Spectral theory of hyperbolic surfaces

Claire Burrin
(Spring 2021, ETH Zürich)

Contents

Chapter 1. Elements of Riemannian geometry 5
1.1. Euclid's postulates 5
1.2. Intrinsic geometry of curved surfaces 6
1.3. The Theorema Egregium 8
1.4. Riemannian manifolds 9
1.5. Topological classification of closed surfaces and Gauss–Bonnet 11
1.6. Geodesics and isometries 12

Chapter 2. Spectral theory of $\mathbf{T}^n$ 15
2.1. Some connections with arithmetic and geometry 16
2.2. Interlude: elements of functional and Fourier analysis 17
2.3. Poisson summation formula 19

Chapter 3. Geometry of the hyperbolic plane 23
3.1. Length, angles, area, curvature 24
3.2. Isometries 25
3.3. Geodesics 27
3.4. Classification of motions 28
3.5. Metric space properties 30

Chapter 4. Fuchsian groups 31
4.1. Interlude: Poincaré and Fuchsian functions 31
4.2. Fuchsian groups and hyperbolic surfaces: uniformization 32
4.3. Fundamental domains 33
4.4. Some arithmetic examples of Fuchsian groups 35
4.5. Geometric and algebraic features of fundamental domains 36
4.6. Some geometric examples of Fuchsian groups 37

Chapter 5. The spectral theorem for compact hyperbolic surfaces 39
5.1. Spectral theorem for compact operators 40
5.2. Hilbert–Schmidt operators 40
5.3. Point-pair invariants 41
5.4. Spherical functions 43
5.5. The heat equation 45

Chapter 6. Selberg's trace formula 51
6.1. The trace formula for compact hyperbolic surfaces 51
6.2. Length spectrum and Huber's theorem 53

6.3. Weyl's law 55

Chapter 7. Interlude: the hyperbolic circle problem 57

Chapter 8. On λ_1 61
8.1. Arbitrarily small spectral gap 61
8.2. Selberg's eigenvalue conjecture 64
8.3. Relation to expander graphs 65

Bibliography 67

CHAPTER 1

Elements of Riemannian geometry

This chapter contains an overview of some notions of non-Euclidean and Riemannian geometry. This will serve as an excuse to contextualise, mathematically and historically, the main object of this course: hyperbolic surfaces.

1.1. Euclid's postulates

In the first volume of *The Elements*, Euclid (300 BC) takes an inventory of planar geometry starting with a list of definitions — what is... a point, a line, parallel lines, a circle, etc — and the following five postulates, based on the use of ruler and compass:

(1) A straight line segment can be drawn joining any two distinct points.
(2) Any straight line segment can be extended indefinitely in a straight line.
(3) Given any straight line segment, a circle can be drawn having the segment as radius and one endpoint as center.
(4) All right angles are congruent.
(5) If two lines are drawn which intersect a third in such a way that the sum of the inner angles on one side is less than two right angles, then the two lines inevitably must intersect each other on that side if extended far enough.

The last postulate is equivalent to what is known as the Parallel Postulate:

(5') (Parallel Postulate) Given a line L and a point p, there exists exactly one line parallel to L that passes through p.

Unlike the first four, the fifth Postulate looks like a proposition one could actually prove, and many prominent thinkers across the centuries tried yet failed. So what happens when we omit the parallel postulate? (5') indicates that the postulate can fail in one of two ways; either L has no parallel through p or it has more than one. On the other hand, it is perhaps easier to see from the formulation of (5) that we are in effect making an assumption on the curvature of the underlying space: the plane is "flat". Here are two model-examples in which (5') indeed fails.

Example 1 (Absence of parallels)**.** *Consider the unit sphere*

$$\mathbf{S}^2 = \{(x, y, z) \in \mathbf{R}^3 : x^2 + y^2 + z^2 = 1\}.$$

We define a line on the sphere to be a great circle. The existence of parallels is impossible in this model, yet one can develop a logically consistent geometry on the sphere from the four other postulates, different from our usual Euclidean geometry. In this model, the sum of the inner angles of a triangle is $> \pi$.

Example 2 (Infinitely many parallels)**.** *Consider the upper half of the complex plane,*

$$\mathbf{H} = \{z = x + iy \in \mathbf{C} : y > 0\}.$$

In this model, we define lines to be either vertical half-lines or semicircles orthogonal to the x-axis. Then each hyperbolic line has infinitely many parallels and the sum of the inner angles of a triangle is $< \pi$.

Lobachevsky, Bolyai, and Gauss were the first to understand, independently, that such non-Euclidean geometries were logically consistent, even if they feel far removed from our intuitive perception of geometric objects. It is only with the development of the geometry of curved spaces, and Riemannian geometry in particular, that non-Euclidean geometry was accepted.

1.2. Intrinsic geometry of curved surfaces

The introduction of Cartesian coordinates in the 17th century allowed to describe geometric shapes with algebraic equations. Then with the development of differential and integral calculus, one could compute various geometric data (such as length and area) that allowed to move from the study of the classical solids to curved surfaces in space. We take as starting point a definition that is more general than that of a parametric surface.

Definition 3. *A connected Hausdorff space $M \subset \mathbf{R}^3$ is a* **smooth surface in $\mathbf{R}^3$** *if each point has a neighborhood that is diffeomorphic to an open subset of $\mathbf{R}^2$.*

Let M be a smooth surface in $\mathbf{R}^3$. Let $U \subset M$ be a neighborhood of $p \in M$ that is diffeomorphic via $f : U \to V$ to a subset $V \subset \mathbf{R}^2$. The inverse $F = f^{-1} : V \to M$ is called a **smooth local parametrization near** p. There is a simple criterium to check whether a smooth injective map $F : V \to M$ defines a local parametrization.

Lemma 4. *Let $V \subset \mathbf{R}^2$ be an open subset. A smooth injective map $F : V \to M$ is a smooth local parametrization if and only if $\partial_x F$ and $\partial_y F$ are linearly independent at each point of V.*

PROOF. Follows by the implicit function theorem. □

Definition 5. *Let F be a smooth local parametrization of M near p. The* **tangent space to M at** *p is the linear space*

$$T_pM = \operatorname{span}\{\partial_x F|_p, \partial_y F|_p\} \cong \mathbf{R}^2,$$

Lemma 6. *The linear space T_pM is independent of the choice of local parametrization.*

PROOF. Let F_1, F_2 be two local parametrizations at p. Show that $DF_2^{-1} \circ DF_1$ is a linear isomorphism. □

Let γ be a parametrized smooth curve on M. We write γ locally as $\gamma = F(\gamma_{\text{loc}}(t))$, where $\gamma_{\text{loc}}(t) = (x(t), y(t)) \in \mathbf{R}^2$. Then

$$\gamma'(t) = \frac{d}{dt}(F \circ \gamma_{\text{loc}})(t) = DF \cdot \gamma'_{\text{loc}}(t)$$

is a vector in $T_{\gamma(t)}M$ and DF is the 2×2 matrix

$$DF = D_{\gamma_{\mathrm{loc}}(t)}F = \begin{pmatrix} \partial_x F|_{\gamma_{\mathrm{loc}}(t)} & \partial_y F|_{\gamma_{\mathrm{loc}}(t)} \end{pmatrix}.$$

Proposition 7. *Given a smooth parametrized curve $\gamma : [a,b] \to M$, its length*

$$L(\gamma) = \int_a^b \|\gamma'(t)\| dt,$$

where $\|x\|^2 = x_1^2 + x_2^2 + x_3^2$, does not depend on the particular choice of parametrization.

PROOF. Consider the reparametrization $\psi : [c,d] \to [a,b]$ with ψ smooth (strictly) monotone and show that $L(\gamma \circ \psi) = L(\gamma)$. □

We now compute the length of γ. We write $\gamma_{\mathrm{loc}}'(t) = (\frac{dx}{dt}(t), \frac{dy}{dt}(t)) = (dx, dy)$. Then

$$\begin{aligned} \|\gamma'(t)\| &= \sqrt{\gamma'(t) \cdot \gamma'(t)} \\ &= \sqrt{(dx, dy)(DF^T DF)\begin{pmatrix} dx \\ dy \end{pmatrix}} \\ &= \sqrt{\|\partial_x F\|^2 dx^2 + 2\partial_x F \cdot \partial_y F dxdy + \|\partial_y F\|^2 dy^2}. \end{aligned}$$

(Beware that our shorthand notation masks the fact that each single term depends on t.) The quantity

$$ds = \sqrt{\|\partial_x F\|^2 dx^2 + 2(\partial_x F \cdot \partial_y F) dxdy + \|\partial_y F\|^2 dy^2}$$

is called the **line element**, so that

$$L(\gamma) = \int_\gamma ds = \int_a^b \|\gamma'(t)\| dt.$$

Its square

$$ds^2 = \|\partial_x F\|^2 dx^2 + 2(\partial_x F \cdot \partial_y F) dxdy + \|\partial_y F\|^2 dy^2$$

is called the **local first fundamental form** and it is completely determined by the symmetric matrix

$$DF^T DF = \begin{pmatrix} \partial_x F \cdot \partial_x F & \partial_x F \cdot \partial_y F \\ \partial_y F \cdot \partial_x F & \partial_y F \cdot \partial_y F \end{pmatrix},$$

which is called the **metric tensor**.

Exercise 8. *Given two local parametrizations F_1 and F_2 with metric tensors A_1 and A_2, let $\phi = F_2^{-1} \circ F_1$ and check that $A_1 = D\phi^T A_2 D\phi$.*

Definition 9. *Let M_1 and M_2 be two smooth surfaces in $\mathbf{R}^3$. A diffeomorphism $\phi : M_1 \to M_2$ is called a* **(local) isometry** *if it (locally) preserves the length of all curves. That is, if for each smooth curve $\gamma : [a,b] \to V \subset M_1$, $L(\phi \circ \gamma) = L(\gamma)$.*

We say that a geometric property is **intrinsic** if it is invariant under local isometries. Colloquially, we understand an isometry to be a transformation that may bend the surface but not stretch it. A geometric property of a smooth surface is intrinsic if it does not depend on the way the surface is embedded in $\mathbf{R}^3$.

THEOREM 1. *Two smooth surfaces M_1 and M_2 are locally isometric near $p_1 \in M_1$, $p_2 \in M_2$ if and only if there are local parametrizations F_1, F_2 near p_1 and p_2 respectively that yield the same local fundamental form.*

PROOF. The only if direction is immediate from the local computation of length. Conversely, let ϕ be a local isometry $M_1 \to M_2$ such that $\phi(p_1) = p_2$. Let F_1 be a local parametrization near p. Then $F_2 = \phi \circ F_1$ is a smooth local parametrization near $\phi(p_1) = p_2$. Since ϕ is an isometry, $L(F_2 \circ \gamma_{\text{loc}}) = L(F_1 \circ \gamma_{\text{loc}})$. To conclude, we show that length determines the first fundamental form locally. Consider the parametrized curve γ, locally given by $\gamma = F \circ \gamma_{\text{loc}}(t)$, where $\gamma_{\text{loc}}(t) = p + \left(\begin{smallmatrix} t \\ 0 \end{smallmatrix}\right)$. Then $\gamma'_{\text{loc}}(t) = e_1$ and

$$\frac{d}{d\varepsilon} L(\gamma|_{[0,\varepsilon]}) = \frac{d}{d\varepsilon} \int_0^{\varepsilon} \|\partial_x F\| dt = \|\partial_x F|_{p+\varepsilon e_1}\|.$$

Taking $\varepsilon = 0$, we conclude that length locally determines $\|\partial_x F|_p\|$. Similarly, using $\gamma_{\text{loc}}(t) = p + \left(\begin{smallmatrix} 0 \\ t \end{smallmatrix}\right)$ and $\gamma_{\text{loc}}(t) = p + t\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$, one determines $\|\partial_y F|_p\|$ and $\partial_x F|_p \cdot \partial_y F|_p$ respectively. □

1.2.1. Angles and areas. If γ_1 and γ_2 are two smooth (oriented) curves on M that intersect at a point p, then we measure their angle of intersection as the angle between their tangent vectors $u = \gamma_1'$, $v = \gamma_2'$ are p, and

$$\cos\theta = \frac{u \cdot v}{\|u\|\|v\|}.$$

Consider a region near a point p locally parametrized by F. We can approximate the region by infinitesimal parallelograms with edges $(\partial_x F)dx$, $(\partial_y F)dy$, where dx, dy are thought of as infinitesimal increments. Each parallelogram has area

$$(\|\partial_x F\|dx)(\|\partial_y F\|dy)|\sin\theta| = \|\partial_x F \times \partial_y F\|dxdy = \sqrt{\det A}dxdy.$$

The area of a region $\Omega = F(U)$ is then given by

$$\text{Area}(\Omega) = \int_U \sqrt{\det(A)}dxdy.$$

One can check that this definition is independent of the choice of parametrization.

1.3. The Theorema Egregium

Recall that the curvature of a differentiable curve at p is the curvature of the circle that best approximates the curve near p. So $\kappa = R^{-1}$, where R is the radius of the osculating circle. There are other, equivalent and more practical, definitions in terms of the parametrization of the curve; e.g., curvature is defined as the magnitude of the acceleration of a particle moving along the curve at unit speed. The curvature of a surface is necessarily a more complicated concept since a surface can curve differently in different directions. We quickly describe (informally) three notions of curvature: normal, principal, and Gaussian.

Normal curvature is the (signed) curvature of the curve (with respect to the chosen unit normal) obtained by intersecting the plane passing through the tangent and normal directions with the surface. All curves on a surface with the same tangent vector at p have the same normal curvature.

Considering all tangent vectors at p, the maximum κ_1 and minimum κ_2 values of the normal curvature at p are called the **principal curvatures**, and their directions (indicating at each point where the surface will curve the more and the least) are called the **principal directions**.

Aside 10. *Remarkably, principal directions always correspond to orthogonal planes (Euler, 1760). From a modern perspective, this is seen as an application of the spectral theorem for finite-dimensional inner product spaces over* $\mathbf{R}$*. The argument goes as follows. Let* M *be a smooth surface in* $\mathbf{R}^3$*. The shape operator (also called Weingarten map, or second fundamental tensor) if the linear operator* $S : T_pf \to T_pf$ *given by* $S(v) = D_v\nu$*, where* D_v *is the directional derivative, and* ν *is the Gauss map. It is selfadjoint with respect to the standard inner product on* T_pf*, and as such it is diagonalizable by an orthonormal basis of eigenvectors. These eigenvectors are precisely the principal directions and the associated eigenvalues are the principal curvatures.*

The **Gaussian curvature** of a smooth surface at a point p is the product of the principal curvatures at this point, i.e., $K = \kappa_1\kappa_2$. Gaussian curvature can be seen to determine the local shape of the surface. E.g., if $K > 0$, the surface is locally convex, while if $K < 0$, the surface is locally saddle-shaped.

Theorem 2 (Gauss' Theorema Egregium (1827)). *On a smooth surface in* $\mathbf{R}^3$*, the Gaussian curvature at any point is completely determined by the first fundamental form. In particular, two smooth surfaces in* $\mathbf{R}^3$ *that are locally isometric have the same Gaussian curvature at each point.*

1.4. Riemannian manifolds

Is there a more general theory in which higher dimensional geometric shape can be studied intrinsically? Riemannian geometry is Riemann's answer to Gauss' question of the existence of a higher-dimensional framework of the intrinsic geometry of curved surfaces. Riemann (who was a student of Gauss) proposed what is our modern framework in his Habilitation thesis. His proposition builds on three ingredients:

(1) the notion of a smooth n-dimensional manifold;
(2) the introduction of a Riemannian metric — Riemann's proposition being that this is all that suffices to study the intrinsic geometry of a smooth manifold;
(3) non-Euclidean geometry.

Definition 11. *A* **topological** n**-dimensional manifold** *is a second countable connected Hausdorff space* M *for which each point* $p \in M$ *has an open neighborhood* $U \subset M$ *that is homeomorphic to an open subset* $V \subset \mathbf{R}^n$*. This homeomorphism* $\phi : U \to V$ *is called a (coordinate) chart. An atlas is a family* $\{(U_\alpha, \phi_\alpha)\}_\alpha$ *of charts covering* M*.*

Remark 12. *The condition that* M *be second countable guarantees the existence of a countable atlas.*

Definition 13. *A smooth atlas on a manifold* M *is an atlas such that whenever* $U_\alpha \cap U_\beta \neq \emptyset$*, the transition map*

$$\phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$$

is a diffeomorphism. We say that M is a smooth manifold if it admits a maximal smooth atlas.

Definition 14. *A* **smooth n-dimensional manifold** *is a topological n-dimensional manifold M where each point has an open neighborhood that is homeomorphic to an open subset of $\mathbf{R}^n$ equipped with a maximal smooth atlas.*

Our former definition — Definition 3 — of a smooth surface was limiting: it forced us to consider only surfaces sitting smoothly in $\mathbf{R}^3$. This rules out, for example, the Klein bottle, which is smooth but can not be embedded in $\mathbf{R}^3$ without self-intersections.[1] There are also other benefits to working with abstract surfaces. Take the example of the torus; it can be seen as a surface of revolution in $\mathbf{R}^3$ — obtained as a surface of revolution by rotating the circle $x^2 + z^2 = b^2$ with center $(a, 0, 0)$ about the z-axis —, parametrized by

$$T^2 = \{((a + b\cos\psi)\cos\theta, (a + b\cos\psi)\sin\theta, b\sin\psi) : \theta, \psi \in [0, 2\pi]\}$$

(with $a > b > 0$), but it is often more convenient to look at it abstractly as the quotient

$$\mathbf{T}^2 = \mathbf{R}^2/\mathbf{Z}^2 = \{(x \pmod 1, y \pmod 1) : x, y \in \mathbf{R}\}$$

(which clearly does not sit in $\mathbf{R}^3$). Under this identification, two points in $\mathbf{R}^2$ are equivalent if and only if they differ by a $\mathbf{Z}^2$-translation; i.e., $(x, y) \sim (x + m, y + n)$ for all $m, n \in \mathbf{Z}$. A smooth function $f : \mathbf{T}^2 \to \mathbf{R}$ is now simply a smooth function $f : \mathbf{R}^2 \to \mathbf{R}$ that is invariant under translations by $\mathbf{Z}^2$.

Definition 15. *Let M be a smooth manifold. A* **Riemannian metric** *on M is a family $(g_p)_{p\in M}$ of inner products $g_p : T_pM \times T_pM \to \mathbf{R}$ that vary differentiably in p. A* **Riemannian manifold** (M, g) *is a smooth manifold equipped with a Riemannian metric.*

This suffices to compute the length of a smooth (in fact, even piecewise C^1) curve $\gamma : [a, b] \to M$ via

$$L(\gamma) = \int_a^b \sqrt{g_{\gamma(t)}(\gamma'(t), \gamma'(t))}dt$$

and thus also area and curvature. A Riemannian metric $(g_p)_{p\in M}$ on M is usually described by the local expression

$$ds^2 = \sum_{i,j} g_{ij} dx_i dx_j.$$

Theorem 3. *Every smooth manifold admits a Riemannian metric.*

Proof. This is a nice (standard) application of partition of unity building on the canonical Euclidean metric on $\mathbf{R}^n$. □

[1] On the other hand, Whitney's embedding theorem tells us that any smooth n-dimensional manifold can be smoothly embedded in $\mathbf{R}^{2n}$, and so one could in principle study smooth surfaces embedded in $\mathbf{R}^4$...

Example 16. *The upper half-plane* **H** *(cf. Example 2) equipped with the Riemannian metric* $g_z(u,v) = \frac{u \cdot v}{y^2}$ *— equivalently* $ds^2 = y^{-2}(dx^2 + dy^2)$ *— has constant Gaussian curvature* $K = -1$*. As a consequence of the following famous result of Hilbert (1901),* **H** *cannot be isometrically embedded in* $\mathbf{R}^3$*.*

THEOREM 4. *There is no complete regular surface of constant negative curvature that can be immersed in* $\mathbf{R}^3$*.*

We will admit the following result.

THEOREM 5. *Each smooth surface equipped with a Riemannian metric of constant Gaussian curvature* $K = 0, 1, -1$ *is locally isometric to* $\mathbf{R}^2$*,* $\mathbf{S}^2$*, and* $\mathbf{H}^2$ *respectively.*

Definition 17. *A* **hyperbolic surface** *is a smooth surface equipped with a Riemannian metric of constant Gaussian curvature* $K = -1$*.*

— End of class #1 —

1.5. Topological classification of closed surfaces and Gauss–Bonnet

In this section, we give a topological description of closed hyperbolic surfaces, based on the Gauss–Bonnet theorem, which relates the geometry and topology of Riemannian manifolds. In the setting we are in, this is

THEOREM 6 (Gauss–Bonnet). *Let* M *be a compact Riemannian surface. Then*

$$\int_M K dA = 2\pi\chi(M),$$

where $\chi(M)$ *is the Euler characteristic of* M*.*

Since for a hyperbolic surface we have $K = -1$, the above identity implies that

$$\text{area}(M) = -2\pi\chi(M).$$

Recall that if M and M' are homeomorphic, then $\chi(M) = \chi(M')$. Hence the area of a compact hyperbolic surface is invariant under homeomorphism.

