ξ. To guarantee that ξ' is a cycle modulo Γ, we must make various choices in the course of the reduction Γ-equivariantly across the boundary of ξ. This can be done by first choosing 2×2 integral matrices for each sub-0-sharbly of ξ. We refer to [17] for more details and discussion. For the present exposition, we merely remark that we always view a 1-sharbly $\mathbf{u} = [v_1, v_2, v_3]$ as a triangle with vertices labelled by the v_i and with a given (fixed) choice of lifts for each edge (Figure 1). If two edges $\mathbf{v}, \mathbf{v}'$ satisfy $\gamma \cdot \mathbf{v} = \mathbf{v}'$, then we choose the corresponding lifts to satisfy $\gamma M = M'$. The point is that we can then work individually with 1-sharblies enriched with lifts; we don't have to know explicitly the matrices in Γ that glue the 1-sharblies into a cycle modulo Γ.

We emphasize that the lift matrices for any given 1-sharbly in the support of ξ are essentially forced on us by the requirement that ξ be a cycle modulo Γ. There is almost no flexibility in choosing them. Such matrices form an essential part of the input data for our algorithm.
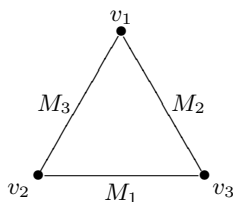

Fig. 1. A 1-sharbly with lifts

3.3 Reducing points

Definition 5. *Let $\mathbf{v}$ be a 0-sharbly with spanning vectors $\{x, y\}$. Assume $\mathbf{v}$ is not Voronoĭ reduced. Then $u \in \mathcal{O}^2 \smallsetminus \{0\}$ is a* reducing point *for $\mathbf{v}$ if the following hold:*

1. $R(u) \neq R(x), R(y)$.
2. $L(u)$ is a vertex of the unique Voronoĭ cone σ (not necessarily top-dimensional) containing the ray $R(x + y)$.
3. If $x = ty$ for some $t \in F^{\times}$, then u is in the span of x.
4. Of the vertices of σ, the point u minimizes the sum of the sizes of the 0-sharblies $[x, u]$ and $[u, y]$.

Given a non-Voronoĭ reduced 0-sharbly $\mathbf{v} = [x, y]$ and a reducing point u, we apply the relation

$$[x, y] = [x, u] + [u, y] \qquad (7)$$

in the hopes that the two new 0-sharblies created are closer to being Voronoĭ reduced. Note that choosing u uses the geometry of the Voronoĭ decomposition instead of (a variation of) the continued fraction algorithms of [8, 12, 23]. Unfortunately we cannot guarantee that the new 0-sharblies on the right of (7) are better than $\mathbf{v}$, but this is true in practice.

3.4 Γ-invariance

The reduction algorithm proceeds by picking reducing points for non-Voronoĭ reduced edges. We want to make sure that this is done Γ-equivariantly; in other words that if two edges $\mathbf{v}$, $\mathbf{v}'$ satisfy $\gamma \cdot \mathbf{v} = \mathbf{v}'$, then if we choose u for $\mathbf{v}$ we want to make sure that we choose γu for $\mathbf{v}'$.

We achieve this by making sure that the choice of reducing point for $\mathbf{v}$ only depends on the lift matrix M that labels $\mathbf{v}$. The matrix is first put into *normal form*, which is a unique representative M_0 of the coset $\mathrm{GL}_2(\mathcal{O}) \backslash M$. This is an analogue of Hermite normal form that incorporates the action of the units of $\mathcal{O}$. There is a unique 0-sharbly associated to M_0; We choose a reducing point u for this 0-sharbly and translate it back to obtain a reducing point for $\mathbf{v}$. Note that u need not be unique. However we can always make sure that the same u is chosen any time a given normal form M_0 is encountered, for instance by choosing representatives of the Voronoĭ cones modulo $\mathrm{GL}_2(\mathcal{O})$ and then fixing an ordering of their vertices.

