VENTOTENE MINICOURSE: COUNTING, EQUIDISTRIBUTION, AND SPECTRAL GAP

These notes are a transcript of the minicourse given in Ventotene, with some added references. I have not adapted them to the formal format of a written survey but I added references to the relevant literature. It goes without saying that all mistakes are my own; thanks for communicating me any typo/suggestion!

1. Introduction

We start with some (very) classical examples of lattice point counting problems coming from number theory before moving on to the hyperbolic circle problem, with the goal of illustrating how counting and equidistribution come to be related. This first section is introductory and can be skipped by the more advanced reader.

We then explain how equidistribution implies a regular growth asymptotic for counting along lattice orbits in well-rounded families of sets, following Duke, Rudnick, and Sarnak [DRS93], and how the equidistribution of circles on the modular surface (related to the hyperbolic circle problem) is a consequence of the mixing of the geodesic flow via Margulis' thickening argument — a strategy that was vastly extended by Eskin and McMullen [EM93].

Harmonic analysis provides another toolbox to study counting problems. The spectral expansion of the counting function for the hyperbolic circle problem highlights the role of spectral gap as a measure of how well equidistributed lattice points are. We use coverings and their Galois groups to explain two results relative to the spectral gap for a lattice in $\mathrm{SL}_2(\mathbb{R})$; that the spectral gap in general can be arbitrarily small, but that for certain families of arithmetic lattices there is a uniform lower bound, leading to Selberg's eigenvalue conjecture.

The best known lower bound in this direction is due to Kim and Sarnak [KS03], and builds on results towards Langlands' functoriality conjecture. We prove instead a weaker bound, following the argument of Sarnak–Xue and Gamburd [SX91, Gam02]; whereas one usually relies on spectral data to gather information towards effective equidistribution/counting, this is one instance where counting can in turn be used to say something about spectral data.

2. Lattice point counting problems

2.1. **Some lattice point counting problems in the work of Gauss.** Some classical examples of lattice point counting problems arise from taking averages of arithmetic functions. This is the case for instance for the sum of squares function

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

that counts the number of ways in which n can be represented as a sum of two squares. The fluctuations of $r_2(n)$ make it hard to predict, but its average behaves much more regularly:

Proposition 2.1. *As $N \to \infty$, we have*

$$\sum_{n=1}^{N} r_2(n) = \#\{\xi \in \mathbb{Z}^2 : \|\xi\|^2 \leq N\} = \#(\mathbb{Z}^2 \cap B_{\sqrt{N}}) = \pi N + O(\sqrt{N}).$$

Date: October 29, 2021.

Proof. Tessellate the plane by unit squares: $\mathbb{R}^2 = \bigcup_{\xi\in\mathbb{Z}^2}([-1/2,1/2)+\xi)$, whereby each translate of the fundamental domain $[-1/2,1/2)$ accounts for a single point in $\mathbb{Z}^2$. The difference $\left|\#(\mathbb{Z}^2\cap B_{\sqrt{N}})-\pi N\right|$ is then bounded above by the area of a (sufficiently large but bounded) annulus around the boundary circle of radius of $\sqrt{N}$. Its growth order is thus comparable to the circumference of the circle. □

It is clear that the estimate obtained for the discrepancy is very crude; one expects the over- and undercounting across the boundary circle to showcase some cancellation. (It is relevant here that we count within a circle. If one is to replace the circle by a square of area N, centered at the origin, then there is no further powersaving in the error term to be expected.)

Let's say we expect the lattice points to be in fact randomly distributed across the boundary circle. Heuristically, if we model this situation with $n = \lceil\sqrt{N}\rceil$ iid random variables $X_1,\dots,X_n$ with uniform distribution in $(-1,1)$, then the sum $\overline{X} = X_1+\cdots+X_n$ is known to have expectancy 0 and standard deviation $\gg N^{1/4}$. The Gauss' circle problem conjectures that the lattice points are indeed as close to randomly distributed across the circle as possible:

Conjecture 2.2 (Gauss' circle problem)**.**

$$\#(\mathbb{Z}^2\cap B_{\sqrt{N}}) = \pi N + O\left(N^{1/4+\varepsilon}\right),$$

for any $\varepsilon>0$.

Our second example of lattice point counting problem concerns the average of the class number for negative discriminants. Let $Q(x,y)=ax^2+bxy+cy^2\in\mathbb{Z}[x,y]$ be an integral binary quadratic form. What numbers does the form Q represent? The discriminant $D=b^2-4ac$ determines whether Q is a product of linear forms (in which case, its study is not particularly interesting, and this corresponds to D being a perfect square), is indefinite, i.e., takes both positive and negative values ($D>0$), is positive definite, i.e., Q takes on only positive values ($D<0$ and $a>0$), or is negative definite ($D<0$ and $a<0$).