Consider M as a topological surface. In topology, one constructs surfaces by gluing. The sphere amounts to glueing adjacent edges of a square, the torus to gluing opposite edges of a square, the cylinder to gluing only one pair of opposite edges, and the Möbius band by gluing one pair of opposite edges as well but in opposite directions. The latter two examples have boundary, and the Möbius band is the prototypical nonorientable surface. Given two topological surfaces M and N, their connected sum $M\#N$ is obtained by removing a small disk of each and glueing them along the boundaries. These realizations allow to compute the Euler characteristic $\chi = V - E + F$ easily. The topological classification theorem for compact surfaces states that the above constructions are the only topologically distinct constructions possible, and that thus compact orientable surfaces are classified by their genus, i.e., their number of handles.

THEOREM 7 (Topological classification of compact surfaces). *Each compact orientable topological surface is homeomorphic to either the sphere, the torus or a connected sum* Σ_g *(with* $g \geq 2$*) of tori.*

Exercise 18. *Check that* $\chi(\Sigma_g) = 2 - 2g$*.*

1.6. Geodesics and isometries

Fact 19. *Let M be a Riemannian manifold. The map*

$$d : M \times M \to \mathbf{R}_{\geq 0}, \quad d(p,q) = \inf_{\gamma} L(\gamma),$$

taken over all smooth curves in M that join p to q, is a distance function on M. Its associated topology is the topology of M.

Definition 20. *Let (X,d) be a metric space, and let $\gamma : [a,b] \to X$ be a curve. Its image in X is called a* **(metric) geodesic** *if there is a constant $\lambda > 0$ such that $d(\gamma(t), \gamma(t+\varepsilon)) = \lambda \cdot \varepsilon$ for each small $\varepsilon > 0$.*

Geodesics on Riemannian manifolds are usually defined as the solution (via calculus of variation) of a minimization problem, and the following equivalence is an implication of this definition. This equivalence shows that the notions of geodesic in metric geometry and Riemannian geometry are consistent.

Definition 21. *Let $\gamma : [a,b] \to M$ be a curve that is proportionally parametrized by arclength (i.e., $L(\gamma) = \lambda|b-a|$ for some $\lambda > 0$). Its image in M is a* **(Riemannian) geodesic** *if and only if*

$$L(\gamma|_{[t,t+\varepsilon]}) = d(\gamma(t), \gamma(t+\varepsilon))$$

for each small $\varepsilon > 0$.

Examples 22.

- *Let $M = \mathbf{R}^n$, and $x \in \mathbf{R}^n$. The tangent space T_xM can be defined as the set of all vectors $\gamma'(0)$, where γ is a smooth curve on M with $\gamma(0) = x$. Hence $T_x\mathbf{R}^n \cong \mathbf{R}^n$. The standard metric corresponds to the usual dot product $g_p : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$, $(u,v) \mapsto u \cdot v$, and we recover the usual length of curves in $\mathbf{R}^3$, and $ds^2 = dx^2 + dy^2$. The geodesics of this metric are the straight lines in $\mathbf{R}^n$.*
- *Let $M = \mathbf{S}^n$, and $p \in \mathbf{S}^n$. If γ is a curve on $\mathbf{S}^n$ with $\gamma(0) = p$, then $\gamma(t)\cdot\gamma(t) = 1$, and taking $\frac{d}{dt}\big|_{t=0}$, we find $\gamma'(0)\cdot p = 0$. Hence $T_p\mathbf{S}^n$ is the hyperplane tangent to p, and $T_p\mathbf{S}^n \cong \mathbf{R}^n$. The usual dot product in $\mathbf{R}^n$ induces the standard round metric on $\mathbf{S}^2$ and geodesics are segments of great circles.*

Observe at the hand of this last example that geodesics are **locally** distance minimizing and not globally: given two close by points p and q on a sphere, there is a unique great circle passing through them, giving two great circle segments from p to q. These are both geodesics, although only one is length minimizing.

Definition 23. *Let (M,g) and (N,h) be Riemannian manifolds. A diffeomorphism $f : M \to N$ is a* **(Riemannian) isometry** *if it preserves the lengths of all curves.*

Locally, there exists a function $\widetilde{f}$ such that $f = \phi_\beta^{-1} \circ \widetilde{f} \circ \phi_\alpha$. We say that f is a diffeomorphism at p if $\widetilde{f}$ is a diffeomorphism at $\phi_\alpha(p)$, i.e., $\widetilde{f}^{-1}$ exists and both $\widetilde{f}$, $\widetilde{f}^{-1}$ are C^∞. One can check that this is independent of the particular choice of coordinate charts. If $f : M \to N$ is an isometry, then for each curve γ on M,

$$L(f \circ \gamma) = \int_a^b \sqrt{h_{f(\gamma(t))}(Df \cdot \gamma'(t), Df \cdot \gamma'(t))}dt = \int_a^b \sqrt{g_{\gamma(t)}(\gamma'(t), \gamma'(t))}dt = L(\gamma).$$

In other words, a diffeomorphism f is an isometry of Riemannian manifold if and only it preserves the metrics. As follows from the definition in metric geometry given below, Riemannian isometries are moreover metric isometries.

Definition 24. *Let (X, d_X) and (Y, d_Y) be metric spaces. A bijection $f : X \to Y$ is a* **(metric) isometry** *if it preserves distances, i.e., if $d_Y(f(x), f(y)) = d_X(x, y)$ for all pairs $(x, y) \in X \times X$.*

We can immediately conclude that (Riemannian/metric) isometries preserve (Riemannian/metric) geodesics. The set $\mathrm{Isom}(M, g)$ of all isometries $f : M \to M$ is a group[2] with respect to the composition of functions.

Examples 25.

- $\mathrm{Isom}(\mathbf{R}^n) \cong \mathrm{O}(n) \ltimes \mathbf{R}^n$
- $\mathrm{Isom}(\mathbf{S}^n) \cong \mathrm{O}(n+1)$
- $\mathrm{Isom}(\mathbf{H}^n) \cong \mathrm{PO}(n, 1)$

For instance: Observe that an isometry of $\mathbf{R}^n$ is in $O(n)$ if and only if $f(0) = 0$. Indeed, by definition, isometries preserve the dot product (hence are orthogonal) and preserve lines in $\mathbf{R}^n$. Hence, f preserves all lines passing through the origin if and only if $f(0) = 0$. Up to a translation, every isometry can be made to fix the origin. This shows that isometries of $\mathbf{R}^n$ are generated by elements of $O(n)$ and translations by vectors of $\mathbf{R}^n$.

The Laplacian on $\mathbf{R}^n$ is given by

$$\Delta = \mathrm{div} \circ \mathrm{grad} = \nabla \cdot \nabla = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}.$$

The divergence, gradient, and Laplacian have analogues on a Riemannian manifold (M, g), expressed in local coordinates by

$$\nabla_g = g^{ij} \frac{\partial}{\partial x_j}$$

and

$$\Delta_g = \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x_i} \left(\sqrt{|g|} g^{ij} \frac{\partial}{\partial x_j} \right).$$

The proof of the following proposition reduces to a computation in local coordinates.

Fact 26. *Let (M, g), (N, h) be Riemannian manifolds. A diffeomorphism $f : M \to N$ is an isometry if and only for all $\phi \in C_c^\infty(N)$,*

$$\Delta_g(\phi \circ f) = \Delta_h \phi \circ f.$$

[2]In fact, by a theorem of Myers and Steenrod, it is a Lie group.

CHAPTER 2

Spectral theory of $\mathbf{T}^n$

Let again $\mathbf{T}^n = \mathbf{R}^n/\mathbf{Z}^n = \{x + \mathbf{Z}^n : x \in \mathbf{R}^n\}$. Each $x \in \mathbf{T}^n$ has a unique representative in the hypercube $[0,1)^n$. We view functions on $\mathbf{T}^n$ as functions on $\mathbf{R}^n$ that are invariant under $\mathbf{Z}^n$-translations, i.e., $f(x+\xi) = f(x)$ for all $x \in \mathbf{R}^n$ and $\xi \in \mathbf{Z}^n$. The standard Euclidean metric on $\mathbf{R}^n$ induces a metric on $\mathbf{T}^n$ by letting the projection $\pi : \mathbf{R}^n \to \mathbf{R}^n/\mathbf{Z}^n$ be a local isometry. We can view $\mathbf{T}^n$ a compact metric space for the distance function

$$d(x + \mathbf{Z}^n, y + \mathbf{Z}^n) = \min_{\xi \in \mathbf{Z}^n} \|x - y + \xi\|,$$

where $\|\cdot\|$ is the standard norm in $\mathbf{R}^n$. The (geometric) Laplacian on $\mathbf{R}^n$

$$\Delta = -\left(\frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}\right)$$

commutes with $\mathbf{Z}^n$-translations (which are isometries of $\mathbf{R}^n$) and descends to $C^\infty(\mathbf{T}^n)$.

Remark 27. *The minus sign guarantees that the eigenvalues of Δ are nonnegative, a convenient feature for the spectral theory.*

It is easy to check that the Laplacian commutes with translations and as such it descends to $\mathbf{T}^n$, and acts linearly on a (dense) domain[1] of the Hilbert space $L^2(\mathbf{T}^n)$. Solutions in $L^2(\mathbf{T}^n)$ to the spectral problem $\Delta\varphi = \lambda\varphi$ are easy to come by. We use the shorthand notation $e(x) := e^{2\pi i x}$. For each $\xi \in \mathbf{Z}^n$,

$$\varphi_\xi(x) = e(x \cdot \xi)$$

satisfies

$$\Delta\varphi_\xi = 4\pi\|\xi\|^2\varphi_\xi,$$

and $\{\varphi_\xi\}_{\xi\in\mathbf{Z}^n}$ is an orthonormal family in $L^2(\mathbf{T}^n)$, as indeed,

$$\langle\varphi_\xi, \varphi_\eta\rangle = \int_{\mathbf{T}} e(x_1(\xi_1 - \eta_1))dx_1 \cdots \int_{\mathbf{T}} e(x_n(\xi_n - \eta_n))dx_n = \begin{cases} 1 & \text{if } \xi = \eta, \\ 0 & \text{otherwise.} \end{cases}$$

An application of the (complex version of the) Stone–Weierstrass theorem implies that this family is dense in $L^2(\mathbf{T}^n)$; we have found all the eigenfunctions of Δ.

Fourier analysis provides us with the spectral decomposition of $L^2(\mathbf{T}^n) = \oplus E_{\lambda_i}$; each $f \in C^\infty(\mathbf{T}^n)$ admits the Fourier series expansion

$$f(x) = \sum_{\xi\in\mathbf{Z}^n} \widehat{f}(\xi)\varphi_\xi(x), \tag{2.1}$$

[1]The space of smooth functions $C^\infty(\mathbf{T}^2)$ is dense in $L^2(\mathbf{T}^n)$ with respect to the compact-open topology.

where

$$\widehat{f}(\xi) = \langle f, \varphi_\xi \rangle = \int_{\mathbf{T}^n} f(x)\overline{\varphi_\xi(x)}dx.$$

The uniform convergence of (2.1) can be deduced from the rapid decay of Fourier coefficients:

Proposition 28. *For each $\xi \in \mathbf{Z}^n$ and $k \geq 1$, there exists a constant $C_{k,f} > 0$ such that*

$$|\widehat{f}(\xi)| \leq C_{k,f} \|\xi\|_\infty^{-k}.$$

Proof. Let $|\xi_i| = \|\xi\|_\infty$. By integration by parts,

$$\widehat{f}(\xi) = \iint_0^1 \left(\int_0^1 f(x)\overline{\varphi_\xi(x)}dx_i \right) dx_1 \cdots dx_{i-1}dx_{i+1} \cdots dx_n = \frac{1}{(2\pi|\xi_i|)^k} \left\langle \frac{\partial^k}{\partial x_i^k} f, \varphi_\xi \right\rangle.$$

□

2.1. Some connections with arithmetic and geometry

We restrict to $n = 2$ for these illustrations. Observe that each eigenvalue $4\pi^2 n$ of the Laplacian in the previous discussion appears with multiplicity

$$r_2(n) := \#\{\xi \in \mathbf{Z}^2 : \|\xi\|^2 = n\} = \#\{(a, b) \in \mathbf{Z}^2 : a^2 + b^2 = n\},$$

which is precisely the number of representations of $n \in \mathbf{N}$ as a sum of two squares.

Aside 29. *Every odd number is either of the form $4k+1$ or of the form $4k+3$. It was observed early on that all small primes of the form $4k+1$ (at least as far as one could compute then) could be written as the sum of two squares, while that was not the case of primes of the form $4k+3$. (The latter observation is not so surprising; it is an easy exercise to show that no odd number of the form $4k+3$ can be written as a sum of two squares.) Fermat proved that in fact, every prime $p \equiv 1$ (mod 4) can be written as the sum of two squares. Lagrange later showed that every positive integer can be written as the sum of four squares.*

A precise formula for $r_2(n)$ was given by Gauss;

$$r_2(n) = 4\sum_{d|n} \chi(d), \qquad \text{where } \chi(d) = \begin{cases} 1 & d \equiv 1 \ (4), \\ -1 & d \equiv 1 \ (4), \\ 0 & \text{otherwise}, \end{cases}$$

where the sum is taken over all positive divisors d of n. The function $r_2(n)$ fluctuates irregularly, but is trivially bounded by the divisor function $d(n)$ for which we know that for any $\varepsilon > 0$, $d(n) \ll_\varepsilon n^\varepsilon$[2]. Hence this tells us that the dimension of the eigenspace for $4\pi^2 n$ has multiplicity $\ll_\varepsilon n^\varepsilon$.

Geodesics on $\mathbf{T}^2$ are the projections on $\mathbf{T}^2$ of lines in $\mathbf{R}^2$ (since geodesics are preserved by isometries, and hence by translations by $\mathbf{Z}^2$). Explicitly, each geodesic on $\mathbf{T}^2$ can be parametrized as $t \in \mathbf{R} \mapsto x + ty + \mathbf{Z}^2$ for some fixed $x, y \in \mathbf{R}^n$.

[2]The Vinogradov notation $\ll_\varepsilon$ means that there exist constants $C_1(\varepsilon)$ and $C_2(\varepsilon)$ such that when $n \geq C_1(\varepsilon)$ then $d(n) \leq C_2(\varepsilon)n^\varepsilon$.

THEOREM 8. *Every geodesic on* $\mathbf{T}^2$ *is either periodic or becomes equidistributed on* $\mathbf{T}^2$, *i.e., for any* $f \in C^\infty(\mathbf{T}^2)$,

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T f(x+ty)dt = \int_{\mathbf{T}^2} f(x)dx.$$

In other words, either a geodesic trajectory is periodic, or it fills the whole torus uniformly.

Remark 30. *This is a special case of a more general phenomenon for certain flat surfaces, called the Veech dichotomy, and that can be used to study the dynamics of mathematical billiards.*

PROOF. Each geodesic on $\mathbf{T}^2$ lifts to a line in $\mathbf{R}^2$. Fix $L = \{x + ty : t \in \mathbf{R}\} \subset \mathbf{R}^2$, $x, y \in \mathbf{R}^2$. Assume first that the slope of $y = (y_1, y_2)$ is rational. That is, $\frac{y_2}{y_1} = \frac{p}{q}$, where $p, q \in \mathbf{Z}$, $q > 0$, $(p, q) = 1$. We may assume that $y_1 > 0$. Then choosing $t = \frac{q}{y_1}$,

$$x + ty = x + t\begin{pmatrix} q \\ p \end{pmatrix} \equiv x \pmod{\mathbf{Z}^2}.$$

Suppose now that the slope of y is not rational. Then $\xi \cdot y = 0$ ($\xi \in \mathbf{Z}^2$) if and only if $\xi = 0$. Taking the Fourier expansion of $f \in C^\infty(\mathbf{T}^2)$,

$$\frac{1}{T}\int_0^T f(x+ty)dt = \widehat{f}(0) + \sum_{\xi\neq 0} \widehat{f}(\xi)\varphi_\xi(x)\frac{1}{T}\int_0^T \varphi_\xi(ty)dt \tag{2.2}$$

$$= \int_{\mathbf{T}^2} f(x)dx + \sum_{\xi\neq 0} \widehat{f}(\xi)\varphi_\xi(x)\frac{(\varphi_\xi(Ty)-1)}{2\pi i T\xi\cdot y}. \tag{2.3}$$

To bound the sum, we use the fast decay of Fourier coefficients (see Proposition 28) and the Cauchy–Scwharz inequality. For each $k \geq 1$,

$$\begin{aligned}
\sum_{\xi\neq 0} \frac{|\widehat{f}(\xi)|}{|\xi\cdot y|} &= \sum_{\xi_1\neq 0,\xi_2} \frac{|\widehat{f}(\xi)|}{|\xi\cdot y|} + O(1) \\
&\leq \left(\sum_{\xi_1\neq 0,\xi_2} |\widehat{\partial_k f}(\xi)|^2\right)^{1/2} \left(\sum_{\xi_1\neq 0,\xi_2} \frac{1}{(2\pi\xi_1)^{2k}|\xi\cdot y|^2}\right)^{1/2} + O(1) \\
&= \|\partial_k f\|_2 \left(\sum_{\xi_1\neq 0,\xi_2} \frac{1}{(2\pi\xi_1)^{2k}|\xi\cdot y|^2}\right)^{1/2} + O(1) = O(1),
\end{aligned}$$

where we used Parseval identity. Dividing by T and letting $T \to \infty$ concludes. □

— End of class #2 —

2.2. Interlude: elements of functional and Fourier analysis

The Laplacian on $\mathbf{R}^n$ is well-defined on the dense subspace $C_c^\infty(\mathbf{R}^n) \subset L^2(\mathbf{R}^n)$, where $L^2(\mathbf{R}^n)$ is the Hilbert space of (classes of) square-integrable functions on $\mathbf{R}^n$. A linear operator T on a Hilbert space is called symmetric if $\langle Tf, g\rangle = \langle f, Tg\rangle$ for all f, g in its domain.

Proposition 31. *The geometric Laplacian on $C_c^\infty(\mathbf{R}^n)$ is linear, symmetric, and nonnegative. In particular, each eigenvalue is nonnegative.*

PROOF. Let Δ denote the standard Laplacian (i.e., without the negative sign). By integration by parts, for any $f, g \in C_c^\infty(\mathbf{R}^n)$,

$$\langle \Delta f, g\rangle = -\int_{\mathbf{R}^n} \nabla f(x)\overline{\nabla g(x)}dx = \langle f, \Delta g\rangle$$

and this implies immediately $\langle -\Delta f, f\rangle = \|\nabla f\|^2 \geq 0$. □

Remark 32. *We will now use Δ to denote the geometric Laplacian, hoping this raises no confusion.*

The following standard theorem of functional analysis guarantees that Δ extends to a linear selfadjoint operator on $L^2(\mathbf{R}^n)$.

THEOREM 9 (Friedrich). *A symmetric nonnegative operator on a dense subspace of a Hilbert space $\mathcal{H}$ admits a (unique) selfadjoint extension to $\mathcal{H}$.*

Exercise 33. *More generally, let T be a linear self-adjoint operator on a Hilbert space. Verify that its eigenvalues are real, and that eigenfunctions corresponding to distinct T-eigenvalues are orthogonal.*

However, the linear operator Δ is unbounded and so this does not suffice for a spectral theorem (— the very subject of functional analysis). From our construction of eigenfunctions on $\mathbf{T}^n$, we quickly realize that for each $\xi \in \mathbf{R}^n$, we also have $\Delta\varphi_\xi = 4\pi^2\|\xi\|^2\varphi_\xi$. One might note that $\varphi_\xi \notin L^2(\mathbf{R}^n)$, nonetheless, a spectral argument concludes that the spectrum of Δ is as large as possible, spanning the whole half-line $[0, \infty)$. We cannot hope for a spectral decomposition of $L^2(\mathbf{R}^n)$ along the lines of (2.1), yet Fourier analysis provides us with a good approximation. Each $f \in L^1(\mathbf{R}^n)$ admits the Fourier transform

$$\widehat{f}(u) = \int_{\mathbf{R}^n} f(x)\overline{\varphi_u(x)}dx$$

and if $\widehat{f}$ is itself integrable, we have

$$f(x) = \int_{\mathbf{R}^n} \widehat{f}(u)\varphi_u(x)du.$$

The condition that both f and its Fourier transform be L^1 is cumbersome, and we shall replace it with the (weaker) condition that f be a Schwartz function, that is a smooth function for which all partial derivatives have rapid decay.

Definition 34. *A smooth function $f : \mathbf{R}^n \to \mathbf{C}$ is Schwartz if for all multi-indices*[3] *$\alpha, \beta \in \mathbf{N}^n$,*

$$\sup_{x\in\mathbf{R}^n} |x^\alpha D^\beta f(x)| < \infty.$$

We denote the space of all Schwartz functions by $\mathscr{S}(\mathbf{R}^n)$.

[3]For a multi-index $\alpha = (\alpha_1, \dots, \alpha_n)$, we have $x^\alpha = x_1^{\alpha_1}\cdots x_n^{\alpha_n}$ and $D^\alpha f = \partial_{x_1}^{\alpha_1}\cdots\partial_{x_n}^{\alpha_n} f$.

Exercise 35. *Show that...*

(1) *...any smooth function of compact support is in $\mathscr{S}(\mathbf{R}^n)$.*
(2) *...the infinite series on the RHS of the Poisson summation formula (see below) converges.*

2.3. Poisson summation formula

A fundamental tool to study connections between harmonic analysis, geometry, and number theory is provided by so called trace formulas. In the case of $\mathbf{T}^n$, this builds on the Poisson summation formula.

