DYNAMICS OF GEODESICS, AND MAASS CUSP FORMS

ANKE POHL AND DON ZAGIER

Abstract. The correspondence principle in physics between quantum mechanics and classical mechanics suggests deep relations between spectral and geometric entities of Riemannian manifolds. We survey—in a way intended to be accessible to a wide audience of mathematicians—a mathematically rigorous instance of such a relation that emerged in recent years, showing a dynamical interpretation of certain Laplace eigenfunctions of hyperbolic surfaces.

1. Introduction

Suppose we have a huge space, such as the earth or a billiard table, and a small marble sitting on this space. We give this marble an initial push and observe its trajectory as it travels over the space. As we experienced from a very young age on, the marble goes straight until it hits an obstacle, e. g., the boundary of the billiard table, of which it reflects with outgoing angle equal to incoming angle, and then continues its straight path until the next obstacle where the same game restarts.

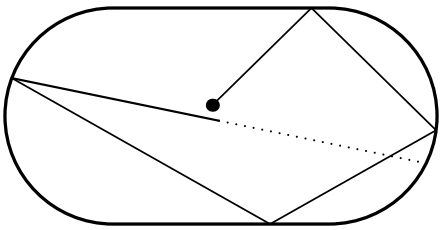

Figure 1. Trajectory on a stadium-shaped billiard table.

In Figure 1 this situation is depicted for a flat stadium-shaped billiard table. In Figure 2 it is shown for a disk with a bump in the middle, indicating that 'straight path' here means 'path of minimal resistance' or 'path of minimal effort'.

In terms of physics, the motion of the marble is predicted by the laws of classical mechanics. In such a description, moving objects are often modeled as point particles, that is, as objects without size or dimension, identifying the object with its center of mass.

In reality, any real-world object has a non-zero size, and the idealization as a point is not always desirable or correct. If we consider a very small marble which is almost a point, say of the size of an electron, or if we zoom in into our previous marble and try to describe the trajectory of a single electron of it then we notice

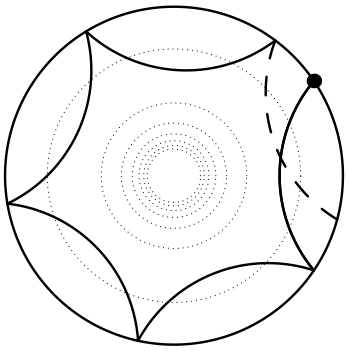

FIGURE 2. Trajectory on a disk with a bump in the middle. Height level curves are indicated by dotted circles.

that the classical mechanics model is not accurate on this subatomic level. One of the obstacles is the impossibility to determine simultaneously with absolute precision the position and momentum of the considered particle, as expressed by Heisenberg's famous uncertainty principle. Thus, the classical mechanical principles of determinism and time reversibility are not valid anymore. On such small scale, a more accurate model is provided by quantum mechanics, which describes the probability with which the particle attains a specific position-momentum combination.

The correspondence principle in physics states that, in the limit of passing to large scale, the predictions of quantum mechanics reproduce those of classical mechanics. However, the precise relation between classical and quantum mechanics is not yet fully understood, and its investigation gives rise to many interesting mathematical questions.

In terms of mathematics, the classical mechanical aspects of the motion of the marble considered above translate to properties of the geodesic flow on a Riemannian manifold X, whereas the quantum mechanical description relates to the Laplace operator on X and its (L^2-)eigenvalues and eigenfunctions. The correspondence principle then suggests an intimate relation between geometric-dynamical aspects of X on the one hand, and its spectral aspects on the other hand:

physics		**mathematics**
classical mechanics	$\leftsquigarrow$	geometric entities: {periodic geodesics; lengths of periodic geodesics}
quantum mechanics	$\leftsquigarrow$	spectral entities: {Laplace eigenfunctions; Laplace eigenvalues}

During the last century, many results showing relations between geometric-dynamical and spectral properties of Riemannian manifolds have been obtained.

In Section 2 below we discuss—as an appetizer—the flat 1-torus where a clear relation between the lengths of periodic geodesics ('classical mechanical objects') and the Laplace eigenvalues ('quantum mechanical objects') appears.

The main aim of this article is to present a much deeper relation between periodic geodesics and Laplace eigenfunctions that has emerged in recent years, but now for a class of hyperbolic surfaces.

In a nutshell, this goes as follows. A well-chosen discretization of the flow along the periodic geodesics gives rise to a one-parameter family of *transfer operators*, which are evolution operators that are reminiscent of weighted graph Laplacians and that also may be thought of as discretizations of the hyperbolic Laplacian. As such, these operators are simultaneously objects of classical and quantum mechanical nature, and therefore can serve as mediators between the dynamical and spectral entities of the hyperbolic surface under consideration. In our case, highly regular, rapidly decaying eigenfunctions (called *period functions*) of eigenvalue 1 of the transfer operator with parameter s are in bijection with rapidly decaying Laplace eigenfunctions (called *Maass cusp forms*) with spectral parameter s. This provides a purely dynamical characterization of the Maass cusp forms (not just their eigenvalues), shows a close dependence between periodic geodesics and these Laplace eigenfunctions, and provides a deep-lying mathematical realization of an instance of the correspondence principle.

The modular surface was the first hyperbolic surface for which such a result could be established, through combination of work by E. Artin [1], Series [21], Mayer [13, 14], Lewis [10], Bruggeman [2], Chang–Mayer [5], and Lewis–Zagier [11, 12]. Taking advantage of the constructions involved, an extension to a class of finite covers of the modular surface was achieved in the combination of [5, 7, 8]. An alternative proof for the modular surface was provided in [15, 4]. The recent development of a new type of discretizations for geodesic flows on hyperbolic surfaces [19] and of a cohomological interpretation of the Maass cusp forms [3] allowed to prove such a relation between periodic geodesics and Laplace eigenfunctions for a large class of hyperbolic surfaces far beyond the modular surface and in a very direct way [16, 18, 17].

In Sections 3–7 below we survey this new approach, although in an informal way and restricting for simplicity to the modular surface. We attempt to provide sufficiently precise definitions and enough details to keep the exposition as understandable as possible without introducing too much technical material. As a general principle we invite all readers to rely on their intuitive understanding of the geometry and dynamics of Riemannian manifolds, to use the many figures as a support, and to ignore the exact expressions of all formulas.

Acknowledgements. AP wishes to thank the Max Planck Institute for Mathematics in Bonn for hospitality and excellent working conditions during the preparation of this manuscript. Further, she acknowledges support by the DFG grants PO 1483/2-1 and PO 1483/2-2.

2. An appetizer

In this section we will treat the 'baby case' of the flat 1-*torus*

$$\mathbb{T} \;=\; \mathbb{R}/\mathbb{Z} \;=\; [0,1]/\{0{=}1\},$$

and show an intimate and very clear relation between geometric and spectral entities, and hence a mathematical rigorous instance of the correspondence principle.

Of course, this specific one-dimensional Riemannian manifold is much too simple to be representative of the general situation. However, it allows us to provide—without too much technical effort—a first instance of the relation between the geometry and the spectrum as motivated by the considerations from physics. We will also use this 'baby example' to carefully introduce the relevant geometrical and spectral concepts, whose counterparts in the situation of hyperbolic surfaces will be treated with less details.

2.1. **The flat 1-torus.** For a pictorial, but rather sketchy construction of the flat 1-torus $\mathbb{T}$ we may imagine the set $\mathbb{R}$ of real numbers as a number line, and glue together this line at any two points that are separated by an integer distance. The glueing process can be visualized as rolling up the line to a unit circle. (See Figure 3.) Alternatively, we may take the interval $[0,1]$ and glue together its two endpoints 0 and 1.

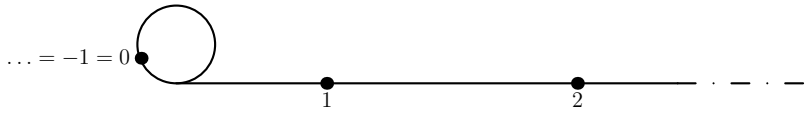

Figure 3. Rolling up $\mathbb{R}$ to form $\mathbb{T}$.

Both these geometric constructions indicate that $\mathbb{T}$ carries more structure than just being a set. In particular, measuring distances on $\mathbb{T}$ is possible, and a notion of derivatives exists.

In order to be able to formulate such additional structures in precise terms and to work with them, we use a formula-based definition of $\mathbb{T}$. For that, we identify any two points of $\mathbb{R}$ that differ by an integer only. Thus, for each $t \in \mathbb{R}$, all points in the set

$$\{t + m \mid m \in \mathbb{Z}\} \tag{1}$$

are unified to a single element, which we denote by $[t]$. The torus $\mathbb{T}$, as a set, consists of all these elements. The glueing process in the pictorial construction is a visualization of the *projection map*

$$\pi_{\mathbb{T}} \colon \mathbb{R} \to \mathbb{T}, \quad t \mapsto [t]. \tag{2}$$

This map is *locally injective*, which means that for any $t \in \mathbb{R}$ we find a small $\varepsilon > 0$ such that the restriction of $\pi_{\mathbb{T}}$ to the interval $(t-\varepsilon, t+\varepsilon)$ is injective. In rough terms, small pieces of the torus $\mathbb{T}$ look exactly like small pieces of $\mathbb{R}$. It is precisely this property which allows us to push certain structures of $\mathbb{R}$ to $\mathbb{T}$.

2.2. **Geometric entities.** We define the *distance* between two points $x, y \in \mathbb{T}$ to be the minimal distance between any two of their representatives in $\mathbb{R}$, hence

$$d_{\mathbb{T}}(x,y) \;\coloneqq\; \min\big\{d_{\mathbb{R}}(t_x,t_y) \;\big|\; [t_x]=x,\; [t_y]=y\big\},$$

where

$$d_{\mathbb{R}}(t_x,t_y) \;\coloneqq\; |t_x - t_y|$$

is the usual euclidean distance on $\mathbb{R}$. A *straight path* or *geodesic* in $\mathbb{T}$ is—roughly said—a path such that for any two nearby points on the path no shorter way between them exists than the path itself.