We now describe how M_0 is constructed from M. Let Ω_* be a fundamental domain for the action of $(\mathcal{O}^\times, \cdot)$ on $F^\times$. For $t \in \mathcal{O}$, let $\Omega_+(t)$ be a fundamental domain for the action of $(t\mathcal{O}, +)$ on F.

Definition 6. *A nonzero matrix $M \in \mathrm{Mat}_2(F)$ is in* normal form *if M has one of the following forms:*

1. $\begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix}$, *where $b \in \Omega_*$.*

2. $\begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix}$, *where $a \in \Omega_*$ and $b \in F$.*

3. $\begin{bmatrix} a & b \\ 0 & d \end{bmatrix}$, *where $a, d \in \Omega_*$ and $b \in \Omega_+(d)$.*

It is easy to check that the normal form for M is uniquely determined in the coset $\mathrm{GL}_2(\mathcal{O}) \cdot M$. To explicitly put $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ in normal form, the first step is to find $\gamma \in \mathrm{GL}_2(\mathcal{O})$ such that $\gamma \cdot M$ is upper triangular. Such a γ can be found after finite computations as follows. Let $N \colon F \to \mathbb{R}$ be defined by $N(\alpha) = |\mathrm{Norm}_{F/\mathbb{Q}}(\alpha)|$. If

$$0 < N(c) < N(a),$$

then let $\alpha \in \mathcal{O}$ be an element of smallest distance from a/c. Let

$$\gamma' = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -\alpha \\ 0 & 1 \end{bmatrix}.$$

Then $\gamma' \in \mathrm{GL}_2(\mathcal{O})$ and $\gamma' M = \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}$ with $N(c') < N(c)$ and $N(a') < N(a)$.
Repeating this procedure will yield the desired result.

After a reducing point is selected for $\mathbf{v}$ and the relation (7) is applied, we must choose lifts for the 0-sharblies on the right of (7). This we do as follows:

Definition 7. *Let $[v_1, v_2]$ be a non-reduced 0-sharbly with lift matrix M and reducing point u. Then the* inherited lift $\hat{M}_i$ *for $[v_i, u]$ is the matrix obtained from M by keeping the column corresponding to v_i and replacing the other column by u.*

3.5 The algorithm

Let $T = [v_1, v_2, v_3]$ be a non-degenerate sharbly. Let M_i be the lifts of the edges of T as shown in Figure 1. The method of subdividing the interior depends on the number of edges that are Voronoĭ reduced. After each subdivision, lift data is attached using inherited lifts for the exterior edges. The lift for each interior edge can be chosen arbitrarily as long as the same choice is made for the edge to which it is glued. We note that steps (I), (II), and (III.1) already appear in [17], but (III.2) and (IV) are new subdivisions needed to deal with the complications of the units of $\mathcal{O}$.

(I) Three non-reduced edges. If none of the edges are Voronoĭ reduced, then we split each edge by choosing reducing points u_1, u_2, and u_3. In addition, form three additional edges $[u_1, u_2], [u_2, u_3]$, and $[u_3, u_1]$. We then replace T by the four 1-sharblies

$$[v_1, v_2, v_3] \longmapsto [v_1, u_3, u_2] + [u_3, v_2, u_1] + [u_2, u_1, v_3] + [u_1, u_2, u_3].$$

(II) Two non-reduced edges. If only one edge is Voronoĭ reduced, then we split the other two edges by choosing reducing points u_1 and u_3. We form two additional edges $[u_1, u_3]$ and ℓ, where ℓ is taken to be either $[v_1, u_1]$ or $[v_3, u_3]$, whichever has smaller size. More precisely:

1. If $\mathrm{Size}([v_1, u_1]) \leq \mathrm{Size}([u_3, v_3])$, then we form two additional edges $[u_1, u_3]$ and $[v_1, u_1]$, and replace T by the three 1-sharblies

$$[v_1, v_2, v_3] \longmapsto [v_1, u_3, u_1] + [u_3, v_2, u_1] + [v_1, u_1, v_3].$$

