THIN GROUPS AND APPLICATIONS

EDGAR ASSING

1. Introduction

In these lectures we are interested in thin groups and (some of) their applications. Even though thin groups have been around for a long time the term thin group has only been coined recently by Peter Sarnak. This re-branding goes hand in hand with a surge of progress in the area, which led to many interesting results and applications. Here we are interested in some specific applications to number theory following a series of works by Alex Kontorovich and Jean Bourgain. However, this forces us to fill our toolbox with methods and results from many branches of mathematics. Many of the results discussed on the way are interesting and important in their own right and we will give full proofs when ever possible.

We take inspiration from several lectures and mini-courses by Alex Kontorovich on the topic. (All mistakes, misunderstandings and typos are of course only due to the author of these notes!)

1.1. What is a thin group.

This section bares the same name as the nice little paper [11]. We start directly with the definition of a thin group.

Definition 1.1. Let Γ be a (finitely generated) subgroup of $\mathrm{GL}_n(\mathbb{Z})$ and let $G = \mathrm{Zcl}(\Gamma)$ be its Zariski closure. We say that Γ is a *thin group* if the index of Γ in $G(\mathbb{Z})$ is infinite.

We now give a series of examples and non-examples for the case $n = 2$:

(1) Let $\Gamma = \langle T, S\rangle$ for $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. It is well known that $\Gamma = \mathrm{SL}_2(\mathbb{Z})$. Therefore we have $G = \mathrm{Zcl}(\Gamma) = \mathrm{SL}_2$, so that $G(\mathbb{Z}) = \Gamma$. In particular, this group is **not thin**. (It is the basic example of an arithmetic group!)

(2) Let $\Gamma = \langle T^2, -ST^{-2}S\rangle$. One can compute that

$$\Gamma = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \colon a \equiv d \equiv 1 \bmod 4,\ 2 \mid b, c \right\}.$$

Again we find that the Zariski closure is $G = \mathrm{SL}_2$. Since $[G(\mathbb{Z})\colon \Gamma] = 12$ this is again **not thin**. (It is a classical congruence subgroup.)

(3) Let $\Gamma = \langle T^2 \rangle = \begin{pmatrix} 1 & 2\mathbb{Z} \\ 0 & 1 \end{pmatrix}$. The Zariski closure of Γ is the algebraic group G given by the equations

$$(a, b, c, d) \colon ad - bc - 1 = a - 1 = d - 1 = c = 0.$$

(This is a unipotent group which is strictly smaller than SL_2 and will be called U for later reference.) Of course $G(\mathbb{Z}) = \begin{pmatrix} 1 & \mathbb{Z} \\ 0 & 1 \end{pmatrix}$, so that $[G(\mathbb{Z}) \colon \Gamma] = 2$. Therefore Γ is **not thin** (despite it having infinite index in $\mathrm{SL}_2(\mathbb{Z})$).

(4) Take $\Gamma = \langle A \rangle$ with $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$. One computes inductively that

$$A^n = \begin{pmatrix} f_{2n+1} & f_{2n} \\ f_{2n} & f_{2n-1} \end{pmatrix},$$

where f_n is the nth Fibonacci number (i.e. $f_{n+1} = f_n + f_{n-1}$ and $f_0 = f_1 = 1$). Let $K = \mathbb{Q}(\frac{1+\sqrt{5}}{2})$. Then there is a matrix $g \in \mathrm{SL}_2(K)$ such that

$$g\Gamma g^{-1} = \Gamma_1 = \left\{ \begin{pmatrix} \left(\frac{1+\sqrt{5}}{2}\right)^{2n} & 0 \\ 0 & \left(\frac{1+\sqrt{5}}{2}\right)^{-2n} \end{pmatrix} : n \in \mathbb{Z} \right\}.$$

The Zariski closure of Γ_1 is the algebraic torus T given by the equations

$$(a, b, c, d) \colon ad - bc - 1 = b = c = 0.$$

Conjugating back we find that $G = g^{-1}Tg = \mathrm{Zcl}(\Gamma)$. Again we find $G(\mathbb{Z}) = \Gamma$. Again Γ is **not thin** (even though again $[\mathrm{SL}_2(\mathbb{Z}) \colon \Gamma] = \infty$).

(5) Let $\Gamma = \langle T^4, S \rangle$. Counter intuitively one can show that $[\mathrm{SL}_2(\mathbb{Z}) \colon \Gamma] = \infty$. Furthermore the only subvarieties of SL_2 that are also groups are up to conjugation T, U and UT. One checks that Γ is not contained in any of these, so that the Zariski closure is given by $G = \mathrm{SL}_2$. Therefore we have found our first **thin group**.

We have seen 5 groups out of which only one was thin. Note that this numbers do not represent reality very well and are due to our (poor) choices. Indeed it can be shown that random subgroups of arithmetic groups are thin (in a precise way). However it remains non-trivial in general to verify that a given group is thin (or not). This problem gets worse in higher rank.

(6) Consider $\Gamma = \langle A, B \rangle$ with

$$A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 2 & 4 \\ 0 & -1 & -1 \\ 0 & 1 & 0 \end{pmatrix}.$$

One can verify that the Zariski closure of Γ is $G = \mathrm{SL}_3$. A harder task is to see that the generators have the relations $A^3 = B^3 = (AB)^4 = 1$ and no others. (Γ is a faithful representation of the hyperbolic triangle group $(3, 3, 4)$.) This allows one to conclude that Γ has infinite index in $G(\mathbb{Z})$ and therefore is **thin**.

Already for $n = 3$ it is easy to give examples of finitely generated Γ with full Zariski closure where it is not known if they are thin or not.

One can define the notion of thinness in greater generality. Roughly speaking an integer set is called thin (i.e. *thin integer set*) if it has density zero in the integer points of its Zariski closure. This can be precisely formulated and one sees that when the integer set turns out to be a subgroup of a linear group then one recovers the notion of thin group discussed above. We will need this more general notion only for one explicit example which we will work out in detail now.

Fix $A \in \mathbb{N}$ and consider the semigroup

$$\Gamma_A = \left\langle \begin{pmatrix} a & 1 \\ 1 & 0 \end{pmatrix} : 0 \leq a \leq A \right\rangle^+ \cap \mathrm{SL}_2(\mathbb{Z}).$$

Note that for $A = 1$ we essentially recover example (4) above which wasn't thin. Therefore it makes sense to exclude this case. The following lemma establishes that Γ_A is thin making the imprecise definition of thin integer sets precise for this particular example.

Lemma 1.1. *Let $A \geq 2$ and let G be the Zariski closure of Γ_A. Then we have*

$$\frac{\sharp(\Gamma_A \cap B_X)}{\sharp(G(\mathbb{Z}) \cap B_X)} = o(1),$$

where $B_X = \{\gamma \in \mathrm{SL}_2(\mathbb{R}) \colon \|\gamma\| \leq X\}$.