More precisely, a *path* on $\mathbb{T}$ is a differentiable map $p\colon I \to \mathbb{T}$, where $I \subseteq \mathbb{R}$ is an interval. The set I should be thought of as a time interval, and $p(t)$ as the position where we are at time t if we travel along the path p. The *speed* of p is given by its derivative p'. The path p is said to be of *unit speed* if $|p'(t)| = 1$ for all $t \in I$. A path $p\colon I \to \mathbb{T}$ of unit speed is *straight* if for any $t \in I$ there exists $\varepsilon > 0$ such that for all $t_1, t_2 \in (t-\varepsilon, t+\varepsilon) \cap I$ we have

$$d_{\mathbb{T}}\big(p(t_1),p(t_2)\big) \;=\; \left|\int_{t_1}^{t_2} |p'(t)|\,dt\right| \;=\; \big|t_1 - t_2\big| \;=\; d_{\mathbb{R}}(t_1,t_2).$$

That is, the distance between $p(t_1)$ and $p(t_2)$ equals the length of the path between $p(t_1)$ and $p(t_2)$, which here also equals the euclidean distance between t_1 and t_2. From now on, 'geodesic' will always mean a *unit speed, complete geodesic*, i. e., a straight path of unit speed with time interval $I = \mathbb{R}$.

In everyday language, the notion of path usually does not refer to the motion, i. e., to a map $p\colon I \to \mathbb{T}$, but rather to the static object, i. e., to the image $p(I)$ of p. The orientation, however, is important: 'the path from a to b'. We too will use the notion of geodesic more flexibly and apply it to refer to either

(G1) a geodesic $p\colon \mathbb{R} \to \mathbb{T}$ defined as above as a path, or
(G2) the oriented image of such a geodesic, or—more precisely—its equivalence class when we identify any two such geodesics that differ only by a shift in their arguments.

The motivation for the second usage is that we are typically not interested in the specific time parametrization of a geodesic. The context should always clarify which version is being used.

In our one-dimensional 'baby example' there are only two geodesics in the sense of (G2), namely those represented by the two geodesics in the sense of (G1) given by

$$p_{\pm}\colon \mathbb{R} \to \mathbb{T}, \quad t \mapsto [\pm t].$$

(See Figure 4.) Both these geodesics are *periodic*, that is, they 'close up', or in rigorous terms, there exists $t_0 > 0$ such that for all $t \in \mathbb{R}$,

$$p_{\pm}(t) \;=\; p_{\pm}(t+t_0).$$

The minimal such t_0 is called the *(primitive) period* or *(primitive) length* $\ell(p_{\pm})$ of the geodesic $p_{\pm}$, which here is $\ell(p_{\pm}) = 1$ in both cases. Periodicity and lengths are invariants under the equivalence of geodesics, and hence an intrinsic notion for geodesics in the sense of (G2).

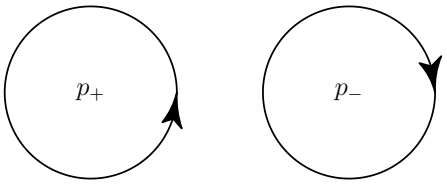

FIGURE 4. The two periodic geodesics on $\mathbb{T}$.

The *geometric entity* of $\mathbb{T}$ or, from the standpoint of the Introduction, the *classical mechanical object* we are interested in, is the *(primitive) geodesic length spectrum* $L_{\mathbb{T}}$, defined as the multiset (= set with multiplicities) of lengths of the periodic geodesics in the sense of (G2). In our case, this is

$$L_{\mathbb{T}} \;=\; \{\text{lengths of periodic geodesics}\} \;=\; \{1,1\}.$$

2.3. **Spectral entities.** The *spectral entity* or the *quantum mechanical object* is the Laplace spectrum of $\mathbb{T}$, which we now explain. The local injectivity of the projection map $\pi_{\mathbb{T}}$ from (2) allows us to transfer all local notions from $\mathbb{T}$ to $\mathbb{R}$. In particular, a function $f\colon \mathbb{T} \to \mathbb{C}$ is *differentiable* if

$$F \;:=\; f \circ \pi_{\mathbb{T}} \colon \mathbb{R} \to \mathbb{C}$$

is differentiable. The *derivative* of f at $[t] \in \mathbb{T}$ is then the derivative of F at $t \in \mathbb{R}$. The *Laplace operator* on $\mathbb{T}$ is

$$\Delta_{\mathbb{T}} \;:=\; -\frac{d^2}{dt^2},$$

and a basis for its L^2-eigenfunctions is given by the family

$$f_k \colon \mathbb{T} \to \mathbb{C}, \quad f_k\big([t]\big) := e^{2\pi i k t} \qquad (k \in \mathbb{Z}).$$

A straightforward calculation shows that

$$\Delta_{\mathbb{T}} f_k \;=\; (2\pi k)^2 f_k.$$

Thus, the *Laplace spectrum* of $\mathbb{T}$ is the multiset

$$\sigma(\mathbb{T}) \;=\; \{\text{Laplace eigenvalues}\} \;=\; \{(2\pi k)^2 \mid k \in \mathbb{Z}\}.$$

2.4. **Relation between geometric and spectral entities.** A rather astonishing observation is that the geodesic length spectrum $L_{\mathbb{T}}$ of $\mathbb{T}$ and the Laplace spectrum $\sigma(\mathbb{T})$ almost determine each other. To see this we consider the *dynamical zeta function*

$$\zeta_{\mathbb{T}}(s) \;:=\; \prod_{\ell \in L_{\mathbb{T}}} \left(1 - e^{-s\ell}\right) \;=\; \left(1 - e^{-s}\right)^2.$$

Then

$$\zeta_{\mathbb{T}}(s) = 0 \qquad \Longleftrightarrow \qquad s = 2\pi i k \quad \text{for some } k \in \mathbb{Z},$$

and the order of each zero is 2. In order words,

$$\zeta_{\mathbb{T}}(s) = 0 \qquad \Longleftrightarrow \qquad (is)^2 \in \sigma(\mathbb{T}), \tag{3}$$

and the order of s as a zero corresponds to the order of $(is)^2$ as eigenvalue, except for $s = 0$, where the order of the Laplace eigenvalue $(is)^2 = 0$ is 1, whereas the order of the zero $s = 0$ of $\zeta_{\mathbb{T}}$ is 2.

Thus knowing the geodesic length spectrum $L_{\mathbb{T}}$, and hence the dynamical zeta function $\zeta_{\mathbb{T}}$, we can deduce all Laplace eigenvalues, and even their multiplicities up to the difficulty at $s = 0$. Conversely, if we are given the Laplace spectrum $\sigma(\mathbb{T})$ (with multiplicities), and hence all zeros of $\zeta_{\mathbb{T}}$ with almost all multiplicities, then we can easily deduce the exact formula of $\zeta_{\mathbb{T}}$ and thus the geodesic length spectrum.

This ends the 1-dimensional 'appetizer'. In the rest of the paper we will study a 2-dimensional case, again describing first the geometric side, then the spectral side, and then the relation between them. Of course, this case is much more involved, but we have tried to introduce the concepts in the torus case in such a way that they generalize naturally.

3. Geometric and spectral sides of the modular surface

In the previous section we considered the torus $\mathbb{T}$, which is a quotient of the flat 1-manifold $\mathbb{R}$ by a discrete group action. From now on, we will consider hyperbolic surfaces, which are orbit spaces of the *hyperbolic plane* by discrete groups of isometries. For concreteness we will discuss only the *modular surface* $X = \mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}$, even though the results hold for a much larger class. We will provide precise definitions further below in this section.

In the course of the following four sections we will survey—as already mentioned in the Introduction—a rather deep relation between the geodesic flow on X and the Maass cusp forms for the modular group $\mathrm{PSL}_2(\mathbb{Z})$, resulting in a dynamical characterization of Maass cusp forms, or from a physics point of view, a description of certain quantum mechanical wave functions using only tools and objects from classical mechanics. The proof of this relation is split into three major steps:

(I) A cohomological interpretation of Maass cusp forms, which we will explain in Section 4 below. Representing Maass cusp forms faithfully as cocycle classes in suitable cohomology spaces provides an interpretation of these forms in a rather algebraic way which here simplifies to relate them to further objects.

(II) A well-chosen discretization of the geodesic flow on X, which we will construct in Section 5 below. This discretization extracts those geometric and dynamical properties from the geodesic flow on X that are crucial for the relation to Maass cusp forms, and it discards all the other additional properties. This condensed, discrete version of the geodesic flow is also of a rather algebraic nature.

(III) A connection between the discretization of the geodesic flow and the cohomology spaces, as discussed in Section 6 below. The central object mediating between these objects is the evolution operator (with specific weights, adapted to the spectral parameter of Maass cusp forms; a *transfer operator*) of the action map in the discrete version of the geodesic flow. We will see that the highly regular eigenfunctions of the evolution operator with parameter s

are building blocks for the cocycle classes of the Maass cusp forms with spectral parameter s, and will establish an explicit bijection between these eigenfunctions and the Maass cusp forms.

The first two steps are independent of each other, and also the corresponding sections can be read independently. The third step necessarily takes advantage of the results from Sections 4 and 5, however only the final results are needed, not the information how to achieve these. In Section 7 below we will provide a brief overview of these steps.

In the remainder of this section we introduce the geometric and spectral objects that we will need further on.

3.1. **The hyperbolic plane.** The *hyperbolic plane* is a certain two-dimensional manifold with Riemannian metric in which Euclid's parallel axiom fails: on the hyperbolic plane, for every straight line L (infinitely extended in both directions) and any point p not on L there are infinitely many lines $\tilde{L}$ passing through p that do not intersect L.