2. Otherwise, we form two additional edges $[u_1, u_3]$ and $[v_3, u_3]$, and replace T by the three 1-sharblies

$$[v_1, v_2, v_3] \longmapsto [v_1, u_3, v_3] + [u_3, v_2, u_1] + [u_3, u_1, v_3].$$

(III) One non-reduced edge. If two edges are Voronoĭ reduced, then we split the other edge by choosing a reducing point u_1. The next step depends on the configuration of $\{v_1, v_2, v_3, u_1\}$.

1. If $[v_2, u_1]$ or $[u_1, v_3]$ is not Voronoĭ reduced or $v_2 = tv_1$ for some $v \in F$, then we form one additional edge $[v_1, u_1]$ and replace T by the two 1-sharblies

$$[v_1, v_2, v_3] \longmapsto [v_1, v_2, u_1] + [v_1, u_1, v_3].$$

2. Otherwise, a central point w is chosen. The central point w is chosen from the vertices of the top cone containing the barycenter of $[v_1, v_2, v_3, w]$ so that it maximizes the number of Voronoĭ reduced edges in the set

$$S = \{[v_1, w], [v_2, w], [v_3, w], [u_1, w]\}.$$

We do not allow v_1, v_2 or v_3 to be chosen as a central point. We form four additional edges $[v_1, w], [v_2, w], [u_1, w]$, and $[v_3, w]$ and replace T by the four 1-sharblies

$$[v_1, v_2, v_3] \longmapsto [v_1, v_2, w] + [w, v_2, u_1] + [w, u_1, v_3] + [w, v_3, v_1].$$

(IV) All edges Voronoĭ reduced. If all three edges are Voronoĭ reduced, but T is not Voronoĭ reduced, then a central point w is chosen. The central point w is chosen from the vertices of the top cone containing the barycenter of $[v_1, v_2, v_3]$ so that it maximizes the sum $\#E + \#P$, where E is the set of Voronoĭ reduced edges in $\{[v_1, w], [v_2, w], [v_3, w]\}$ and P is the set of Voronoĭ reduced triangles in $\{[v_1, v_2, w], [v_2, v_3, w], [v_3, v_1, w]\}$. We do not allow v_1, v_2 or v_3 to be chosen as a central point. We form three additional edges $[v_1, w], [v_2, w]$, and $[v_3, w]$ and replace T by the three 1-sharblies

$$[v_1, v_2, v_3] \longmapsto [v_1, v_2, w] + [w, v_2, v_3] + [w, v_3, v_1].$$

4 Comments

First, we emphasize that the reducing point u of Definition 5 works in practice to shrink the size of a 0-sharbly $\mathbf{v}$, but we have no proof that it will do so. The difficulty is that Definition 5 chooses u using the geometry of the Voronoĭ polyhedron Π and not the size of $\mathbf{v}$ directly. Moreover, our experience with examples shows that this use of the structure of Π is essential to reduce the original 1-sharbly cycle (cf. §5.2).

Next, as mentioned in §3.1, case (IV) is necessary: there are 1-sharblies T with all three edges Voronoĭ reduced, yet T is itself not Voronoĭ reduced. An example is given in the next section. The point is that in $\bar{\mathcal{C}}$ the points $L(v)$ and $L(\varepsilon v)$ are different if ε is not a torsion unit, but after passing to the Hilbert modular surface $L(v)$ and $L(\varepsilon v)$ define the same cusp. This means one can take a geodesic triangle Δ in the Hilbert modular surface with vertices at three cusps

that by any measure should be considered reduced, and can lift Δ to a 3-cone in the GL_2-symmetric space that is far from being Voronoĭ reduced.