Proof. Since $A \geq 2$, the Zariski closure of Γ_A is $G = \mathrm{SL}_2$. It is an easy exercise to check that

$$\sharp(\mathrm{SL}_2(\mathbb{Z}) \cap B_X) \asymp X^2.$$

Further we will show below, that there is a number $1 > \delta_A > \frac{1}{2}$ so that

$$\sharp(\Gamma_A \cap B_X) \asymp X^{2\delta_A}.$$

This completes the proof so far. □

1.2. **A local to global conjecture.** Let $\Gamma \subset M_{2\times 2}(\mathbb{Z})$ be a thin integer set. (We can think of the two cases $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$ being a thin group or Γ being the semigroup Γ_A for $A \geq 2$.) let $F \colon M_{2\times 2}(\mathbb{Z}) \to \mathbb{C}$ be an (affine) linear map taking integer values on Γ. Then we are interested in properties of the image $F(\Gamma)$. For $n \in \mathbb{Z}$ put

$$\mathrm{mult}_{\Gamma,F}(n) = \sharp\{\gamma \in \Gamma \colon F(\gamma) = n\} \text{ and } \mathrm{mult}_{\Gamma,F,X}(n) = \sharp\{\gamma \in \Gamma \cap B_X \colon F(\gamma) = n\}.$$

(If F and Γ are clear from the context we may drop the subscripts.) We call n *admissible* if $n \in F(\Gamma) \bmod q$ for all $q \in \mathbb{N}$. (This rules out all congruence

obstructions, but a priori requires the verification of infinitely many congruences.) Now given

$$\sharp(\Gamma \cap B_X) \asymp X^{2\delta}$$

one might expect that in favorable situations one has

$$\mathrm{mult}_{\Gamma,F,X}(n) \asymp X^{2\delta-1-o(1)},$$

for $n \asymp X$ admissible. This is a slightly imprecise but a very general local-global conjecture for thin orbits.

For thin groups $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$ progress towards the local-global conjecture above can be made. Given $\mathbf{v}_0, \mathbf{w}_0 \in \mathbb{Z}^2$ we consider $F(\gamma) = \langle \mathbf{v}_0 \cdot \gamma, \mathbf{w}_0 \rangle$. Then

$$F(\Gamma) = S = \langle \mathbf{v}_0 \cdot \Gamma, \mathbf{w}_0 \rangle = \{ \langle \mathbf{v}_0 \cdot \gamma, \mathbf{w}_0 \rangle \,|\, \gamma \in \Gamma \} \subset \mathbb{Z}.$$

If $n \in S$ we say that it is represented by the triple $(\Gamma, \mathbf{v}_0, \mathbf{w}_0)$. In general we are interested in understanding the structure of the set S. Recall that $n \in \mathbb{Z}$ is admissible if

$$n \in S + q\mathbb{Z} \text{ for all } q.$$

In other words, n is locally represented everywhere. We write $\mathcal{A} \subset \mathbb{Z}$ for the set of all admissible numbers. Of course the set S must be contained in $\mathcal{A}$. If $\mathcal{A} = S$, we would say that a local to global principle holds for S. This is maybe to optimistic in general. However one can proof the following *almost every local to global principle* in quite some generality.

Theorem 1.2 (Bourgain-Kontorovich 2010)**.** *Let Γ be thin, free, finitely generated with no parabolic elements and assume that the Poincare series*

$$\mathcal{P}_\Gamma(s) = \sum_{\gamma \in \Gamma} \|\gamma\|^{-2s} \tag{1}$$

converges absolutely for $\mathrm{Re}(s) > 1 - 5 \times 10^{-5}$. (The last condition is rather technical but it can be thought of as Γ being not to thin.) Then there is $\eta_0 > 0$ so that

$$\frac{\sharp(S \cap [1, N])}{\sharp(\mathcal{A} \cap [1, N])} = 1 + O(N^{-\eta_0}).$$

This can be read as (a quantitative version of) for almost every n we have that n is represented (by $(\Gamma, \mathbf{v}_0, \mathbf{w}_0)$) if and only if n is admissible.

The proof will also give a lower bound of the expected size for the multiplicities $\mathrm{mult}_{\Gamma,F,X}(n)$ for almost all admissible $n \asymp X$.

We now turn towards the case $\Gamma = \Gamma_A$. In this case we formulate the following formal conjecture

Conjecture 1.1 (Bourgain-Kontorovich, The Local-Global Conjecture)**.** *Assume $A \geq 2$ (so that Γ_A is Zariski dense in SL_2) and that the image under the map*

F is infinite (i.e. the Zariski closure of $F(\Gamma_A)$ is the affine line). For a growing parameter X and an admissible integer $n \asymp X$ we have

$$\mathrm{mult}_{\Gamma_A,F,X}(n) \asymp X^{2\delta_A-1-o(1)}.$$

1.3. **Applications.** We will now continue to give (real world) applications of this conjecture. These should motivate the rest of this lecture in which we will give full proves of as many of the facts presented in this introduction as possible. We partly follow [10].

1.3.1. *Zaremba's conjecture.* Given $x \in (0,1)$ we can consider the *continued fraction expansion*

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}}.$$

We will write this as $x = [a_1, a_2, \ldots]$ and call the numbers $a_j \in \mathbb{N}$ *partial quotients* of x. For $A \geq 1$ we define

$$\mathfrak{R}_A = \left\{ \frac{b}{d} = [a_1, \ldots, a_k] \,\middle|\, (b,d) = 1 \text{ and } a_j \leq A \text{ for all } j \right\} \text{ and}$$

$$\mathfrak{D}_A = \left\{ d \geq 1 \,\middle|\, \exists b \text{ with } (b,d) = 1 \text{ and } \frac{b}{d} \in \mathfrak{R}_A \right\}.$$

Exercise 1.1. *Show that $\mathcal{D}_1$ is the Fibonacci sequence and that $\{2^j : j \in \mathbb{N}\} \subset \mathcal{D}_3$.*

Conjecture 1.2 (Zaremba)**.** *Every number is the denominator of a reduced fraction whose partial quotients are absolutely bounded. In other words there is $A \geq 1$ so that*

$$\mathfrak{D}_A = \mathbb{N}.$$

(Possibly $A = 5$ suffices.)[1]

It is well known that $\frac{b}{d} = [a_1, \ldots, a_k]$ if and only if

$$\begin{pmatrix} d & b \\ \star & \star \end{pmatrix} = \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_k & 1 \\ 1 & 0 \end{pmatrix}.$$

This brings the semigroup

$$\widetilde{\Gamma}_A = \left\langle \begin{pmatrix} a & 1 \\ 1 & 0 \end{pmatrix} : a \leq A \right\rangle^+ \tag{2}$$

into the picture. Obviously we get

$$\mathcal{D}_A = \langle \mathbf{v}_0 \cdot \widetilde{\Gamma}_A, \mathbf{w}_0 \rangle \tag{3}$$

[1]It was also conjectured by Niederreiter (1978) and Hensley (1996) that $\mathcal{D}_3$ and $\mathcal{D}_2$ contain all sufficiently large integers. Hensley also generalized Zaremba's conjecture by replacing the set $\{1, \ldots, A\}$ by some general alphabet $\mathfrak{A}$. This is the denominators allowed in the partial fraction decomposition must be contained in $\mathfrak{A}$. As observed by J. Bourgain and A. Kontorovich this conjecture needs to modified implementing the notion of admissibility.

for $\mathbf{v}_0 = \mathbf{w}_0 = (1, 0)$. One can now establish a weak form of Zaremba's conjecture as soon as one has an analogue of Theorem 1.2 for semigroups:

Theorem 1.3 (Bourgain-Kontorovich 2011)**.** *For $A = 50$ we have*

$$\frac{\sharp(\mathcal{D}_A \cap [1, N])}{N} \to 1 \text{ as } N \to \infty.$$

In other words, almost every number is the denominator of a reduced fraction whose partial quotients ar bounded by 50.[2]

Zaremba's conjecture is a very beautiful and completely elementary conjecture (requiring only Euclid's algorithm). The partial solution stated above requires (surprisingly) heavy machinery. Of course the full Zaremba Conjecture follows from the Local-Global-Conjecture in the previous subsection with

$$F\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = a.$$

(The difference between Γ_A and $\widetilde{\Gamma}_A$ is not essential.) In this case one can even make the more precise conjecture

$$\mathrm{mult}_{\Gamma_A, F}(n) \sim 2\delta_A \frac{\sharp(\Gamma_A \cap B_n)}{n} \cdot \frac{\pi^2}{6} \cdot \prod_{p|n} (1 - \frac{1}{p}).$$

The almost all version of Zaremba's conjecture stated above actually follows directly from the following result in the thin orbit setting.