THEOREM 10 (Poisson summation formula). *Let $f \in \mathscr{S}(\mathbf{R}^n)$. Then*

$$\sum_{\xi \in \mathbf{Z}^n} f(\xi) = \sum_{\xi \in \mathbf{Z}^n} \widehat{f}(\xi).$$

PROOF. Define the average

$$F(x) = \sum_{\xi \in \mathbf{Z}^n} f(x+\xi).$$

Since $f \in \mathscr{S}(\mathbf{R}^n)$, the average F converges absolutely and uniformly on compacta. Moreover F is smooth and descends to a function on $\mathbf{T}^n$. As such, it admits the Fourier expansion

$$F(x) = \sum_{\xi \in \mathbf{Z}^n} \widehat{F}(\xi)\varphi_\xi(x).$$

"Taking the trace" leaves us with

$$\sum_{\xi \in \mathbf{Z}^n} f(\xi) = \sum_{\xi \in \mathbf{Z}^n} \widehat{F}(\xi),$$

where by "folding/unfolding,"

$$\begin{aligned}
\widehat{F}(\xi) &= \int_{\mathbf{T}^n} F(x)\overline{\varphi_\xi(x)}dx \\
&= \sum_{\eta \in \mathbf{Z}^n} \int_{[0,1)^n} f(x+\eta)\overline{\varphi_\xi(x)}dx \\
&= \sum_{\eta \in \mathbf{Z}^n} \int_{[0,1)^n+\eta} f(x)\overline{\varphi_\xi(x)}dx \ = \int_{\mathbf{R}^n} f(x)\overline{\varphi_\xi(x)}dx = \widehat{f}(\xi).
\end{aligned}$$

□

The Poisson summation formula is important for at least the following two reasons. It is a precursor of more general phenomena (trace formulas) in spectral analysis that play a fundamental role linking spectral and geometric information of a Riemannian manifold. In this direction, one of the highlight of this course will be Selberg's trace formula for (compact) hyperbolic surfaces. Secondly, it implies the famous Jacobi inversion theorem, which we will spend the rest of this class discussing. (Another important domain of application, which we won't have time to discuss, is the geometry of numbers; the Poisson summation formula is of routine use in estimating the number of lattice points contained in a bounded domain in space.)

In general, to a Riemannian manifold M with discrete spectrum $\{\lambda_k\}$, we associate the **spectral partition function** (or spectral generating function)

$$Z_M(t) = \sum_{k\geq 0} e^{-t\lambda_k},$$

which converges uniformly for $t \in \mathbf{R}_{>0}$. We rewrite this function

$$Z_M(t) = \sum_{k\geq 0} \mu_k e^{-t\tilde{\lambda}_k},$$

where $\tilde{\lambda}_k < \tilde{\lambda}_{k+1}$ and μ_k is the multiplicity of $\tilde{\lambda}_k$. We claim that the spectrum $\{\lambda_k\}$ is completely determined by Z_M. In fact, we can extract the eigenvalues and their multiplicities from Z_M using the following relations

$$\lim_{t\to\infty} e^{rt} Z_M(t) = \lim_{t\to\infty} \sum_{k\geq 0} \mu_k e^{t(r-\tilde{\lambda}_k)} = \begin{cases} 0 & r < \tilde{\lambda}_0, \\ \mu_0 & r = \tilde{\lambda}_0, \\ \infty & r > \tilde{\lambda}_0. \end{cases},$$

followed by

$$\lim_{t\to\infty} e^{rt}\left(Z_M(t) - e^{-\tilde{\lambda}_0 t}\right) = \lim_{t\to\infty} \sum_{k\geq 1} \mu_k e^{t(r-\tilde{\lambda}_k)} = \begin{cases} 0 & r < \tilde{\lambda}_1, \\ \mu_1 & r = \tilde{\lambda}_1, \\ \infty & r > \tilde{\lambda}_1 \end{cases}$$

and so on.

For the n-dimensional torus $\mathbf{T}^n$, the spectral partition function is given by

$$Z_n(t) = \sum_{\xi\in\mathbf{Z}^n} e^{-4\pi^2\|\xi\|^2 t} = \sum_{k\geq 0} r_n(k) e^{-4\pi^2 kt},$$

where $r_n(k)$ is the number of ways to represent k as a sum of n squares. The function Z_n is closely related to Jacobi's theta series

$$\theta(z) = \sum_{n\in\mathbf{Z}} e(n^2 z),$$

where $z \in \{x + iy \in \mathbf{C} : y > 0\}$ (otherwise the series wouldn't converge). In fact,

$$\theta(z)^2 = \left(\sum_{m\in\mathbf{Z}} e(m^2 z)\right)\left(\sum_{n\in\mathbf{Z}} e(n^2 z)\right) = \sum_{k\geq 0} r_2(k) e(kz)$$

and more generally

$$\theta(z)^n = \sum_{k\geq 0} r_n(k) e(kz).$$

Hence $Z_n(t) = \theta(2\pi i t)^n$. A famous application of the Poisson summation formula is the following identity, which takes cleanest form when introducing

$$\widetilde{\theta}(t) = \sum_{n\in\mathbf{Z}} e^{-\pi n^2 t},$$

i.e., $\widetilde{\theta}(t) = \theta(it/2)$.

THEOREM 11 (Jacobi's inversion formula). *For the modified theta function $\widetilde{\theta}(t)$, we have*

$$\widetilde{\theta}\,(t) = \frac{1}{\sqrt{t}}\widetilde{\theta}\left(\frac{1}{t}\right)$$

for all $t > 0$.

PROOF. We apply Poisson summation to $f(x) = e^{-\pi x^2 t}$, where $t > 0$. We easily see that $f \in \mathscr{S}(\mathbf{R})$. As a consequence, we can differentiate under the integral sign so that

$$\begin{aligned}\frac{d}{dx}\widehat{f}(x) &= -2\pi i \int_{\mathbf{R}} u f(u)\overline{\varphi_x(u)}du \\ &= \frac{i}{t}\int_{\mathbf{R}} \left(\frac{d}{du}e^{-\pi u^2 t}\right)\overline{\varphi_x(u)}du \\ &= -\frac{2\pi x}{t}\widehat{f}(x),\end{aligned}$$

where the last equality follows from integrating by parts. This ODE has unique solution $\widehat{f}(x) = Ce^{-\pi x^2/t}$, where the constant is evaluated in terms of the Gaussian integral

$$C = \widehat{f}(0) = \int_{\mathbf{R}} e^{-\pi u^2 t} du = \frac{1}{\sqrt{t}}.$$

□

Corollary 36 (Trace formula for the torus)**.** *For any $t > 0$,*

$$\sum_{\xi\in\mathbf{Z}^n} \exp(-4\pi^2 t\|\xi\|^2) = \frac{1}{(4\pi t)^{n/2}} \sum_{\xi\in\mathbf{Z}^n} \exp(-\|\xi\|^2/(4t)).$$

PROOF. On the LHS, we recognize $Z_n(t)$. Now we use that $Z_n(t) = \widetilde{\theta}(4\pi t)^n$ and apply Jacobi's inversion formula. □

The LHS is the spectral partition function for $\mathbf{T}^n$ and can also be viewed as the trace of the heat operator $e^{-t\Delta}$, $t > 0$. We can assign a geometric interpretation to the RHS. Let $n = 2$ and recall that a geodesic on $\mathbf{T}^2$ is the projection of a line segment in $\mathbf{R}^2$, hence determined by a point x and a slope σ. If this slope is rational, the geodesic closes up and its length is of the form $\|\xi\|^2$ for some $\xi \in \mathbf{Z}^2$. Any translation on the torus shifts x but preserves σ, and thus yields a closed geodesic of same length. These geodesics differ by a translation, that is, by an isometry; geometrically, they are not distinct. We may restrict our attention to closed geodesics up to free homotopy.[4] From this point of view, the RHS of the trace formula can be interpreted as a sum over all closed geodesics on $\mathbf{T}^n$ up to free homotopy.

— End of class #3 —

[4]Notice in particular that each free homotopy class in $\pi_1(\mathbf{T}^2) = \mathbf{Z}^2$ contains infinitely many closed geodesics of equal length. This is very different from the situation for Riemannian manifolds of *negative* curvature; there, each free homotopy class of curves contains a *unique* minimizing closed geodesic.

CHAPTER 3

Geometry of the hyperbolic plane

During the rest of the semester, we will study the geometry of the hyperbolic plane and the spectral theory of its Laplacian explicitly, from the ground up. Let

$$\mathbf{H} = \{x + iy : y > 0\}$$

denote the complex upper half-plane. For each $z \in \mathbf{H}$, the tangent space $T_z\mathbf{H}$ is isomorphic to $\mathbf{C}$, and we equip $\mathbf{H}$ with the Riemannian metric given by

$$g_z(u, v) = \frac{u\overline{v}}{y^2} \quad \text{or equivalently} \quad ds^2 = \frac{dx^2 + dy^2}{y^2},$$

where y is the imaginary part of $z = x + iy$. The Riemannian metric g_z induces the norm

$$|v|_z = \frac{|v|}{y}.$$

on $T_z\mathbf{H} \cong \mathbf{C}$. We will first examine what this means for the induced notions of length, angle, area, curvature, the Laplacian, and isometries.

Remark 37. *For computations, it is at times convenient to work in a different model. The Cayley transform*

$$\mathbf{H} \to \mathbf{D}, \; z \mapsto \frac{z - i}{z + i},$$

is a biholomorphism mapping the upper half-plane to the open unit disk in $\mathbf{C}$. *It induces the hyperbolic metric*

$$h_z(u, v) = \frac{4u\overline{v}}{(1 - |z|^2)^2} \quad \textit{or equivalently} \quad ds^2 = \frac{4(dx^2 + dy^2)}{(1 - |z|^2)^2}$$

on the unit disk $\mathbf{D}$. *By construction, the Cayley transform is a bijective isometry between* $(\mathbf{H}, g)$ *and* $(\mathbf{D}, h)$.

Proof. The inverse Cayley transform is given by $f^{-1}(z) = \frac{-i(z+1)}{z-1}$. The induced metric is given by

$$h_z(u, v) = g_{f^{-1}(z)}(df^{-1}u, df^{-1}v) = \frac{|df^{-1}|^2}{(\mathfrak{Im} f^{-1}(z))^2} u\overline{v}.$$

Direct computations give $df^{-1}(z) = 2i/(z-1)^2$ and $\mathfrak{Im} f^{-1}(z) = (1-|z|^2)/|1-z|^2$. $\square$

3.1. Length, angles, area, curvature

3.1.1. Length. The hyperbolic length of a parametrized piecewise C^1 curve $\gamma : [a,b] \to \mathbf{H}$, $\gamma(t) = x(t) + iy(t)$, is given by

$$L(\gamma) = \int_a^b \frac{\sqrt{x'(t)^2 + y'(t)^2}}{y(t)} dt.$$

Examples 38. *Let $a < b$. The hyperbolic length of the segment $[a + iy, b + iy]$ is*

$$\frac{b-a}{y}.$$

The closer the segment is to the real axis, the longer it is. The hyperbolic length of the segment $[x + ia, x + ib]$ is

$$\int_a^b \frac{dt}{t} = \log \frac{b}{a}.$$

3.1.2. Angles. The angle θ between two vectors u, v in $T_z\mathbf{H} \cong \mathbf{C}$ is given by

$$\cos\theta = \frac{g_z(u,v)}{|u|_z|v|_z} = \frac{u\overline{v}}{|u||v|}$$

and so coincides with the usual Euclidean angle.

3.1.3. Area. We compute the hyperbolic area from the hyperbolic metric tensor $A = \operatorname{diag}(1/y^2, 1/y^2)$ (see Section 1.2.1); the *hyperbolic area* of $A \subset \mathbf{H}$ is given by

$$\operatorname{area}(A) = \iint_A d\mu(z) = \iint_A \frac{dxdy}{y^2}$$

when this integral exists.

Proposition 39. *A hyperbolic disk of radius R has area $4\pi \sinh(\frac{R}{2})^2$ and circumference $2\pi \sinh(R)$.*

Proof. Let C be the circle of (Euclidean) radius $r \in (0,1)$ and (Euclidean) center 0 in $\mathbf{C}$. We will compute in the disk model. The (hyperbolic) circumference of C is given by

$$\int_C ds = 2\int_C \frac{\sqrt{dx^2+dy^2}}{1-|z|^2} = 2\int_0^{2\pi} \frac{r}{1-r^2}dt = \frac{4\pi r}{1-r^2}.$$

We want to express r in terms of the hyperbolic radius of C. The hyperbolic length of the segment $[0, ir]$ in $\mathbf{D}$ is

$$R = \int_0^r \frac{2}{1-t^2}dt = \log\frac{1-r}{1+r}$$

hence

$$r = \frac{1-e^R}{1+e^R}$$

and the circumference is

$$\frac{4\pi r}{1-r^2} = 4\pi\frac{e^R - e^{-R}}{4} = 2\pi\sinh(R).$$

The area is

$$4\int_0^r\int_0^{2\pi}\frac{tdtd\theta}{(1-t^2)^2}=4\pi\int_{1-r^2}^1\frac{du}{u^2}=\frac{4\pi r^2}{1-r^2}$$

Plugging once more $r=\frac{1-e^R}{1+e^R}$, conclude that

$$\text{area}=4\pi\sinh(R/2)^2.$$

$\square$

Note that this situation is very different from Euclidean geometry: here, area and length of a circle are comparable. In other words, hyperbolic area is concentrated close to the boundary.

3.1.4. Curvature and Laplacian. The hyperbolic metric also determines the Gaussian curvature ($K=-1$) and (geometric) Laplacian for $\mathbf{H}$, given by

$$\Delta=-y^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right).$$

3.2. Isometries

Definition 40. *A Möbius transformation of $\widehat{\mathbf{C}}=\mathbf{C}\cup\{\infty\}$ is a biholomorphic rational function of the form*

$$z\mapsto\frac{az+b}{cz+d},$$

where $a,b,c,d\in\mathbf{C}$, $ad-bc\neq 0$ and $\infty\mapsto\frac{a}{c}$, $-\frac{d}{c}\mapsto\infty$.

The set of all Möbius transformations forms a group that can be identified with $\mathrm{PGL}_2(\mathbf{C})=\mathrm{GL}_2(\mathbf{C})/\mathbf{C}^\times$, via

$$\begin{pmatrix}a&b\\c&d\end{pmatrix}\mapsto\left(z\mapsto\frac{az+b}{cz+d}\right).$$

A generalized circle in $\widehat{\mathbf{C}}$ is either a circle in $\mathbf{C}$ or a straight line passing through ∞. A central geometric feature of Möbius transformations is that

Proposition 41. *Möbius transformations preserve generalized circles in $\hat{\mathbf{C}}$.*

Proof. A Möbius transformation can be decomposed into the composition of a finite number of inversions $z\mapsto z^{-1}$, scalings $z\mapsto\tau\cdot z$, and translations $z\mapsto z+w$. Indeed, we can decompose $z\mapsto\frac{az+b}{cz+d}$ into

$$z\mapsto cz+d\mapsto\frac{1}{cz+d}+t\mapsto\frac{ctz+dt+1}{cz+d}\mapsto\frac{az+(\frac{ad}{c}+\frac{a}{ct})}{cz+d}=\frac{az+b}{cz+d}$$

by choosing $t=\frac{-a}{ad-bc}$. (This is assuming that $c\neq 0$; we leave it to the reader to show that this decomposition is also possible when $c=0$.) Scalings and translations clearly preserve generalized circles. The inversion $z\mapsto z^{-1}$ is the composition of the circle inversion[1] $i(z) = \overline{z}^{-1}$ with respect to the unit circle and complex conjugation. Both preserve generalized circles. □

The Möbius transformations of $\hat{\mathbf{C}}$ that preserve the real axis $\mathbf{R}$ must have real coefficients, i.e., correspond to elements of $\mathrm{PGL}_2(\mathbf{R})$. The upper half-plane $\mathbf{H}$ is preserved by the subgroup $\mathrm{PGL}_2^+(\mathbf{R})$ of elements of $\mathrm{PGL}_2(\mathbf{R})$ with positive determinant, since

$$\mathfrak{Im}\left(\frac{az+b}{cz+d}\right) = \mathfrak{Im}\left(\frac{(az+b)(c\overline{z}+d)}{|cz+d|^2}\right) = \frac{ad-bc}{|cz+d|^2}y.$$

There is a canonical isomorphism $\mathrm{PGL}_2^+(\mathbf{R}) \cong \mathrm{PSL}_2(\mathbf{R})$ via $g \mapsto (\det(g))^{-1/2}g$.

Proposition 42. $\mathrm{PSL}_2(\mathbf{R}) \subset \mathrm{Isom}(\mathbf{H})$.

Proof. We use that

$$g'(z) = \frac{1}{(cz+d)^2}$$

to see that

$$L(g \circ \gamma) = \int_a^b \frac{|(g.\gamma)'(t)|}{\mathfrak{Im}(g.\gamma(t))}dt = \int_a^b \frac{|g'.(\gamma(t))|}{\mathfrak{Im}(g.\gamma(t))}|\gamma'(t)|dt = \int_a^b \frac{|\gamma'(t)|}{\mathfrak{Im}(\gamma(t))}dt = L(\gamma).$$

□

In particular, the group $\mathrm{PSL}_2(\mathbf{R})$, acting on the extended half-plane $\overline{\mathbf{H}} = \mathbf{H} \cup \mathbf{R} \cup \{\infty\}$, contains the following motions:

- translations by $t \in \mathbf{R}$,

$$n_t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} : z \mapsto z + x,$$

  which translate every point in $\mathbf{H} \cup \mathbf{R}$ along horizontal lines and fixes the point at ∞;
- dilations by $t \in \mathbf{R}_{>0}$,

$$a_t = \begin{pmatrix} t^{1/2} & 0 \\ 0 & t^{-1/2} \end{pmatrix} : z \mapsto t \cdot z,$$

  which dilate along half-rays starting at 0 and fixes the points at 0 and ∞;

[1]Inversion with respect to the Euclidean circle C of radius r and center z_0 is given by

$$\iota_C(z) = \frac{r^2}{\overline{\overline{z - z_0}}} + z_0.$$

Here is a pot pourri of its properties:

- Any line through z_0 is preserved;
- A line tangent to C is reflected to a circle inside C passing through z_0 and tangent to C;
- Circles that pass through z_0 are reflected to lines;
- Circles that don't pass through z_0 are reflected to circles;
- Circles that are orthogonal to C are preserved.

- the involution
$$k = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} : z \mapsto -z^{-1},$$
which inverts points across the unit semicircle $\mathbf{S}^1 \cap \mathbf{H}$ and fixes the point at i.

In fact, this is enough to decompose $\mathrm{SL}_2(\mathbf{R})$:

THEOREM 12 (Bruhat decomposition). *Each* $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbf{R})$ *can be written*
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & \frac{a}{c} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1/c & 0 \\ 0 & c \end{pmatrix} \begin{pmatrix} 1 & \frac{d}{c} \\ 0 & 1 \end{pmatrix}$$
if $c \neq 0$ *and*
$$\begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \begin{pmatrix} 1 & ba \\ 0 & 1 \end{pmatrix}$$
otherwise.

There is an important class of hyperbolic isometries that does not belong to $\mathrm{PSL}_2(\mathbf{R})$: reflections. To define hyperbolic reflections, we first need to describe hyperbolic lines.

3.3. Geodesics

Proposition 43. *Geodesics in* $\mathbf{H}$ *are either straight lines or semicircles that meet the real axis orthogonally.*

PROOF. Assume first that z_1 and z_2 lie on a vertical half-line in $\mathbf{H}$, i.e., $z_1 = x + ia$ and $z_2 = x + ib$, where we may assume that $a < b$. Then any curve γ joining z_1 to z_2 verifies
$$L(\gamma) = \int \frac{\sqrt{x'(t)^2 + y'(t)^2}}{y(t)} dt \geq \int \frac{|y'(t)|}{y(t)} dt = \int_a^b \frac{dy}{y} = \log \frac{b}{a},$$
which is precisely the hyperbolic length of the segment $[x + ia, x + ib]$.

Suppose now that z_1 and z_2 are in general position. Then z_1 and z_2 lie on a semicircle L with center α in $\mathbf{R}$. There exists a Möbius transformation $g \in \mathrm{PSL}_2(\mathbf{R})$ that maps L to a vertical half-line in $\mathbf{H}$ (choosing g such that $g(\alpha) = \infty$). We have showed that the length-minimzing path from $g(z_1)$ to $g(z_2)$ is the straight-edge segment connecting them. Since g is an isometry it preserves the length of curves and we conclude that the path from z_1 to z_2 along L is distance-minimizing. □

The above proof relies on an idea that we will use again: up to conjugation by some g, we may assume that a geodesic is in preferred position. In fact, suppose γ is a geodesic in $\mathbf{H}$ with endpoints $\overline{\alpha} < \alpha \in \mathbf{R}$. Then the image of $A = \frac{1}{\sqrt{\alpha - \overline{\alpha}}} \left(\begin{smallmatrix} -1 & \overline{\alpha} \\ 1 & -\alpha \end{smallmatrix}\right)$ in $\mathrm{PSL}_2(\mathbf{R})$ can be seen as the isometry mapping γ to the vertical geodesic from 0 to ∞. In particular, given two geodesics γ_1, γ_2 in $\mathbf{H}$, there exists $g \in \mathrm{PSL}_2(\mathbf{R})$ that maps γ_1 to γ_2.