Finally, the reduction algorithm can be viewed as a two stage process. When a 1-sharbly T has 2 or 3 non-reduced edges or 1 non-reduced edge and satisfies the criteria for case 1, then in some sense T is "far" from being Voronoĭ reduced. One tries to replace T by a sum of 1-sharblies that are more reduced in that the edges have smaller size. However, this process will not terminate in Voronoĭ reduced sharblies. In particular, if T is "close" to being Voronoĭ reduced, then one must use the geometry of the Voronoĭ cones more heavily. This is why we need the extra central point w in (III.2) and (IV).

For instance, suppose $T = [v_1, v_2, v_3]$ is a 1-sharbly with 1 non-reduced edge such that the criteria for (III.2) are satisfied when the reducing point is chosen. One can view choosing the central point and doing the additional split as first moving the bad edge to the interior of the triangle, where the splitting no longer needs to be Γ-invariant. The additional freedom allows one to make a better choice. Indeed, without the central point chosen wisely, this does lead to some problems. In particular, there are examples where $[v_1, u_1]$ is not Voronoĭ reduced, and the choice of the reducing point for this edge is v_2, leading to a repeating behavior. Thus the distinction had to be made.

5   The case $F = \mathbb{Q}(\sqrt{2})$

5.1

Let $F = \mathbb{Q}(\sqrt{2})$ and let $\varepsilon = 1 + \sqrt{2}$, a fundamental unit of norm -1. Computations of H. Ong [24, Theorem 4.1.1] with positive definite binary quadratic forms over F allow us to describe the Voronoĭ polyhedron Π and thus the Voronoĭ decomposition of $\mathcal{C}$:

Proposition 1 ([24, Theorem 4.1.1]). *Modulo the action of $\mathrm{GL}_2(\mathcal{O})$, there are two inequivalent top Voronoĭ cones. The corresponding facets of Π have 6 and 12 vertices, respectively.*

We fix once and for all representative 6-dimensional cones A_0 and A_1. To describe these cones, we give sets of points $S \subset \mathcal{O}^2$ such that the points $\{L(v) \mid v \in S\}$ are the vertices of the corresponding face of Π. Let e_1, e_2 be the canonical basis of $\mathcal{O}^2$. Then we can take A_0 to correspond to the 6 points

$$e_1, e_2, e_1 - e_2, \bar{\varepsilon}e_1, \bar{\varepsilon}e_2, \bar{\varepsilon}(e_1 - e_2),$$

and A_0 to correspond to the 12 points

$$e_1, e_2, \bar{\varepsilon}e_1, \bar{\varepsilon}e_2, e_1 - e_2, e_1 + \bar{\varepsilon}e_2, e_2 + \bar{\varepsilon}e_1, \bar{\varepsilon}(e_1 + e_2), \alpha, \beta, \bar{\varepsilon}\alpha, \bar{\varepsilon}\beta,$$

where $\alpha = e_1 - \sqrt{2}e_2$, $\beta = e_2 - \sqrt{2}e_1$. Since A_1 is not a simplicial cone, there exist basis sharblies that are Voronoĭ reduced but do not correspond to Voronoĭ cones.

Now we consider cones of lower dimension. Modulo $\mathrm{GL}_2(\mathcal{O})$, every 2-dimensional Voronoĭ cone either lies in $\bar{\mathcal{C}} \smallsetminus \mathcal{C}$ or is equivalent to the cone corresponding to $\{e_1, e_2\}$. The $\mathrm{GL}_2(\mathcal{O})$-orbits of 3-dimensional Voronoĭ cones are represented by $\{e_1, e_2\} \cup U$, where U ranges over

$$\{e_1 - e_2\}, \{\bar{\varepsilon} e_1\}, \{\bar{\varepsilon}(e_1 - e_2)\}, \{e_1 - \sqrt{2} e_2, e_2 - \sqrt{2} e_1\}, \{e_1 + \bar{\varepsilon} e_2\}.$$

Note that all but one of the 3-cones are simplicial.