Theorem 1.4 (Bourgain-Kontorovich 2011)**.** *There is a constan $c < \infty$ so that for $A \geq 50$ and all sufficiently large N we have*

$$\sharp\{n \leq N : \exists \gamma \in \Gamma_A \text{ with } \gamma_{11} = n\} = N\left(1 + O\left(e^{-c\sqrt{\log(N)}}\right)\right).$$

For those who don't find Zaremba's conjecture itself very motivating we will now give some more (or less) practical applications.

Quasi Monte Carlo numerical integration: Given a finite sequence $X = (x_i)_{i \leq N} \subset [0, 1]^s$ we define the *discrepancy* as

$$D(X) = \max_{\substack{I \subset [0,1]^s, \\ \text{box}}} |\mathrm{Vol}(I) - \frac{1}{N} \sharp\{j \leq N : x_j \in I\}|.$$

The notion of discrepancy appears in many contexts. In many applications it is desirable to choose sequences X minimizing $D(X)$. (Often one has also practical

[2]The lower bound on A was later improved to $A \geq 5$ and probably the techniques can be pushed to $A \geq 4$.

constraints concerning the computability of X.) Indeed in the context of numerical integration on has the Hlawka-Koksma inequality stating that

$$|\int_{[0,1]^s} f(x)dx - \frac{1}{N}\sum_{j=1}^{N} f(x_j)| \ll V(f)\cdot D(X),$$

for functions f of bounded variation $V(f) = \max_{\alpha\subset\{1,\cdots,s\}} \|\partial^\alpha f\|_{L^1} < \infty$.

One has the following result due to Schmidt (1972):

$$D(X) \gg \frac{\log(N)}{N} \text{ for any sequence } X \subset [0,1]^2.$$

It seems desirable to come as close as possible to this lower bound.

Theorem 1.5 (Zaremba)**.** *Take $(b,d) = 1$ with $\frac{b}{d} \in \mathcal{R}_A$. Consider the sequence*

$$X = \{x_j = (\frac{j}{d}, \frac{bj}{d} \ mod\ 1) \colon 1 \le j \le d\} \subset [0,1]^2.$$

Then we have

$$D(X) < \left(\frac{4A}{\log(A+1)} + \frac{4A+1}{\log(d)}\right)\frac{\log(d)}{d}.$$

Thus taking $A = 5$ Zaremba's conjecture allows us to find for each (length) d a corresponding b such that $\frac{b}{d} \in \mathcal{R}_5$. The theorem then implies that the corresponding sequence X essentially optimizes the discrepancy. Theorem 1.4 allows one to consider almost all lengths d.

Pseudo Random Numbers: The easiest pseudo random number generator is the linear congruential number generator

$$x \mapsto bx + c \bmod d.$$

For simplicity we take $x_0 = 1$, $c = 0$, d prime, and b a primitive root modulo d. Then the quality of the pseudo random numbers generated depend on the statistical properties of the sequence

$$X_1 = \{\frac{b^j}{d} \bmod 1 \colon 1 \le j \le d\}.$$

(Note that by Fermat's Little Theorem $\frac{b^d}{d} \bmod 1 = 0$.) Empirically the sequence X_1 behaves nicely for large d (**Exercise**). The next simplest quality test is to look at the serial correlation of pairs

$$X_2 = \{(\frac{b^j}{d}, \frac{b^{j+1}}{d}) \bmod 1 \colon 1 \le j \le d-1\}.$$

Note that X_2 is essentially the sequence defined by Zaremba. We can now make the following observation

Corollary 1.6 (Bourgain-Kontorovich 2011)**.** *There are infinitely many fractions $\frac{b}{d} \in \mathcal{R}_{51}$ with d prime and b a primitive root modulo d. In particular, for such (b,d) the discrepancy of X_2 is essentially best possible.*

Proof. Let $\mathcal{P} = \mathcal{P}_N$ be the set of primes up to N so that

(1) $p \equiv 3 \bmod 4$ and
(2) $(p-1)/2$ is a 10-almost-prime.

Using a standard sieve argument one can show (**Exercise**) that

$$\sharp\mathcal{P} \gg \frac{N}{\log(N)^2}.$$

Note that the cardinality of the exceptional set is much smaller, so that $\mathcal{D}_A \cap \mathcal{P}$ is unbounded as $N \to \infty$ (as long as $\delta_A > \delta_0$). If $p = d$ appears in this intersection then the multiplicity L is at least

$$N^{2\delta_A - 1.001} > N^{10/11}.$$

Thus there are $b_1, \dots, b_L$ distinct numbers so that $\frac{b_j}{d} \in \mathfrak{R}_A$ for $1 \le j \le L$. Take any primitive root r modulo d and write

$$b_j \equiv r_j^k \bmod d.$$

The set of exponents will be denoted by $K = \{k_1, \dots, k_L\}$. We will use the elementary fact, that b_j is a primitive root modulo d if and only if $(k_j, d-1) = 1$.

Let $K' \subset K$ be the subset of $k \in K$ with $(k, d-1) > 2$. Since $d \in \mathcal{P}$ each such k has a prime factor fo size $N^{\frac{1}{10}}$. We conclude that $\sharp K' \ll N^{9/10}$. Thus we obtain a nonempty set $K'' = K \setminus K'$.

Take $k \in K''$ and consider $b \equiv r^k \bmod d$. If $(k, d-1) = 1$ we are done since b is a primitive root. Therefore let us assume $(k, d-1) = 2$. Then b is a square modulo d and we set $b' = d - b$. In particular $b' \equiv -r^k \bmod d$ and since $d \equiv 3 \bmod 4$ we conclude that b' is a primitive root. It remains to verify that $\frac{b'}{d} = 1 - \frac{b}{d} \in \mathfrak{R}_{A+1}$. This can be elementary deduced from $\frac{b}{d} \in \mathfrak{R}_A$. □

Of course Zaremba's conjecture gives a stronger result that for every prime d there is a primitive root b modulo d, which is good in the sense that the sequence X_2 has essentially best possible discrepancy by the theorem of Schmidt.

The Lusztig Conjecture: We now enter the realm of (geometric) representation theory. (We will be brief. For more details see [8, 12, 13].) Let $\overline{\mathbb{F}}_p$ be the algebraic closure of $\mathbb{F}_p$. Let G be a connected and simply connected algebraic group over $\overline{\mathbb{F}}_p$. (We can regard G as a universal Chevalley group. For us it is enough to keep the case $G = \mathrm{SL}_n$ in mind.) Fix a maximal torus $T \subset G$ and assume that the corresponding root system of G agrees with the root system of a Lie algebra $\mathfrak{g}$. Write $\mathfrak{g}_{\mathbb{Z}}$ for the $\mathbb{Z}$-span of a Chevalley basis for $\mathfrak{g}$. Then we can identify the Lie algebra $\mathfrak{g}_p$ of G with $\mathfrak{g}_{\mathbb{Z}} \otimes_{\mathbb{Z}} \overline{\mathbb{F}}_p$. As usual a G-module is a $\overline{\mathbb{F}}_p$ vector space together

with an homomorphism of $G \to \mathrm{GL}(V)$ of algebraic groups. The weight spaces are defined by

$$V_\lambda = \{v \in V \colon tv = \lambda(t)v \text{ for all } t \in T\}.$$

If V is finite dimensional one defines the formal character $\mathrm{ch}(V) \in \mathbb{Z}[X]$ by setting

$$\mathrm{ch}(V) = \sum_{\lambda \in X} \dim(V_\lambda) e(\lambda).$$

(Here $\{e(\lambda)\}_{\lambda \in X}$ is the standard basis of the group ring $\mathbb{Z}[X]$. Note that we identify the group of characters of T with the group of integral weights X.) This character is W-invariant.