Let $n\in\mathbb{N}$ and let $\mathcal{Q}_n$ denote the set of positive definite quadratic forms of discriminant $-n$, i.e.,

$$\mathcal{Q}_n = \{Q : D=-n, a>0\}.$$

One can define an equivalence class on this set given by $Q_1\sim Q_2$ if $Q_1\circ\gamma = Q_2$ for some $\gamma\in \mathrm{SL}_2(\mathbb{Z})$. (In fact this is true for $\gamma\in\mathrm{GL}(2,\mathbb{Z})$ but it will be useful later on to narrow down to $\mathrm{SL}_2(\mathbb{Z})$). The value set of Q is class invariant since we consider only invertible linear transformations. We set $N(n)$ to be the class number of $\mathcal{Q}_n$.

Proposition 2.3. *Understanding $N(n)$ amounts to a lattice point counting problem. In fact,*

$$N(n) = \#\{(a,b,c)\in\mathbb{Z}^3 : 4ac-b^2=n, -a<b\le a<c \text{ or } 0\le b\le a=c\}.$$

This implies in particular that $N(n)<+\infty$.

Proof. Let $Q\in\mathcal{Q}_n$ and consider $Q(x,1)=ax^2+bx+c=0$. This quadratic equation has a single solution in the upper half plane $\mathbb{H}$, given by

$$z_Q = \frac{-b+i\sqrt{n}}{2a}.$$

One can check that if $Q'=Q\circ\gamma$ then $z_{Q'}=\gamma^{-1}z_Q$ under the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{H}$ by Möbius transformation. The latter action of the modular group is discontinuous and we can

tesselate the upper half plane $\mathbb{H}$ accordingly. So to each class $[Q]$ we associate a point z_Q in the fundamental domain

$$\mathcal{F} = \{z = x + iy : |z| > 1, -1/2 < x \leq 1/2 \text{ or } |z| = 1, 0 \leq x \leq 1/2\}. \tag{2.1}$$

This forces

$$-a < b \leq a < c \text{ or } 0 \leq b \leq a = c.$$

□

We can restrict further to $\mathcal{Q}_n^*$, the subset of primitive classes, i.e., taking $(a, b, c) = 1$. The number h_{-n} of primitive classes of positive definite forms with discriminant $-n$, is called the (narrow) class number. Gauss indicated (without proof but with extensive numerical check) that

$$\sum_{n=1}^{N} h_{-n} \sim \frac{\pi}{18\zeta(3)} N^{3/2}$$

as $N \to \infty$. The study of the error term was only much later initiated by Vinogradov.

2.2. **The hyperbolic circle problem.** We now introduce a lattice point counting problem that combines aspects from these two classical problems. From here on, we equip $\mathbb{H}$ with the hyperbolic metric

$$ds^2 = \frac{dx^2 + dy^2}{y^2}.$$

Consider $B_R = \{z \in \mathbb{H} : \rho(i, z) < R\}$, where ρ is the distance function induced by the hyperbolic metric, and set $\mathcal{O} = \mathrm{SL}_2(\mathbb{Z})i$. The orbit $\mathcal{O}$ is discrete, and in particular $|\mathcal{O} \cap B_R|$ is finite. (Actually, this is equivalent to count all $A \in \mathrm{SL}_2(\mathbb{Z})$ with bounded Frobenius norm $\|A\| = \sqrt{a^2 + b^2 + c^2 + d^2}$, via the identity $\|\gamma\|^2 = 2\cosh \rho(\gamma i, i)$.) More generally, it is still true that $|\mathcal{O} \cap B_R|$ is finite if we replace $\mathrm{SL}_2(\mathbb{Z})$ by any lattice in $\mathrm{SL}_2(\mathbb{R})$.

What happens if we try to reproduce Gauss' argument for the Euclidean circle problem? We have a tessellation of $\mathbb{H}$ by isometric copies of the fundamental domain $\mathcal{F}$ given by (2.1). We run into a problem as each of these copies has a point at infinity. What if we replace $\mathrm{SL}_2(\mathbb{Z})$ by a cocompact discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$? Then the fundamental domain is a (hyperbolic) polygon without points at infinity; we again recover the trivial bound

$$\left| \#(\mathcal{O} \cap B_R) - \frac{\mathrm{area}(B_R)}{\mathrm{area}(\mathcal{F})} \right| \ll \mathrm{circ}(B_R).$$

Now we face a new problem, due this time to the fact that we are in constant negative curvature: area and length have comparable size. In fact,

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

This means that if we want to prove the kind of regular growth observed in the Euclidean setting, we need to establish some nontrivial cancellation along the boundary. To guide us, it is useful to repeat our earlier heuristic; if we think of the lattice points has being statistically independent, then we expect some error term of the form $O(e^{R/2})$.

3. From equidistribution to orbital counting

3.1. **Equidistribution and mixing.** Another way to measure the statistic independence of the lattice points across the boundary circle is via equidistribution. Let S_R denote the boundary circle of B_R, i.e., $S_R = \{z \in \mathbb{H} : \rho(z,i) = R\}$. Instead of counting the number of copies of $\mathcal{F}$ that cover the interior of S_R, we will "fold" S_R onto $\mathcal{F}$.