Abstractly, the hyperbolic plane is the unique two-dimensional connected, simply connected, complete Riemannian manifold with constant sectional curvature -1. There are many models for the hyperbolic plane. We use its *upper half plane model*[1]

$$\mathbb{H} \coloneqq \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\},$$

where the *line element of the Riemannian metric* is given by

$$ds^2_{x+iy} \coloneqq \frac{dx^2 + dy^2}{y^2}. \tag{4}$$

Informally, the Riemannian metric allows us to measure distances and angles. Angles in hyperbolic geometry are identical to the euclidean angles in $\mathbb{H}$. Distances between points however are changed in hyperbolic geometry when compared to euclidean geometry. From a euclidean point of view, hyperbolic distances between two points increase when these move nearer to the real axis $\mathbb{R}$.

In the upper half plane model of the hyperbolic plane, the (G2)-version of geodesics, i. e., infinite paths that are straight with respect to this metric, are the (oriented) semi-circles with center on $\mathbb{R}$ or the vertical rays based on the real axis. (See Figure 5.)

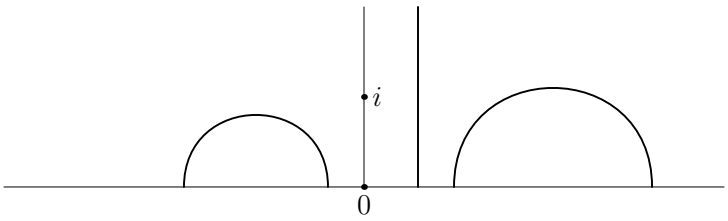

FIGURE 5. Geodesics on $\mathbb{H}$.

[1]Another widely known model for the hyperbolic plane is the Poincaré disk model, which prominently features in several of M. C. Escher's pictures.

A *Riemannian isometry* is a bijective map on $\mathbb{H}$ which preserves the distance between any two points. In particular, any Riemannian isometry maps geodesics to geodesics. The group of *orientation-preserving Riemannian isometries* on the hyperbolic plane is isomorphic to the (projective) matrix group

$$G \coloneqq \mathrm{PSL}_2(\mathbb{R}) \coloneqq \mathrm{SL}_2(\mathbb{R})/\{\pm \mathrm{id}\}.$$

The element $g \in G$ represented by the matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{R})$ is denoted by $g = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, with square brackets. It then has one other representative in $\mathrm{SL}_2(\mathbb{R})$, namely $\left(\begin{smallmatrix} -a & -b \\ -c & -d \end{smallmatrix}\right)$. The *action* of G on $\mathbb{H}$ is given by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot z \coloneqq \frac{az+b}{cz+d}. \tag{5}$$

Occasionally, we will omit the dot $\cdot$ in the notation.

3.2. **The modular surface.** A subgroup of G of particular importance is the *modular group*

$$\Gamma \coloneqq \mathrm{PSL}_2(\mathbb{Z}).$$

It acts on $\mathbb{H}$ preserving the tesselation by triangles as indicated in Figure 6. The

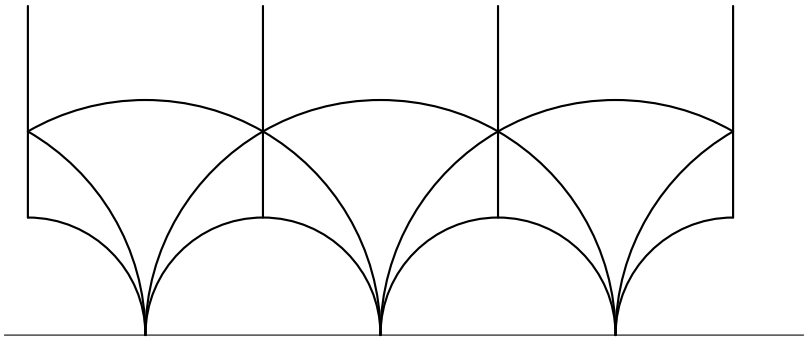

FIGURE 6. Tesselation of $\mathbb{H}$ by triangles.

modular surface is the orbit space

$$X \coloneqq \Gamma\backslash\mathbb{H},$$

that is, the space we obtain if we identify any two points of $\mathbb{H}$ that are mapped to each other by some element of Γ. A model is given by the (closed) *fundamental domain*

$$\mathcal{F}_0 \coloneqq \left\{ z \in \mathbb{H} \;\middle|\; |z| \geq 1,\ |\operatorname{Re} z| \leq \tfrac{1}{2} \right\}$$

(see Figure 7). It contains at least one point of any Γ-orbit. Only points in the boundary of $\mathcal{F}_0$ can be identified under the action of Γ, namely the left vertical boundary is mapped to the right one by the element

$$T \coloneqq \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \tag{6}$$

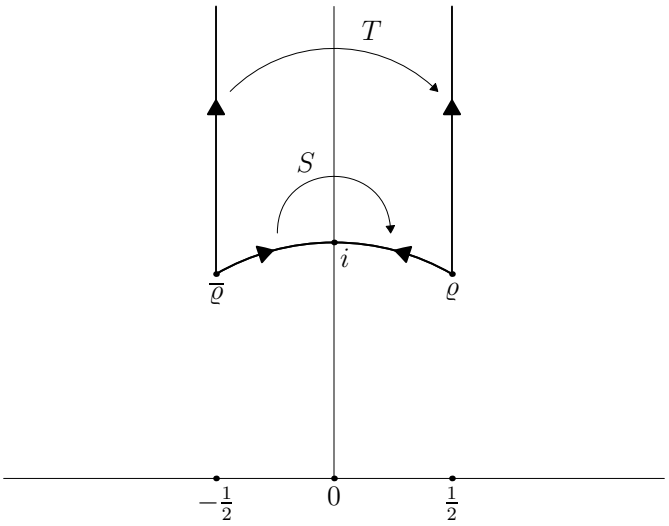

FIGURE 7. Fundamental domain $\mathcal{F}_0$ for Γ.

and the left bottom boundary (from $\overline{\varrho}$ to i) is mapped to the right bottom boundary (from ϱ to i) by

$$S := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{7}$$

If we glue $\mathcal{F}_0$ together according to these boundary identifications then we obtain the modular surface X, as illustrated in Figure 8. This is just like what we did when we represented $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ as $[0,1]/\{0 = 1\}$.

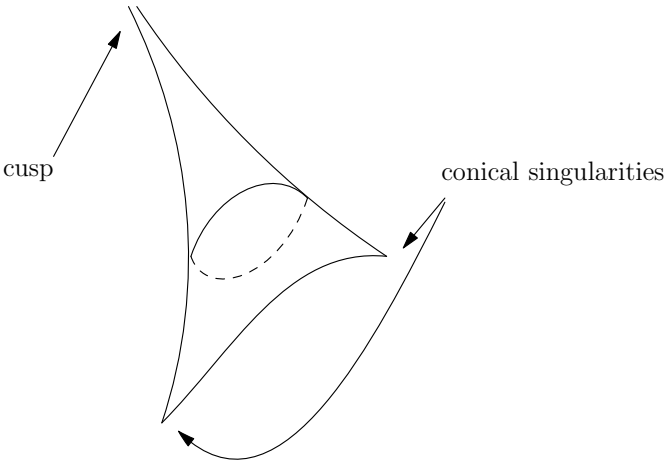

FIGURE 8. The modular surface $X = \Gamma\backslash\mathbb{H}$.

Clearly, there is more than one fundamental domain for the modular surface. Another fundamental domain is, e. g.,

$$\mathcal{F} \;\coloneqq\; \big\{ z \in \mathbb{H} \;\big|\; |z-1| \geq 1,\; 0 \leq \operatorname{Re} z \leq \tfrac{1}{2} \big\}$$

(see Figure 9). It arises from $\mathcal{F}_0$ by cutting off the left half $\mathcal{F}_L \coloneqq \mathcal{F}_0 \cap \{\operatorname{Re} z < 0\}$

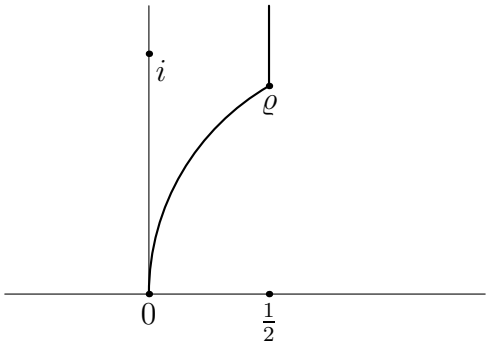

FIGURE 9. Fundamental domain $\mathcal{F}$ for Γ.

from $\mathcal{F}_0$ and gluing $S{\cdot}\mathcal{F}_L$ to the right half of $\mathcal{F}_0$. Thus,

$$\mathcal{F} \;=\; S{\cdot}\mathcal{F}_L \;\cup\; (\mathcal{F}_0 \smallsetminus \mathcal{F}_L).$$

For our constructions in Section 5 below, the fundamental set $\mathcal{F}$ is more convenient than $\mathcal{F}_0$.

The modular surface has an infinite 'end' of finite volume, called the *cusp*. In the fundamental domain $\mathcal{F}_0$ it is represented by the strip going to ∞. In terms of Γ, the presence of the element T in Γ caused the presence of this cusp. As we will see, this cusp and hence the element T play a special role throughout.

For completeness we remark that the modular surface is not a hyperbolic surface in the strict sense because it is not a Riemannian manifold but rather an orbifold. It has the two *conical singularities* at i and ϱ (see Figure 7 or 9). At these points the structure of the quotient space $X = \Gamma\backslash\mathbb{H}$ is not smooth. The non-smoothness, however, does not influence any step in our argumentations.