A (characteristic) feature of hyperbolic geometry is that the sum of the inner angles of a triangle is $< \pi$. This follows from

Proposition 44 (Gauss defect)**.** *If T is a hyperbolic triangle with angles α, β, γ, then*
$$\mathrm{area}(T) = \pi - \alpha - \beta - \gamma.$$

PROOF. Let T_0 be a hyperbolic triangle with a vertex at ∞. By applying isometries, we may assume that the two other vertices lie along the geodesic $\mathbf{S}^1 \cap \mathbf{H}$. The angle at the vertex at ∞ is 0, and we let α, β be the inner angles at the vertices along $\mathbf{S}^1 \cap \mathbf{H}$. Then

$$\begin{aligned}\operatorname{area}(T_0) &= \int_{\cos(\beta)}^{\cos(\pi-\alpha)} \int_{\sqrt{1-x^2}}^{\infty} \frac{dxdy}{y^2} \\ &= \int_{\sin(\frac{\pi}{2}-\beta)}^{\sin(-\frac{\pi}{2}+\alpha)} \frac{dx}{\sqrt{1-x^2}} \\ &= \frac{\pi}{2} - \beta + \frac{\pi}{2} - \alpha = \pi - \alpha - \beta.\end{aligned}$$

Let now T be a hyperbolic triangle in general position with inner angles α, β, γ. We may extend one of the geodesic segments to a point x on the real axis. The area of T is the difference of the areas of the two triangles T' and T'' with vertex at x. Applying isometries, both T' and T'' are isometric to a triangle of the form T_0. Comparing the areas, we have

$$\operatorname{area}(T) = \operatorname{area}(T') - \operatorname{area}(T'') = \pi - (\alpha + \delta) - \beta - (\pi - \delta - (\pi - \gamma)) = \pi - \alpha - \beta - \gamma.$$

□

An easy generalization of the Gauss defect for triangles allows to immediately compute the area of any hyperbolic polygon with n sides.

Proposition 45. *Let $\mathcal{P}$ be a hyperbolic n-gon with inner angles $\alpha_1, \ldots, \alpha_n$. Then*

$$\operatorname{area}(\mathcal{P}) = (n-2)\pi - (\alpha_1 + \cdots + \alpha_n).$$

PROOF. Any polygon with n sides can be decomposed into $n-2$ adjacent triangles. □

We can now define hyperbolic reflections. Let γ be a geodesic in $\mathbf{H}$, and denote by ι_γ the reflection across γ. If $\gamma = i\mathbf{R}_{>0}$, then $\iota_\gamma(z) = -\overline{z}$. By conjugation, any hyperbolic reflection can be written in this form. One easily checks that ι_γ is an isometry and $\iota_\gamma^2 = I$. The latter implies that $\det \iota_\gamma = -1$ and so $\iota_\gamma \notin \mathrm{PSL}_2(\mathbf{R})$. (In particular, ι is orientation-reversing.)

— End of class #4 —

3.4. Classification of motions

Recall that each orientation-preserving isometry of the plane $\mathbf{R}^2$ is either a translation, a rotation, or the identity. We set to establish a similar classification for

$G = \mathrm{PSL}_2(\mathbf{R})$. Let

$$N = \left\{ n_x = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} : x \in \mathbf{R} \right\},$$
$$A = \left\{ a_y = \begin{pmatrix} y^{1/2} & 0 \\ 0 & y^{-1/2} \end{pmatrix} : y > 0 \right\},$$
$$K = \left\{ k_\theta = \begin{pmatrix} \cos\frac{\theta}{2} & \sin\frac{\theta}{2} \\ -\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix} : \theta \in [0, 2\pi) \right\}.$$

Proposition 46. *The group $\mathrm{PSL}_2(\mathbf{R})$ acts transitively on $\mathbf{H}$. More precisely, for each $z \in \mathbf{H}$, there exists $g \in \mathrm{PSL}_2(\mathbf{R})$ such that $g(i) = z$.*

Proof. Let $z = x + iy$ and choose $g = n_x a_y$. Then $g(i) = n_x(iy) = x + iy$. □

The stabilizer subgroup $\mathrm{Stab}(i) = \{g \in G : g(i) = i\}$ corresponds to K. Indeed, $\frac{ai+b}{ci+d} = i$ implies $a = d$, $c = -b$ and we must have $a^2 + b^2 = 1$. Hence we have the identification $\mathbf{H} = G/K$.

Theorem 13. *The isometries in $\mathrm{PSL}_2(\mathbf{R})$ can be classified as follows. Let $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \neq \pm I$ be a matrix in $\mathrm{SL}_2(\mathbf{R})$. Then its image in $\mathrm{PSL}_2(\mathbf{R})$ is...*

type	*alg. criterion*	*geom. criterion*	*conjugate to*
...elliptic	$\lvert a+d\rvert < 2$	*1 fixed point in* $\mathbf{H}$	k_θ
...parabolic	$\lvert a+d\rvert = 2$	*1 fixed point in* $\partial_\infty \mathbf{H}$	n_x
...hyperbolic	$\lvert a+d\rvert > 2$	*2 fixed point in* $\partial_\infty \mathbf{H}$	a_y

Here $\partial_\infty \mathbf{H} = \mathbf{R} \cup \{\infty\}$ denotes the "boundary at infinity."

Proof. Consider the quadratic equation

$$(*) \qquad g(z) = \frac{az+b}{cz+d} = z \iff cz^2 + (d-a)z - b = 0.$$

Its discriminant is $D = (d-a)^2 + 4bc = (a+d)^2 - 4$. The following are equivalent:

$$\begin{aligned} |a+d| > 2 &\iff D > 0 \iff (*) \text{ has two solutions in } \mathbf{R} \cup \{\infty\} \\ |a+d| = 2 &\iff D = 0 \iff (*) \text{ has a unique solution in } \mathbf{R} \cup \{\infty\} \\ |a+d| < 2 &\iff D < 0 \iff (*) \text{ has a unique solution in } \mathbf{H}. \end{aligned}$$

In the first case, we may assume, up to a conjugation, that the two fixed points are at 0 and ∞. The motions in $\mathrm{PSL}_2(\mathbf{R})$ that preserve these points must also preserve the geodesic connecting them and is hence the set of diagonal matrices. In the second case, we may assume, up to conjugation, that the fixed point is at ∞. The motions in $\mathrm{PSL}_2(\mathbf{R})$ that preserve this point can be reduced to the unipotents. In the last case, we conjugate the fixed point to i, and use that $\mathrm{Stab}(i) = K$. □

Theorem 14. $\mathrm{Isom}^+(\mathbf{H}) = \mathrm{PSL}_2(\mathbf{R})$.

Proof. We have already seen that each $g \in \mathrm{PSL}_2(\mathbf{R})$ is an orientation-preserving isometry with respect to the hyperbolic metric. In the other direction, let $\varphi \in \mathrm{Isom}^+(\mathbf{H})$ and let γ denote the vertical geodesic from 0 to ∞. Then $\varphi(\gamma)$ is again a geodesic. Given three points ia, ib, ic (with $0 < a < b < c < \infty$) along γ, there exists a unique[2] Möbius transformation g that maps $\varphi(ia), \varphi(ib), \varphi(ic)$ to ia, ib, ic. Then $g\circ\varphi$ is an isometry fixing the three points ia, ib, ic. Hence $g \circ \varphi$ fixes every point on γ. It is thus either the identity of a hyperbolic reflection across γ. Since both g and φ are orientation-preserving, their composition is as well and we conclude that $\varphi = g^{-1}$. □

3.5. Metric space properties

The notion of lengths of curves with respect to the hyperbolic metric yields a distance function on $\mathbf{H}$. (And on $\mathbf{D}$ as well). We have the following formulas. For z_1, $z_2 \in \mathbf{H}$, we have

$$\cosh d_{\mathbf{H}}(z_1, z_2) = 1 + \frac{|z_1 - z_2|^2}{2y_1y_2}.$$

For z_1, $z_2 \in \mathbf{D}$, we have

$$\tanh \frac{1}{2} d_{\mathbf{D}}(z_1, z_2) = \left| \frac{z_1 - z_2}{1 - z_1\overline{z_2}} \right|.$$

Remark 47. *We endow* $\mathrm{SL}_2(\mathbf{R})$ *with the matrix topology induced by*

$$\|g\|^2 = a^2 + b^2 + c^2 + d^2$$

for $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbf{R})$*. Note that*

$$\|g\|^2 = 2\cosh d_{\mathbf{H}}(i, gi).$$

This provides us with a direct relation between the topology of subgroups of $\mathrm{SL}_2(\mathbf{R})$ *and of their orbits in* $\mathbf{H}$*.*

Proposition 48. *The spaces* $(\mathbf{H}, d_{\mathbf{H}})$ *and* $(\mathbf{D}, d_{\mathbf{D}})$ *are complete metric spaces.*

Proof. Since the two spaces are isometric and Riemannian isometries are metric isometries, it suffices to prove the statement for $(\mathbf{H}, d_{\mathbf{H}})$. Let $(z_n) \subset \mathbf{H}$ be a Cauchy sequence with respect to $d_{\mathbf{H}}$. It must converge to some $z \in \overline{\mathbf{H}}$. We claim that in fact $z \in \mathbf{H}$. Suppose instead that $z_n = x_n + iy_n$ with $y_n \to \infty$. Using that

$$\cosh d_{\mathbf{H}}(z_n, z_m) = \frac{(x_n - x_m)^2}{2y_my_n} + \frac{y_n}{2y_m} + \frac{y_m}{2y_n},$$

we see that the sequence cannot be Cauchy, a contradiction. □

More generally, one has

Theorem 15 (Hopf–Rinow). *For a connected Riemannian manifold* (M, g)*, the following are equivalent:*

1. M *is a complete metric space;*
2. M *is geodesically complete;*
3. *closed and bounded subsets of* M *are compact.*

We say that M is geodesically complete if every geodesic can be extended infinitely. This also implies that given two points, there exists a length-minimizing geodesic that connects them.

[2]This can be proven directly, but can also be seen as a special case of the main theorem of projective geometry.

CHAPTER 4

Fuchsian groups

Definition 49. *A subgroup* $\Gamma < \mathrm{SL}_2(\mathbf{R})$ *is a Fuchsian group if it is discrete with respect to the induced matrix topology.*

We will see that the corresponding subgroups in $\mathrm{PSL}_2(\mathbf{R})$ are precisely the uniformizing groups of hyperbolic surfaces.

4.1. Interlude: Poincaré and Fuchsian functions

We owe the notion of Fuchsian group to Poincaré (1854–1912). In the 1870s, Fuchs studied families of second-order linear differential equations of the form

$$f'' + A(z)f' + B(z) = 0, \tag{4.1}$$

where A, B are functions on the Riemann sphere $\hat{\mathbf{C}}$ with finitely many isolated singularities. Fuchs was interested in local solutions of such equations. At $z = z_0$, a point where A and B are both holomorphic, Eq. (4.1) has two linearly independent solutions f_1, f_2. If z_0 is in the neighborhood of an isolated singularity x, and we take a continuous loop based at z_0 that goes around x once, then we may analytically continue Eq. (4.1). As a consequence, we're also continuously deforming in the space of solutions. After a full revolution, we will have deformed (f_1, f_2) into $(\tilde{f}_1, \tilde{f}_2)$. Since we are back at z_0, $\tilde{f}_1$ and $\tilde{f}_2$ must also be solutions of Eq. (4.1) and linearly independent. In particular, they must be linear combinations of f_1, f_2. To make things explicit, let

$$\tilde{f}_1 = af_1 + bf_2, \;\; \tilde{f}_2 = cf_1 + bf_2$$

for some complex coefficients $a, b, c, d \in \mathbf{C}$. Then

$$\frac{\tilde{f}_1}{\tilde{f}_2} = \frac{a\frac{f_1}{f_2} + b}{c\frac{f_1}{f_2} + d}$$

is a Möbius transform of the ratio f_1/f_2. That $\tilde{f}_1$ and $\tilde{f}_2$ are linearly independent is equivalent to $ad - bc \neq 0$, and the set

$$\Gamma = \{A \in \mathrm{PGL}_2(\mathbf{C}) \text{ obtained in this way}\}$$

is a group that depends on the initial data (z_0, f_1, f_2) only up to conjugation in $\mathrm{PSL}_2(\mathbf{C})$. We have found a canonical group associated to the differential equation Eq. (4.1). This result was exciting enough that in 1878 the *Académie des Sciences* in Paris held a prized essay contest on the topic of differential equations to encourage French mathematicians to further study the works of Fuchs. Poincaré, participating in the contest, was particularly intrigued by the investigations of Fuchs into the local properties of the set-theoretic inverse $F = f^{-1}$ of $f : z \mapsto \frac{f_1}{f_2}(z) = w$. Up to switching f_1 and f_2, we may assume that $\mathfrak{Im}(w) \geq 0$ and restrict $\mathrm{PGL}_2(\mathbf{C})$ to $\mathrm{PSL}_2(\mathbf{R})$. The task becomes:

Find $F : \mathbf{H} \to \mathbf{C}$ holomorphic and Γ-invariant for the action of a subgroup $\Gamma < \mathrm{SL}_2(\mathbf{R})$ on $\mathbf{H}$ by Möbius transformation.

Here Poincaré recognized that the transformation in question were precisely the isometries of the hyperbolic metric. This is, incidentally, what led him to develop the model of the upper half-plane for hyperbolic geometry as we know it now. Further, Poincaré notes that this question is only interesting for groups Γ whose action on $\mathbf{H}$ has discrete orbits, i.e., that is to say that no Γ-orbit should have an accumulation point in $\mathbf{H}$. Indeed, otherwise the only invariant holomorphic functions are the constant functions. This is what led to the above definition of Fuchsian group. We will admit the following equivalent characterizations.

Fact 50. *The following are equivalent.*

(1) Γ *is a Fuchsian group;*
(2) $\Gamma \curvearrowright \mathbf{H}$ *has discrete orbits, i.e. no orbit* Γz *has accumulation points;*
(3) $\Gamma \curvearrowright \mathbf{H}$ *is properly discontinuous, i.e., for any compact subset* $K \subset \mathbf{H}$*,* $K \cap \gamma K = \emptyset$ *except for at most finitely many* $\gamma \in \Gamma$*;*
(4) $\Gamma \curvearrowright \mathbf{H}$ *is wandering, i.e., for any* $z \in \mathbf{H}$*, there exists a neighborhood* U *of* z *in* $\mathbf{H}$ *such that* $U \cap \gamma U \neq \emptyset$ *implies that* $\gamma \in \mathrm{Stab}_\Gamma(z)$ *and the latter group is finite.*

(Remark that these equivalent characterizations are far from true for more general group actions on a topological space.)

4.2. Fuchsian groups and hyperbolic surfaces: uniformization

Group actions that are free (i.e., with trivial stabilizers) and properly discontinuous are closely related to covering spaces. Let X and Y be topological spaces. We say that a continuous surjective map $p : X \to Y$ is a **covering projection** if for every $y \in Y$, there is an open neighborhood U of y in Y such that $p^{-1}(U)$ is a disjoint union of open subsets V_α and $p(V_\alpha)$ is homeomorphic to U. There is a natural action on the fibers $p^{-1}(U)$ by permutations. In fact, the set of all homeomorphisms $f : X \to X$ such that $p \circ f = f$ forms a group, called the **group of deck transformations**. If X is connected, then the action of the group of deck transformations on X is free and properly discontinuous. The following proposition says that this is essentially how such actions arise.

Proposition 51. *Let* X *be a locally compact Hausdorff space. If* Γ *is a group acting properly discontinuously on* X*, then* $\Gamma\backslash X$ *is Hausdorff. If the action is moreover free, then* $X \to \Gamma\backslash X$ *is a covering projection.*

Proof. Let $x, y \in X$ be points in distinct Γ-orbits (so that $\Gamma x \neq \Gamma y$). Let K be a compact neighborhood of x. Since the Γ-action is properly discontinuous, we may assume that K is disjoint from any point in Γy. Then $\cup \gamma K$ is a neighborhood of Γx that is disjoint from Γy. Similarly we can show that y admits a compact neighborhood that is disjoint form K. Suppose now that the action of Γ is moreover free, then we may assume there exists an open neighborhood U of x that contains no other point in Γx. Then U is homeomorphic to its projection in $\Gamma\backslash X$. $\square$

If X has moreover smooth/Riemannian/hyperbolic structure, this additional structure is transported by p to $\Gamma\backslash X$. In the latter two cases, p is then a local isometry. By Theorem 13, a Fuchsian group Γ acts freely on $\mathbf{H}$ if and only if it has no elliptic elements. The quotient space $\Gamma\backslash\mathbf{H}$ is then a hyperbolic surface.

A remarkable feature of covering theory is that it tells us that every hyperbolic surface can be obtained this way. Recall that if a covering space X is connected, then the action of the group of deck transformations is free and properly discontinuous. If X is moreover simply connected (X is called a universal cover in this case), then the group of deck transformation is further isomorphic to $\pi_1(Y, y)$ and Y is realized as $Y = \pi_1(Y, y)\setminus X$. Let M be a hyperbolic surface. It is known that such a space admits a universal cover, which we denote by $\widetilde{M}$. Then we may transport the hyperbolic structure of M to $\widetilde{M}$ via the covering projection. As a result, the fundamental group of M acts on $\widetilde{M}$ by isometries. The following important theorem tells us that $\widetilde{M}$ is isometric to $\mathbf{H}$ and thus $\pi_1(M)$ is isomorphic to a Fuchsian group Γ with no elliptic elements.

THEOREM 16. *Up to isometry and global rescaling, the only smooth complete simply connected surfaces of constant Gaussian curvature are* $\mathbf{R}^2$ *(with* $K = 0$*),* $\mathbf{S}^2$ *(with* $K = 1$*), and* $\mathbf{H}^2$ *(with* $K = -1$*).*

— End of class #5 —

4.3. Fundamental domains

The discussion in the previous section is obviously reminiscent of the study of the torus $\mathbf{T}^2$ obtained by letting the integer group of translations $\mathbf{Z}^2 \subset \mathrm{Isom}(\mathbf{R}^2)$ act properly discontinuously and freely on $\mathbf{R}^2$. To study curves or dynamics of flows on the torus, we rely on its fundamental domain $[0, 1)^2$ in $\mathbf{R}^2$. In this section, we identify fundamental domains for $\Gamma\backslash\mathbf{H}$.

Definition 52. *A connected set* $\mathcal{F} \subset \mathbf{H}$ *is a fundamental domain for* Γ *if*

(1) $\bigcup_{\gamma\in\Gamma} \gamma\overline{\mathcal{F}} = \mathbf{H}$*, where* $\overline{\mathcal{F}}$ *is the closure of* $\mathcal{F}$*;*

(2) $\gamma_1\mathring{\mathcal{F}} \cap \gamma_2\mathring{\mathcal{F}} = \emptyset$ *for all* $\gamma_1 \neq \gamma_2$*, where* $\mathring{\mathcal{F}}$ *is the interior of* $\mathcal{F}$*.*

Example 53. *Let* $\Gamma = \langle n_t\rangle = \left(\begin{smallmatrix} 1 & t\mathbf{Z} \\ 0 & 1 \end{smallmatrix}\right)$*. We obtain a fundamental domain by taking the cylinder* $\mathcal{C} = (0, t) \times (0, \infty)$ *in* $\mathbf{H}$*. Clearly this is not the unique choice. Identifying the two vertical sides of* $\mathcal{C}$ *via the translation* n_t*, we obtain an infinite cylinder, with a* **cusp** *at one end, corresponding to the point at* ∞ *and a* **funnel** *at the opposite one, corresponding to the segment* $(0, t) \subset \mathbf{R}$*.*

Example 54. *Let now* $\Gamma = \langle k_\theta\rangle$*. For* Γ *to be discrete, observe that* θ *must be a rational multiple of* π*. In particular,* Γ *is a finite group. The corresponding surface is an open funnel with a conical point, corresponding to the fixed point* i*.*

In both of these examples, the fundamental domain has infinite area. Indeed,

$$\text{area}([0,t]\times(0,\infty)) = \int_0^\infty \int_0^t \frac{dxdy}{y^2} = \infty.$$

The infinite volume comes from the funnel; we easily see that $\text{area}([0,t]\times[A,\infty)) < \infty$. In this course, we will only consider hyperbolic surfaces of finite area. In this case, the area of the fundamental domain is a numerical invariant of the group:

Proposition 55. *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two fundamental domains for Γ with boundary of null measure and such that* $\text{area}(\mathcal{F}_1) < \infty$*. Then* $\text{area}(\mathcal{F}_1) = \text{area}(\mathcal{F}_2)$.

Proof.

$$\text{area}(\mathcal{F}_1) \geq \text{area}\left(\overline{\mathcal{F}_1} \cap \bigcup_\gamma \gamma \mathring{\mathcal{F}}_2\right) = \text{area}\left(\bigcup_\gamma \overline{\mathcal{F}_1} \cap \gamma \mathring{\mathcal{F}}_2\right)$$

$$= \sum_\gamma \text{area}(\gamma\overline{\mathcal{F}_1} \cap \mathring{\mathcal{F}}_2) \geq \text{area}\left(\bigcup_\gamma \gamma\overline{\mathcal{F}_1} \cap \mathring{\mathcal{F}}_2\right) = \text{area}(\mathcal{F}_2).$$

We can now exchange $\mathcal{F}_1$ and $\mathcal{F}_2$ to obtain equality. □

Here is one way to construct a fundamental domain, called a **Dirichlet domain**. Fix $p \in \mathbf{H}$ (choosing p such that it is not fixed by a nontrivial element of Γ) and set

$$\mathcal{D} = \bigcap_{\substack{\gamma\in\Gamma \\ \gamma\neq\pm I}} \left\{z \in \mathbf{H} : d_{\mathbf{H}}(z,p) < d_{\mathbf{H}}(z,\gamma p)\right\}.$$

In other words the Dirichlet domain $\mathcal{D}$ is the intersection of infinitely many hyperbolic planes, and as such a connected convex (in the hyperbolic sense) region whose boundary is a union of geodesic arcs.