5.2

Now we consider reducing a 1-sharbly T. Let us represent T by a 2×3 matrix whose columns are the spanning vectors of T. We take T to be

$$T = \begin{bmatrix} \sqrt{2} + 3 & 4\sqrt{2} + 4 & 3\sqrt{2} - 4 \\ \sqrt{2} & 5\sqrt{2} - 1 & -3\sqrt{2} - 5 \end{bmatrix},$$

and we choose arbitrary initial lifts for the edges of T. This data is typical of what one encounters when trying to reduce a 1-sharbly cycle modulo Γ.

The input 1-sharbly T has 3 non-reduced edges with edge sizes given by the vector $[5299, 529, 199]$. The first pass of the algorithm follows (I) and splits all 3 edges, replacing T by the sum $S_1 + S_2 + S_3 + S_4$, where

$$S_1 = \begin{bmatrix} \sqrt{2} + 3 & -\sqrt{2} - 1 & 1 \\ \sqrt{2} & -\sqrt{2} & 0 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 4\sqrt{2} + 4 & 0 & -\sqrt{2} - 1 \\ 5\sqrt{2} - 1 & -\sqrt{2} - 1 & -\sqrt{2} \end{bmatrix},$$

$$S_3 = \begin{bmatrix} 3\sqrt{2} - 4 & 1 & 0 \\ -3\sqrt{2} - 5 & 0 & -\sqrt{2} - 1 \end{bmatrix}, \quad S_4 = \begin{bmatrix} 0 & 1 & -\sqrt{2} - 1 \\ -\sqrt{2} - 1 & 0 & -\sqrt{2} \end{bmatrix}.$$

We compute that $\mathrm{Size}(S_1) = [2, 2, 8]$, $\mathrm{Size}(S_2) = [1, 1, 16]$, $\mathrm{Size}(S_3) = [1, 2, 7]$, and $\mathrm{Size}(S_4) = [2, 1, 1]$. Notice that the algorithm replaces T by a sum of sharblies with edges of significantly smaller size. This kind of performance is typical, and looks similar to the performance of the usual continued fraction algorithm over $\mathbb{Z}$. Note also that S_4, which is the 1-sharbly spanned by the three reducing points of the edges T, also has edges of very small size. This reflects our use of Definition 5 to choose the reducing points; choosing them without using the geometry of Π often leads to bad performance in the construction of this 1-sharbly.

Now S_4 has 3 Voronoĭ reduced edges, but is itself not Voronoĭ reduced. The algorithm follows (IV), replaces S_4 by $R_1 + R_2 + R_3$, and now each R_i is Voronoĭ reduced.

The remaining 1-sharblies S_1, S_2, and S_3 have only 1 non-reduced edge. They are almost reduced in the sense that they satisfy the criteria for (III.2). The algorithm replaces S_1 by $O_1 + O_2 + O_3 + O_4$, where O_1 and O_2 are degenerate and O_3 and O_4 are Voronoĭ reduced. The 1-sharbly S_2 is replaced by a $P_1 + P_2 + P_3 + P_4$, and each P_i is Voronoĭ reduced. S_3 is replaced by $Q_1 + Q_2 + Q_3 + Q_4$,

where Q_1 and Q_2 are degenerate, Q_3 is Voronoĭ reduced, and Q_4 is not Voronoĭ reduced. This 1-sharbly is given by

$$Q_4 = \begin{bmatrix} -\sqrt{2}+1 & 0 & 3\sqrt{2}-4 \\ 2\sqrt{2}+3 & -\sqrt{2}-1 & -3\sqrt{2}-5 \end{bmatrix}$$

and has 3 Voronoĭ reduced edges. Once again the algorithm is in case (IV), and replaces Q_4 by a sum $N_1 + N_2 + N_3$ of Voronoĭ reduced sharblies.