It can be shown that any simple G-module is finite dimensional. Furthermore, for any dominant $\mu \in X_+$ there is a (unique up to isomorphism) simple G-module $L_p(\mu)$ such that $\dim L_p(\mu)_\mu = 1$ and such that $L_p(\mu)_\nu \neq \{0\}$ implies $\nu \leq \mu$. We call $L_p(\mu)$ simple G-module with highest weight μ. Every simple G-module is isomorphic to $L_p(\mu)$ for exactly one $\mu \in X_+$.

One can also construct simple modules using reduction modulo p. Let $V(\mu)$ be the (classical) simple $\mathfrak{g}$-module of highest weight μ. We choose a minimal admissible lattice $V_{\mathbb{Z}}(\mu)$ and define $V_p(\mu) = V_{\mathbb{Z}}(\mu) \otimes_{\mathbb{Z}} \overline{\mathbb{F}}_p$. The latter can be regarded as a G-module. One calls $V_p(\mu)$ the Weyl module with highest weight μ. Of course we have

$$\mathrm{ch}(V_p(\mu)) = \mathrm{ch}(V(\mu)) \text{ for all } \mu \in X_+$$

In particular, this character can be computed using Weyl's character formula. The unique simple quotient of $V_p(\mu)$ is $L_p(\mu)$. It can be shown that the composition factors of $V_p(\mu)$ have the form $L_p(\nu)$ with $\nu \leq \mu$ and $L_p(\mu)$ has multiplicity one. Thus there are coefficients $b_{\mu,\nu}$ with $b_{\mu,\mu} = 1$ so that

$$\mathrm{chL_p}(\mu) = \sum_\nu b_{\mu,\nu} \mathrm{ch}(V(\nu)).$$

Knowing these coefficients would allow us to compute all the characters. Lusztig stated a conjecture for the coefficients, which we will state now. Let h be the Coexter number (i.e. $h = \max_{\alpha \in R^+}\{\rho(h_\alpha) + 1\}$, where ρ is the half sum of positive roots, $R^+ \subset R$ is a set of positive roots and $h_\alpha \in [\mathfrak{g}_\alpha, \mathfrak{g}_{-\alpha}]$ is the unique element with $\alpha(h_\alpha) = 2$) of the rootsystem of $\mathfrak{g}$. (For $G = \mathrm{SL}_n(\overline{\mathbb{F}}_p)$ we have $h = n$.) We define

$$\mathcal{M}_p = \{\mu \in X_+ \colon (\mu + \rho)(h_\alpha) \leq p(p - h + 2) \text{ for all } \alpha \in R^+\}.$$

We have $\mathcal{M}_p = \emptyset$ for $p \leq h - 2$. Let W_a be the affine Weyl group of R. There is an action $\circ_p$ of W_a on X (For $w \in W$ we just have $w \circ_p \nu = w \circ \nu$ but if w is the translation by $\gamma \in \mathbb{Z}R$ then $w \circ_p \nu = \nu + p\gamma$). Let

$$C^- = \{\lambda \in X \colon -p \leq (\lambda + \rho)(h_\alpha) \leq 0 \text{ for all } \alpha \in R^+\}.$$

This is a fundamental domain for the $\circ_p$-action of W_a on X. We can write $\mu = w \circ_p \lambda$ for $\lambda \in C^-$. In a long series of papers (written by subsets of the named mathematicians and maybe others) the following result was established:

Theorem 1.7 (Andersen, Jantzen, Soergel, Kashiwara, Tanisaki, Kazhdan, Lusztig)**.** *Let p be sufficiently large*[3] *and $\mu = w \circ_p \lambda \in \mathcal{M}_p$ for w with minimal length. Then*

$$\mathrm{ch}(\mu) = \sum_{\substack{x \in W_a, \\ x \circ_p \lambda \in X_+}} \det(wx) P_{x,w}(1) \mathrm{ch}(V(x \circ_p \lambda)),$$

where $P_{x,w}$ is the Kazhdan-Lusztig polynomial attache to w and x.

This theorem establishes Lusztig's conjecture for sufficiently large p. Of course it is desirable to understand what this means. First, for $p < h$ the statement must fail and there seems to be no conjecture that predicts character formula for $L_k(\mu)$ in this case. However, some low dimensional evidence and maybe intuition leads to the following question

Question. Does Lusztig's conjecture hold for all $p > 0$ (or more conservatively for all $p > 2h - 2$)?

In spectacular fashion the answer to this question turns out to be no. First the answer no was given conditional on the existence of Fibonacci primes. However, using our results towards Zaremba's conjecture we can give a definite unconditional answer.

Theorem 1.8 (Kontorovich, McNamara, Williamson)**.** *Let $G = \mathrm{SL}_n(\overline{\mathbb{F}}_p)$. There is no sub-exponential function $f(n)$ such that Lusztig's conjecture holds for all $p \geq f(n)$.*

We use the following black boxes:

- If the torsion of SL_n grows exponentially in n, then no sub-exponential (lower) bound for p (in n) is sufficient to ensure the validity of Lusztig's conjecture.
- Let $\Gamma = \langle T, -ST^{-1}S \rangle^+$. Given $\gamma \in \Gamma$ we write $l(\gamma)$ for the wordlength of γ in the generators of Γ. Suppose p divides any coefficient γ_{ij} of any matrix $\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \in \Gamma$, then p occurs as torsion in SL_{2l+5} with $l = l(\gamma)$.

This reduces Theorem 1.8 to the following statement

Theorem 1.9. *There are absolute constants $\tau > 0$ and $c > 1$ so that for all sufficiently large L there is $\gamma \in \Gamma$ of wordlength $l(\gamma) \leq L$ and top left entry $\gamma_{11} = p$ prime with $p > \tau c^L$. Even more*

$$\sharp\{p > \tau c^L \colon \exists \gamma \in \Gamma \text{ with } l(\gamma) \leq L \text{ and } \gamma_{11} = p\} \gg \frac{c^L}{L}.$$

[3]Large means gigantic. For example if $G = \mathrm{SL}_n$ it means at least $p \gg n^{n^2}$.