3.3. **Geometric entity: geodesics.** Just as in the case of the torus, the 'geometric entities' for the modular surface are the periodic geodesics and their lengths. A geodesic on X is the image under the projection map

$$\pi \colon \mathbb{H} \to X = \Gamma\backslash\mathbb{H} \tag{8}$$

of a geodesic on $\mathbb{H}$, as illustrated in Figure 10. Geodesics on $\mathbb{H}$ are infinitely long, but geodesics on X can be either infinitely long or else periodic and of finite length. The (primitive) geodesic length spectrum L_X of X is by definition the multiset of the lengths of periodic geodesics. The periodic geodesics on X are closely related to those elements $g \in \Gamma$ with $|\operatorname{tr}(g)| > 2$, the *hyperbolic elements*: For every periodic geodesic $\widehat{\gamma}$ on X and any representing geodesic γ of $\widehat{\gamma}$ on $\mathbb{H}$ (i. e., $\pi(\gamma) = \widehat{\gamma}$) there

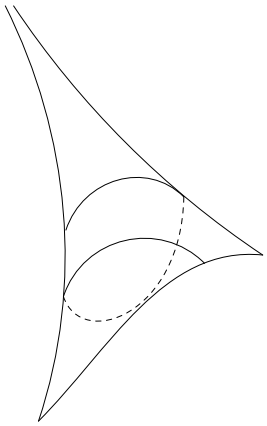

FIGURE 10. A geodesic on the modular surface.

exists a hyperbolic element $g \in \Gamma$ such that $g.\gamma$ is a time-shifted version of γ, i. e., there exists $t_g \in \mathbb{R}$ such that

$$g.\gamma(t) = \gamma(t + t_g) \qquad \text{for all } t \in \mathbb{R}. \tag{9}$$

(Note that $t_g \neq 0$.) If in (9) the value t_g is positive and minimal unter all positive choices for $g \in \Gamma$, then g is *primitive hyperbolic*. An equivalent characterization is that g is hyperbolic and not of the form h^n with $h \in \Gamma$ and $n > 1$.

Conversely, whenever γ is a geodesic on $\mathbb{H}$ and there exists $g \in \Gamma$ and $t_g \in \mathbb{R}$, $t_g \neq 0$ such that (9) holds, then g is hyperbolic and $\pi(\gamma)$ is a periodic geodesic on X. Furthermore, every hyperbolic element in Γ time-shifts a unique geodesic on $\mathbb{H}$. Under this assigment of primitive hyperbolic elements in Γ to periodic geodesics on X, the set of periodic geodesics on X is bijective to the set of conjugacy classes of the primitive hyperbolic elements in Γ, and the (primitive) geodesic length spectrum of X is

$$L_X \;=\; \left\{ 2\operatorname{arcosh}\left(\tfrac{|\operatorname{tr}(g)|}{2}\right) \;\middle|\; g \in \mathrm{HP} \right\},$$

where HP is any set of representatives for the conjugacy classes of primitive hyperbolic elements in Γ. The smallest element in L_X is

$$2\operatorname{arcosh}\left(\tfrac{3}{2}\right) \;=\; 2\log\left(\tfrac{3+\sqrt{5}}{2}\right),$$

and by investigating the set of possible traces of the elements in Γ one can find all elements in L_X (with multiplicities) up to any given bound.

The set L_X is also closely related to the class numbers of indefinite binary quadratic forms. We refer the interested reader to [22, Exercises 18-20 in Section 3.7, and the paragraph below them] and omit any discussion of this relation here.

3.4. **Spectral entity: Laplace eigenfunctions.** We now introduce the spectral objects we are interested in: the Maass wave forms for Γ, and the more restrictive Maass cusp forms.

The Laplacian on $\mathbb{H}$, the *hyperbolic Laplacian*, is

$$\Delta \;\coloneqq\; -y^2\big(\partial_x^2 + \partial_y^2\big) \qquad (z = x + iy).$$

It is the differential operator on $\mathbb{H}$ that commutes with all elements of the group $G = \mathrm{PSL}_2(\mathbb{R})$ of orientation-preserving Riemannian isometries; the factor y^2 corresponds to the factor y^{-2} in the formula of the line element of the Riemannian metric in (4).

Now let $u\colon \mathbb{H} \to \mathbb{C}$ be a Γ-invariant eigenfunction of Δ, that is, a function satisfying $u(g{\cdot}z) = u(z)$ for all $g \in \Gamma$ and all $z \in \mathbb{H}$, and

$$\Delta u \;=\; s(1-s)u \tag{10}$$

for some $s \in \mathbb{C}$. Further below we will see that it is more convenient to work with the *spectral parameter* s rather than with the eigenvalue $s(1-s)$ itself. We do not need to specify *a priori* the regularity of u: since the Laplace operator is elliptic with real-analytic coefficients, the function u is automatically real-analytic.

The invariance of u under the element $T \in \Gamma$ from (6) shows that u is 1-periodic, and hence has a Fourier expansion of the form

$$u(x+iy) \;=\; \sum_{n\in\mathbb{Z}} a_n(y)\, e^{2\pi i n x}.$$

By separation of variables in (10) we see that each function a_n is a solution of a second-order differential equation (depending on s), a modified Bessel equation. This equation has two independent solutions, one exponentially big and one exponentially small as $y \to \infty$, except if $n = 0$, where the two solutions are y^s and y^{1-s} for $s \neq \frac{1}{2}$, and $y^{1/2}$ and $y^{1/2}\log y$ for $s = \frac{1}{2}$. Therefore, if we assume in addition that u has polynomial growth at infinity, in which case u is called a *Maass wave form* for Γ, then the Fourier expansion becomes

$$u(x+iy) \;=\; c_1 y^s + c_2 y^{1-s} + y^{\frac{1}{2}} \sum_{\substack{n\in\mathbb{Z}\\ n\neq 0}} A_n\, K_{s-\frac{1}{2}}(2\pi|n|y)\, e^{2\pi i n x},$$

where the first two terms must be replaced by $c_1 y^{1/2} + c_2 y^{1/2}\log y$ if $s = \frac{1}{2}$. Here K_ν denotes the modified Bessel function of the second kind with index $\nu \in \mathbb{C}$, whose precise definition plays no role here and is therefore omitted, and the A_n are complex numbers that automatically have polynomial growth.

If we further assume that u has *rapid decay* at infinity then $c_1 = c_2 = 0$, and

$$u(x+iy) \;=\; y^{\frac{1}{2}} \sum_{\substack{n\in\mathbb{Z}\\ n\neq 0}} A_n\, K_{s-\frac{1}{2}}(2\pi|n|y)\, e^{2\pi i n x}.$$

In this case, u is called a *Maass cusp form* with spectral parameter s. It is known that the real part of s then always lies between 0 and 1. Since any Maass wave form u is Γ-invariant, we can also consider u as a true function on $X = \Gamma\backslash\mathbb{H}$, and characterize Maass cusp forms as eigenfunctions of Δ on X having rapid decay as their argument tends to the cusp.

The Γ-invariant L^2-eigenfunctions of Δ on $\mathbb{H}$ are the constant functions (with eigenvalue 0) and the Maass cusp forms, whose eigenvalues are positive and tend

to infinity, giving an L^2-Laplace spectrum

$$\sigma(X) \;=\; \big\{0,\; 91.141\cdots,\; 148.432\cdots,\; 190.131\cdots,\; \ldots\big\}$$

whose elements are known numerically to high precision, but not in closed form.

3.5. **Dynamical zeta function.** The analogue of the dynamical zeta function $\zeta_{\mathbb{T}}$ of the torus is the Selberg zeta function Z_X, which has an Euler product given by lengths of periodic geodesics and a Hadamard product in terms of Laplace eigenvalues. More precisely, $Z_X(s)$ is defined for $\operatorname{Re} s > 1$ by

$$Z_X(s) \;=\; \prod_{\ell\in L_X}\prod_{k=0}^{\infty}\big(1-e^{-(s+k)\ell}\big),$$

and the analogue of (3) is Selberg's theorem that this function extends meromorphically to $\mathbb{C}$ and vanishes if s is a spectral parameter.

4. The cohomological interpretation of Maass cusp forms

We now turn to the first step in the passage from geodesics on the modular surface X to Maass cusp forms for Γ: The interpretation of Maass cusp forms in terms of *parabolic* 1*-cohomology* as provided in [3].

The essential part of this cohomological interpretation, of which we take advantage here, is that every Maass cusp form u with spectral parameter s is characterized by a vector $(c_g^u)_{g\in\Gamma}$ of functions given by integrals of the form

$$c_g^u(t) \;=\; \int_{g^{-1}\infty}^{\infty} \omega_s(u,t) \qquad (t\in\mathbb{R}),$$

where $\omega_s(u,\cdot)$ is a certain closed 1-form on $\mathbb{H}$ defined below and where the integration is along any path in $\mathbb{H}\cup\mathbb{R}\cup\{\infty\}$ from $g^{-1}\infty$ to ∞, with at most finitely many points in

$$\mathbb{P}^1(\mathbb{R}) \;\coloneqq\; \mathbb{R}\cup\{\infty\}.$$

The functions $(c_g^u)_{g\in\Gamma}$ satisfy certain relations among each other, so-called cocycle relations, showing that a suitable cohomology theory is the natural home of this setup.

For completeness of exposition and for the convenience of the reader we provide a rather detailed definition of this cohomology (specialized to the modular group Γ), even though these details will not be needed further on. Readers who want to proceed faster to the final result are invited to skip the remaining part of this section after having read Theorem 4.1 below. They should interpret the space $H^1_{\mathrm{par}}(\Gamma;\mathcal{V}_s^{\omega^*,\infty})$ as a vector space whose elements are equivalence classes of maps from Γ to the space of highly regular functions on $\mathbb{R}$ (or rather on $\mathbb{P}^1(\mathbb{R})$), where the notion of 'highly regular' depends on the parameter s. Theorem 4.1 then states that the assignment of Maass cusp forms u with spectral parameter s to the equivalence class of the vector (c_g^u) is bijective and linear.

For the detailed description we start with a few preparations. The parabolic cohomology will then be seen a refinement of the standard group cohomology in order to account for the cusp of the modular surface and the rapid decay of the

Maass cusp forms towards this cusp. The name *parabolic* alludes to the fact that elements in G that stabilize a single point in $\mathbb{P}^1(\mathbb{R})$, such as T, are called parabolic.