To summarize, the final output of the reduction algorithm applied to T is a sum

$$N_1 + N_2 + N_3 + O_3 + O_4 + P_1 + P_2 + P_3 + P_4 + Q_3 + R_1 + R_2 + R_3, \quad \text{where}$$

$$N_1 = \begin{bmatrix} -\sqrt{2}+1 & 0 & 0 \\ 2\sqrt{2}+3 & -\sqrt{2}-1 & -2\sqrt{2}-3 \end{bmatrix},$$

$$N_2 = \begin{bmatrix} 0 & 3\sqrt{2}-4 & 0 \\ -\sqrt{2}-1 & -3\sqrt{2}-5 & -2\sqrt{2}-3 \end{bmatrix},$$

$$N_3 = \begin{bmatrix} 3\sqrt{2}-4 & -\sqrt{2}+1 & 0 \\ -3\sqrt{2}-5 & 2\sqrt{2}+3 & -2\sqrt{2}-3 \end{bmatrix}, \quad O_3 = \begin{bmatrix} -\sqrt{2}-1 & -\sqrt{2}-1 & 1 \\ -1 & -\sqrt{2} & 0 \end{bmatrix},$$

$$O_4 = \begin{bmatrix} -\sqrt{2}-1 & 1 & \sqrt{2}+3 \\ -1 & 0 & \sqrt{2} \end{bmatrix}, \quad P_1 = \begin{bmatrix} 2\sqrt{2}+3 & 4\sqrt{2}+4 & 1 \\ \sqrt{2}+2 & 5\sqrt{2}-1 & -2\sqrt{2}+2 \end{bmatrix},$$

$$P_2 = \begin{bmatrix} 2\sqrt{2}+3 & 1 & 0 \\ \sqrt{2}+2 & -2\sqrt{2}+2 & -\sqrt{2}-1 \end{bmatrix}, \quad P_3 = \begin{bmatrix} 2\sqrt{2}+3 & 0 & -\sqrt{2}-1 \\ \sqrt{2}+2 & -\sqrt{2}-1 & -\sqrt{2} \end{bmatrix},$$

$$P_4 = \begin{bmatrix} 2\sqrt{2}+3 & -\sqrt{2}-1 & 4\sqrt{2}+4 \\ \sqrt{2}+2 & -\sqrt{2} & 5\sqrt{2}-1 \end{bmatrix}, \quad Q_3 = \begin{bmatrix} -\sqrt{2}+1 & 1 & 0 \\ 2\sqrt{2}+3 & 0 & -\sqrt{2}-1 \end{bmatrix},$$

$$R_1 = \begin{bmatrix} -\sqrt{2}+1 & 0 & 0 \\ 2\sqrt{2}+3 & -\sqrt{2}-1 & -2\sqrt{2}-3 \end{bmatrix},$$

$$R_2 = \begin{bmatrix} 0 & 3\sqrt{2}-4 & 0 \\ -\sqrt{2}-1 & -3\sqrt{2}-5 & -2\sqrt{2}-3 \end{bmatrix}, \quad \text{and}$$

$$R_3 = \begin{bmatrix} 3\sqrt{2}-4 & -\sqrt{2}+1 & 0 \\ -3\sqrt{2}-5 & 2\sqrt{2}+3 & -2\sqrt{2}-3 \end{bmatrix},$$

and each of the above is Voronoĭ reduced. Some of these 1-sharblies correspond to Voronoĭ cones and some don't. In particular, one can check that the spanning vectors for P_3, P_4, R_1, and N_1 do form Voronoĭ cones, and all others don't. However, the spanning vectors of O_3 and O_4 almost do, in the sense that they are subsets of 3-dimensional Voronoĭ cones with four vertices.

References

1. A. Ash, *Deformation retracts with lowest possible dimension of arithmetic quotients of self-adjoint homogeneous cones*, Math. Ann. **225** (1977), 69–76.

2. _____, *Unstable cohomology of $SL(n, \mathcal{O})$*, J. Algebra **167** (1994), no. 2, 330–342.

3. A. Ash, D. Grayson, and P. Green, *Computations of cuspidal cohomology of congruence subgroups of $SL_3(\mathbf{Z})$*, J. Number Theory **19** (1984), 412–436.