Proof. Define

$$S_1 = \{p > \tau c^L \colon \exists \gamma \in \Gamma \text{ with } l(\gamma) \leq L \text{ and } \gamma_{11} = p\}.$$

Take a parameter A and write $l_A(\gamma)$ to be the wordlength of $\gamma \in \Gamma_A$ in the generators of Γ_A. Observe

$$\begin{pmatrix} a & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} b & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix}.$$

Thus Γ_A is a sub-semigroup of Γ. We also have the wordlength relation

$$l(\gamma) \leq 2A l_A(\gamma).$$

For the purpose of obtaining a lower bound we can therefore decrease the set S_1 to

$$S_2 = \{p > \tau c^L \colon \exists \gamma \in \Gamma_A \text{ with } l(\gamma) \leq \frac{L}{2A} \text{ and } \gamma_{11} = p\}.$$

It is easy to see inductively that $\|\gamma\|_\infty = \gamma_{11}$ for $\gamma \in \Gamma_A$. Define

$$\varphi = \frac{1+\sqrt{5}}{2} \text{ and } \overline{\varphi} = \frac{1-\sqrt{5}}{2}.$$

These are the eigenvalues of $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$. For any

$$\gamma = \prod_{i=1}^{n} \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} b_i & 1 \\ 1 & 0 \end{pmatrix} \in \Gamma_A$$

we have

$$\|\gamma\|_\infty = \begin{pmatrix} 1 & 0 \end{pmatrix} \gamma \begin{pmatrix} 1 \\ 0 \end{pmatrix} \geq \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^{2n} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = F_{2n+1}.$$

It is a well known fact that the mth Fibonacci number satisfies

$$F_m = (\varphi^m - \overline{\varphi}^m)/\sqrt{5}.$$

For all $\gamma \in \Gamma_A$ we obtain

$$\|\gamma\|_\infty \geq d\varphi^{2l_A(\gamma)} \text{ for } d = \varphi/\sqrt{5}.$$

We define $N = d\varphi^{L/A}$ and further decrease S_2 to

$$\begin{aligned} S_3 &= \{p > \tau c^L \colon \exists \gamma \in \Gamma_A \text{ with } \|\gamma\| \leq N \text{ and } \gamma_{11} = p\} \\ &= \{\tau c^L < p \leq N \colon \exists \gamma \in \Gamma_A \text{ with } \gamma_{11} = p\}. \end{aligned}$$

We are done since the prime number theorem together with Theorem 1.4 implies

$$\sharp\{p \in (\theta N, N] \colon \exists \gamma \in \Gamma_A \text{ with } \gamma_{11} = p\} = (1-\theta)\frac{N}{\log(N)}(1+o(1)).$$

Thus we get

$$\sharp S_1 \geq \sharp S_2 \geq \sharp S_3 \gg \frac{N}{\log}(N) \gg c^L L^{-1}$$

and are done. □

1.3.2. *A Question by Einsiedler, Lindenstrauss, Michel and Venkatesh.* Let us raise the question straight away and then explain some terminology:

Does there exist a compact subset $Y \subset T^1(\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H})$ of the unit tangent bundle of the modular surface which contains infinitely many fundamental closed geodesics?

In order to explain the question we will have to take a glimpse at hyperbolic geometry. Let $\mathbb{H}$ be the upper half plane and denote the tangent bundle by $T\mathbb{H}$ (resp. $T^1\mathbb{H}$ the unit tangent bundle). The group of isometries of $\mathbb{H}$ is $\mathrm{PSL}_2(\mathbb{R})$ and it acts on the unit tangent bundle by

$$T^1\mathbb{H} \ni (z,\zeta) \mapsto \left(\frac{az+b}{cz+d}, \frac{\zeta}{(cz+d)^2}\right).$$

Here we of course have $\zeta \in S^1 \subset \mathbb{C}$ and $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{R})$. The invariant measure on $T^1(\mathbb{H})$ is

$$d\mu = \frac{dxdyd\theta}{y^2} \text{ for } \theta = \arg(\zeta).$$

The geodesics in $\mathbb{H}$ are semi-circles orthogonal to $\partial\mathbb{H}$. (This includes vertical half lines!) A point $(z,\zeta) \in T^1\mathbb{H}$ determines a geodesic through z in direction ζ. Following the geodesic flow for time t moves the point z along this geodesic to a the point at distance t from z. Taking t to ∞ brings us to the virtual point on the boundary $\partial\mathbb{H}$. (One can identify $T^1\mathbb{H}$ with $\mathrm{PSL}_2(\mathbb{R})$. The geodesic flow in PSL_2 becomes right multiplication by the diagonal subgroup $A = \{a_t = \mathrm{diag}(e^{t/2}, e^{-t/2})\}$.)

We are interested in the quotient $X = \mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}$. The unit tangent bundle T^1X can be identified with $\mathrm{PSL}_2(\mathbb{Z})\backslash \mathrm{PSL}_2(\mathbb{R})$. A closed geodesic starts at some point $\mathrm{PSL}_2(\mathbb{Z}) \cdot g$ and returns after a least time $l > 0$ to the same point including the tangent vector. We can write this as

$$\mathrm{PSL}_2(\mathbb{Z}) \cdot ga_l = \mathrm{PSL}_2(\mathbb{Z}) \cdot g \text{ or } ga_l = \pm Mg \text{ for some } M \in \mathrm{PSL}_2(\mathbb{Z}).$$

Of course $M = ga_lg^{-1}$, has eigenvalues $e^{\pm l/2}$ and has trace

$$\mathrm{Tr}(M) = 2\cosh(l/2).$$

Since M is only determined up to conjugation by $\mathrm{PSL}_2(\mathbb{Z})$ we conclude that primitive closed geodesics correspond to primitive hyperbolic conjugacy classes $[M]$ in Γ. (Primitive means $[M]$ is not of the form $[N^k]$ for $k \geq 2$.) The visual point α from g is fixed by M and given by

$$\alpha = \lim_{t\to\infty} a_t \cdot i = \frac{a-d+\sqrt{\mathrm{Tr}(M)^2-4}}{2c} \text{ for } M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

The Galois conjugate of α is the visual point of the backwards geodesic flow.

Recall that $\mathrm{PSL}_2(\mathbb{Z})$ is generated by T and S and let

$$\mathcal{F} = \underbrace{\{\mathrm{Re}(z) > -\frac{1}{2}\}}_{=\mathcal{F}_T} \cap \underbrace{\{\mathrm{Re}(z) < \frac{1}{2}\}}_{=\mathcal{F}_{T^{-1}}} \cap \underbrace{\{|z| > 1\}}_{=\mathcal{F}_S}$$

be the standard fundamental domain. If we follow the geodesic flow through $\mathcal{F}$, we will leave the fundamental domain at some point passing through a wall of some $\mathcal{F}_L$ for $L \in \{T, T^{-1}, S\}$. It is clear that L is also the transformation that needs to be applied to re-enter in $\mathcal{F}$. We record this letter L. Therefore, given a starting point $(z, \zeta) \in T^1X$ we obtain the corresponding cutting sequence, which is a sequence of letters L, by following the geodesic through $\mathcal{F}$ as just described. The sequence consists of a number of T's followed by one S followed by a number of T^{-1}'s followed by S and so on. It suffices to record the number of T's and T^{-1}'s. For example

$$TTST^{-1}T^{-1}T^{-1}STST^{-1}T^{-1}S\cdots \rightsquigarrow 2, 3, 1, 2, \cdots .$$

It (miraculously) turns out that this sequence is closely related to the partial fraction expansion of the visual point α.

We call a quadratic irrational α reduced if it and its Galois conjugate α' satisfy the inequalities $-1 < \alpha' < 0 < 1 < \alpha$. A representative M of the conjugacy class $[M]$ is called reduced if its visual point α is. One can show that α is reduced if and only if its continued fraction is exactly periodic. For reduced M, the coding of the geodesic flow corresponds exactly to the sequence giving the partial fraction decomposition of α.