Proposition 56. *The Dirichlet domain $\mathcal{D}$ is a fundamental domain.*

Proof. Let $z \in \mathbf{H}$. We show that $z \in \Gamma z_0$ for some $z_0 \in \mathcal{D}$. Since Γz is a discrete orbit, there exists $z_0 \in \Gamma z$ such that

$$d_{\mathbf{H}}(z_0,p) < d_{\mathbf{H}}(\gamma z_0,p) = d_{\mathbf{H}}(z_0,\gamma^{-1}p)$$

for all nontrivial $\gamma \in \Gamma$. Let now $z_1, z_2 \in \Gamma z$; in particular, $z_1 = \gamma z_2$ for some $\gamma \in \Gamma$. Hence if $z_1 \in \mathcal{D}$, then $d_{\mathbf{H}}(z_1,p) < d_{\mathbf{H}}(z_2,p)$. But if z_2 also lies in $\mathcal{D}$, the same reasoning implies that $d_{\mathbf{H}}(z_2,p) < d_{\mathbf{H}}(z_1,p)$, a contradiction. □

We use this construction to build a fundamental domain for $\Gamma = \mathrm{SL}_2(\mathbf{Z})$. Fix $p = 2i$. (Note that i is an elliptic fixed point but $2i$ is not.) The group Γ contains the matrices $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}0&-1\\1&0\end{smallmatrix}\right)$. We deduce that the Dirichlet domain for Γ at $p = 2i$ is a subset of the hyperbolic triangle

$$\mathcal{T} = \{z \in \mathbf{C} : |z| > 1,\ |\mathfrak{Re}(z)| < 1/2\}.$$

In fact, we can show that this is precisely the whole Dirichlet domain.

Exercise 57. *Show that for each $z \in \mathcal{T}$ and $\gamma \in \Gamma$ not unipotent, $\mathfrak{Im}(\gamma z) < \mathfrak{Im}(z)$. Deduce that $\mathcal{T} = \mathcal{D}$, the Dirichlet domain for Γ at $p = 2i$.*

Having a 'nice' fundamental domain allows to visualize the action of the Fuchsian group on $\mathbf{H}$, and to select representative of the orbits (see next section). The unipotent matrix $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$ identifies the two vertical sides, while the involution $\left(\begin{smallmatrix}0&-1\\1&0\end{smallmatrix}\right)$ identifies the two circular segments across the vertical line passing through i. The resulting surface has one cusp and one conical point. From the Gauss defect formula,

$$\operatorname{area}(\mathcal{T}) = \pi - 2\frac{\pi}{3} = \frac{\pi}{3}.$$

We say that Γ is geometrically finite, if (any of) its fundamental domain is a convex polygon with finitely many sides.

Theorem 17 (Siegel). *If Γ is a Fuchsian group such that $\operatorname{area}(\Gamma\backslash\mathbf{H}) < \infty$, then Γ is geometrically finite and finitely generated.*

4.4. Some arithmetic examples of Fuchsian groups

Arithmetically, $\mathrm{SL}_2(\mathbf{Z})\backslash\mathbf{H}$ is a very nice object, but geometrically, it has conical points and hence no smooth structure, and it has a cusp hence is not compact. We obtain a larger family of examples with modular arithmetic as follows. For $N \geq 1$, let $\Gamma(N)$ be the kernel of the projection

$$\mathrm{SL}_2(\mathbf{Z}) \to \mathrm{SL}_2(\mathbf{Z}/N\mathbf{Z}),$$

that is $\Gamma(N)$ is the group of all matrices $A \in \mathrm{SL}_2(\mathbf{Z})$ such that $A \equiv I \pmod{N}$. That is, if $\left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right) \in \Gamma(N)$, then $a \equiv d \equiv 1 \pmod{N}$ and $b \equiv c \equiv 0 \pmod{N}$. We say that Γ is a **congruence subgroup** if Γ contains some $\Gamma(N)$ as a subgroup. The smallest such N is called the level of Γ, and $\Gamma(N)$ is called a **principal congruence subgroup**.

Proposition 58. *If $N > 1$, the group $\Gamma(N)$ contains no elliptic elements.*

Proof. By the classification of motions, every elliptic element of $\Gamma(1)$ has trace 0 or ± 1, and is thus conjugate to either of $\pm\left(\begin{smallmatrix}0&-1\\1&0\end{smallmatrix}\right)$, $\pm\left(\begin{smallmatrix}1&-1\\1&0\end{smallmatrix}\right)$, $\pm\left(\begin{smallmatrix}0&-1\\1&1\end{smallmatrix}\right)$. None of these matrices is congruent to $I \pmod{N}$ if $N > 1$. □

We use the following general proposition to examine the form of fundamental domains for $\Gamma(N)$.

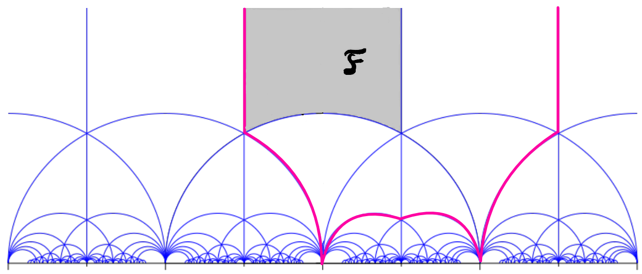
Proposition 59. *Let Γ be a Fuchsian group, with fundamental domain $\mathcal{F}$, and let $\Gamma' < \Gamma$ be a finite-index subgroup. Fix a finite set $\{\gamma_i\} \subset \Gamma$ of coset representatives such that $\Gamma = \cup \Gamma'\gamma_i$, and denote $\{\overline{\gamma}_i\}$ the resulting set of distinct elements in $\mathrm{PSL}_2(\mathbf{R})$ Then*

$$\mathcal{F}' = \bigcup \overline{\gamma_i}\mathcal{F} \tag{4.2}$$

is a fundamental domain for Γ'.

Proof. Left as exercise. □

Below is a fundamental domain for $\Gamma(2)$ obtained from the union of $[\overline{\Gamma(1)} : \overline{\Gamma(2)}] = [\Gamma(1) : \Gamma(2)] = 6$ copies of the standard fundamental domain for $\Gamma(1)$ (in grey).



— End of class #6 —

4.5. Geometric and algebraic features of fundamental domains

From the examples encountered thus far, we observe that some geometric (or algebraic) features of the surface (or its uniformizing Fuchsian group) can be read off a fundamental domain. We have the following bijections.

Proposition 60. *Let Γ be a Fuchsian group. There is a bijection between the set of cusps for $M = \Gamma\backslash\mathbf{H}$ and the set of Γ-orbits of parabolic fixed points of Γ.*

PROOF. Exercise. □

Corollary 61. *Let Γ be a cofinite Fuchsian group. Then $M = \Gamma\backslash\mathbf{H}$ is compact if and only if Γ contains parabolic elements.*

Proposition 62. *There is a bijection between the set of conical singularities on $M = \Gamma\backslash\mathbf{H}$ and the set of Γ-orbits of elliptic fixed points of Γ.*

Proposition 63. *There is a bijection between the set of closed geodesics on $M = \Gamma\backslash\mathbf{H}$ and the set of Γ-conjugacy classes of hyperbolic elements in Γ.*

PROOF. Let γ be a closed geodesic on M. Pick a lift $\widetilde{\gamma}$ of γ in $\mathbf{H}$; this is a geodesic of the hyperbolic plane. Up to a conjugation, we may assume that $\widetilde{\gamma}$ is the vertical geodesic joining 0 and ∞. Then

$$\mathrm{Stab}_G(\widetilde{\gamma}) = \pm A \cong \mathbf{R}$$

(see Section 3.4 for the definition of A) and hence

$$\mathrm{Stab}_\Gamma(\widetilde{\gamma}) = (\Gamma \cap A)\{\pm I\} \cong \mathbf{Z}$$

by discreteness. The cyclic generator γ_0 of this subgroup is hyperbolic. If we pick a different lift $\hat{\gamma}$ of γ, then there exists some $\gamma' \in \Gamma$ such that $\gamma'\hat{\gamma} = \widetilde{\gamma}$, and

$$\mathrm{Stab}_\Gamma(\hat{\gamma}) = \left\langle \gamma'^{-1}\gamma_0\gamma' \right\rangle .$$

We can now check that this map is bijective. □

A choice of fundamental domain corresponds to a choice of generating set for Γ, which can be read off from side-pairing.

Proposition 64. *Let $\mathcal{F}$ be a fundamental domain for Γ. Then Γ is generated by the set*

$$S = \{\gamma \in \Gamma : \overline{\mathcal{F}} \cap \gamma\overline{\mathcal{F}} \neq \emptyset,\ \gamma \neq \pm I\}.$$

PROOF. Let $\Gamma^* = \langle S \rangle$. To each $z \in \mathbf{H}$, we can associate an element $\gamma \in \Gamma$ such that $\gamma z \in \overline{\mathcal{F}}$. If there is a distinct element $\gamma' \in \Gamma$ for which $\gamma' z \in \overline{\mathcal{F}}$, then

$$\gamma' z \in \overline{\mathcal{F}} \cap \gamma'\gamma^{-1}\overline{\mathcal{F}} \neq \emptyset.$$

Thus $\gamma'\gamma^{-1} \in \Gamma^*$ and $\Gamma^*\gamma' = \Gamma^*\gamma$. We thus have a mapping $\Phi : \mathbf{H} \to \Gamma^* \backslash \Gamma$. Since $\mathbf{H}$ is connected and locally compact, it suffices to show that Φ is constant on compact neighborhoods to conclude that $\Gamma^* = \Gamma$. Let K be a compact neighborhood of $z \in \mathbf{H}$. Then K is contained in a finite union of translates $\gamma_i\overline{\mathcal{F}}$, $i \in I$. Upon taking K to be sufficiently small, we may assume that z lies in each $\gamma_i\overline{\mathcal{F}}$. Then for any $z' \in K$, we have that $z' \in \gamma_i\overline{\mathcal{F}}$ for some $i \in I$ and thus $\Phi(z') = \Gamma^*\gamma_i^{-1} = \Phi(z)$. ☐

Corollary 65. *The group $\mathrm{SL}_2(\mathbf{Z})$ is generated by $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.* ☐

4.6. Some geometric examples of Fuchsian groups

We can also use fundamental domains to construct Fuchsian groups. Consider a hyperbolic triangle T with sides A, B, C. Let ι_A denote the hyperbolic reflection in the geodesic passing through A, and consider $\Gamma^* = \langle \iota_A, \iota_B, \iota_C \rangle$ the group generated by reflections across the sides of T.

Exercise 66. *Check that T is a fundamental domain for Γ^*, and that $\Gamma^*(T)$ tessellates $\mathbf{H}$.*

On the other hand, since reflections have determinant -1, $\Gamma^* \not\subset \mathrm{PSL}_2(\mathbf{R})$. To obtain a Fuchsian group, we consider instead the index 2 subgroup $\Gamma = \Gamma^* \cap \mathrm{PSL}_2(\mathbf{R})$.

Lemma 67. *We have $\Gamma^* = \Gamma \cup \Gamma\iota_A$.*

PROOF. Suppose $\gamma \in \Gamma^*$ but $\gamma \notin \Gamma$. This implies that $\det\gamma = -1$. Hence $\gamma\iota_A \in \Gamma^* \cap \mathrm{PSL}_2(\mathbf{R}) = \Gamma$. ☐

Let $z \in T$. Then $\Gamma z \subset \Gamma^* z$. Since $\Gamma^*(T)$ tessellates $\mathbf{H}$, we conclude that Γz is discrete, and Γ is Fuchsian. By Proposition 59, we have a new fundamental domain for Γ, given by the union of T and its reflection across A. One can check that the four sides can be identified via $\iota_A\iota_B$ and $\iota_B\iota_C$ and that the inner angles of T must be of the form $\alpha = \frac{\pi}{a}$, $\beta = \frac{\pi}{b}$, $\gamma = \frac{\pi}{c}$, with $2 \leq a, b, c \leq \infty$ and

$$\frac{1}{a} + \frac{1}{b} + \frac{1}{c} < 1.$$

The resulting surface is a sphere with three conical singularities.

Remark 68. *A convex hyperbolic polygon with an even number of sides is not necessarily the fundamental domain of a Fuchsian group. Poincaré clarified what conditions are required for this to hold; this is the Poincaré polygon theorem. Using its construction, one can show that every surface of fixed genus* $g \geq 2$ *can be equipped with a hyperbolic structure and that the set of all such structures is a complex manifold of dimension* $3g - 3$*.*

Remark 69. *Observe that the modular group* $\mathrm{SL}_2(\mathbf{Z})$ *is an example of (non-cocompact) triangle group. Most triangle groups are however not arithmetic groups. In fact, Takeuchi (1977) classified all arithmetic triangle groups; up to* $\mathrm{SL}_2(\mathbf{R})$*-conjugacy, there are only finitely many. Thompson (1980) showed that up to conjugacy, there are only finitely many congruence subgroups of fixed genus. In view of the preceding remark, we conclude that such groups (and the resulting hyperbolic surfaces) are very special objects.*

CHAPTER 5

The spectral theorem for compact hyperbolic surfaces

In this chapter, we take $M = \Gamma\backslash\mathbf{H}$ to be a compact hyperbolic surface. We think of functions on M as Γ-invariant functions $f : \mathbf{H} \to \mathbf{C}$. The (geometric) Laplacian on $\mathbf{H}$ with respect to the hyperbolic metric is given by

$$\Delta = -y^2 \left(\frac{d^2}{dx^2} + \frac{d^2}{dy^2} \right)$$

and commutes with the action of Γ.

Exercise 70. *Check this claim.*

The spectral problem we seek to solve is the following: find square-integrable functions $\varphi : \mathbf{H} \to \mathbf{C}$ such that

$$\begin{cases} \Delta\varphi = \lambda\varphi, \\ \varphi(\gamma z) = \varphi(z), \\ \displaystyle\int_M |\varphi(z)|^2 d\mu(z) < \infty. \end{cases}$$

The space $L^2(M)$ (which contains $C^\infty(M)$ as a dense subset) is a separable[1] Hilbert space with respect to the inner product

$$\langle f, g\rangle = \int_M f(z)\overline{g(z)} d\mu(z).$$

By the Friedrich's extension theorem (see Chapter 2.2), the geometric Laplacian on M admits a selfadjoint extension, which we will again denote by Δ. We want to exhibit a complete orthonormal basis $\{\varphi_k\}_{k\geq 0}$ of Δ-eigenfunctions such that each $f \in C^\infty(M)$ admits a "spectral expansion"

$$f(z) = \sum_{k\geq 0} \langle f, \varphi_k\rangle \, \varphi_k(z)$$

to play the part of Fourier analysis on M. We note here that the Laplacian is an elliptic operator; by the elliptic regularity theorem, its eigenfunctions are automatically C^∞.

It is hard (and in some cases open) to find explicit solutions to this spectral problem. Thus, what we are after is an existence theorem. The Laplacian Δ is an unbounded symmetric operator on $L^2(M)$, with domain $C^\infty(M)$, so that the usual spectral theorems for Hilbert spaces don't apply directly. It was Selberg's insight that the spectral

[1]This follows from the fact that $\mathbf{H}$ is separable, and implies that $L^2(M)$ admits a countable orthonormal basis.

theory of Δ can be formulated in terms of point-pair invariants and the Hilbert–Schmidt operators they induce.

— End of class #7 —

5.1. Spectral theorem for compact operators

A linear operator on a finite-dimensional inner product space is diagonalizable if and only if it is self-adjoint (over $\mathbf{R}$) or normal (over $\mathbf{C}$). In that case, the underlying vector space admits an orthonormal basis of eigenvectors. This is the spectral decomposition of the space. For a separable Hilbert space $\mathcal{H}$, the simplest spectral theorem is the following.

THEOREM 18 (Spectral theorem for compact operators). *Let $\mathcal{H}$ be a separable Hilbert space, and let T be a compact self-adjoint (or normal) operator on $\mathcal{H}$. Then $\mathcal{H}$ admits an orthonormal basis $\{\varphi_k\}_{k\geq 0}$ of eigenvectors of T, so that $T\varphi_k = \lambda_k\varphi_k$, whereby $\lambda_k \to 0$ as $k \to \infty$.*

PROOF. We only prove the very last statement, which implies that if λ is a nonzero eigenvalue, the corresponding eigenspace is finite-dimensional. Suppose by contradiction that there is $\varepsilon > 0$ and a subsequence (φ_j) of the above orthonormal basis of eigenfunctions such that $|\lambda_j| > \varepsilon$ for each j. Since (φ_j) is bounded and T is compact, $(T\varphi_j)$ must have a convergent subsequence. However, for each $j \neq k$, we have

$$\begin{aligned}\|T\varphi_j - T\varphi_k\|^2 &= \|\lambda_j\varphi_j - \lambda_k\varphi_k\|^2 \\ &= \langle \lambda_j\varphi_j - \lambda_k\varphi_k, \lambda_j\varphi_j - \lambda_k\varphi_k\rangle = |\lambda_j|^2 + |\lambda_k|^2 > 2\varepsilon^2,\end{aligned}$$

which shows that the convergent subsequence $(T\varphi_j)$ is not Cauchy, a contradiction. □

5.2. Hilbert–Schmidt operators

Let X be a locally compact space equipped with a positive Borel measure μ. A linear integral operator on $L^2(X,\mu)$ is provided by integration against a kernel $K \in L^2(X \times X)$. More explicitly, we have the integral operator $T_K : L^2(X) \to L^2(X)$,

$$T_K f(x) = \int_X K(x,y) f(y) d\mu(y).$$

Observe that

- $\|T_K f\| \leq \|K\|\|f\|$, and
- $\langle T_K f, g\rangle = \langle f, T_{K^*} g\rangle$, where $K^*(x,y) = \overline{K(y,x)}$.

Let $\mathcal{H}$ be a separable Hilbert space. In particular, $\mathcal{H}$ admits a countable orthonormal basis $\{e_i\}_{i\geq 1}$. The **Hilbert–Schmidt norm** of a linear operator $T : \mathcal{H} \to \mathcal{H}$ is given by

$$\|T\|_{\mathrm{HS}}^2 := \sum_{i\geq 1} \|Te_i\|_{\mathcal{H}}^2.$$

The definition does not depend on the particular choice of basis. If $\mathcal{H}$ is finite-dimensional, this is the usual Frobenius norm. If $\|T\|_{\mathrm{HS}} < \infty$, we call T a **Hilbert–Schmidt operator**.

Exercise 71. *Show that* $\|T_K\|_{HS} = \|K\|_2$.

A linear operator T on $\mathcal{H}$ is said to be compact if it maps bounded sets into compact sets. Since $\mathcal{H}$ is separable, a subset of $\mathcal{H}$ is compact if and only if it is sequentially compact, so that T is compact if and only if for every sequence $x_n \in \mathcal{H}$ of unit vectors, there is a subsequence such that $T(x_{n_k})$ is convergent.

THEOREM 19 (Hilbert–Schmidt). *Every Hilbert–Schmidt operator T on a separable Hilbert space is compact.*

SKETCH OF PROOF. Let $\{e_i\}$ be an orthonormal basis of $\mathcal{H}$. Define a sequence of linear operators T_n via

$$T_n e_i = \begin{cases} Te_i & \text{if } i \leq n, \\ 0 & \text{otherwise.} \end{cases}$$

Then each T_n has finite rank and is therefore compact. Let $v = \sum a_i e_i \in \mathcal{H}$. Using the Cauchy–Schwartz inequality, we have

$$\|(T - T_n)v\| \leq \sum_{i>n} |a_i| \|Te_i\| \leq \|v\| \|T\|_{\mathrm{HS}} < \infty.$$

Hence for every $\varepsilon > 0$, there is a compact operator T_ε such that $\|T - T_\varepsilon\| < \varepsilon$. One concludes using that any bounded linear operator that is ε-close to a compact operator must be compact. □

5.3. Point-pair invariants

Definition 72. *A* **point-pair invariant** *is a function* $k : \mathbf{H} \times \mathbf{H} \to \mathbf{C}$ *such that*

$$k(gx, gy) = k(x, y)$$

for each $g \in \mathrm{Isom}(\mathbf{H})$.