Actually, to make the distinction between the interior and exterior of S_R precise and help parametrize the expansion of S_R as $R \to \infty$, we pass to the unit tangent bundle $T^1\mathbb{H}$ of $\mathbb{H}$ equipped with the Riemannian metric. Elements of $T^1\mathbb{H}$ are points $p = (z, v)$, where $z \in \mathbb{H}$ represents the position of p and $v \in T_z\mathbb{H} \cong \mathbb{C}$, with $\|v\|_z = \frac{|v|}{y} = 1$, represents the direction of p. The action of Γ on $\mathbb{H}$ by Möbius transformation extends to $T^1\mathbb{H}$ by taking derivatives:

$$\gamma(z,v) = (\gamma z, \gamma'(z)v) = \left(\frac{az+b}{cz+d}, \frac{v}{(cz+d)^2}\right). \tag{3.1}$$

We fix $\tilde{S}_R \subset T^1\mathbb{H}$ to be the boundary circle with unit outward normal vector at each position. We use the geodesic flow to expand $\tilde{S}_R$ in R. We have the following convenient algebraic parametrization. Let

$$\begin{aligned} G &= \mathrm{SL}_2(\mathbb{R}), \\ N &= \left\{n_x = \left(\begin{smallmatrix} 1 & x \\ 0 & 1 \end{smallmatrix}\right) : x \in \mathbb{R}\right\}, \\ A &= \left\{a_t = \left(\begin{smallmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{smallmatrix}\right) : t \in \mathbb{R}\right\}, \\ K &= SO(2) = \mathrm{Stab}_G(i). \end{aligned}$$

The Iwasawa decomposition $G = NAK$ yields the identification $G(i, \uparrow) = T^1\mathbb{H}$, implying $\Gamma\backslash G = \Gamma\backslash T^1\mathbb{H}$, where on the left the action is by left multiplication and on the right the action is given by (3.1). Further, we have the parametrization $\tilde{S}_R = Ka_R$. (Check that $a_t(i, \uparrow) = (e^t i, \uparrow)$ and $\rho(a_t i, i) = t$.)

Theorem 3.1. *The expanding orbits $(\Gamma \cap K)\backslash Ka_R$ become equidistributed in $\Gamma\backslash G$ as $R \to \infty$. That is, for every $f \in C_c(\Gamma\backslash G)$, we have*

$$\lim_{R\to\infty} \fint_{\Gamma\cap K\backslash K} f(ka_R)dk = \fint_{\Gamma\backslash G} f(g)dg.$$

(The barred integrals are used to denote that we integrate against the normalized probability Haar measure on the designated quotient.)

Proof. We will show that this follows from the (strong) mixing of the geodesic flow; namely that for any $f_1, f_2 \in L^2(\Gamma\backslash G)$, we have

$$\fint_{\Gamma\backslash G} f_1(ga_R) f_2(g)dg \to \fint_{\Gamma\backslash G} f_1 \fint_{\Gamma\backslash G} f_2.$$

To use mixing to prove equidistribution, we apply the following "thickening" trick (that goes back to Margulis' thesis). Let $f \in C_c(\Gamma\backslash G)$ and let $\varepsilon > 0$. By the uniform continuity of f, there exists a small neighborhood U of $e \in G$ such that

$$|f(g) - f(gu)| < \varepsilon.$$

We use that the geodesic flow normalizes the horocycle flow; explicitly, this is given by the algebraic relation $a_t n_x a_{-t} = n_{xe^t}$. This allows to choose a small enough neighborhood $V \subset$

AN such that

$$KVa_R \subset Ka_RU \tag{3.2}$$

for all $R > 0$. Then for every $k \in K$, $v \in V$, and $R > 0$, we have

$$|f(kva_R) - f(ka_R)| < \varepsilon.$$

Hence

$$\left| \fint_{\Gamma\cap K\backslash KV} f(kva_R)dkdv - \fint_{\Gamma\cap K\backslash K} f(ka_R)dk \right| \leq \fint_V \fint_{\Gamma\cap K\backslash K} |f(gkva_R) - f(ka_R)|dkdv < \varepsilon.$$

We now apply mixing to the annulus $\Gamma \cap K\backslash KV$. Let χ denote the characteristic function supported on V. For R sufficiently large, we have

$$\fint_{\Gamma\cap H\backslash KV} f(kva_R)dkdv = \frac{\mu(\Gamma\backslash G)}{\mu(\Gamma KV)} \fint_{\Gamma\backslash G} f(ga_R)\chi(g)dg = \fint_{\Gamma\backslash G} f(g)dg + O(\varepsilon).$$

We conclude by letting $\varepsilon \to 0$. □

The relation (3.2) is a simple instance of what is called the wavefront lemma in action. Observe this is the only geometric input in the proof. Eskin and McMullen established the wavefront lemma in the context of affine symmetric spaces, by building on an extension of the Cartan decomposition $G = KAK$ for Riemannian symmetric spaces [EM93].