The upper half plane $\mathbb{H}$ has a dynamically defined boundary consisting of all 'infinite endpoints' of its geodesics. Considering Figure 5, this boundary is given by $\mathbb{P}^1(\mathbb{R})$. The action of G on $\mathbb{H}$, as defined in (5), extends continuously to an action on $\mathbb{H} \cup \mathbb{P}^1(\mathbb{R})$ in the obvious way, replacing the right-hand side of (5) by a/c if $z = \infty$ (and $c \neq 0$) and by ∞ if $z = -d/c$. For any $s \in \mathbb{C}$, we define an action of G on locally-defined functions on $\mathbb{P}^1(\mathbb{R})$ by setting

$$\tau_s(g^{-1})f(t) \;\coloneqq\; \big(g'(t)\big)^s f(g{\cdot}t) \tag{11}$$

(sometimes also denoted $f|_{2s}g$) wherever it is defined.

Let $\mathcal{V}_s^{\omega^*,\infty}$ (called the space of *smooth, semi-analytic vectors of the principal series representation with spectral parameter s in the line model*) denote the space of smooth (C^∞) functions $\varphi\colon \mathbb{P}^1(\mathbb{R}) \to \mathbb{C}$ that are real-analytic on $\mathbb{R}$ up to a finite set that may depend on φ, with the action (11). Smoothness at the point ∞ here means that the map

$$\tau_s(S)\varphi\colon t \mapsto |t|^{-2s}\varphi\big(-\tfrac{1}{t}\big)$$

extends smoothly to the point 0 (recall the element S from (7)). The vector space $Z^1_{\mathrm{par}}(\Gamma; \mathcal{V}_s^{\omega^*,\infty})$ of *parabolic* 1*-cocycles* is then the space of maps $c\colon \Gamma \to \mathcal{V}_s^{\omega^*,\infty}$ such that

- for all $g, h \in \Gamma$, we have

$$c_{gh} \;=\; \tau_s(h^{-1})c_g + c_h, \tag{12}$$

  where c_g denotes the function $c(g)$, and
- there exists $\varphi \in \mathcal{V}_s^{\omega^*,\infty}$ such that

$$c_T \;=\; \tau_s(T^{-1})\varphi - \varphi.$$

  (For general discrete subgroups we would need a similar condition for representatives of each conjugacy class of parabolic elements.)

The subspace $B^1_{\mathrm{par}}(\Gamma; \mathcal{V}_s^{\omega^*,\infty})$ of 1*-coboundaries* consists of the maps $c\colon \Gamma \to \mathcal{V}_s^{\omega^*,\infty}$ for which there exists $\varphi \in \mathcal{V}_s^{\omega^*,\infty}$ such that

$$c_g \;=\; \tau_s(g^{-1})\varphi - \varphi$$

for every $g \in \Gamma$. The quotient

$$H^1_{\mathrm{par}}(\Gamma; \mathcal{V}_s^{\omega^*,\infty}) \;\coloneqq\; Z^1_{\mathrm{par}}(\Gamma; \mathcal{V}_s^{\omega^*,\infty})/B^1_{\mathrm{par}}(\Gamma; \mathcal{V}_s^{\omega^*,\infty})$$

is called the *space of parabolic* 1*-cohomology classes* with values in $\mathcal{V}_s^{\omega^*,\infty}$.

For any two real-analytic functions u, v on $\mathbb{H}$ we define the *Green's form* to be the real-analytic 1-form

$$[u, v] \;\coloneqq\; \frac{\partial u}{\partial z} \cdot v \cdot dz \;+\; u \cdot \frac{\partial v}{\partial \overline{z}} \cdot d\overline{z},$$

which is easily seen to be closed if u and v are eigenfunctions of Δ with the same eigenvalue. For any $s \in \mathbb{C}$ and any $t \in \mathbb{R}$ the function $R(t;\cdot)^s\colon \mathbb{H} \to \mathbb{C}$, where

$$R(t; z) \;\coloneqq\; \operatorname{Im} \frac{1}{t - z}\,,$$

is a Δ-eigenfunction with eigenvalue $s(1-s)$. Therefore, if u is a Maass cusp form with spectral parameter s, then for any $t \in \mathbb{R}$ the 1-form

$$\omega_s(u,t) \coloneqq \left[u, R(t;\cdot)^s\right]$$

is closed. From this it follows that, for any $g \in \Gamma$, the integral

$$c_g^u(t) \coloneqq \int_{g^{-1}\infty}^{\infty} \omega_s(u,t) \tag{13}$$

is independent of the chosen path from $g^{-1}\infty$ to ∞. The integral is convergent due to the rapid decay of u at the cusp. The regularities of u and $R(\cdot\,;\cdot)^s$ yield $c_g^u \in \mathcal{V}_s^{\omega^*,\infty}$. Furthermore, the Γ-invariance of u implies the transformation formula

$$\tau_s(g) \int_a^b \omega_s(u,t) \;=\; \int_{g\cdot a}^{g\cdot b} \omega_s(u,t) \qquad (g \in \Gamma,\ a,b \in \mathbb{P}^1(\mathbb{R})) \tag{14}$$

and from this one easily deduces that the map c^u satisfies the cocycle relation (12) and hence is a parabolic cocycle. Then we have:

Theorem 4.1 ([12, 3])**.** *For $s \in \mathbb{C}$, $\operatorname{Re} s \in (0,1)$, the map $u \mapsto [c^u]$ defines a bijection*

$$\{\textit{Maass cusp forms with spectral parameter } s\} \xrightarrow{\sim} H^1_{\mathrm{par}}(\Gamma; \mathcal{V}_s^{\omega^*,\infty}).$$

5. Discretization of geodesics

In this section we will discuss the second step in the passage from geodesics on the modular surface X to Maass cusp forms for Γ: The construction of a discretization of the motion along the geodesics on X.

We will show that the *discrete dynamical system*

$$F\colon (0,\infty) \smallsetminus \mathbb{Q} \to (0,\infty) \smallsetminus \mathbb{Q}$$

given by the two branches

$$\begin{cases} (0,1) \smallsetminus \mathbb{Q} \xrightarrow{\sim} (0,\infty) \smallsetminus \mathbb{Q}, & x \mapsto T_1^{-1}x = \frac{x}{1-x} \\ (1,\infty) \smallsetminus \mathbb{Q} \xrightarrow{\sim} (0,\infty) \smallsetminus \mathbb{Q}, & x \mapsto T_2^{-1}x = x-1 \end{cases} \tag{15}$$

can be thought of as a discrete version of the geodesic flow on X: The map F and its iterates capture the essential geometric and dynamical properties of the geodesic flow that will be needed for establishing the relation between the geodesics on X and the Maass cusp forms for Γ. In particular, the orbits of the map F describe the future behavior of (almost all) geodesics on X, and periodic geodesics on X correspond to points $x \in (0,\infty) \smallsetminus \mathbb{Q}$ with periodic (i. e., finite) orbits under F.[2]

The construction of F from the geodesic flow on X proceeds in several steps: We first choose a 'good' cross section (in the sense of Poincaré) for the geodesic flow on X, i. e., a subset $\widehat{C}$ of the unit tangent bundle of X that is intersected by all periodic geodesics at least once, and each intersection between any geodesic on X and $\widehat{C}$ is discrete. We refer to the discussion below for precise definitions.

[2]We remark that the formula for F is identical to the map Φ given in [6, Section 1.1, Lemma] in connection with the so-called rational period functions.

The choice of $\widehat{C}$ yields a first return map, which is the map that assigns to each element $\widehat{v} \in \widehat{C}$ the next intersection between $\widehat{C}$ and the geodesic on X starting at time 0 in the direction $\widehat{v}$. The first return map provides a first discretization of the geodesic flow on X.

Then we choose a 'good' set of representatives for $\widehat{C}$, i. e., a subset C^* of the unit tangent bundle of $\mathbb{H}$ that is bijective to $\widehat{C}$ with respect to the canonical quotient map. The specific properties of C^* will allow us to semi-conjugate the first return map to a map on $(0,\infty) \smallsetminus \mathbb{Q}$, which is precisely the map F.

The construction we will present below is a special case of the algorithm in [19] for finding good discretizations for geodesic flows on much much general hyperbolic surfaces. We refer to [19] for further details and all omitted proofs.

As in Section 5, readers who want to proceed faster to the final result are invited to skip the remaining part of this section. In Section 6 only the map F will be needed, not the details of its construction.

5.1. **Geodesics.** While in Section 3 we used the notion of geodesics in the sense of (G2) (adapted to the hyperbolic plane and the modular surface in place of the real line and the torus), we now also need geodesics in the sense of (G1).

A geodesic γ on $\mathbb{H}$ in the sense of (G1) is completely determined by requiring that it passes through a given point $z \in \mathbb{H}$ at time $t = 0$ in a given direction. Recall that we consider only geodesics of unit speed, so that the speed in the given direction does not form another parameter. Therefore we may identify geodesics in the sense of (G1) with the set of all unit length direction vectors at all points of $\mathbb{H}$, thus, with the *unit tangent bundle* $S\mathbb{H}$ of $\mathbb{H}$.