It follows that a point-pair invariant depends only on the hyperbolic distance $d_{\mathbf{H}}(z, w)$. To construct such a function, we choose a "nice" function $k : \mathbf{R} \to \mathbf{C}$ and set

$$k(z, w) := k(d_{\mathbf{H}}(z, w)).$$

For the moment, we take "nice" to mean that $k \in C_c^\infty(\mathbf{R})$, is real-valued, and even. Let $M = \Gamma \backslash \mathbf{H}$. We now introduce the **automorphic kernel**

$$K(z, w) = \sum_{\gamma \in \Gamma} k(z, \gamma w). \tag{5.1}$$

If M is compact, our assumptions on k guarantee that the sum converges. We consider the associated integral operator

$$T_K : L^2(M) \to L^2(M), \quad T_K f(z) = \int_M K(z, w) f(w) d\mu(z).$$

Remark 73 (The noncompact case)**.** *Suppose for simplicity that* $\Gamma = \mathrm{SL}_2(\mathbf{Z})$*, and that* z, w *are two points in the standard fundamental domain for* Γ*. Let* $z = x + iy$*,* $w = u + iv$ *and recall the hyperbolic distance formula*

$$\cosh d_{\mathbf{H}}(z,w) = 1 + \frac{|z-w|^2}{2yv} = \frac{(x-u)^2}{2yv} + \frac{y}{2v} + \frac{v}{2y}.$$

Suppose that y *and* v *go towards the cusp at the same rate, i.e.,* $C^{-1}v \leq y \leq Cv$ *for some constant* $C > 1$*. Then the RHS above is* $\leq C$ *as* $y, v \to \infty$*, so that for* C *small enough,* $d_{\mathbf{H}}(z,w)$ *is contained in the support of a "nice" kernel* k*. Clearly, this is true as well for* $d_{\mathbf{H}}(z, w+n)$ *for each* $n \in \mathbf{Z}$*. Hence*

$$K(z,w) = \sum_{\gamma \in \mathrm{SL}_2(\mathbf{Z})} k(z, \gamma w) \geq \sum_{n \in \mathbf{Z}} k(z, w+n).$$

Using Poisson summation, one can show that

$$K(z,w) = \sum_{\gamma \in \mathrm{SL}_2(\mathbf{Z})} k(z, \gamma w) \geq \sum_{n \in \mathbf{Z}} k(z, w+n) \sim \sqrt{yv}$$

as $y \asymp v \to \infty$*. Hence if* M *is noncompact, a nice point-pair invariant does not guarantee the convergence of (5.1).*

Proposition 74. *Let* $M = \Gamma \backslash \mathbf{H}$ *be a compact hyperbolic surface and let* k *be a "nice" point-pair invariant. Then*

(1) K *is bi-*Γ*-invariant, i.e.,* $K(\gamma_1 z, \gamma_2 w) = K(z,w)$ *for all* $\gamma_1, \gamma_2 \in \Gamma$*;*
(2) K *is symmetric, i.e.,* $K(z,w) = K(w,z)$*;*
(3) T_K *is self-adjoint and compact;*
(4) $T_K f = T_k f$ *for any* Γ*-invariant square-integrable function* $f : \mathbf{H} \to \mathbf{C}$*;*
(5) $\Delta T_K = T_K \Delta$*.*

Remark 75. *If we drop the assumption that* k *is real-valued, then* T_K *is nevertheless normal.*

Proof. The first two points follow immediately from the definition. The self-adjointness follows from the fact that k is symmetric and real-valued. By assumption again, the automorphic kernel is in $L^2(M \times M)$ so that (by Exercise 72) $\|T_K\|_{\mathrm{HS}} = \|K\|_2 < \infty$, and T_K is compact by the Hilbert–Schmidt theorem. The fourth point follows from "folding/unfolding": fix a fundamental domain $\mathcal{F}$ for Γ, then

$$\begin{aligned}\int_{\mathcal{F}} K(z,w) f(w) d\mu(w) &= \sum_{\gamma \in \Gamma} \int_{\mathcal{F}} k(z, \gamma w) f(w) d\mu(w) \\ &= \sum_{\gamma \in \Gamma} \int_{\gamma \mathcal{F}} k(z,w) f(\gamma^{-1} w) d\mu(\gamma^{-1} w) \\ &= \sum_{\gamma \in \Gamma} \int_{\gamma \mathcal{F}} k(z,w) f(w) d\mu(w) = \int_{\mathbf{H}} k(z,w) f(w) d\mu(w).\end{aligned}$$

By (4), point (5) follows immediately if $\Delta_z k(z,w) = \Delta_w k(z,w)$. Observe that for $w \in \mathbf{H}$ fixed, $k(z,w)$ is radially symmetric about w, since it depends only on $d_{\mathbf{H}}(z,w)$.

Expressing $\Delta_z k(z,w)$ in geodesic polar coordinates about w, we recover the same expression as if we express $\Delta_w k(z,w)$ in geodesic polar coordinates about z. □

THEOREM 20. *If M is a compact hyperbolic surface, then there exists a complete orthonormal basis $\{\varphi_k\}_{k\geq 0}$ for $L^2(M)$ composed of Δ-eigenfunctions.*

PROOF. Let Σ be the set of all orthonormal subsets of $L^2(M)$ composed of Δ-eigenfunctions and ordered by inclusion. By Zorn's lemma, there exists a maximal such family $S \in \Sigma$. Let $\mathcal{H} = \overline{\text{span}(S)}$. Clearly, $\mathcal{H}$ is Δ-invariant, i.e. for every $f \in \mathcal{H}$, $\Delta f \in \mathcal{H}$. If its orthogonal complement is trivial, i.e., $\mathcal{H}^\perp = \{0\}$, we are done. Suppose for contradiction, $\mathcal{H}^\perp \neq \{0\}$. Fix a nice point-pair invariant and consider the resulting integral operator T_K. Since Δ and T_K commute, the subspace $\mathcal{H}^\perp$ is also T_K-invariant. The linear restriction of T_K to $\mathcal{H}^\perp$ is compact and self-adjoint so that by the spectral theorem, $\mathcal{H}^\perp$ decomposes into a direct sum of T_K-eigenspaces E_λ, where each E_λ is finite-dimensional. Once again, since T_K and Δ commute, each E_λ is also Δ-invariant. (Indeed, $\lambda\Delta v = \Delta T v = T\Delta v$ so that $\Delta v \in E_\lambda$). Every linear operator on a finite dimensional linear space has a nontrivial eigenvector, hence E_λ contains a nontrivial Δ-eigenfunction, contradicting the maximality of S. □

Remark 76. *The theory of point-pair invariants was developed by Selberg in the more general setting of weakly symmetric spaces. A symmetric space is a connected Riemannian manifold M for which geodesic inversion in any point is a global isometry. This definition guarantees that the (Lie) group of isometries of a symmetric space acts transitively on M. A weakly symmetric space M is a connected Riemannian manifold M with a (Lie) group of isometries G acting transitively and an isometry σ that normalizes G and such that for any pair $(x,y) \in M$, there is an isometry $g \in G$ such that $\sigma(x) = g(y)$ and $g(y) = \sigma(x)$. This condition is the weakest to guarantee that the operators T_K will commute with Δ.*

Remark 77. *Selberg's 1957 seminal paper is inscribed in the study of the Laplacian (and related heat and wave equations) for compact Riemannian manifolds that was developed starting in 1949 with papers of Minakshisunderam–Pleijel and by Maass in relation to the study of automorphic forms. Until then, the study of the Laplacian had mostly been restricted to Euclidean domains.*

— End of class #8 —

5.4. Spherical functions

We used in the previous section that $k(z,w)$ can be seen as a radial function about w. Conversely, a radial function yields a point-pair invariant, so that the study of point-pair invariants amounts to the study of radial functions. We build on this idea (still following Selberg). The first observation is that every function $f : \mathbf{H} \to \mathbf{C}$ can be "radialized" about a point $z_0 \in \mathbf{H}$ as follows. First let $z_0 = i$. Then

$$f_i(z) = \frac{1}{2\pi}\int_0^{2\pi} f(k_\theta z)d\theta$$

is K-invariant and $f_i(i) = f(i)$. More generally, if $z_0 = g.i$,

$$f_{z_0}(z) = \frac{1}{2\pi}\int_0^{2\pi} f(gk_\theta g^{-1}z)d\theta$$

is $gKg^{-1} = \mathrm{Stab}_G(z_0)$-invariant and $f_{z_0}(z_0) = z_0$. We say that a function $f : \mathbf{H} \to \mathbf{C}$ is **spherical** if $\Delta f = \lambda f$ and f is radial about a point $z_0 \in \mathbf{H}$.

Lemma 78. *The space of spherical functions about z_0 with eigenvalue λ is 1-dimensional. In fact, there exists a unique function $\omega_\lambda(z; z_0)$ such that*

$$\begin{cases} \Delta_z \omega_\lambda(z; z_0) = \lambda\omega_\lambda(z; z_0) \\ \omega_\lambda(z_0; z_0) = 1. \end{cases}$$

Sketch of proof. For any radial function f about z_0, we have in geodesic polar coordinates about z_0

$$\Delta f(u) - \lambda f(u) = u(u + 1)f''(u) + (2u + 1)f'(u) - \lambda f(u) = 0.$$

This differential equation has two linear independent solutions $F_\lambda(u)$, $G_\lambda(u)$, which can be explicitly constructed. Then $f(u) = aF_\lambda(u) + bG_\lambda(u)$. We use that as $u \to 0$, $F_\lambda(u) \to 1$ and $G_\lambda(u) \to \infty$ to conclude that $b = 0$ and $a = f(z_0)$. □

Theorem 21. *Let k be a "nice" point-pair invariant. If f is a Δ-eigenfunction with eigenvalue λ, then*

$$T_k f(z) = \int_{\mathbf{H}} k(z, w)f(w)d\mu(w) = \widehat{k}(\lambda)f(z),$$

where $\widehat{k}(\lambda)$ is the Selberg transform

$$\widehat{k}(\lambda) = \int_{\mathbf{H}} k(z_0, w)w_\lambda(w; z_0)d\mu(w)$$

(which is independent of z_0).

Proof. We first replace f by its radialization. One easily sees that if $\Delta f = \lambda f$, then $\Delta f_z = \lambda f_z$. Then f_z is spherical and by Lemma 79, it must be a scalar multiple of $\omega_\lambda(\cdot, z)$. In fact, since $f_z(z) = f(z)$ and $\omega_\lambda(z; z) = 1$, we have that $f_z(w) = f(z)\omega_\lambda(w; z)$. On the other hand,

$$\begin{aligned} T_k f_z(z) &= \int_{\mathbf{H}} k(z, w)f_z(w)d\mu(w) \\ &= \frac{1}{2\pi}\int_0^{2\pi}\int_{\mathbf{H}} k(gk_\theta g^{-1}z, w)f(w)d\mu(w)d\theta \\ &= \int_{\mathbf{H}} k(z, w)f(w)d\mu(w) = T_k f(z), \end{aligned}$$

where we have used that $gKg^{-1} = \mathrm{Stab}(z)$. To conclude, pick $z_0 \in \mathbf{H}$. There is $g \in G$ such that $z = g.z_0$, and by change of variables

$$\int_{\mathbf{H}} k(z,w)\omega_\lambda(w;z)d\mu(w) = \int_{\mathbf{H}} k(z_0, g^{-1}.w)\omega_\lambda(g^{-1}w; z_0)d\mu(w)$$
$$= \int_{\mathbf{H}} k(z_0, w)\omega_\lambda(w; z_0)d\mu(w).$$

□

Remark 79. *In the statement above, we only need the point-pair invariant k to be sufficiently "nice" as to guarantee the convergence of the integrals.*

Can one recover the point-pair invariant k from the Selberg transform? We will state without proof the following formulas. Beforehand, we introduce the following parametrization for eigenvalues of Δ in $\mathbf{H}$

$$\lambda = \frac{1}{4} + r^2, \tag{5.2}$$

where $r \in \mathbf{C}$. We will more often write $\widehat{k}(r)$ in place of $\widehat{k}(\lambda)$.

Fact 80. *We have the following relations:*

$$\widehat{k}(r) = \int_{-\infty}^{\infty} g(u)e^{iru}du,$$
$$g(u) = \sqrt{2}\int_{|u|}^{\infty} \frac{k(t)\sinh(t)}{\sqrt{\cosh(t) - \cosh(u)}}dt,$$
$$k(u) = -\frac{1}{\pi\sqrt{2}}\int_{u}^{\infty} \frac{g'(t)}{\sqrt{\cosh(t) - \cosh(u)}}dt.$$

5.5. The heat equation

We will apply these formulas to solve the heat equation on compact hyperbolic surfaces. That is, we want to find a function $u : M \times \mathbf{R}_{\geq 0} \to \mathbf{R}$ (the temperature) such that $t \mapsto u(\cdot, t)$ is C^1 and

$$\begin{cases} \dfrac{\partial u}{\partial t} + \Delta u(x,t) = 0, \\ u(x,0) = f(x) \end{cases}$$

for a given continuous function f on M (the initial condition). Recall that $\Delta u(x, \cdot)$ can be understood as the deviation from the average value of u in a neighborhood of x with its precise value at x. Hence the heat equation states that the rate of temperature change at x is determined by the heat in the neighborhood of x. The solution of the heat equation for $M = \mathbf{R}$ is due to Fourier (1811) and his development of Fourier series.

We will assume that f is an eigenfunction of Δ and look for a solution of the heat equation (HE) on $\mathbf{H}$ of the form

$$u(z,t) = \int_{\mathbf{H}} p_t(z,w) f(w) d\mu(w),$$

where p_t is a point-pair invariant to which Theorem 21 applies; i.e., $u(z,t) = \widehat{p_t}(\lambda) f(z)$. That is, we look for a point-pair invariant p_t such that the above integral converges and that satisfies

$$\begin{cases} \dfrac{\partial \widehat{p_t}}{\partial t} + \lambda \widehat{p_t}(\lambda) = 0, \\ \widehat{p_0}(\lambda) = 1. \end{cases}$$

This system has the unique solution $\widehat{p_t}(\lambda) = e^{-\lambda t}$. We can now use the above formulas to derive a candidate for p_t. In fact, one obtains that the **heat kernel**

$$p_t(z,w) = \frac{\sqrt{2}}{(4\pi t)^{3/2}} e^{-t/4} \int_{d_{\mathbf{H}}(z,w)}^{\infty} \frac{u e^{-u^2/4t}}{\sqrt{\cosh(u) - \cosh d(z,w)}} du$$

yields a solution $u(z,t)$ of (HE) on $\mathbf{H}$ for all $t > 0$. Details can be found in [**1**, Chapter 3.5] including a proof that

$$p_t(z,w) = O\left(\frac{e^{-d_{\mathbf{H}}(z,w)^2/(8t)}}{t} \right). \tag{5.3}$$

The following lemma allows to derive a solution of the heat equation on M.

Lemma 81. *Let Γ be a Fuchsian group without parabolic elements (such that $M = \Gamma \backslash \mathbf{H}$ is compact). The automorphic kernel*

$$P_t(z,w) = \sum_{\gamma \in \Gamma} p_t(z, \gamma w)$$

converges absolutely.

Proof. Following (5.3), there exists a constant $C > 0$ such that

$$|P_t(z,w)| \leq C t^{-1} \sum_{\gamma \in \Gamma} e^{-d_{\mathbf{H}}(z,\gamma w)^2/(8t)} \leq C t^{-1} \sum_{n \geq 0} \#\{z' \in \Gamma w : n \leq d_{\mathbf{H}}(z,z') < n+1\} e^{-n^2/(8t)}.$$

To estimate the inner count, we build on the standard geometric argument for the Euclidean circle problem.[2] Let $\mathcal{F}$ be a fundamental domain for Γ. Observe that

$$\#\{z' \in \Gamma w : d(z,z') < T\} = \frac{\operatorname{area}(\bigcup \gamma\mathcal{F})}{\operatorname{area}(\mathcal{F})},$$

where the union is taken over all $\gamma \in \Gamma$ such that $d(z, \gamma w) < T$. This union is contained in the hyperbolic disk $B_{T+\operatorname{diam}\mathcal{F}}$ centered at z and with radius $T + \operatorname{diam}\mathcal{F}$. We have $\operatorname{area}(B_{T+\operatorname{diam}\mathcal{F}}) \leq C_1 e^T$ for some constant C_1. We conclude that

$$|P_t(z,w)| \leq C_2 \frac{1}{t} \sum_{n\geq 0} e^{n-n^2/(8t)} < \infty.$$

□

Remark 83. *This shows that for the hyperbolic circle problem, $N_T = \#\{z' \in \Gamma w : d_{\mathbf{H}}(z,z') < T\} = O(e^T)$. Gauss' argument (i.e., Proposition 83 in the footnote) does not directly yield the stronger result $|\Omega_T| \sim e^T$ (as $T \to \infty$). This is because in the hyperbolic case, both area and perimeter of a hyperbolic disk of radius T are of the same order e^T, and hence one obtains no cancellation by bounding the contribution of the boundary circle by an annulus.*

— End of class #9 —

Proposition 84. *If M be a compact hyperbolic surface, then the solution of the heat equation is unique.*

PROOF. Let u_1 and u_2 be two solutions to (HE) and set $v = u_1 - u_2$. Then

$$\frac{\partial}{\partial t}\|v(\cdot,t)\|^2 = 2\left\langle v(\cdot,t), \frac{\partial v}{\partial t}(\cdot,t)\right\rangle = -2\left\langle \Delta v(\cdot,t), v(\cdot,t)\right\rangle \leq 0.$$

---

[2]Consider

$$N_T = \#\{\xi \in \mathbf{Z}^2 : \|\xi\| < T\},$$

with $\|\xi\|^2 = u^2 + v^2$, $\xi = \left(\begin{smallmatrix} u \\ v \end{smallmatrix}\right)$. The classical Euclidean circle problem consists in estimating precisely the size of Ω_T. It was originally motivated by the study of $r_2(n) = \#\{(a,b) \in \mathbf{Z}^2 : n = a^2 + b^2\}$, the number of representations of n as a sum of squares. The function $r_2(n)$ fluctuates a lot, but its average

$$\sum_{n\leq N} r_2(n) = \#\{(a,b) \in \mathbf{Z}^2 : a^2 + b^2 \leq N\} = \Omega_{\sqrt{N}}.$$

is better behaved.

Proposition 82. *As $T \to \infty$, $N_T = \pi T^2 + O(T)$.*

PROOF. Consider the tessellation of $\mathbf{R}^2$ by unit squares with centers at each $\xi \in \mathbf{Z}^2$. The area of the union of all squares with center in within T of the origin is exactly N_T. In this way, one sees that the difference $|N_T - \pi T^2|$ is bounded above by the area of the annulus with center 0 and radius $T - \frac{\sqrt{2}}{2} \leq t \leq T + \frac{\sqrt{2}}{2}$. □

This estimate is very rough; indeed, there is a lot of cancellation in the over- vs. undercounting of unit squares around the boundary of the disk of radius T. Numerical computations indicate that the discrepancy $|N_T - \pi T^2|$ grows as T^θ for some θ. The conjecture (Hardy, 1917), the so-called Gauss' circle problem, is that $N_T = \pi T^2 + O(T^\theta)$ with $\theta = 1/2 + \varepsilon$ for any $\varepsilon > 0$. The current world record goes to Bourgain and Watt (2017) with $\theta \approx 0.63$.

Since $\lim_{t\to 0^+} v(\cdot,t)=0$, we conclude that $v=0$. □

Lemma 85. *The heat operators*

$$\mathcal{P}_t f(z)=\int_M P_t(z,w)f(w)d\mu(w),$$

$t>0$, have the following properties

(1) *$\mathcal{P}_t$ is continuous in t for all $t>0$;*
(2) *$\mathcal{P}_t\to 1$ as $t\to 0$;*
(3) *$\mathcal{P}_t$ is compact and self-adjoint;*
(4) *$\mathcal{P}_s\circ\mathcal{P}_t=\mathcal{P}_{s+t}=\mathcal{P}_t\circ\mathcal{P}_s$.*

PROOF. Exercise. □

Proposition 86. *There exists a complete orthonormal basis $\{\varphi_j\}_{j\geq 0}\subset C^\infty(M)$ in $L^2(M)$ such that $\mathcal{P}_t\varphi_j=\eta_j\varphi_j$ and where*

$$1=\eta_0>\eta_1\geq\cdots\to 0$$

and $\eta_j>0$ for all $j\geq 0$.

PROOF. Let first $t=1$. The existence of $\{\varphi_j\}$ with $\mathcal{P}_1\varphi_j=\eta_j\varphi_j$ follows from the spectral theorem for $\mathcal{P}_t$. Let now $q\in\mathbf{N}$ and $t=\frac{1}{q}$. Once again by the spectral theorem for compact operators, $\mathcal{P}_{1/q}$ admits a complete orthonormal basis of eigenfunctions. Suppose that φ is an eigenfunction of $\mathcal{P}_{1/q}$, $q\in\mathbf{N}$, with $\mathcal{P}_{1/q}\varphi=\mu\varphi$. Then

$$\mathcal{P}_{1/q}^q\varphi=\mu^q\varphi=\mathcal{P}_1\varphi,$$

that is φ is also an eigenfunction of $\mathcal{P}_1$. Since both families are complete, we conclude that $\mathcal{P}_{1/q}$ and $\mathcal{P}_1$ have the same eigenfunctions. This proves the claim for all $t\in\mathbf{Q}_{>0}$ considering that $\mathcal{P}_{p/q}\varphi_j=\mathcal{P}_{1/q}^p\varphi_j=\eta_j^{p/q}\varphi_k$. By the continuity of the heat kernels in t, this extends to all $t\in\mathbf{R}_{>0}$.

Observe that $\lim_{t\to 0}\mathcal{P}_t=1$ implies that $\eta_j^t\to 1$ as $t\to 0^+$. This proves that $\eta_j>0$ for all j's. On the other hand, for $\mathcal{P}_t\varphi=\eta^t\varphi$, we have

$$\frac{\partial}{\partial t}\langle\mathcal{P}_t\varphi,\varphi\rangle=-\langle\Delta\mathcal{P}_t\varphi,\varphi\rangle=-\eta^t\langle\Delta\varphi,\varphi\rangle=\log\eta\eta^t\|\varphi\|^2.$$

Hence

$$\log\eta=-\frac{\Delta\varphi,\varphi}{\|\varphi\|^2}=-\frac{\|\nabla\varphi\|^2}{\|\varphi\|^2}\leq 0,$$

which shows $\eta\leq 1$, and $\eta=1$ only if φ is constant. We show that $\eta_0=1$ is realized. Let φ be a constant function. Then $\tilde{\varphi}(\cdot,t)=\varphi$ is a solution to the heat equation with initial condition φ. This is also the case of $\mathcal{P}_t\varphi$ so that by the uniqueness of the solution to the heat equation, we have $\mathcal{P}_t\varphi=\tilde{\varphi}(\cdot,t)=\varphi$. □

Since $\mathcal{P}_t$ and Δ commute, $\{\varphi_j\}_{j\geq 0}$ as above is also a complete orthonormal basis for Δ. Moreover, we have

$$0=\left(\frac{\partial}{\partial t}+\Delta\right)\mathcal{P}_t\varphi_j=(\log\eta_j)\eta_j^t\varphi_j+\eta_j^t\varphi_j,$$

implying $\Delta\varphi_j = -\log\eta_j\varphi_j$.