3.2. **A more general setup.** Let G be a locally compact group and let H be a closed subgroup of G. Recall that if G contains a lattice Γ, it is necessarily unimodular.[1] We assume that $\Gamma \cap H$ is a lattice in H for the same reason.

Take $(B_T)_{T>0} \subset H\backslash G$ to be a continuous family of compact well-rounded sets such that $\mathrm{vol}(B_T) \to \infty$ as $T \to \infty$. Well-rounded has the following precise meaning.

Definition 3.2. *We say that the sets $(B_T)_{T>0} \subset H\backslash G$ are well-rounded if for every $\varepsilon > 0$ there is a small enough neighborhood U of $e \in G$ such that for*

$$B_T^+ = \bigcup B_T g \quad B_T^- = \bigcap_{g\in U} B_T g$$

we have that

$$\frac{\mathrm{vol}(B_T^+ \setminus B_T^-)}{\mathrm{vol}(B_T)} < \varepsilon.$$

Theorem 3.3. *Let $x = H \in H\backslash G$. If the orbits $(\Gamma \cap H)\backslash Hg$ become equidistributed as $g \to \infty$, we have*

$$N_T = \#(x\Gamma \cap B_T) \sim \frac{\mu(\Gamma \cap H\backslash H)}{\mu(\Gamma\backslash G)} \mathrm{vol}(B_T)$$

as $T \to \infty$.

[1]The proof is as follows: Since Γ is discrete, we have $\Gamma \subset \ker \Delta_G \subset G$. Then the Haar measure of $\ker \Delta_G\backslash G \subset \Gamma\backslash G$ is finite and this implies that $\ker \Delta_G\backslash G$ is compact and as such isomorphic to a compact subgroup of $\mathbb{R}_+$. We conclude that $G = \ker \Delta_G$.

Proof. Let $N_T(g) = \#(x\Gamma g \cap B_T)$. The Haar measures obey Fubini-type relations so that we can fold and unfold the following integral against any test-function $\varphi \in C_c(\Gamma\backslash G)$:

$$\begin{aligned}\int_{\Gamma\backslash G} N_T(g)\varphi(g)dg &= \int_{\Gamma\backslash G} \sum_{\gamma\in\Gamma\cap H\backslash\Gamma} 1_{B_T}(x\gamma g)\varphi(g)dg \\ &= \int_{\Gamma\cap H\backslash G} 1_{B_T}(xg)\varphi(g)dg \\ &= \int_{H\backslash G}\int_{H\cap\Gamma\backslash H} 1_{B_T}(xhg)\varphi(hg)dhdg \\ &= \int_{H\backslash G} 1_{B_T}(xg)\left(\int_{H\cap\Gamma\backslash H}\varphi(hg)dh\right)dg.\end{aligned}$$

If the orbits $H\cap\Gamma\backslash Hg$ become equidistributed as $g\to\infty$ (leaving every compact set), then by dividing both sides by $\mathrm{vol}(B_T)$ and chasing Haar measure normalizations we arrive to

$$\lim_{T\to\infty}\frac{1}{\mathrm{vol}(B_T)}⨍_{\Gamma\backslash G} N_T(g)\varphi(g)dg \;=\; \frac{\mu(\Gamma\cap H\backslash H)}{\mu(\Gamma\backslash G)}⨍_{\Gamma\backslash G}\varphi(g)dg.$$

To derive from this approximation an asymptotic for the precise count $N_T = N_T(e)$, we rely on the definition of well-roundedness. Fix $\varepsilon > 0$. Since (B_T) is a well-rounded family, we can construct set $B_T^{\pm}$ such that

$$\frac{\mathrm{vol}(B_T^+/B_T^-)}{\mathrm{vol}(B_T)} < \varepsilon.$$

Let $N_T^+ = \#(x\Gamma\cap B_T^+)$. By definition, for any $g\in U$, we have

$$N_T \le N_T^+(g)$$

Hence choosing $\varphi\ge 0$, supported on U, we have

$$\frac{N_T}{\mathrm{vol}B_T}⨍_{G/\Gamma}\varphi dg \le \frac{1}{\mathrm{vol}B_T}⨍_{G/\Gamma} N_T^+(g)\varphi(g)dg < (1+\varepsilon)\frac{1}{\mathrm{vol}B_T^+}⨍_{G/\Gamma} N_T^+(g)\varphi(g)dg$$

so that

$$\limsup_{T\to\infty}\frac{N_T}{\mathrm{vol}B_T} \le (1+\varepsilon)\frac{\mu(H/\Gamma\cap H)}{\mu(G/\Gamma)}.$$

One can reproduce the same argument to obtain the lower bound

$$\liminf_{T\to\infty}\frac{N_T}{\mathrm{vol}B_T} \ge (1-\varepsilon)\frac{\mu(H/\Gamma\cap H)}{\mu(G/\Gamma)}.$$

Since $\varepsilon>0$ can be chosen arbitrarily small, we conclude that

$$\lim_{T\to\infty}\frac{N_T}{\mathrm{vol}B_T} = \frac{\mu(H/\Gamma\cap H)}{\mu(G/\Gamma)}.$$

□

Coming back to the hyperbolic circle problem, we find that

$$\#(\mathcal{O}\cap B_R)\sim\frac{\mathrm{area}(B_R)}{\mathrm{area}(\mathcal{F})}$$

as $R\to\infty$.