For $v \in S\mathbb{H}$ we let $\gamma_v \colon \mathbb{R} \to \mathbb{H}$ be the geodesic on $\mathbb{H}$ such that

$$\gamma_v'(0) = v. \tag{16}$$

Both the tangent vector $\gamma_v'(0)$ to γ_v at time $t = 0$ and the element $v \in S\mathbb{H}$ are combinations of position and direction, the position $\gamma_v(0)$ being the *base point* $\mathrm{base}(v) \in \mathbb{H}$. The *geodesic flow* on $\mathbb{H}$ (the motion along geodesics on $\mathbb{H}$) is the map

$$\mathbb{R} \times S\mathbb{H} \to S\mathbb{H}, \quad (t, v) \mapsto \gamma_v'(t). \tag{17}$$

The action of G on $\mathbb{H}$ by Riemannian isometries induces an action of G on $S\mathbb{H}$ by

$$g \cdot v \;\coloneqq\; (g \cdot \gamma_v)'(0) \qquad (g \in G,\ v \in S\mathbb{H}).$$

The *unit tangent bundle* of X is then just the quotient

$$SX \;=\; \Gamma \backslash S\mathbb{H}.$$

We denote the projection map

$$\pi \colon S\mathbb{H} \to SX \tag{18}$$

with the same symbol as the projection map $\mathbb{H} \to X$ of (8). The context always clarifies which one is meant. We typically denote a geodesic on $\mathbb{H}$ by γ and a unit tangent vector in $S\mathbb{H}$ by v, and use $\widehat{\gamma}$ and $\widehat{v}$ for the corresponding geodesic $\pi(\gamma)$

on X and unit tangent vector $\pi(v) \in SX$. In analogy with (16), for any $\widehat{v} \in SX$ we let $\widehat{\gamma}_v$ denote the geodesic on X determined by

$$\widehat{\gamma}_v'(0) = \widehat{v}.$$

Also the *geodesic flow* on X is inherited from the geodesic flow on $\mathbb{H}$ as defined in (17), and hence is the map

$$\mathbb{R} \times SX \to SX, \quad (t, \widehat{v}) \mapsto \widehat{\gamma}_v'(t).$$

5.2. **Cross section.** By a *cross section* we mean (slightly deviating from the standard definition) a subset $\widehat{C}$ of SX such that

(C1) every periodic geodesic on X intersects $\widehat{C}$. In other words, for any periodic geodesic $\widehat{\gamma}$ there exists $t \in \mathbb{R}$ such that $\widehat{\gamma}'(t) \in \widehat{C}$.

(C2) each intersection of any geodesic on X with $\widehat{C}$ is discrete. In other words, for any geodesic $\widehat{\gamma}$ and $t \in \mathbb{R}$ with $\widehat{\gamma}'(t) \in \widehat{C}$ there exists $\varepsilon > 0$ such that

$$\widehat{\gamma}'\big((t-\varepsilon, t+\varepsilon)\big) \cap \widehat{C} = \big\{\widehat{\gamma}'(t)\big\}.$$

We define a *set of representatives* C^* for a cross section $\widehat{C}$ to be a subset of $S\mathbb{H}$ that is bijective to $\widehat{C}$ under the projection map π from (18). (We write C^* rather than C because the latter traditionally denotes the full preimage of $\widehat{C}$ in $S\mathbb{H}$.) Of course, to characterize a cross section $\widehat{C}$ it suffices to provide a set of representatives, but choosing a cross section and a set of representatives that serves our purposes is an art. For the modular surface we will take

$$C^* \;\coloneqq\; \{v \in S\mathbb{H} \mid \operatorname{base}(v) \in i\mathbb{R}^+,\ \gamma_v(\infty) \in (0,\infty) \smallsetminus \mathbb{Q}\}$$

as set of representatives, where

$$\gamma_v(\infty) \;\coloneqq\; \lim_{t\to\infty} \gamma_v(t).$$

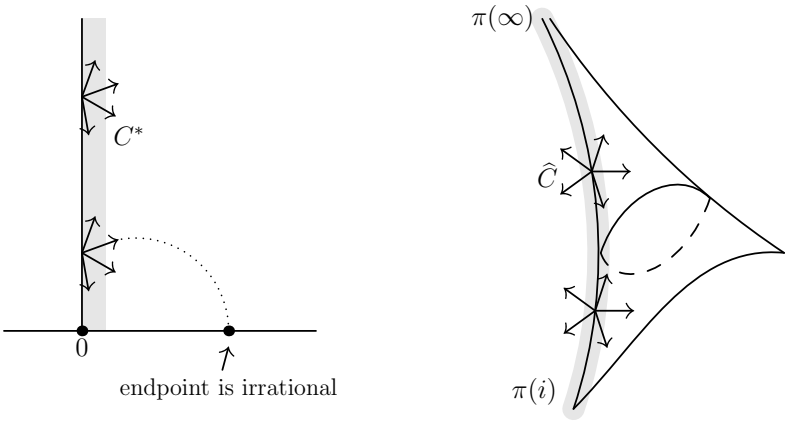

FIGURE 11. The set of representatives C^* and the cross section $\widehat{C}$. The gray shadows indicate the directions of the elements of $\widehat{C}$ and C^*.

The associated cross section

$$\widehat{C} \;\coloneqq\; \pi(C^*)$$

is the set of unit tangent vectors $\widehat{v} \in SX$ sitting on the geodesic from $\pi(i)$ to $\pi(\infty)$ such that the geodesic emanating from $\widehat{v}$ does not converge to the cusp $\pi(\infty)$ in future or past time. A pictorial representation of C^* and $\widehat{C}$ is given in Figure 11.

5.3. **Discretization.** We will now show how to relate the geodesic flow on X to a discrete dynamical system on (a subset of) $\mathbb{R}_{>0}$. In the case of the modular surface, this construction is closely related to continued fractions, more precisely to Farey fractions. The reader interested in this connection may find the articles [1, 20, 21, 9] useful.

Let $\widehat{v} \in \widehat{C}$ be an element of the cross section and consider the associated geodesic $\widehat{\gamma}_v$ on X. By the choice of $\widehat{C}$, the geodesic $\widehat{\gamma}_v$ intersects $\widehat{C}$ again in future time. Let $t_0 > 0$, the *first return time*, be the minimal positive number such that

$$\widehat{w} \;\coloneqq\; \widehat{\gamma}'_v(t_0) \in \widehat{C}.$$

(See Figure 12.) Let $v, w \in C^*$ be the elements in the set of representatives

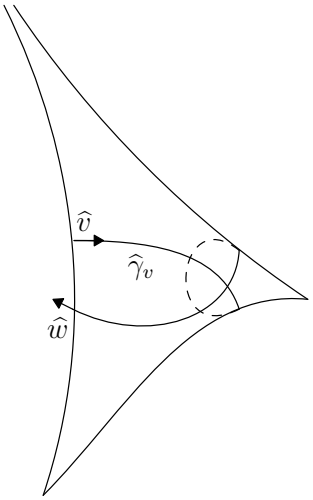

FIGURE 12. The geodesic determined by $\widehat{v}$ and its first return to $\widehat{C}$.

corresponding to $\widehat{v}, \widehat{w}$, and γ_v, γ_w the associated geodesics on $\mathbb{H}$. (See Figure 13.) Since the unit tangent vector $\gamma'_v(t_0) \in S\mathbb{H}$ projects to $\widehat{w}$ under π, that is,

$$\pi\big(\gamma'_v(t_0)\big) \;=\; \widehat{w},$$

there exists a unique element $g \in \Gamma$ such that

$$\gamma'_v(t_0) \;=\; g{\cdot}w\,.$$

This element is characterized by

$$\gamma'_v(t_0) \in g{\cdot}C^*, \tag{19}$$

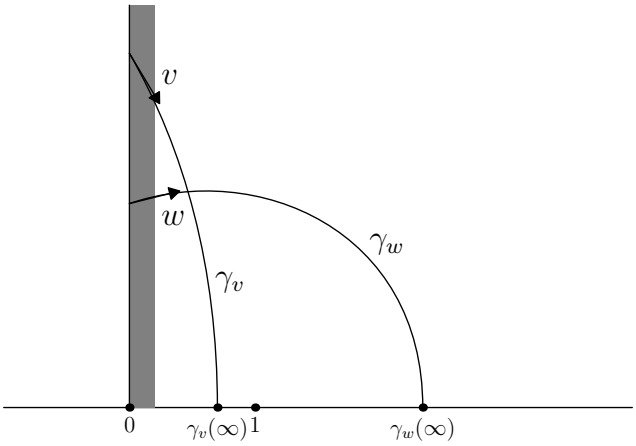

FIGURE 13. Associated geodesics on $\mathbb{H}$.

i. e., by the first intersection of γ_v with some Γ-translate of C^* after passing through v. To find the element g we consider the neighboring translates of the fundamental domain $\mathcal{F}$ and the relevant translates of C^*.

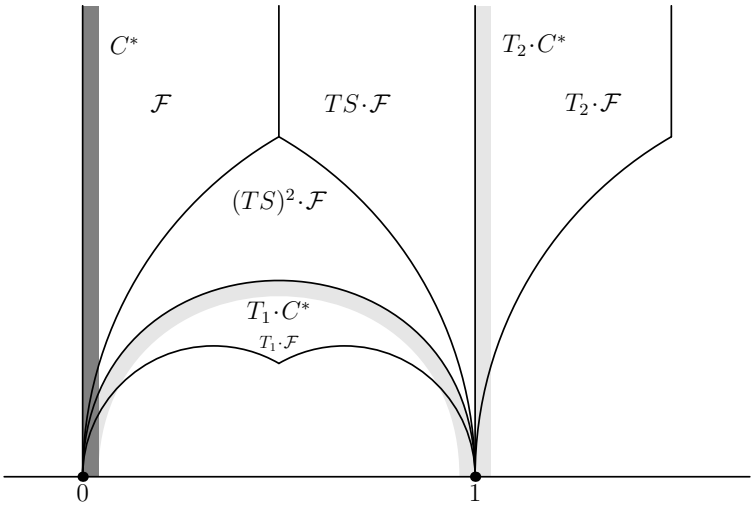

FIGURE 14. Relevant Γ-translates of $\mathcal{F}$ and C^*.