The eigenvalue $\lambda_0 = 0$ corresponds to the normalized constant function $\varphi_0 = \text{area}(M)|^{-1/2}$. Indeed,

$$\|\varphi_0\|^2 = \int_M \varphi_0 = \frac{1}{\text{area}(M)}\int_M = 1.$$

(This also holds when M is noncompact but still of finite area.) The strict inequality that follows implies that $\lambda_0 = 0$ has multiplicity 1. Indeed, if φ is a nonconstant eigenfunction of Δ, then

$$\lambda\|\varphi\|^2 = \langle\Delta\varphi, \varphi\rangle = \|\nabla\varphi\|^2 > 0,$$

which shows that $\lambda_j > 0$ for all $j \geq 1$.

— End of class #10 —

CHAPTER 6

Selberg's trace formula

6.1. The trace formula for compact hyperbolic surfaces

Let M be a compact hyperbolic surface with uniformizing Fuchsian group Γ. Let $\{\varphi_j\}_{j\geq 0}$ be an orthonormal basis for $L^2(M)$, and let $k \in C_c^\infty(\mathbf{R})$ be an even function. Then the automorphic kernel K admits the spectral expansion

$$K(z,w) = \sum_{j\geq 0} \widehat{k}(\lambda_j)\varphi_j(z)\overline{\varphi}_j(w).$$

We wish to "take the trace"

$$\int_M K(z,z)d\mu(z) = \sum_{j\geq 0} \widehat{k}(\lambda_j) \int_M |\varphi_j(z)|^2 d\mu(z) = \sum_{j\geq 0} \widehat{k}(\lambda_j). \tag{6.1}$$

Exercise 87. *Observe that if $k(\rho) = O(e^{-\rho(1+\varepsilon)})$, then*

$$\int_{\mathbf{H}} k(z,i)d\mu(z) = \int_0^\infty k(\rho)\sinh\rho d\rho$$

converges, as does the corresponding trace $\int K(z,z)d\mu(z)$.

We will compute the LHS by grouping the different terms in

$$\int_M K(z,z)d\mu(z) = \int_M \sum_{\gamma\in\Gamma} k(z,\gamma z)d\mu(z)$$

according to the geometric interpretation of the classification of motions. We are assuming here that all nontrivial elements of Γ are hyperbolic. Recall that conjugacy classes of hyperbolic elements correspond to closed geodesics in M. Let

$$\{\gamma\} = \{\tau\gamma\tau^{-1} : \tau \in \Gamma\}$$

denote the Γ-conjugacy class of $\gamma \in \Gamma$. Observe that one is counting γ with multiplicity if one allows for $\tau \in Z(\gamma) = \{\tau \in \Gamma : \tau\gamma = \gamma\tau\}$. By folding/unfolding,

$$\begin{aligned}\int_M K(z,z)d\mu(z) &= \int_M k(0)d\mu(z) + \sum_{\substack{\{\gamma\}\\ \gamma \text{ hyp}}} \int_{\Gamma\backslash\mathbf{H}} \sum_{\tau\in Z(\gamma)\backslash\Gamma} k(z,\gamma z)d\mu(z)\\ &= \operatorname{area}(M)k(0) + \sum_{\substack{\{\gamma\}\\ \gamma \text{ hyp}}} \int_{Z(\gamma)\backslash\mathbf{H}} k(z,\gamma z)d\mu(z). \end{aligned}\tag{6.2}$$

We call (6.1) the **pretrace formula** and (6.2) the **geometric side** of the trace formula. As it turns out, the fundamental domains $Z(\gamma)\backslash\mathbf{H}$ are very easy to describe geometrically.

Proposition 88. *For each hyperbolic γ, let $\widetilde{\gamma}$ denote the unique geodesic in $\mathbf{H}$ that connects the two fixed points of γ. Then $Z(\gamma) = \mathrm{Stab}_\Gamma(\widetilde{\gamma})$. In particular, it is an infinite cyclic group.*

PROOF. Let $\tau \in Z(\gamma)$. Then $\tau\widetilde{\gamma} = \gamma\tau\widetilde{\gamma}$ implies that γ fixes $\tau\widetilde{\gamma}$. By uniqueness, we have $\tau\widetilde{\gamma} = \widetilde{\gamma}$. Recall from the proof of Proposition 63 that $\mathrm{Stab}_\Gamma(\widetilde{\gamma}) = \langle\gamma_0\rangle$ is an infinite cyclic group. In particular, we have $\gamma = \gamma_0^k$ for some k, and $\mathrm{Stab}_\Gamma(\widetilde{\gamma}) \subset Z(\gamma)$. ☐

Definition 89. *We say that $\gamma_0 \in \Gamma$ is* **primitive** *if there is no $\gamma \in \Gamma$ such that $\gamma_0 = \gamma^k$ for some $k \in \mathbf{N}$.*

Suppose that $\widetilde{\gamma} = i\mathbf{R}_{>0}$, and parametrize the cyclic generator $\langle\gamma_0\rangle = Z(\gamma)$ by

$$\gamma_0 = \begin{pmatrix} e^{\ell/2} & 0 \\ 0 & e^{-\ell/2} \end{pmatrix}.$$

Then a fundamental domain for $Z(\gamma)$ is given by the horizontal strip between i and $\gamma_0(i) = e^\ell i$, of width $d_{\mathbf{H}}(i, e^\ell i) = \ell$.

Exercise 90. *The quantity ℓ is called the translation length of γ. More generally, the translation length of $\gamma \in \Gamma$ is defined by*

$$\ell_\gamma = \inf_{z\in\mathbf{H}} d_{\mathbf{H}}(z, \gamma.z).$$

1. *Show that if γ is either elliptic or parabolic, then $\ell_\gamma = 0$, and that if γ is hyperbolic and conjugate to γ_0, then $\ell_\gamma = \ell$.*
2. *Observe that for γ hyperbolic,*

$$\ell_\gamma = 2\cosh^{-1}\left(\frac{|\mathrm{tr}\gamma|}{2}\right).$$

Hence ℓ_γ depends only on the conjugacy class of γ. Conclude that ℓ_γ is the length of the closed geodesic in $\Gamma\backslash\mathbf{H}$ that corresponds to $\{\gamma\}$.

Lemma 91. *The integral*

$$\int_{Z(\gamma)} k(z, \gamma z)\,d\mu(z)$$

depends only on the conjugacy class of γ in $G = \mathrm{SL}_2(\mathbf{R})$.

PROOF. Let $\gamma' = g^{-1}\gamma g$. Then $Z(\gamma') = g^{-1}Z(\gamma)g$ and

$$\int_{Z(\gamma')\backslash\mathbf{H}} k(z, \gamma' z)\,d\mu(z) = \int_{g^{-1}Z(\gamma)g\backslash\mathbf{H}} k(z, g^{-1}\gamma g z)\,d\mu(z) = \int_{Z(\gamma)\backslash\mathbf{H}} k(z, \gamma z)\,d\mu(z).$$

☐

THEOREM 22 (Selberg's trace formula). *For a compact hyperbolic surface M, an even test function $g \in C_c^\infty(\mathbf{R})$ with Fourier transform h, we have*

$$\sum_{j\geq 0} h(r_j) = \frac{\text{area}(M)}{4\pi}\int_{-\infty}^{\infty} h(t)t\tanh(\pi t)dt + \sum_{\substack{\{\gamma_0\} \\ \gamma_0 \text{ prim hyp}}} \frac{\ell_{\gamma_0}}{2}\sum_{n\geq 1}\frac{g(n\ell_{\gamma_0})}{\sinh(\frac{n\ell_{\gamma_0}}{2})}, \tag{6.3}$$

where the sum on the LHS is indexed over $\lambda_j = \frac{1}{4}+r_j^2$.

PROOF. By Fact 81, if $g \in C_c^\infty(\mathbf{R})$ is even, there is a nice kernel function k such that g comes from k, and the Schwartz function $h = \widehat{g}$ is the Selberg transform of k. Consider the geometric side of the pretrace formula. Using the inverse Selberg transform, we have

$$k(0) = -\frac{1}{\sqrt{2\pi}}\int_0^\infty \frac{g'(u)}{\sinh(u/2)}du = \frac{1}{4\pi}\int_{-\infty}^{\infty} h(t)t\tanh(\pi t)dt.$$

Then for any hyperbolic $\gamma = \gamma_0^n \in \Gamma$, we can choose $g \in G$ such that $g\widetilde{\gamma} = i\mathbf{R}_{>0}$ and conclude with Lemma 92 that

$$\int_{Z(\gamma)\backslash\mathbf{H}} k(z,\gamma z)d\mu(z) = \int_{-\infty}^{\infty}\int_1^{e^{n\ell}} k(z,e^{n\ell}z)d\mu(z).$$

Let $U(\cosh d_{\mathbf{H}}(z,e^{n\ell}z)) = k(z,e^{n\ell}z)$. Then

$$\begin{aligned}\int_{-\infty}^{\infty}\int_1^{e^{n\ell}} k(z,e^{n\ell}z)d\mu(z) &= \int_1^{e^{n\ell}}\frac{dy}{y}\int_{\mathbf{R}} U\left(1+2\sinh^2(\tfrac{n\ell}{2})(x^2+1)\right)dx \\ &= \frac{\ell}{\sinh\frac{n\ell}{2}}\int_{\sinh^2\frac{n\ell}{2}}^{\infty}\frac{U(1+2u)}{\sqrt{u-\sinh^2\frac{n\ell}{2}}}du \\ &= \frac{\ell}{\sqrt{2}\sinh\frac{n\ell}{2}}\int_\ell^\infty \frac{k(\rho)\sinh(\rho)d\rho}{\sqrt{\cosh\rho-\cosh(n\ell)}} \\ &= \frac{\ell}{2\sinh\frac{n\ell}{2}}g(n\ell),\end{aligned}$$

where g is the Fourier inverse of the Selberg transform. This concludes the proof of (6.3). □

Exercise 92. *Show that if γ is parabolic, then the domain of integration $Z(\gamma)\backslash\mathbf{H}$ is a vertical strip, and if γ is elliptic, it is a sector based at i.*

— End of class #11 —

6.2. Length spectrum and Huber's theorem

The **length spectrum** $\mathcal{L}$ the collection of lengths of all closed geodesics on M, ordered by size and accounting for multiplicities. This set is discrete, as follows from the following proposition.

Proposition 93. *Let M be a compact hyperbolic surface. Let $\pi(L)$ denote the number of closed geodesics on M with length bounded by L. We have $\pi(L) = O(e^L)$.*

Proof. Let γ be a closed oriented geodesic of length $\leq L$ on M. Fix $p \in M$ and a lift $z \in \mathbf{H}$ with respect to the universal covering $\mathbf{H} \to \Gamma\backslash\mathbf{H}$. Let γ_1 be a freely homotopic closed geodesic passing through p. Its length ℓ_1 is bounded by $\ell_1 \leq L + 2\text{diam}(M)$. It lifts to a geodesic segment in $\mathbf{H}$ connecting z to z' with $d(z, z') \leq L + 2\text{diam}(M)$. We conclude that the number of closed oriented geodesics of length $\leq L$ is bounded above by the number of points in Γz that lie within a (hyperbolic) ball of radius $L + 2\text{diam}(M)$ centered at z. We can now conclude with the argument given in the proof of Lemma 82. □

The right-most term in the trace formula (6.3) can be rewritten

$$\sum_{\ell \in \mathcal{L}} \frac{\ell_0 g(\ell)}{2\sinh(\ell/2)},$$

where ℓ_0 denotes the primitive length. If we apply the trace formula to the heat kernel[1] by choosing $h(r) = e^{-tr^2}$. Its Fourier inverse is

$$g(u) = \frac{1}{\sqrt{4\pi t}} e^{-u^2/(4t)}$$

so that the trace formula is

$$\sum_{j \geq 0} e^{-t\lambda_j} = \frac{\text{area}(M)}{4\pi} e^{-t/4} \int_{-\infty}^{\infty} e^{-tr^2} r \tanh(\pi r) dr + e^{-t/4} \sum_{\ell \in \mathcal{L}} \frac{\ell_0 e^{-\ell^2/(4t)}}{4\sqrt{\pi t}\sinh(\ell/2)}, \qquad (6.4)$$

where on the LHS, we recognize the spectral partition function as seen in 2.3. This is the idea behind Huber's theorem.

Theorem 23 (Huber, 1959). *Two compact hyperbolic surfaces M and M' have the same spectrum of the Laplacian if and only if they have the same area and the same length spectrum.*

Proof. Recall that one may recover each eigenvalue and its multiplicity from the spectral partition function. The same can be done for the length spectrum along the following lines. Let μ_0 be the multiplicity of the shortest geodesic length ℓ_0. Then

$$\lim_{t \to 0} e^{\omega^2/(4t)} \sum_{\ell \in \mathcal{L}} \frac{\ell_0 e^{-\ell^2/(4t)}}{2\sinh(\ell/2)} = \begin{cases} 0 & \omega < \ell_0, \\ \frac{\mu_0 \ell_0}{2\sinh(\ell_0/2)} & \omega = \ell_0, \\ \infty & \omega > \ell_0. \end{cases}$$

Suppose we understand the Laplace spectrum of M. We claim we can recover the area of M and its length spectrum. By the above discussion, it suffices to show that we the Laplace spectrum determines the area. By integration by parts,

$$\int_{-\infty}^{\infty} e^{-tr^2} r \tanh(\pi r) dr = \frac{1}{t} + O(1)$$

[1] Observe from the discussion of Section 5.5 that $h(r)$ is indeed the Selberg transform of the heat kernel.

as $t \to 0$. Denote the hyperbolic contribution in (6.4) by $\sum_{\ell \in \mathcal{L}} \psi(\ell)$. We leave it to the reader to check that $\psi(u) = O(\frac{e^{-c/t}e^{-u}}{\sqrt{t}u \log u})$. It follows that $\sum_{\ell \in \mathcal{L}} \psi(\ell) \to 0$ as $t \to 0$. We conclude that as $t \to 0$,

$$\sum_{j \geq 0} e^{-t\lambda_j} = \frac{\text{area}(M)}{4\pi t} + O(1). \tag{6.5}$$

That is, the spectrum of the Laplacian determines the area. Conversely, suppose that we know the length spectrum of M. That is, we know the function

$$\sum_{j \geq 0} e^{-t\lambda_j} - \frac{\text{area}(M)}{4\pi} e^{-t/4} \sigma(t) = \text{ length spectrum contribution,} \tag{6.6}$$

where $\sigma(t) = O(t^{-1})$. Multiplying by $e^{\omega t}$ with $\omega < 1/4$ and taking $t \to \infty$, the second term goes to 0 and we recover the multiplicity and location of each eigenvalue $\leq 1/4$ (if they exist). Multiplying (6.6) by $e^{t/4}\sigma(t)^{-1}$, we now recover area(M) from the knowledge of the length spectrum and the Laplacian spectrum below 1/4. We proceed in the same way to recover the larger eigenvalues from the spectral partition function. □

The asymptotic relation (6.5) says that the spectrum of $\Delta|_{L^2(M)}$ determines the area of M. By the Gauss–Bonnet Theorem 6, a compact hyperbolic surface has area

$$\text{area}(M) = -2\pi\chi(M) = 4\pi(g-1).$$

Hence: two compact hyperbolic surfaces with the same Laplacian spectrum (same eigenvalues with same multiplicities) are topologically equivalent (i.e., homeomorphic). Vignéras showed that if two compact hyperbolic surfaces have the same Laplacian spectrum, they are however not necessarily geometrically equivalent (i.e., isometric).

6.3. Weyl's law

Weyl's law (1911) is one of the earliest result connecting the spectrum of the Laplacian to the geometry of the underlying space. Let $\Omega \subset \mathbf{R}^2$ be a bounded measurable domain. Consider the Dirichlet problem

$$\begin{cases} \Delta u = \lambda u \text{ on } \Omega \\ u|_{\partial\Omega} = 0. \end{cases}$$

Its solution yields a discrete spectrum $\{\lambda_j\}$. Let $N(\lambda) = \#\{j \geq 0 : \lambda_j \leq \lambda\}$. Then Weyl's law says that

$$\lim_{\lambda \to \infty} \frac{N(\lambda)}{\lambda} = \frac{\text{area}(\Omega)}{4\pi}.$$

Weyl's law was later shown to hold for all compact Riemannian manifolds. For a compact hyperbolic surface M, the asymptotic

$$N(\lambda) \sim \frac{\text{area}(M)}{4\pi}\lambda$$

as $\lambda \to \infty$ is easily deduced from Eq. (6.5) via the following Tauberian theorem, which is a reformulation of a theorem of Hardy and Littlewood (1916).

THEOREM 24. *Given a sequence (a_n) of nonnegative numbers whose power series satisfies*

$$\sum_{n\geq 1} a_n x^n \sim \frac{1}{1-x}$$

as $x \to 1^-$, we have

$$\sum_{n\leq N} a_n \sim N$$

as $N \to \infty$.

PROOF OF WEYL'S LAW. By (6.5), we have that

$$\frac{4\pi}{\mathrm{area}(M)} \sum_{j\geq 0} e^{-\lambda_j t} \sim \frac{1}{t} \sim \frac{1}{1-e^{-t}},$$

as $t \to 0^+$, and the Tauberian theorem of Hardy–Littlewood implies Weyl's law. ☐

— End of class #12 —

CHAPTER 7

Interlude: the hyperbolic circle problem

Let Γ be a Fuchsian group. The hyperbolic circle problem consists in estimating precisely the number

$$N_R = \#\{z' \in \Gamma z : d_{\mathbf{H}}(z', w) < R\},$$

where $z, w \in \mathbf{H}$ are fixed, and $R \to \infty$. We reformulate this count as

$$N_R = \sum_{\gamma \in \Gamma} 1_R(d_{\mathbf{H}}(w, \gamma z)),$$

where 1_R is the characteristic function

$$1_R(t) = \begin{cases} 1 & \text{if } |t| \le R, \\ 0 & \text{if } |t| > R, \end{cases}$$

and consider its (formal) spectral expansion

$$N_R = \sum_{j \ge 0} h_R(r_j) \varphi_j(z) \overline{\varphi_j(w)}, \tag{7.1}$$

where h_R is the Selberg transform of the characteristic function 1_R. (We purposefully leave the question of convergence aside for now.)

The first term in the sum (7.1) corresponds to the (isolated) eigenvalue λ_0 and the corresponding constant eigenfunction $\varphi_0 = (\text{area}(M))^{-1/2}$. (Recall the discussion at the end of Section 5.5.) We have

$$\begin{aligned} h_R &= \int_{\mathbf{H}} 1_R(d_{\mathbf{H}}(z, i)) \omega_0(z; i) d\mu(z) \\ &= \int_{\mathbf{H}} 1_R(d_{\mathbf{H}}(z, i)) d\mu(z) = \int_{B_R(i)} d\mu(z) = \text{area}(B_R) \sim e^R. \end{aligned}$$

Hence the first term in the sum (7.1) is

$$\frac{\text{area}(B_R)}{\text{area}(M)}.$$

Recall that the eigenvalues of Δ are parametrized according to the convention $\lambda_j = \frac{1}{4} + r_j^2$ (with $r_j \in \mathbf{C}$). This leads us to distinguish the so called "small eigenvalues" $\lambda_j < \frac{1}{4}$ corresponding to the complex eigenparameters

$$r_j \in \left[-\tfrac{i}{2}, \tfrac{i}{2}\right] \setminus \{0\} \qquad \frac{1}{2} = |r_0| > |r_1| \ge \cdots \ge |r_k| > 0.$$

from the eigenvalues $\lambda_j \ge \frac{1}{4}$ corresponding to the real eigenparameters

$$(r_j)_{j \ge k+1} \subset \mathbf{R} \qquad 0 \le |r_{k+1}| \le |r_{k+2}| \le \dots$$

Observe that since the spectrum of Δ is discrete, there can only be finitely many "small eigenvalues."