An integral quadratic form is given by

$$Q(x, y) = Ax^2 + Bxy + Cy^2.$$

We call Q primitive if $(A, B, C) = 1$. A number n is said to be represented by Q if there are $x, y \in \mathbb{Z}$ with $n = Q(x, y)$. A linear change of variables

$$Q(x, y) \mapsto Q(ax + by, cx - dy) \text{ for } ad - bc = \pm 1$$

does not change the set of integers represented by Q. We call two quadratic forms Q, Q' (strictly) equivalent if there is $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ such that $Q(x, y) = Q'(\gamma(x, y))$. This gives an equivalence relation $Q \sim Q'$. We define the discriminant of Q by

$$D_Q = B^2 - 4AC.$$

We call Q definite if $D_Q < 0$. (In this case it only takes negative or positive values.) When $D_Q > 0$, we call Q indefinite. Note that $D_Q \equiv 0, 1 \bmod 4$. Furthermore, if $Q \sim Q'$, then $D_Q = D_{Q'}$. Write $\alpha_Q = \frac{-B+\sqrt{D_Q}}{2A}$ for the root of $Q(x, 1)$. We define the class group as

$$\mathcal{C}_D = \{[Q] \colon D_Q = D\} \text{ and write } h_D = \sharp\mathcal{C}_D.$$

Gauß showed that $\mathcal{C}_D$ has a group structure (justifying the name) and that for non-square D one has $1 \leq h(D) < \infty$.

A discriminant is fundamental if it is the discriminant of a (quadratic) field. Such D's are either congruent 1 modulo 4 and squarefree or $D/4$ is squarefree and congruent $2, 3$ modulo 4. Dirichlet's class number formula says that

$$h_D = \sqrt{|D|} L(1, \chi_D) \cdot \begin{cases} (2\pi)^{-1} & \text{for } D \leq -5, \\ \log(\epsilon_D)^{-1} & \text{for } D > 0. \end{cases}$$

The class numbers are very mysterious and important objects. There is lots of work (and conjectures) on them but we will not discuss this in more detail now.

Given a hyperbolic matrix M we want to attach a quadratic form Q_M. This is done essentially by equating α_{Q_M} with the visual point α of the geodesic associated to M. This leads

$$Q_0 = cx^2 + (d-a)xy - by^2.$$

However, this does not need to be primitive. To fix this we set

$$Q_M = \frac{\operatorname{sgn}(\operatorname{Tr}(M))}{s} Q_0 \text{ for } s = \gcd(c, d-a, -b).$$

The sign makes this independent of the choice $\pm M$, which is desirable since we are working with PSL_2. We can invert this map by defining

$$M_Q = \begin{pmatrix} (t - Bs)/2 & -Cs \\ As & (t + Bs)/2 \end{pmatrix} \text{ where } Q = Ax^2 + Bxy + Cy^2$$

and (t, s) is a fundamental solution to the Pellian equation $T^2 - S^2 D_Q = 4$. (Taking a fundamental solution ensures that M_Q is primitive.) One can also write down the inverse map using the continued fraction expansion.

The discriminant of a closed geodesic γ on the modular surface or its corresponding hyperbolic conjugacy class is defined to be that of its associated equivalence class of binary quadratic forms. This is given by

$$D_M = \frac{\operatorname{Tr}(M)^2 - 4}{s^2} \text{ for } s = \gcd(c, d-a, b).$$

If D_M is a fundamental discriminant, then we call M as well as the corresponding closed geodesic fundamental.

Given a fundamental discriminant D we have constructed a correspondence between elements in the class group $\mathcal{C}_D$ and fundamental closed geodesics. We abuse notation and write $\gamma \in \mathcal{C}_D$. This explains the question raised by Einsiedler-Lindenstrauss-Michel-Venkatesh in 2004. Let us briefly explain the motivation behind the question.

Theorem 1.10 (Duke's Theorem)**.** *As $D \to \infty$ through fundamental discriminants*

$$\frac{1}{h_D} \sum_{\gamma \in \mathcal{C}_D} \frac{1}{l(\gamma)} \int_\gamma \mathbb{1}_A ds \to \frac{1}{\operatorname{Vol}(X)} \int_X \mathbb{1}_A \frac{dxdy}{y^2},$$

for domains $A \subset X$.

This raises the question if individual geodesics already equidistribute. (Of course if $h_D = 1$ this is the case.) This will not be the case in general. Even among fundamental closed geodesics one can find examples of sequences where the limiting measure is dy/y instead of $dxdy/y^2$. This happens for examples if we allow the mass of the geodesics to escape towards the cusp. The question now arises when one wants to exclude this phenomenon. This is asking for sequences of low-lying geodesics that not equidistribute.

Let us translate the question into thin orbit language. Recall that $\alpha = [\overline{a_0, \dots, a_l}]$ is fixed by the matrix

$$M = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_l & 1 \\ 1 & 0 \end{pmatrix}$$

and $\alpha \in \mathbb{Q}(\sqrt{\mathrm{Tr}(M)^2 - 4})$. One can see that if $\mathrm{Tr}(M)^2 - 4$ is squarefree, then the corresponding closed geodesic is fundamental. In particular one can deduce the answer of the question from Conjecture 1.1 with the map $F(\gamma) = \mathrm{Tr}(\gamma)$. But one can even give an unconditional answer.

Theorem 1.11 (Bourgain-Kontorovich 2014)**.** *There exist infinitely many low-lying fundamental geodesics. More precisely for each $\epsilon > 0$ there is a compact region $Y = Y(\epsilon) \subset X$ and a set $\mathcal{D} = \mathcal{D}(\epsilon)$ of positive fundamental discriminants, such that*

(1) *for each $D \in \mathcal{D}$, many of the geodesics in the ccorresponding class group are low-lying:*
$$\sharp\{\gamma \in \mathcal{C}_D : \gamma \subset Y\} > h_D^{1-\epsilon};$$
(2) *there are many discriminants in $\mathcal{D}$:*
$$\sharp\mathcal{D} \cap [1, T] > T^{\frac{1}{2}-\epsilon} \text{ as } T \to \infty.$$

We will reduce this statement to the following result

Theorem 1.12 (Bourgain-Kontorovich 2014)**.** *Many elements $\gamma \in \Gamma_A$ have traces satisfying $\mathrm{Tr}(\gamma)^2 - 4$ being square-free. More precisely for any $\eta > 0$ there is $A = A(\eta) < \infty$ such that*

$$\sharp\{\gamma \in \Gamma_A \cap B_X : \mathrm{Tr}(\gamma)^2 - 4 \text{ is square-free}\} > X^{2\delta_A - \eta}.$$

as $X \to \infty$.

Proof of Theorem 1.11. The reduction to Theorem 1.12 is a simple consequence of the discussion above. We set

$$\mathcal{T} = \{t \geq 1 : t^2 - 4 \text{ square-free}\}.$$

For an integer t and $A < \infty$ define the trace multiplicity by

$$\mathcal{M}_A(t) = \sharp\{\gamma \in \Gamma_A : \mathrm{Tr}(\gamma) = t\}.$$

From Theorem 1.12 we deduce that

$$\begin{aligned} N^{2\delta-\eta} &< \sum_{t\in\mathcal{T}\cap[1,N]} \mathcal{M}_{A,N}(t) \\ &= \sum_{t\in\mathcal{T}\cap[1,N]} \mathcal{M}_{A,N}(t)(\mathbb{1}_{\mathcal{M}_{A,N}(t)\geq W} + \mathbb{1}_{\mathcal{M}_{A,N}(t)<W}) \\ &\ll N^{1+\epsilon} \sum_{t\in\mathcal{T}\cap[1,N]} \mathbb{1}_{\mathcal{M}_{A,N}(t)\geq W} + NW. \end{aligned}$$

Setting $W = N^{2\delta-1-2\eta}$ (and renaming constants) yields

$$\sum_{t\in\mathcal{T}\cap[1,N]} \mathbb{1}_{\mathcal{M}_{A,N}(t)\geq W} > N^{2\delta-1-\eta}.$$