4. Spectral expansions and spectral gap

Fix $z, w \in \mathbb{H}$. This time we are going to approach the (extended) hyperbolic circle problem

$$N_R(z,w) = \#\{z' \in \Gamma z : \rho(z',w) \leq R\} = ?$$

using harmonic analysis.

4.1. **Elementary harmonic analysis.** To fix notation and some important ideas, we quickly review some very classical harmonic analysis, namely the decomposition of the periodic functions of $L^2(\mathbb{T})$ into harmonics. We have a countable orthonormal basis for $L^2(\mathbb{T})$ given by $\{\varphi_n\}_{n\in\mathbb{Z}}$, $\varphi_n(x) = e^{2\pi i n x}$. (One easily checks that this is an orthonormal family with respect to the standard inner product on $L^2(\mathbb{T})$, and by an application of Stone–Weierstrass, one can deduce that this family is dense in $L^2(\mathbb{T})$.) This allows to write

$$f = \sum_{n\in\mathbb{Z}} \langle f, \varphi_n\rangle \varphi_n = \sum_{n\in\mathbb{Z}} \widehat{f}(n)\varphi_n,$$

where the equality holds in the L^2 sense. To have a pointwise equality, we need the Fourier coefficients to decay sufficiently. By integration by parts, one immediately has

$$\widehat{f}(n) = \int_0^1 f(x)\overline{\varphi_n(x)}dx = \frac{\widehat{f^{(k)}}(n)}{(2\pi i n)^k}$$

for each $k \geq 0$, meaning that the rate of decay of $\widehat{f}$ depends on the rate of decay of the derivatives of f. Ideally, we would want f to be a smooth function with rapidly decreasing derivatives, such as a Schwartz function. In fact, the space of Schwartz functions is preserved by the Fourier transform.

4.2. **Asymptotics for the hyperbolic circle problem.** We will admit (purposefully leaving the issue of convergence aside) that on the orbit space $\Gamma\backslash\mathbb{H}$ the counting function $N_R(z,w)$ has spectral expansion

$$N_R(z,w) = \sum_{j\geq 0} h(\lambda_j)\varphi_j(z)\overline{\varphi_j(w)} + \sum \frac{1}{4\pi}\int_{-\infty}^{\infty} h(r)E(z,\tfrac{1}{2}+ir)\overline{E(w,\tfrac{1}{2}+ir)}dr. \tag{4.1}$$

From right to left, we have Eisenstein series $E(\cdot,\frac{1}{2}+it)$, cusp forms φ_j, and the Selberg/Harish-Chandra transform h. We refer the reader to the textbook of Iwaniec [Iwa02] for precise definitions of these terms. For our discussion, we will only need that

- the functions φ_j form a complete orthonormal family in $L^2(\Gamma\backslash\mathbb{H})$ of solutions to the spectral problem $(\Delta + \lambda_j)\varphi_j = 0$, where Δ denotes the hyperbolic Laplacian

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

  and where the eigenfunctions φ_j are ordered according to

  $$0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \to \infty.$$

  The eigenvalue $\lambda_0 = 0$ is realized since constant functions belong to $L^2(\Gamma\backslash\mathbb{H})$, and it can be shown, by the maximum principle, to have multiplicity 1; this is encoded by the first strict inequality. The size of λ_1 is accordingly called the *spectral gap*.

- the Selberg/Harish-Chandra transform can be computed explicitly from the LHS of (4.1). More precisely, it is expressed as the inverse Fourier transform of a function that is completely determined by the LHS, something like Fourier coefficients in our previous model. The point is that the pointwise convergence of the RHS depends on the regularity of the function on the LHS.

Here the function on the LHS is an average over a characteristic function:

$$N_R(z,w) = \sum_{\gamma\in\Gamma} 1_R(\rho(\gamma z, w)), \qquad \text{where } 1_R(\rho) = \begin{cases} 1 & \rho \le R, \\ 0 & \rho > R. \end{cases}$$

As a result, if we were to compute the Selberg transform h of $N_R(z,w)$ explicitly, the spectral expansion (4.1) would yield an expression of the form

$$N_R(z,w) = \frac{\text{area}(B_R)}{\text{area}(M)} + \sum_{0<\lambda_j<\frac{1}{4}} c_j \varphi_j(z)\overline{\varphi_j(w)} e^{R(\frac{1}{2}+\sqrt{\frac{1}{4}-\lambda_j})} + O(e^{R/2}S)$$

where the constants c_j are explicit and nonzero, and the term S, uniform in R, is composed of an infinite sum and an integral that diverge as a consequence of Weyl's law. It is nonetheless interesting to note that as $\text{area}(B_R) \sim \pi e^R$ (as $R \to \infty$), the next growth term is not given by $e^{R/2+o(1)}$ but by $e^{R(1/2+\sqrt{1/4-\lambda_1})}$ (provided that the spectral gap satisfies $0 < \lambda_1 < 1/4$ and that $\varphi_1(z)\overline{\varphi_1(w)} \neq 0$). We come back to this point in the next section.