Computing[1] the Selberg transform h_R of 1_R, we obtain from (7.1) an expression of the form

$$N_R = \frac{\text{area}(B_R)}{\text{area}(M)} + \sum_{\lambda_j < \frac{1}{4}} c_j e^{R(\frac{1}{2}+|r_j|)} + O\left(\sum_{\lambda_j \geq \frac{1}{4}} \frac{1}{|r_j|^{3/2}} e^{R/2}\right) + O\left(e^{R/2}\right), \tag{7.2}$$

where $c_j \in \mathbf{C}$ are constants depending on $|r_j|, \varphi_j(z)\overline{\varphi_j(w)}$. Ideally the sum indexed by $\lambda_{\geq \frac{1}{4}}$ would converge such that the two last terms amount to $O(e^{R/2})$ and we could conclude that

$$N_R \sim \frac{\text{area}(B_R)}{\text{area}(M)} \tag{7.3}$$

as $R \to \infty$, and in fact more precisely that, for R large, if $\lambda_1 < \frac{1}{4}$,

$$\left|\frac{N_R}{\text{area}(B_R)} - \frac{1}{\text{area}(M)}\right| \ll e^{R(|r_1|-1/2)}, \tag{7.4}$$

while if $\lambda_1 \geq \frac{1}{4}$,

$$\left|\frac{N_R}{\text{area}(B_R)} - \frac{1}{\text{area}(M)}\right| \ll e^{-R/2}. \tag{7.5}$$

However, we do not have convergence. Indeed,

$$\sum_{\lambda_j \geq \frac{1}{4}} \frac{1}{|r_j|^{3/2}} > \sum_{\lambda_j \geq \frac{1}{4}} \frac{1}{\lambda_j^{3/4}}$$

and upon considering the partial sums

$$\sum_{\frac{1}{4} \leq \lambda_j \leq T} \frac{1}{\lambda_j^{3/4}} \geq T^{-3/4} \sum_{\frac{1}{4} \leq \lambda_j \leq T} 1,$$

Weyl's law implies that the sum diverges. In practice, to attack the hyperbolic circle problem, one replaces the characteristic function 1_R by a smooth approximation. This yields stronger estimates for the Selberg transform h_R from which one recovers (7.3) and (7.4), but this comes at the cost of a weaker version of (7.5), namely that if $\lambda_1 \geq \frac{1}{4}$,

$$\left|\frac{N_R}{\text{area}(B_R)} - \frac{1}{\text{area}(M)}\right| \ll e^{-R/3}.$$

This result was derived by various mathematicians in the 1950/60s (Selberg, Huber, Delsarte...), and has not yet been improved upon. The (open) conjecture is that

$$\left|\frac{N_R}{\text{area}(B_R)} - \frac{1}{\text{area}(M)}\right| \ll e^{R(-\frac{1}{2}+\varepsilon)}$$

[1]This is done, e.g., in Lemma 2.4 of Fernando Chamizo, *Some applications of large sieve in Riemann surfaces*, Acta Arithmetica, 1996.

for any $\varepsilon > 0$.

CHAPTER 8

On λ_1

By now, we know that the trivial eigenvalue $\lambda_0 = 0$ comes associated with area(M). From the hyperbolic circle problem, we saw that if M has small eigenvalues $\lambda_j < \frac{1}{4}$, then the size of λ_1, called the **spectral gap**, controls the speed of convergence in (7.3).

Proposition 94. *We have*

$$\lambda_1(M) = \inf \frac{\int_M |\nabla f|^2 d\mu}{\int_M f^2 d\mu},$$

where the infinimum is taken over all functions $f \in \mathrm{Dom}(\Delta)$ such that $\int_M f(z)d\mu(z) = 0$.

Exercise 95. *We have the spectral decomposition $L^2(M) = \mathbf{C}\cdot 1 \oplus (\mathbf{C}\cdot 1)^\perp$, where $\mathbf{C}\cdot 1$ denotes the eigenspace of constant functions to the base eigenvalue $\lambda_0 = 0$. Show that $f \in (\mathbf{C}\cdot 1)^\perp$ if and only if $\int_M f(z)d\mu(z) = 0$.*

Proof. Since $f \in (\mathbf{C}\cdot 1)^\perp$, its spectral expansion reduces to

$$f(z) = \sum_{j\geq 1} \langle f, \varphi_j \rangle \varphi_j(z).$$

We can now check that

$$\int_M |f(z)|^2 d\mu(z) = \sum_{j\geq 1} | \langle f, \varphi_j \rangle |^2$$

and

$$\int_M |\nabla f|^2 d\mu(z) = \sum_{j\geq 1} \lambda_j | \langle f, \varphi_j \rangle |^2 \geq \lambda_1 \int_M |f(z)|^2 d\mu(z).$$

Finally, if $\Delta f = \lambda_1 f$, then by integration by parts, one has $\lambda_1 \|f\|_2 = \langle \Delta f, f \rangle = \|\nabla f\|^2$. ☐

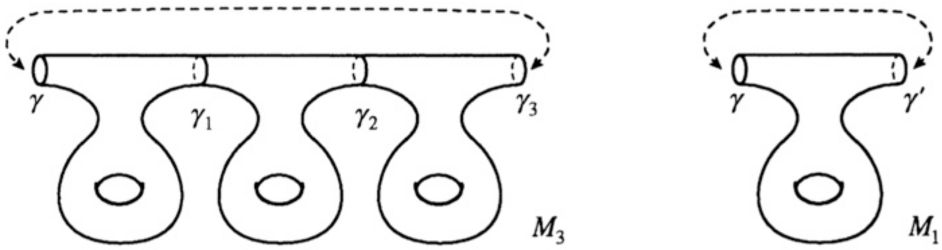
8.1. Arbitrarily small spectral gap

One sees from (7.4) and (7.5) that it is undesirable that λ_1 be too small. This next result shows that one can always find a surface with arbitrarily small spectral gap.

Theorem 25. *For any $\varepsilon > 0$, there exists a compact hyperbolic surface M such that $\lambda_1(M) < \varepsilon$.*

We will sketch two proofs of Theorem 25, a geometric one due to Cheeger [4] and an analytic one due to Selberg (unpublished).

Sketch of proof of Theorem 25, after Cheeger. Let $M_1 = \Sigma_2$ and choose a simple (no selfintersections) closed geodesic on M_1 such that $M_1 \setminus \gamma$ is still connected. Let M_i be the surface obtained by gluing i copies of M_1 along γ as follows (figure taken from [3]).



Decompose $M_{2i} = A \cup B$ such that $\text{area}(A) = \text{area}(B)$ and A meets B in two copies of γ. Let U be a small neighborhood of $A \cap B$ and set

$$f = \begin{cases} 1 & \text{on } A - U, \\ -1 & \text{on } B - U, \\ \text{linear transition} & \text{on } U. \end{cases}$$

Then $\int_M f = 0$, $\int_M |\nabla f|^2$ is bounded, and

$$\int_M |f|^2 \approx \text{area}(M_{2i}) = 2i \cdot \text{area}(M).$$

By Proposition 95, we have $\lambda_1(M_i) \to 0$ as $i \to \infty$. □

This construction led Cheeger to introduce the **Cheeger constant** with the aim of bounding λ_1 from below.

Definition 96. *Let M be a compact connected Riemannian manifold. The Cheeger constant is*

$$h(M) := \inf \frac{\text{vol}(\partial D)}{\text{vol}(D)},$$

where the infinimum is taken over all domains (bounded regular sets) $D \subset M$ such that $\text{vol}(D) \leq \frac{1}{2}\text{vol}(M)$.

Theorem 26 (Cheeger's inequality).

$$\lambda_1(M) \geq \frac{h(M)^2}{4}.$$

In the words of Brooks [2], surfaces that are "short and fat" thus have larger spectral gap. Cheeger's inequality also shows that to obtain an arbitrarily small spectral gap, one needs surfaces that are "long and thin" as in the construction above. A converse was obtained by Buser, who showed that there is a constant C such that

$$\lambda_1(M) \leq C(h(M)^2 + h(M)).$$

Remark 97. *A full proof of Theorem 25 building on Buser's inequality is given* [1, Chapter 3.10.1].

Before passing on to Selberg's proof, a few preliminary comments on characters are required. A character on a group G is a group homomorphism $\chi : G \to \mathbf{S}^1$. The set of all characters on G forms the group $\widehat{G} = \mathrm{Hom}(G, \mathbf{S}^1)$. For a Fuchsian group, we have the following identifications (without proof)

$$\widehat{\Gamma} = \mathrm{Hom}(\Gamma, \mathbf{S}^1) = \mathrm{Hom}(\Gamma/[\Gamma,\Gamma], \mathbf{S}^1) \cong \mathrm{Hom}(H_1(M), \mathbf{S}^1) \cong \mathrm{Hom}(\mathbf{Z}^{2g}, \mathbf{S}^1) \cong \mathbf{T}^{2g},$$

where g is the genus of M. (Geometrically, this corresponds to assigning to each generating cycle on the surface a value $\theta \in \mathbf{S}^1 \cong \mathbf{T} = \mathbf{R}/\mathbf{Z}$.) In other words, the characters on Γ are parametrized by the $2g$-dimensional torus $\mathbf{T}^{2g}$. The trivial character $\chi \equiv 1$ is in one-to-one correspondence with the vector $0 \in \mathbf{T}^{2g}$.

Sketch of proof, after Selberg. Let $M = \Gamma\backslash\mathbf{H}$ be a compact hyperbolic surface and fix $\Theta \in \mathbf{T}^{2g}$. Let χ_Θ denote the associated character. We consider the following modified spectral problem: find solutions $f : \mathbf{H} \to \mathbf{C}$ of

$$\begin{cases} \Delta f = \lambda f \\ \displaystyle\int_M |f|^2 < \infty \\ f(\gamma z) = \chi_\Theta(\gamma) f(z). \end{cases}$$

(Note that although f does not descend to a function on M, $|f|$ does, and hence $\int_M |f|^2$ is well-defined.) We will rely on the following facts.

- The modified spectral problem admits a complete resolution with spectrum
$$0 \le \lambda_1(\Theta) \le \lambda_2(\Theta) \le \lambda_3(\Theta) \le \dots;$$
- The bottom eigenvalue $\lambda_0(\Theta)$ is continuous in Θ. Hence for any $\varepsilon > 0$, there exists $\delta > 0$ such that $\|\Theta\| < \delta$ implies $\lambda_0(\Theta) < \varepsilon$;
- $\lambda_0(\Theta) = 0$ if and only if $\Theta = 0$.

We only prove this last statement. If $\Theta = 0$, the associated character is the trivial character, and the spectral problem we are considering is our usual spectral problem, for which we know that the bottom eigenvalue is $\lambda_0(0) = 0$. Conversely, for f an eigenfunction for $\lambda_0(\Theta)$,

$$\lambda_0(\Theta) = \frac{\int_M |\nabla f|^2}{\int_M |f|^2} = 0$$

and this only holds if $|\nabla f| = 0$, i.e., if f is constant. Since $f(\gamma z) = \chi_\Theta(\gamma) f(z)$ for all $\gamma \in \Gamma$, we conclude that $\chi_\Theta \equiv 1$, which is equivalent to $\Theta = 0$.

We now choose $0 \neq \|\Theta\| < \delta$ with $\Theta \in (\mathbf{Q}/\mathbf{Z})^{2g}$, and let $\Gamma_\Theta = \ker(\chi_\Theta : \Gamma \to \mathbf{S}^1)$. Since $\Theta \in (\mathbf{Q}/\mathbf{Z})^{2g}$, the image of χ_Θ is finite, consisting of N-th roots of unity. Let f be an eigenfunction for $\lambda_0(\Theta)$. By construction, f is Γ_Θ-invariant, and thus $0 < \lambda_0(\Theta) < \varepsilon$ is an eigenvalue of Δ for the usual spectral problem on $\Gamma_\Theta\backslash\mathbf{H}$. Since $\lambda_0(\Theta) \neq 0$, it is bounded below by $\lambda_1(\Gamma_\Theta\backslash\mathbf{H})$, and we conclude that $\lambda_1(\Gamma_\Theta\backslash\mathbf{H}) < \varepsilon$. □

Remark 98. *Although these two proofs use different languages, in both cases one ends up constructing large enough cyclic covers to force small enough λ_1. An extension of the argument of Selberg also allows to show that for any $k > 0$ there exists a compact hyperbolic surface M with k small eigenvalues (i.e., $\lambda_j < \frac{1}{4}$, $j = 0, \ldots, k$). This latter result was also obtained by Randol (1974) by use of the Selberg trace formula.*

— End of class #13 —

8.2. Selberg's eigenvalue conjecture

Let $X_N = \Gamma(N)\backslash\mathbf{H}$. In 1965, Selberg stated the following conjecture, which is still open: for every $N \geq 1$,

$$\lambda_1(X_N) \geq \frac{1}{4}.$$

In the same article, Selberg proved that the lower bound

$$\lambda_1(X_N) \geq \frac{3}{16}$$

holds for each $N \geq 1$. This is referred to as **Selberg's 3/16 theorem**. Selberg's conjecture has been verified for all $N < 857$ (Booker–Strömbergsson, 2007).

Looking at the fundamental domains for $\Gamma(N)$, one sees that the congruence surfaces X_N are "hedgehog shaped," and in particular "short and fat." Nonetheless, it is remarkable that the spectral gap would be uniform in N. For comparison,

Proposition 99. *For each odd prime p,*

$$\operatorname{area}(X(p)) = \frac{\pi}{6} p(p^2 - 1).$$

Proof. Let $\mathcal{F}$ be the standard fundamental domain for $\Gamma(1) = \mathrm{SL}_2(\mathbf{Z})$ and let $\mathcal{F}_p$ be a fundamental domain for $X(p)$ obtained as in Proposition 59. In particular, it follows from Proposition 59 that

$$\operatorname{area}(\mathcal{F}_p) = [\overline{\Gamma(1)} : \overline{\Gamma(p)}]\operatorname{area}(\mathcal{F}).$$

Note that $-I \in \Gamma(1)$ and $-I \notin \Gamma(p)$, hence $[\overline{\Gamma(1)} : \overline{\Gamma(p)}] = \frac{1}{2}[\Gamma(1) : \Gamma(p)] = \frac{1}{2}|\mathrm{SL}_2(\mathbf{Z}_p)|$. Since $\mathrm{SL}_2(\mathbf{Z}_p) = \ker(\det : \mathrm{GL}_2(\mathbf{Z}_p) \to \mathbf{Z}_p^*)$, we have

$$|\mathrm{SL}_2(\mathbf{Z}_p)| = \frac{|\mathrm{GL}_2(\mathbf{Z}_p)|}{p - 1}.$$

The group $\mathrm{GL}_2(\mathbf{Z}_p)$ is in bijection with the set of ordered bases for the vector space $\mathbf{Z}_p \times \mathbf{Z}_p$ over $\mathbf{Z}_p$. Hence $|\mathrm{GL}_2(\mathbf{Z}_p)| = (p^2 - 1)(p^2 - 1 - (p - 1)) = (p^2 - 1)p(p - 1)$. □

Contrarily to the surfaces we have studied so far, X_N is noncompact. Indeed, for each $N \geq 1$, the parabolic element $\left(\begin{smallmatrix} 1 & N \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma(N)$. What does the spectrum of a noncompact hyperbolic surface look like ? Recall that the spectrum of $\Delta|_{L^2(\mathbf{R}^n)}$ is the full interval $[0, \infty)$; it is only by passing to $\mathbf{R}^n/\mathbf{Z}^n$ that we gain discrete spectrum $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \ldots$. This is somewhat analoguous to what happens in the hyperbolic setting. One can show that the spectrum of $\Delta|_{L^2(\mathbf{H})}$ is $[\frac{1}{4}, \infty)$. Upon quotienting by a cocompact Fuchsian group Γ (i.e., such that $\Gamma\backslash\mathbf{H}$ is compact) we have discrete spectrum. If however Γ is noncocompact (i.e., Γ contains at least one parabolic element), then the spectrum of $\Delta|_{L^2(\Gamma\backslash\mathbf{H})}$ contains the whole $[\frac{1}{4},\infty)$. In the case of X_N, $\Delta|_{L^2(X_N)}$ moreover has a discrete spectrum

$$0=\lambda_0<\lambda_1\leq\lambda_2\leq\lambda_3\leq\dots\qquad\lambda_j\to\infty.$$

(This follows from the Weyl's law one can deduce from the Selberg trace formula for noncompact surfaces, and in fact, this was Selberg's motivation to develop the trace formula.) The eigenvalue $\lambda_0=0$ is once again realized because constant functions on X_N are L^2. Moreover, we easily see that

$$\lambda_0=\inf\frac{\int_M|\nabla f|^2}{\int_M|f|^2},$$

where the infinimum is taken over all $f\in\mathrm{Dom}(\Delta)$, and this shows that $\lambda_0=0$ only if f is constant. Consequently, λ_0 has multiplicity 1. Eigenvalues in the discrete spectrum that are larger than $\frac{1}{4}$ are embedded in the continuous spectrum, and we usually do not have much control on them.

8.3. Relation to expander graphs

To a graph $X=(V,E)$, one can associate a combinatorial Cheeger constant as follows.

Definition 100. *The Cheeger constant of a finite graph $X=(G,E)$ is*

$$h(X)=\inf\frac{|\partial D|}{|D|},$$

where $D\subset V$ is a subset of $0<|D|\leq\frac{|G|}{2}$ vertices.

"Short and fat" in this context correspond to the graph being sparse and highly connected, which is the colloquial definition of an expander graph.

Definition 101. *Let $(X_m)_{m\geq1}$ be a family of finite, connected, k-regular graphs with $|V_m|\to\infty$ as $m\to\infty$. We say it is a family of expanders if there is $\varepsilon>0$ such that $h(X_m)\geq\varepsilon$ for each $m\geq1$.*

It is easy to construct stand-alone examples of expander graphs. In fact, by probabilistic arguments (Pinsker, 1973), random regular graphs are expanders. But it was no so obvious to showcase a family of expanders. The first example was constructed by Margulis using the recent introduction of property (T) in group theory by Kazhdan (1967). It is given by $V_m=\mathbf{Z}_m\times\mathbf{Z}_m$ with each vertex $\{x,y\}$ connected to $\{x+y,y\}$, $\{x-y,y\}$, $\{x,x+y\}$, $\{x,x-y\}$. In the rest of this section, we show how Selberg's $3/16$ theorem yields another family of expanders.

Definition 102 (Cayley graph)**.** *Let G be an abstract group, and $S\subset G$ a finite nonempty subset that is symmetric, i.e., $S=S^{-1}$. The Calyley graph $\mathcal{G}$ is the graph with vertex set $V=G$ and edge set*

$$E=\{\{x,y\}:x,y\in G,\ y=xs\text{ for some }s\in S\}.$$

Remark 103. *Because we require S to be symmetric, the adjacency relation for vertices is also symmetric; Cayley graphs are undirected.*

Exercise 104. *Let $\mathcal{G} = \mathcal{G}(G, S)$ be a Cayley graph. Show that $\mathcal{G}$ is a simple $|S|$-regular graph, and that $\mathcal{G}$ is connected if and only if $\langle S \rangle = G$.*

Examples 105.

(1) *Let $G = \mathbf{Z}/6\mathbf{Z}$ and $S = \{1, 5\}$. The corresponding Cayley graph is the cyclic graph C_6.*
(2) *Let $G = \mathbf{Z}/6\mathbf{Z}$ and $S = \{3\}$. The corresponding Cayley graph is ...*
(3) *Let $G = \mathrm{SL}_2(\mathbf{Z})$ and $S = \{\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right), \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\}$. The Cayley graph $\mathcal{G}(G, S)$ is again the cyclic graph C_6. Hence nonisomorphic groups may have the same Cayley graph.*
(4) *Let C_n be a cyclic graph with n vertices. Then its Cheeger constant is*

$$h(C_n) = \frac{2}{\lfloor \frac{n}{2} \rfloor} \approx \frac{4}{n}.$$

In particular, the family $\{C_n\}_{n \geq 1}$ is not a family of expanders.

The relation between the spectral theory of hyperbolic surfaces and the spectral theory of graphs was extensively developed by Brooks and Burger. A special case of this dictionary is the following theorem, which builds on covering theory. More precisely, Γ/Γ_i are the Galois group of the coverings $\Gamma_i\backslash\mathbf{H} \to \Gamma\backslash\mathbf{H}$ acting by isometries on $\Gamma_i\backslash\mathbf{H}$.

THEOREM 27. *Let Γ be a finitely generated Fuchsian group, with a finite symmetric set S of generators. Let $\{\Gamma_i\}$ be a family of normal subgroups of Γ with finite index. Then the Cayley graphs $\mathcal{G}_i = \mathcal{G}_i(\Gamma/\Gamma_i, S)$ are expanders if and only if there is $\varepsilon > 0$ such that $\lambda_1(\Gamma_i\backslash\mathbf{H}) > \varepsilon$ for every $i \geq 1$.*

Then by Selberg's 3/16-theorem,

Corollary 106. *The family*

$$(\mathcal{G}_N)_{N\geq 1} = (\mathcal{G}_i(\Gamma/\Gamma(N), S))_{N\geq 1},$$

where S is the image of $\{\pm\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right), \pm\left(\begin{smallmatrix} 1 & \pm 1 \\ 0 & 1 \end{smallmatrix}\right)\}$ in $\overline{\Gamma}(1)/\overline{\Gamma}(N)$, forms a family of expanders.

Bibliography

[1] Nicolas Bergeron, The Spectrum of Hyperbolic Surfaces. Springer, 2016.
[2] Robert Brooks, *Some relations between spectral geometry and number theory,* 1992.
[3] Peter Buser, Geometry and spectra of compact Riemann surfaces. Springer, 2010.
[4] Jeff Cheeger, *A lower bound for the smallest eigenvalue of the Laplacian.*
[5] S. Gallot, D. Hulin, J. Lafontaine, Riemannian Geometry. Springer, 2004.
[6] Henryk Iwaniec, Introduction to the Spectral Theory of Automorphic Forms. Revista matematica iberoamericana, 1995.
[7] Peter Sarnak, *Spectra of Hyperbolic Surfaces.* Bulletin of the American Mathematical Society vol. 40, num. 4, 2003, p. 441-478.
[8] Atle Selberg, *Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces witn applications to Dirichlet series.* Journal of the Indian Mathematical Society, vol. 20, 1956, p. 47-87.