4. A. Ash, P. E. Gunnells, and M. McConnell, *Cohomology of congruence subgroups of $SL_4(\mathbf{Z})$ II*, submitted to J. Number Theory.

5. _____, *Cohomology of congruence subgroups of $SL_4(\mathbf{Z})$*, J. Number Theory **94** (2002), 181–212.

6. A. Ash and M. McConnell, *Experimental indications of three-dimensional Galois representations from the cohomology of $SL(3, \mathbf{Z})$*, Experiment. Math. **1** (1992), no. 3, 209–223.

7. A. Ash, R. Pinch, and R. Taylor, *An $\widehat{A_4}$ extension of $\mathbf{Q}$ attached to a non-selfdual automorphic form on $GL(3)$*, Math. Ann. **291** (1991), 753–766.

8. A. Ash and L. Rudolph, *The modular symbol and continued fractions in higher dimensions*, Invent. Math. **55** (1979), 241–250.

9. A. Ash, *Unstable cohomology of $SL(n, \mathcal{O})$*, J. Algebra **167** (1994), no. 2, 330–342.

10. J. Bygott, *Modular forms and modular symbols over imaginary quadratic fields*, Ph.D. thesis, Exeter, 1999.

11. J. E. Cremona, *Hyperbolic tessellations, modular symbols, and elliptic curves over complex quadratic fields*, Compositio Math. **51** (1984), no. 3, 275–324.

12. _____, *Algorithms for modular elliptic curves*, second ed., Cambridge University Press, Cambridge, 1997.

13. J. E. Cremona and E. Whitley, *Periods of cusp forms and elliptic curves over imaginary quadratic fields*, Math. Comp. **62** (1994), no. 205, 407–429.

14. L. Dembélé, *Explicit computations of Hilbert modular forms on $\mathbf{Q}(\sqrt{5})$*, Experiment. Math. **14** (2005), no. 4, 457–466.

15. _____, *Quaternionic Manin symbols, Brandt matrices, and Hilbert modular forms*, Math. Comp. **76** (2007), no. 258, 1039–1057 (electronic).

16. J. Franke, *Harmonic analysis in weighted L_2-spaces*, Ann. Sci. École Norm. Sup. (4) **31** (1998), no. 2, 181–279.

17. P. E. Gunnells, *Computing Hecke eigenvalues below the cohomological dimension*, Experiment. Math. **9** (2000), no. 3, 351–367.

18. P. E. Gunnells and D. Yasaki, *Computing Hecke operators on modular forms over real quadratic and complex quartic fields*, in preparation.

19. P. E. Gunnells, *Modular symbols for $\mathbf{Q}$-rank one groups and Voronoĭ reduction*, J. Number Theory **75** (1999), no. 2, 198–219.

20. M. Koecher, *Beiträge zu einer Reduktionstheorie in Positivitätsbereichen. I*, Math. Ann. **141** (1960), 384–432.

21. R. Lee and R. H. Szczarba, *On the homology and cohomology of congruence subgroups*, Invent. Math. **33** (1976), no. 1, 15–53.

22. M. Lingham, *Modular forms and elliptic curves over imaginary quadratic fields*, Ph.D. thesis, Nottingham, 2005.

23. Y.-I. Manin, *Parabolic points and zeta-functions of modular curves*, Math. USSR Izvestija **6** (1972), no. 1, 19–63.

24. H. E. Ong, *Perfect quadratic forms over real-quadratic number fields*, Geom. Dedicata **20** (1986), no. 1, 51–77.

25. W. Stein, *Modular forms, a computational approach*, Graduate Studies in Mathematics, vol. 79, American Mathematical Society, Providence, RI, 2007, With an appendix by P. E. Gunnells.

26. B. van Geemen, W. van der Kallen, J. Top, and A. Verberkmoes, *Hecke eigenforms in the cohomology of congruence subgroups of $SL(3, \mathbf{Z})$*, Experiment. Math. **6** (1997), no. 2, 163–174.