We observe that the unit tangent vector $\gamma_v'(t_0)$ can be only in $T_1 \cdot C^*$ or $T_2 \cdot C^*$, where

$$T_1 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \qquad \text{and} \qquad T_2 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

as shown in Figure 14. Explicitly, this unit tangent vector is in $T_1 \cdot C^*$ if and only if $\gamma_v(\infty) \in (0, 1)$, and it is in $T_2 \cdot C^*$ if and only if $\gamma_v(\infty) \in (1, \infty)$. In Figure 15 we

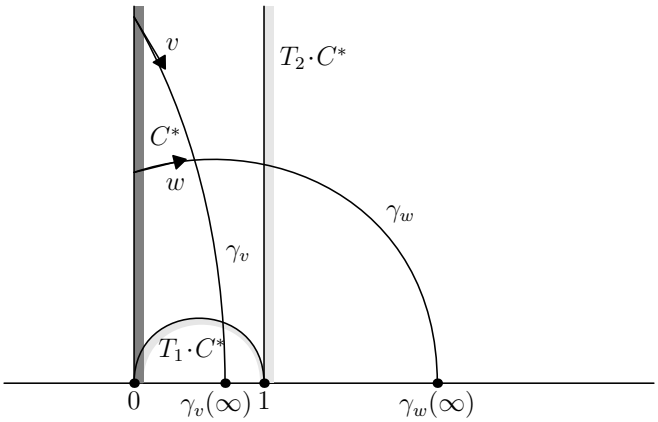

FIGURE 15. Next intersection.

have $g = T_1$, so that here

$$w \;=\; T_1^{-1}\gamma_v'(t_0), \qquad \gamma_w(\infty) \;=\; T_1^{-1}\gamma_v(\infty).$$

We further observe that for every point $x \in (0,\infty) \smallsetminus \mathbb{Q}$, no matter which $v \in C^*$ with $\gamma_v(\infty) = x$ we consider, we find the same value for the element $g \in \Gamma$ in (19). In other words, g only depends on x, not on the specific element $v \in C^*$ with $\gamma_v(\infty) = x$. Therefore the procedure just described induces a *discrete dynamical system*

$$F\colon (0,\infty) \smallsetminus \mathbb{Q} \to (0,\infty) \smallsetminus \mathbb{Q}, \tag{20}$$

where for each $x \in (0,\infty) \smallsetminus \mathbb{Q}$, we pick $v \in C^*$ such that $\gamma_v(\infty) = x$, let g be the element in Γ such that $\gamma_v'(t_0) \in g \cdot C^*$ and set

$$F(x) \;\coloneqq\; g^{-1} \cdot x.$$

Theorem 5.1 ([19])**.** *The set $\widehat{C}$ is a cross section for the geodesic flow on X, and C^* is a set of representatives for $\widehat{C}$. The induced discrete dynamical system (as in* (20)*) is the map F as given in* (15)*.*

6. Transfer operators and Maass cusp forms

In this section we carry out the third and final step in the passage from geodesics on the modular surface X to Maass cusp forms for Γ: Tie together the discrete dynamical system F from Section 5 and the parabolic interpretation of Maass cusp forms from Section 4.

The mediating object between both sides is the *transfer operator family* $(\mathcal{L}_s)_{s\in\mathbb{C}}$ associated to F. The *transfer operator* $\mathcal{L}_s$ with parameter s is the operator

$$\mathcal{L}_s f(t) \;\coloneqq\; \sum_{w \in F^{-1}(t)} |F'(w)|^{-s} f(w), \tag{21}$$

acting on functions $f\colon (0,\infty) \to \mathbb{C}$. This operator has its origin in the thermodynamic formalism of statistical mechanics. It is a generalization of the transfer

matrix for lattice–spin systems, which is used to find equilibrium distributions. The weight, in particular its s-dependence, is motivated within this framework, where s serves as an inverse Boltzmann constant and temperature. From a purely mathematical point of view, this operator can be seen as an evolution operator or as a graph Laplacian on a somewhat generalized graph, in both cases with appropriate weights. The explicit expression for F allows us to evaluate (21) in our special case to

$$\mathcal{L}_s f(t) \;=\; f(t+1) \;+\; (t+1)^{-2s} f\Big(\frac{t}{t+1}\Big), \qquad t>0,$$

or, using (11), to

$$\mathcal{L}_s \;=\; \tau_s(T_1^{-1}) \;+\; \tau_s(T_2^{-1}).$$

(This simple formula is for the modular group only. For other groups one can have a vector of more complicated finite sums.)

The correspondence that we have been aiming at is a bijection between the eigenfunctions of $\mathcal{L}_s$ with eigenvalue 1 and the Maass cusp forms with spectral parameter s. More precisely, we have the following theorem.

Theorem 6.1 ([16, 17])**.** *Let $s\in\mathbb{C}$, $1>\operatorname{Re} s>0$. Then the Maass cusp forms with spectral parameter s are bijective to the real-analytic eigenfunctions f of $\mathcal{L}_s$ for which the map*

$$\begin{cases} f & \text{on } (0,\infty) \\ -\tau_s(S)f & \text{on } (-\infty,0) \end{cases} \tag{22}$$

extends smoothly to 0. If u is a Maass cusp form with spectral parameter s then the associated eigenfunction of $\mathcal{L}_s$ is

$$f(t) \;\coloneqq\; \int_0^\infty \omega_s(u,t). \tag{23}$$

We will now explain the main steps of the proof with an emphasis on intuition and heuristics. Some steps will be omitted, most prominently some discussions of convergence and regularities. We hope to convince the reader that a major part of the proof is encoded in Figure 16 and that the choice of the integral path in (23) and the function in (22) is natural. The exposition will show that the bijection

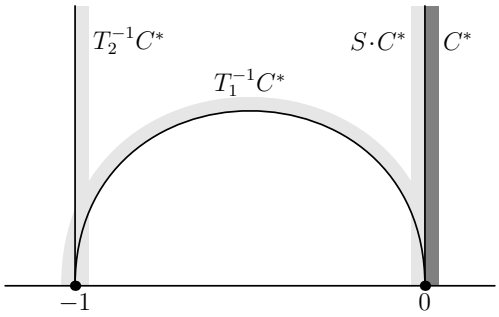

FIGURE 16. Relevant Γ-translates for proof of Theorem.

claimed in the Theorem is not just proven by showing that the dimensions of both spaces are equal; we will provide an explicit map.

Proof (key elements). We present the main part of the proof, split into four steps.

Step 1: Relation between $\mathcal{L}_s$ and C^*. We first reconsider the transfer operator $\mathcal{L}_s$ and its domain. Let $f\colon (0,\infty) \to \mathbb{C}$ be a function in the domain of $\mathcal{L}_s$. We may think of f as being a mass distribution or density on $(0,\infty)$ of which the transfer operator evaluates its s-weighted evolution under one application of F. Recalling that F is a discrete version of the geodesic flow on X, that $\mathcal{L}_s$ is a weighted evolution operator of F, and that the essential ingredient of this discretization is the set C^*, we may intuitively think of f as being a 'shadow' of some function f^* on C^* that is constant on any set of the form

$$E_t \;\coloneqq\; \{v \in C^* \mid \gamma_v(\infty) = t\} \qquad (t \in (0,\infty)).$$

Thus,

$$f(t) \;=\; f^*(v) \qquad \text{for any } v \in E_t.$$

When developing the formula for F we asked where the geodesics determined by the elements in C^* go to. In the expression for $\mathcal{L}_s$, the preimage of F is used. Hence, when building $\mathcal{L}_s$, we may alternatively ask where these geodesics come from. For the modular group Γ, the relevant sets are $T_1^{-1}C^*$ and $T_2^{-1}C^*$. (See Figure 16.)

Step 2: Relation between Maass cusp forms and C^*. Let u be a Maass cusp form with spectral parameter s. We use the characterization of u via a cocycle class in the space $H^1_{\mathrm{par}}(\Gamma; \mathcal{V}_s^{\omega^*,\infty})$ from the Theorem in Section 4, and then use the family of functions $(c_g^u)_{g\in\Gamma}$ from (13) as a representative for this cocycle class. We think of each c_g as being the integral along the geodesic from $g^{-1}\infty$ to ∞, or even better, as an integral over the set of unit tangent vectors to this geodesics. In particular, for $g = S$ we have $S^{-1}\infty = 0$, so that

$$c_S^u(t) \;=\; \int_0^\infty \omega_s(u,t) \qquad (t \in \mathbb{R}) \tag{24}$$

is the integral along the geodesic from 0 to ∞. Thus, in an intuitive way, we may think of c_S^u as an integral over $C^* \cup S{\cdot}C^*$.

Step 3: From Maass cusp forms to eigenfunctions of $\mathcal{L}_s$. Let u be a Maass cusp form with spectral parameter s with associated function vector $(c_g^u)_{g\in\Gamma}$. We want to associate to u in a natural way an eigenfunction f of $\mathcal{L}_s$ with eigenvalue 1. The intuitive way of thinking of c_S^u and any function f as objects on C^* suggests using C^* as linking pin. Staying in this intuition, we should restrict c_S^u to an integral over C^* and use $f^* = c_S^u|_{C^*}$. In terms of the actual objects (and their rigorous definitions) we are led to set

$$f \;\coloneqq\; c_S^u|_{(0,\infty)}, \tag{25}$$

which is precisely (23).

We now show that (25) indeed defines an eigenfunction of $\mathcal{L}_s$ with eigenvalue 1. So far we have used in (24), and hence in (25), the geodesic from 0 to ∞ as path

of integration. Since the 1-form $\omega_s(u,t)$ is closed, we may change the path to be the geodesic from 0 to -1 followed by the geodesic from -1 to ∞:

$$\int_0^\infty \omega_s(u,t) \;=\; \int_0^{-1} \omega_s(u,t) \;+\; \int_{-1}^\infty \omega_s(u,t).$$

Using the transformation formula (14) we now find

$$\begin{aligned} f(t) \;&=\; \int_0^\infty \omega_s(u,t) \\ &=\; \int_{T_1^{-1}0}^{T_1^{-1}\infty} \omega_s(u,t) \;+\; \int_{T_2^{-1}0}^{T_2^{-1}\infty} \omega_s(u,t) \\ &=\; \tau_s(T_1^{-1}) \int_0^\infty \omega_s(u,t) \;+\; \tau_s(T_2^{-1}) \int_0^\infty \omega_s(u,t) \\ &=\; \tau_s(T_1^{-1}) f(t) \;+\; \tau_s(T_2^{-1}) f(t). \end{aligned}$$

Therefore $f = \mathcal{L}_s f$.