Given $\epsilon > 0$ we take $\eta > 0$ sufficiently small and A large enough so that

$$2\delta - 1 - \eta > 1 - \epsilon.$$

To the choice $A > 0$ we have a corresponding compact region

$$Y = Y(\epsilon) \subset X = T^1(\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}).$$

We define

$$\mathcal{D} = \{D = t^2 - 4 \colon t \in \mathcal{T},\, \mathcal{M}_A(t) > t^{2\delta-1-\eta}\}.$$

All $D \in \mathcal{D}$ are square-free so that D is fundamental. Of course the corresponding geodesic is also fundamental. By construction we have

$$\sharp(\mathcal{D}\cap[1,T]) \geq \sharp\{t \in \mathcal{T}\cap[1,\sqrt{T}] \colon \mathcal{M}_A(t) > t^{2\delta-1-\eta}\} > T^{\frac{1}{2}-\epsilon}.$$

This confirms (2). Finally, for each $D = t^2 - 4 \in \mathcal{D}$ the trace multiplicity satisfies

$$\mathcal{M}_A(t) > t^{1-\epsilon} > (\sqrt{D})^{1-\epsilon} \gg (\sharp\mathcal{C}_D)^{1-\epsilon}.$$

Note that not each $\gamma \in \Gamma_A$ corresponds uniquely to a closed geodesic in X. However, the corresponding visual points of the geodesics are all reduced. This implies that any two differ only by a cyclic permutation of their partial quotients. Since there are only $\ll \log(t)$ such permutations we are done. □

1.3.3. *McMullen's (classical) Arithmetic Chaos.* A similar question is known as McMullen's Arithmetic Chaos:

Conjecture 1.3 (Arithmetic Chaos V1)**.** *There is a compact subset $Y \subset X$ such that for all real quadratic fields K, the set of closed geodesics defined over K and lying in Y has positive entropy.*

This can be directly formulated in terms of continued fractions:

Conjecture 1.4 (Arithmetic Chaos V2)**.** *There is $A < \infty$ so that, for any real quadratic field K the set*

$$\{[\overline{a_0, a_1, \dots, a_l}] \in K \colon a_j \leq A\}$$

has exponential growth as $l \to \infty$.

The only observation needed for the reformulation is that a geodesic goes high in the cusp, then the corresponding partial fraction will have large partial quotients. The reason we stated this conjecture is the following result

Proposition 1.13. *Conditionally on Conjecture 1.1 McMullen's Arithmetic Chaos conjecture is true.*

Proof. By assuming $A \geq 2$ we can assume that all integers n are admissible. Taking $F = \mathrm{Tr}$ and $n \asymp X$ the local to global conjecture tells us that

$$\sharp\{\gamma \in B_X \cap \Gamma_A \colon \mathrm{Tr}(\gamma) = n\} \gg X^{\eta},$$

for some $\eta = 2\delta_A - 1 + o(1) > 0$ where we take A and X large enough. Similar reductions as we have seen earlier now reveals that

$$\{[\overline{a_0, a_1, \dots, a_l}] \in K \colon a_j \leq A\} \gg e^{\eta' l}.$$

where $K = K_n = \mathbb{Q}(\sqrt{n^2 - 4})$. Here we again used that $\log(\|\gamma\|) \asymp l_A(\gamma)$. Given a fixed real quadratic field $K = \mathbb{Q}(\sqrt{D})$ it remains to find $n \asymp X$ so that $K_n = K$. This is possible for (in terms of D) large enough X by solving the classical Pell equation $n^2 - d^2 D = 4$. □

1.4. **Odds and Ends.** Towards the end of the introduction we supply some explanations that were skipped above. First recall the notion of Hausdorff dimension. For $A \subset S^1$ the s-dimensional *Hausdorff measure* is given by

$$H^s(A) = \lim_{\epsilon \to 0} \inf\{\sum_j \mathrm{Vol}(I_j)^s \colon A \subset \bigcup_j I_j, \mathrm{Vol}(I_j) < \epsilon\}.$$

(Without changing anything we can treat the situation $A \subset [0, 1]$.) One can see that there is a threshold d such that

$$H^s(A) = \begin{cases} \infty & \text{if } s < d, \\ 0 & \text{if } s > d. \end{cases}$$

We define the *Hausdorff dimension* to be this threshold:

$$\dim_H(A) = d.$$

Some easy examples are

- The set $A = S^1$ has Hausdorff dimension 1.
- The Cantor middle third set has Hausdorff dimension $\frac{\log(2)}{\log(3)}$.

Let us (approximately) compute another Hausdorff dimension. Define the set

$$\mathcal{C}_A = \{[a_1, a_2, \ldots] \colon a_j \leq A \text{ for all } j\}.$$

This can be thought of as the limiting set of Γ_A. Put $\delta_A = \dim_H(\mathcal{C}_A)$. We also define

$$\mathcal{C}_A^{(r)} = \{[a_1, a_2, \ldots, a_r] \colon a_j \leq A \text{ for all } j\} \text{ and } \mathcal{C}_A^{(\infty)} = \bigcup_{r \in \mathbb{N}} \mathcal{C}_A^{(r)}.$$

We will slightly abuse notation and view r-tuples $\mathbf{a} = (a_1, \ldots, a_r) \in [1, A]^r \cap \mathbb{N}^r$ as elements in $\mathcal{C}_A^{(r)}$ by considering the associated continued fraction. We write $l(\mathbf{a})$ for the length of $\mathbf{a}$ and $\mathrm{den}(\mathbf{a})$ for the denominator of $[a_1, \ldots, a_{l(\mathbf{a})}]$. With this notation at hand we define the zeta function

$$\zeta(s, A) = \sum_{\mathbf{a} \in \mathcal{C}_A^{(\infty)}} \mathrm{den}(\mathbf{v})^{-s}.$$

This sum converges for s with sufficiently large real part. Let $D(A)$ denote the abscissa of convergence of $\zeta(s, A)$.

Lemma 1.14 (Cusick)**.** *We have $D(A) = 2\delta_A$.*

Proof. First rewrite the zeta function as

$$\zeta(s, A) = \sum_{n \in \mathbb{N}} \frac{r_A(n)}{n^s} \text{ for } r_A(n) = \sharp\{\mathbf{v} \in \mathcal{C}_A^{\infty} \colon \mathrm{den}(\mathbf{v}) = n\}.$$

Take $\alpha = [a_1, a_2, \ldots] \in \mathcal{C}_A$. Truncating the continued fraction at $k \in \mathbb{N}$ gives us the convergents

$$\frac{p_k}{q_k} = [a_1, \ldots, a_k].$$

These satisfy the approximation

$$|\alpha - \frac{p_k}{q_k}| < \frac{1}{q_k^2}.$$

Having made this observation we can find $\alpha_j = [a_1^{(j)}, \ldots] \in \mathcal{C}_A$ so that $\mathcal{C}_A$ is covered by intervals of length $\frac{2}{(q_k^{(j)})^2}$ for $\frac{p_k^{(j)}}{q_k^{(j)}} = [a_1^{(j)}, \ldots, a_k^{(j)}]$. We get the estimate

$$\sum_j \frac{2}{(q_k^{(j)})^{2\beta}} \leq 2 \sum_{n=1}^{\infty} \frac{r_A(n)}{n^{2\beta}} = 2\zeta(2\beta, A).$$

Now by definition of the Hausdorff dimension the left hand side diverges for $\beta < \delta_A$. But this implies that the zeta function (defined as sum) diverges, so that $2\beta \leq D(A)$ for all $\beta < \delta A$. This implies $2\delta_A \leq D(A)$.