To give sense to such a spectral expansion, one needs to replace the characteristic function 1_R by a smooth approximation. Applying spectral expansion and controlling for the approximation yields

Theorem 4.1 (Delsarte, Huber, Selberg, 1950s)**.**

$$N_R(z,w) = \frac{\text{area}(B_R)}{\text{area}(M)} + \sum_{0<\lambda_j<\frac{1}{4}} c_j \varphi_j(z)\overline{\varphi_j(w)} e^{R(\frac{1}{2}+\sqrt{\frac{1}{4}-\lambda_j})} + O(e^{2R/3}) \tag{4.2}$$

Idea of proof. Fix $\varepsilon > 0$. We replace the characteristic function 1_R by a smooth approximation $\tilde{1}_{R,\varepsilon}$ by convolution with a mollifier. As a result, the Selberg transform is rapidly decaying and the spectral expansion for the mollified counting function $\tilde{N}_{R,\varepsilon}(z,w)$ holds pointwise with an error term on the RHS that depends on R and ε. On the LHS, one can bound the discrepancy between $N_R(z,w)$ and $\tilde{N}_{R,\varepsilon}(z,w)$ again in terms of R and ε. It remains to choose $\varepsilon = \varepsilon(R)$ to optimize the error term. A detailed proof of this process is transcribed in the PhD thesis of Cherubini [Che18]. □

As can be seen from the argument of proof, the constant factor of $\frac{2}{3}$ appearing in the error rate is an artefact of the mollification process; in other words, it is the price to pay for the regularization of the counting function. By analogy to the Euclidean circle problem, one expects the true state of affairs to involve the error term $O(e^{R/2+\varepsilon})$. The asymptotic (4.2) has never been improved.

4.3. **Two theorems on the spectral gap.** Recall our ongoing heuristic: if we believe the lattice points in Γz to be uniformly distributed, then we expect a square-root cancellation leading to an error term of the form $O(e^{R/2+\varepsilon})$. Theorem 4.1 shows this is possible only as so far as we don't have small eigenvalues $0 < \lambda_j < 1/4$. In this sense, the spectral gap can be seen as a measure of the equidistribution of the lattice points in Γz around the boundary circle of B_R. Unfortunately (perhaps), the spectral gap can be arbitrarily small.

Theorem 4.2. *For any $\varepsilon > 0$, there is a lattice $\Gamma < \mathrm{SL}_2(\mathbb{R})$ such that $\lambda_1(\Gamma) < \varepsilon$.*

The idea is to build sufficiently high (finite degree) regular covers

$$\begin{array}{c} \Gamma_n\backslash\mathbb{H} \\ \big\downarrow \\ \Gamma\backslash\mathbb{H} \end{array}$$

Observe that the discrete spectrum of $\Gamma_n\backslash\mathbb{H}$ contains the discrete spectrum of $\Gamma\backslash\mathbb{H}$. (Each L^2 solution f to the eigenvalue problem $(\Delta+\lambda)f = 0$ that is Γ-invariant is automatically Γ_n-invariant. Equivalently, these solutions can be seen as eigenfunctions in $L^2(\Gamma_n\backslash\mathbb{H})$ on which the Galois group Γ/Γ_n of the covering acts trivially.) On the other hand, each L^2 solution to $(\Delta+\lambda)f = 0$ on which the Galois group Γ/Γ_n acts nontrivially produces "new" eigenvalues. The goal is to produce small eigenvalues in this way.

Proof of Theorem 4.2, after Selberg. Choose Γ torsionfree (this can always be achieved via Selberg's lemma) with nonvanishing first Betti number. Hence there is some $n \geq 1$ such that the character group of Γ is

$$\widehat{\Gamma} = \mathrm{Hom}(\Gamma, \mathbb{T}) = \mathrm{Hom}(\Gamma/[\Gamma,\Gamma], \mathbb{T}) \cong \mathrm{Hom}(\mathbb{Z}^n, \mathbb{T}) \cong \mathbb{T}^n.$$

Fix $\Theta \in \mathbb{T}^n$ and let χ_Θ denote the associated character. We consider the following modified eigenvalue problem: find solutions $f : \mathbb{H} \to \mathbb{C}$ for

$$\begin{cases} (\Delta+\lambda)f = 0 \\ f(\gamma z) = \chi_\Theta(\gamma) f(z) \text{ for each } \gamma \in \Gamma \\ \displaystyle\int_{\Gamma\backslash\mathbb{H}} |f(z)|^2 d\mu(z) < \infty. \end{cases} \tag{4.3}$$

We will rely on the following results of Selberg.