Step 4: From eigenfunctions of $\mathcal{L}_s$ to Maass cusp forms. Conversely, let f be an eigenfunction $\mathcal{L}_s$ with eigenvalue 1. We want to associate to f a Maass cusp form u in a way which inverts the mapping from above and which is also natural. Instead of trying to do this directly, we will define a parabolic 1-cocycle $c = c^f$ in $Z^1_{\mathrm{par}}(\Gamma; \mathcal{V}_s^{\omega^*,\infty})$. The Theorem in Section 4 then implies that the cocycle c is indeed of the form $c = c^u$ for a unique Maass cusp form u.

In order to define c we prescribe it on the group elements T and S by setting

$$c_T \;\coloneqq\; 0,$$

which is motivated by (13), and

$$c_S \;\coloneqq\; \begin{cases} f & \text{on } (0,\infty) \\ -\tau_s(S) f & \text{on } (-\infty, 0), \end{cases} \tag{26}$$

according to the heuristic above. The minus sign in the second row is motivated by the fact that S 'changes the direction' of the geodesic from 0 to ∞. In order to define c_S also at 0 (and ∞), we need to require that the right hand side of (26), equivalently (22), extends smoothly to 0.

Since T and S generate all of Γ, the cocycle relation (12) dictates the value of c on all other elements. It remains to show that c is well-defined, which here means that if a combination of T and S equals the identity then the corresponding combination of c_T and c_S vanish. To that end we use the presentation

$$\Gamma \;=\; \Big\langle\; S,\; T \;\Big|\; S^2 = \big(T^{-1}S\big)^3 = \mathrm{id} \;\Big\rangle$$

and show that

$$\tau_s(S) c_S \;+\; c_S \quad \text{and} \quad \big(\tau_s\big((ST)^2\big) + \tau_s(ST) + 1\big)\big(\tau_s(S) c_{T^{-1}} + c_S\big)$$

vanish identically. For the first expression, this follows immediately from (26). For the second expression we use $c_T = 0$ and find

$$\begin{aligned}
&\big(\tau_s\big((ST)^2\big) + \tau_s(ST) + 1\big)\big(\tau_s(S)c_{T^{-1}} + c_S\big) \\
&\qquad = \ \tau_s\big((ST)^2\big)c_S \ + \ \tau_s(ST)c_S \ + \ c_S \\
&\qquad = \ \begin{cases} -\tau_s\big(T_2^{-1}\big)f - \tau_s\big(T_1^{-1}\big)f + f & \text{on } (0,\infty) \\ \tau_s\big(T_1^{-1}S\big)\big[-\tau_s\big(T_1^{-1}\big)f + f - \tau_s\big(T_2^{-1}\big)f\big] & \text{on } (-1,0) \\ \tau_s\big(T^{-1}S\big)\big[f - \tau_s\big(T_2^{-1}\big)f - \tau_s\big(T_1^{-1}\big)f\big] & \text{on } (-\infty,-1), \end{cases}
\end{aligned}$$

which vanishes since $f = \mathcal{L}_s f$. This calculation

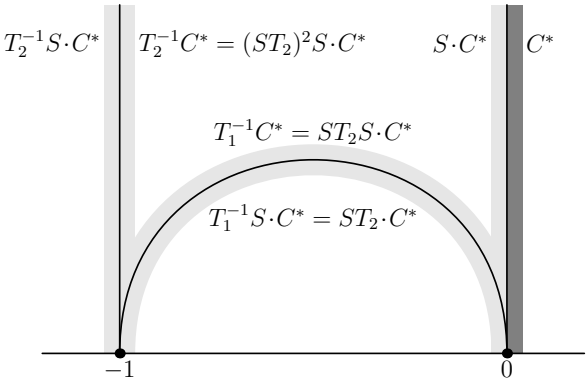

FIGURE 17. Relevant Γ-translates for proof of Theorem.

can also be read off from Figure 17, as the reader can verify. □

7. Recapitulation and closing comments

We have surveyed an intriguing relation between the periodic geodesics on the modular surface $X = \Gamma\backslash\mathbb{H}$ ('classical mechanical objects') and the Maass cusp forms for Γ ('quantum mechanical objects'). For this, we started simultaneously on both ends:

On the geometric side, we developed a discrete version of the (periodic part of the) geodesic flow on the modular surface by means of a cross section in the sense of Poincaré. We realized this discretization as a discrete dynamical system on $(0,\infty)$ by using a well-chosen representation of the cross section on the upper half plane. This step turns the geodesic flow into a discrete and somehow finite object while preserving its essential dynamical features.

On the spectral side, we characterized the Maass cusp forms as cocycle classes in a certain cohomology space. The isomorphism from Maass cusp forms to cocycle classes is given by an integral transform, where a certain 1-form is integrated along certain geodesics. Even though the cocycle classes remain objects of quantum mechanical nature, this characterization of Maass cusp forms constitutes a first and very important step towards the geometry and dynamics of the modular surface.

Connecting these two sides is the family of transfer operators, which from their definition are purely classical mechanical objects but which clearly exhibit a quantum mechanical nature. These transfer operators depend heavily on the choice of the discretization. The proof of the isomorphism between eigenfunctions of the transfer operators and the parabolic 1-cocycles clearly shows that the shape of the set of representatives is crucial. Here, it is the set of (almost) all unit tangent vectors that are based on the geodesic from 0 to ∞ and that point 'to the right'.

This set of representatives and its Γ-translates can be seen as a geometric realization of the cohomology. The transfer operator then encodes the cocycle relation. An eigenfunction with eigenvalue 1 of the transfer operator obeys a geometric variant of the cocycle relation, and hence can be related to an actual cocycle, which in turn characterizes a Maass cusp form.

References

[1] E. Artin, *Ein mechanisches System mit quasiergodischen Bahnen*, Abh. Math. Sem. Univ. Hamburg **3** (1924), 170–175.

[2] R. Bruggeman, *Automorphic forms, hyperfunction cohomology, and period functions*, J. Reine Angew. Math. **492** (1997), 1–39.

[3] R. Bruggeman, J. Lewis, and D. Zagier, *Period functions for Maass wave forms and cohomology*, Mem. Am. Math. Soc. **1118** (2015), iii–v + 128.

[4] R. Bruggeman and T. Mühlenbruch, *Eigenfunctions of transfer operators and cohomology*, Journal of Number Theory **129** (2009), 158–181.

[5] C.-H. Chang and D. Mayer, *The transfer operator approach to Selberg's zeta function and modular and Maass wave forms for* $\mathrm{PSL}(2, \mathbf{Z})$, Emerging applications of number theory (Minneapolis, MN, 1996), IMA Vol. Math. Appl., vol. 109, Springer, New York, 1999, pp. 73–141.

[6] Y. Choie and D. Zagier, *Rational period functions for* $\mathrm{PSL}(2, \mathbb{Z})$, A tribute to Emil Grosswald: number theory and related analysis, Providence, RI: American Mathematical Society, 1993, pp. 89–108.

[7] A. Deitmar and J. Hilgert, *A Lewis correspondence for submodular groups*, Forum Math. **19** (2007), no. 6, 1075–1099.

[8] M. Fraczek, D. Mayer, and T. Mühlenbruch, *A realization of the Hecke algebra on the space of period functions for* $\Gamma_0(n)$, J. Reine Angew. Math. **603** (2007), 133–163.

[9] S. Katok and I. Ugarcovici, *Symbolic dynamics for the modular surface and beyond*, Bull. Amer. Math. Soc. (N.S.) **44** (2007), no. 1, 87–132 (electronic).

[10] J. Lewis, *Spaces of holomorphic functions equivalent to the even Maass cusp forms*, Invent. Math. **127** (1997), 271–306.

[11] J. Lewis and D. Zagier, *Period functions and the Selberg zeta function for the modular group*, The mathematical beauty of physics: A memorial volume for Claude Itzykson. Conference, Saclay, France, June 5–7, 1996, Singapore: World Scientific, 1997, pp. 83–97.

[12] ———, *Period functions for Maass wave forms. I*, Ann. Math. (2) **153** (2001), no. 1, 191–258.

[13] D. Mayer, *On the thermodynamic formalism for the Gauss map*, Comm. Math. Phys. **130** (1990), no. 2, 311–333.

[14] ———, *The thermodynamic formalism approach to Selberg's zeta function for* $\mathrm{PSL}(2, \mathbf{Z})$, Bull. Amer. Math. Soc. (N.S.) **25** (1991), no. 1, 55–60.

[15] D. Mayer, T. Mühlenbruch, and F. Strömberg, *The transfer operator for the Hecke triangle groups*, Discrete Contin. Dyn. Syst. **32** (2012), no. 7, 2453–2484.

[16] M. Möller and A. Pohl, *Period functions for Hecke triangle groups, and the Selberg zeta function as a Fredholm determinant*, Ergodic Theory Dynam. Systems **33** (2013), no. 1, 247–283.

[17] A. Pohl, *A dynamical approach to Maass cusp forms*, J. Mod. Dyn. **6** (2012), no. 4, 563–596.

[18] ———, *Period functions for Maass cusp forms for $\Gamma_0(p)$: A transfer operator approach*, Int. Math. Res. Not. **14** (2013), 3250–3273.
[19] ———, *Symbolic dynamics for the geodesic flow on two-dimensional hyperbolic good orbifolds*, Discrete Contin. Dyn. Syst., Ser. A **34** (2014), no. 5, 2173–2241.
[20] I. Richards, *Continued fractions without tears*, Math. Mag. **54** (1981), no. 4, 163–171.
[21] C. Series, *The modular surface and continued fractions*, J. London Math. Soc. (2) **31** (1985), no. 1, 69–80.
[22] A. Terras, *Harmonic analysis on symmetric spaces and applications. I*, New York etc.: Springer-Verlag, 1985.

AP: University of Bremen, Department 3 – Mathematics, Bibliothekstr. 5, 28359 Bremen, Germany

E-mail address: apohl@uni-bremen.de

DZ: Max Planck Institute for Mathematics, Vivatsgasse 7, 53111 Bonn, Germany and International Centre for Theoretical Physics, Strada Costiera, Trieste, Italy

E-mail address: dbz@mpim-bonn.mpg.de