- The spectral problem (4.3) admits a complete resolution with discrete spectrum
$$0 \leq \lambda_0(\Theta) \leq \lambda_1(\Theta) \leq \lambda_2(\Theta) \leq \dots;$$
- $\lambda_0(\Theta) = 0$ if and only if $\Theta = 0$;
- The bottom eigenvalue $\lambda_0(\Theta)$ is continuous in Θ.

Fix $\varepsilon > 0$. We can choose Θ such that

(1) $\Theta \neq 0$;
(2) Θ is small enough so that $\lambda_1(\Theta) < \varepsilon$;
(3) $\Theta \in (\mathbb{Q}/\mathbb{Z})^n$.

Let $\Gamma_\Theta = \ker(\chi_\Theta : \Gamma \to \mathbb{T})$. By (3), the image of χ_Θ is finite; we have a finite cyclic covering $\Gamma_\Theta\backslash\mathbb{H}$ of $\Gamma\backslash\mathbb{H}$. Let f be a solution of (4.3) for the eigenvalue $\lambda = \lambda_0(\Theta) > 0$. Then f is Γ_Θ-invariant, and thus a solution of the *usual* eigenvalue problem for Γ_Θ. This means that $\lambda_0(\Theta)$ belongs to the discrete spectrum for the usual eigenvalue problem for Γ_Θ, i.e.,

$$0 = \lambda_0 < \lambda_1 \leq \dots \leq \lambda_1(\Theta) < \varepsilon.$$

□

On the other end, one of the most important open problem in the subject is Selberg's eigenvalue conjecture, which states that

Conjecture 4.3 (Selberg, 1965)**.** *For $N \geq 1$, we have $\lambda_1(\Gamma(N)) \geq 1/4$.*

The groups $\Gamma(N)$ are the principal congruence subgroups

$$\begin{aligned}\Gamma(N) &= \ker(\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})) \\ &= \{A \in \Gamma(1) = \mathrm{SL}_2(\mathbb{Z}) : A \equiv I \,(\mathrm{mod}\ N)\}.\end{aligned}$$

The current best bound towards the conjectural $1/4$ is due to Kim and Sarnak [KS03], while $1/4$ was recently verified for $N \leq 226$ by Booker, Lee, and Strömbergsson [BLS20].

We might explain the uniform spectral gap appearing here as follows. Take $N = p$ prime. The regular covering under consideration is

$$\begin{array}{c}\Gamma(p)\backslash\mathbb{H} \\ \big\downarrow \\ \Gamma(1)\backslash\mathbb{H}\end{array}$$

with Galois group $\Gamma(1)/\Gamma(p) \cong \mathrm{SL}_2(\mathbb{F}_p)$. This group has "many more symmetries" than the finite cyclic Galois group that appeared in Selberg's construction. This is more precisely expressed in the language of representation theory: If $\Gamma(1)/\Gamma(p)$ acts nontrivially on an eigenfunction f on $\Gamma(p)\backslash\mathbb{H}$ (meaning that f is not the lift of a function living on $\Gamma(1)\backslash\mathbb{H}$) then the dimension of the eigenspace containing f is $\geq \frac{p-1}{2}$. In other words, the corresponding "new" eigenvalue has multiplicity $\geq \frac{p-1}{2}$. (For comparison, the dimension of each nontrivial representation of a finite cyclic group is 1.) The uniform lower bound for $\lambda_1(\Gamma(N))$ manifests some arithmetic rigidity safeguard against a rapid accumulation of small eigenvalues with high multiplicity as one lets $p \to \infty$.

We close with the proof of the following weaker bound, following Gamburd's execution of the strategy of [SX91] in the noncocompact case [Gam02], which is deduced from counting bounds for the hyperbolic circle problem.

Theorem 4.4. *For all primes p sufficiently large, we have $\lambda_1(\Gamma(p)) \geq 5/36$.*

Sketch of proof, after Gamburd. Let $X(p) = \Gamma(p)\backslash\mathbb{H}$. Let $f \in L^2(X(p))$ satisfy $(\Delta+\lambda_1(p))f = 0$. Suppose first that f is $\Gamma/\Gamma(p)$-invariant. Then $\lambda_1(p)$ belongs to the spectrum of $X(1)$ and we have

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

The latter lower bound[2] was first established by Roelcke using the Fourier coefficients of cusp forms [Roe56].

Suppose instead that $\Gamma/\Gamma(p)$ acts nontrivially on f and that $\lambda_1 \coloneqq \lambda_1(p) < 1/4$. Recall that λ_1 has multiplicity $m_1 \geq \frac{p-1}{2}$. We will use the hyperbolic circle problem to force a lower bound on the spectral gap. If X is compact, we have the L^2-equality

$$\iint_{X\times X} |N_R(z,w)|^2 dzdw = \sum_{j\geq 0} |h(\lambda_j)|^2 + \sum \frac{1}{4\pi}\int_{-\infty}^{\infty} |h(t)|^2 dt \geq\ m_1|h(\lambda_1)|^2.$$

Although here $X(p)$ is noncompact, Gamburd shows that upon replacing $X(p)$ by its compact core we nonetheless recover the geometric upper bound

$$m_1|h(\lambda_1)|^2 = O\left(\mathrm{area}(X_p)e^R \int_0^R e^{-t} N_{e^t} dt\right),$$

[2]It has since been numerically verified that $\lambda_1(1) \approx 91.14$; see [Hej92, BSV06, BS07].

where $N_{e^t} = \#\{\gamma \in \Gamma(p) : \|\gamma\| \leq e^t\}$. We can compute explicitly that $|h(\lambda_1)| \gg e^{Rs}$ with $s = \frac{1}{2} + \sqrt{\frac{1}{4} - \lambda_1(p)}$. We also know that $\mathrm{area}(X(p)) = \mathrm{area}(X(1))[\Gamma : \Gamma(p)] \sim p^3$. However we cannot apply the counting asymptotic (4.2) since it itself depend on the mulitplicity of λ_1. We need an upper bound on N_T that does not involve harmonic analysis. This is here possible as the counting problem reduces to the following "elementary" lattice point counting problem

$$\begin{aligned} N_T &= \#\{\gamma \in \Gamma(p) : \|\gamma\| \leq T\} \\ &\leq \#\{(a,b,c,d) \in \mathbb{Z}^4 : ad - bc = 1,\ a \equiv d \equiv 1,\ b \equiv c \equiv 0 \pmod{p},\ |a|, |b|, |c|, |d| \leq T\}. \end{aligned}$$

Observe that $a + d \equiv 2 \pmod{p^2}$. There are $O(T/p^2)$ choices of $\xi \equiv 1 \pmod{p^2}$, $|\xi| \leq 2T$ and $O(T/p) + 1$ choices of $a \equiv 1 \pmod{p}$ and $|a| \leq T$, so that we have $O(T^2/p^3)$ choices for a and d. This completely determines $\xi = ad - 1$ and the trivial divisor bound yields $d(T^2) = O_\varepsilon(T^{2\varepsilon})$ possibilities for the pair b, c. Hence for p sufficiently large, we have

$$N_T \ll \frac{T^{2(1+\varepsilon)}}{p^3} + \frac{T^{(1+\varepsilon)}}{p^2} + 1.$$

Fixing $T = p^3$, we have, for p large enough,

$$\frac{p-1}{2} \leq m(\lambda_1, \Gamma(p)) \ll p^{6(\frac{1}{2} - \sqrt{\frac{1}{4} - \lambda_1}) + \varepsilon}.$$

Comparing exponents yields the lower bound $\lambda_1 \geq 5/64$. □

References

[Ber16] N. Bergeron, The Spectrum of Hyperbolic Surfaces. Springer Universitext 2016.

[BLS20] A. Booker, M. Lee, A. Strömbergsson, *Twist-minimal trace formulas and the Selberg eigenvalue conjecture*. JLMS, 2020.

[BS07] A. Booker, A. Strömbergsson, *Numerical computations with the trace formula and the Selberg eigenvalue conjecture*. Crelle, 2007.

[BSV06] A. Booker, A. Strömbergsson, A. Venkatesh, *Effective computation of Maass cusp forms*. IMRN, 2006.

[Che18] G. Cherubini, Studies in the hyperbolic circle problem. PhD Thesis, University of Copenhagen, 2018.

[DRS93] W. Duke, Z. Rudnick, P. Sarnak, *Density of integer points on affine homogeneous varieties*. Duke Math. J., 1993.

[EM93] A. Eskin, C. McMullen, *Mixing, counting and equidistribution in Lie groups*. Duke Math. J., 1993.

[Gam02] A. Gamburd, *On the spectral gap for infinite index "congruence" subgroups of* $\mathrm{SL}_2(\mathbb{Z})$. Israel J. Math., 2002.

[Hej92] D. Hejhal, *On eigenvalues of the Laplacian for Hecke triangle groups*. Advanced Studies in Pure Mathematics, Zeta Functions in Geometry, 1992.

[Iwa02] H. Iwaniec, Spectral methods of automorphic forms. Graduate Studies in Mathematics, AMS, 2002.

[KS03] H. Kim, P. Sarnak, Appendix to H. Kim, *Functoriality of the exterior square of* GL_4 *and the symmetric fourth power of* GL_2. JAMS 2003.

[Roe56] W. Roelcke, *Über die Wellengleichung bei Grenkreisgruppen erster Art*. 1956.

[SX91] P. Sarnak, X. Xue, *Bounds for multiplicities of automorphic representations*. Duke Math. J. 1991.