For the reverse inequality we have to work harder. Write

$$f_r(\delta) = \sum_{\mathbf{v} \in \mathcal{C}_A^{(r)}} \mathrm{den}(\mathbf{v})^{-2\delta}.$$

This family of functions satisfies $A^{-2\delta} f_n(\delta) f_m(\delta) < f_{m+n}(\delta) < f_m(\delta) f_n(\delta)$ for $m, n \in \mathbb{N}$ (**Exercise**).

We claim that for $n \geq 2$ there is a unique solution $0 < \sigma_n < 1$ to $f(\delta) = 1$ and that

$$\lim_{n\to\infty} (\sigma_n) = \delta_A.$$

Furthermore, for $m, r \geq 2$, we have $\sigma_{rm} < \sigma_m$, and $\sigma_m > \delta_A$.

We will now show that $\zeta(2\sigma_m, A)$ converges for each $m \geq 2$. This directly implies $D(A) \leq 2\delta_A$ as desired. Put $N = rm$ and look at $n = jN + t$. Then we have

$$f_n(\sigma_m) < f_N(\sigma_m)^j f_t(\sigma_m).$$

With this at hand we can estimate

$$\zeta(2\sigma_m, A) = \sum_{n=1}^{\infty} f_n(\sigma_m) < C \sum_{i=0}^{\infty} f_N(\sigma_m)^i < \infty,$$

for $C = \sup_{0\leq t\leq N} f_t(\sigma_m)$.

The existence of these numbers is straight forward. We find

$$1 = f_{m+n}(\sigma_{m+n}) < f_n(\sigma_{m+n}) f_m(\sigma_{m+n}).$$

Thus one of the factors on the right is bigger than one, which implies $\sigma_{m+n} < \max(\sigma_n, \sigma_m)$. Inductively taking $m = n, 2n, 3n, \ldots$ gives the desired property $\sigma_{rm} < \sigma_m$. We directly obtain the generalization

$$\sigma_{rn+sm} < \max(\sigma_m, \sigma_n).$$

Let $\sigma = \limsup_{n\to\infty} \sigma_n$. Given two primes p, q we take a sufficiently large number n of the form $n = rp + sq$ and obtain $\sigma \leq \max(\sigma_p, \sigma_q)$. Thus there is at most one prime p_e with $p_e < \overline{\sigma}$. Using the properties of the family f_n one obtains

$$f_{rn}(\delta) > A^{-2r} f_n(\delta)^r.$$

By the mean value theorem there is $\sigma_{rn} \leq \delta \leq \sigma_n$ so that

$$\begin{aligned}
1 - A^{-2r} &= 1 - A^{-2r} f_n(\sigma_n) > 1 - f_{rn}(\sigma_n) \\
&= f_{rn}(\sigma_{rn}) - f_{rn}(\sigma_n) = (\sigma_n - \sigma_{rn}) f'_{rn}(\delta) \\
&> (\sigma_n - \sigma_{rn}) 2 \min_{\mathbf{v}} \log(\mathrm{der}(\mathbf{v})) f_{rn}(\delta) \\
&> (\sigma_n - \sigma_{rn}) 2A^{-2r} f_n(\delta) \min_{\mathbf{v}} \log(\mathrm{der}(\mathbf{v})) \\
&> (\sigma_n - \sigma_{rn}) 2A^{-2r} \log(A^{\frac{1}{2}rn}).
\end{aligned}$$

We conclude that $\sigma_n - \sigma_{rn} < C_A \frac{A^{2r}}{rn}$. Suppose there is n with $\sigma_n = \sigma - \epsilon$. Then we find a large prime $p \neq p_e$ so that $C_A \frac{A^{2n}}{pn} < \epsilon$. Thus we have $\sigma_p - \sigma_{pn} < \epsilon$. This gives the contradiction

$$\sigma_p > \sigma = \sigma_n + \epsilon \geq \sigma_{pn} + \epsilon > \sigma_p + \epsilon.$$

Thus we have seen that

$$\liminf_{n\to\infty} \sigma_n \geq \sigma = \limsup_{n\to\infty} \sigma_n,$$

so that $\sigma = \lim_{n\to\infty} \sigma_n$. It is a nice **Exercise** to identify sigma with δ_A. □

Lemma 1.15 (Hensley)**.** *We have* $1 - \frac{5}{2A} < \delta_A < 1 - \frac{1}{10(A+1)}$. *Even more*

$$\delta_A = 1 - \frac{6}{\pi^2 A} + o(A^{-1}).$$

We will omit the proof. Note that it is easier to show $1 - \delta_A \asymp A^{-1}$. The latter suffices to deduce that by choosing A sufficiently large we can make δ_A as close to 1 as desired.

Finally we can complete the estimate for $\sharp(\Gamma_A \cap B_X)$ used earlier.

Lemma 1.16 (Hensley)**.** *Let* $A \geq 2$*, then*

$$\sharp(\Gamma_A \cap B_X) \asymp X^{2\delta_A},$$

where δ_A *is as above.*

Proof. First recall the definitions of Γ_A and $\widetilde{\Gamma}_A$. Of course the count only differs up to constants. Therefore it is enough to show that $\sharp(\widetilde{\Gamma}_A \cap B_X) \asymp X^{2\delta_A}$. We can take the ball B_X with respect to the maximum norm and it is an easy observation that, if $\gamma \in \widetilde{\Gamma}_A$, then the upper left entry is always the largest. Employing an earlier observation we find that

$$\sharp(\Gamma_A \cap B_X) \asymp \sharp(\widetilde{\Gamma}_A \cap B_X) \asymp \sharp\{\frac{b}{d} \in \mathcal{R}_A \colon d \leq X\} \asymp \sharp\{\mathbf{a} \in \mathcal{C}_A^{(\infty)} \colon \mathrm{den}(\mathbf{v}) \leq X\}.$$

Let us denote the number on the right by $F_A(X)$.

Take $\mathbf{u} \in \mathcal{C}_A^{(\infty)}$ with $\mathrm{den}(\mathbf{u}) > X$. Then we can (uniquely) write $\mathbf{u} = (\mathbf{a}, \mathbf{w})$ so that

$$\frac{X}{1+w_1} < \mathrm{den}(\mathbf{a}) \leq X < \mathrm{den}((\mathbf{a}, w_1)).$$

For $s > D(A)$ we estimate

$$\sum_{\mathbf{w}\in\mathcal{C}_A^{(\infty)}} \sum_{\substack{\mathbf{a}\in\mathcal{C}_A^{(\infty)},\\ \frac{X}{1+w_1}<\mathrm{den}(\mathbf{a})\leq X}} \mathrm{den}((\mathbf{a},\mathbf{w}))^{-s} > \frac{1}{4} \sum_{\mathbf{w}\in\mathcal{C}_A^{(\infty)}} \mathrm{den}(\mathbf{w})^{-s} \sum_{\substack{\mathbf{a}\in\mathcal{C}_A^{(\infty)},\\ \frac{X}{1+w_1}<\mathrm{den}(\mathbf{a})\leq X}} \mathrm{den}(\mathbf{a})^{-s}$$

$$> \frac{1}{4}(F_A(X) - F_A(X/2))x^{-s}\zeta(s, A).$$

This leads to

$$\frac{1}{2}\zeta(s, A) > \frac{1}{4}(F_A(X) - F_A(X/2))X^{-s}\zeta(s, A).$$

We infer that $F_A(X) - F_A(X/2) \leq 4X^{D(A)}$. This directly implies $F_A(X) \ll X^{D(A)}$. This completes the proof of the upper bound by the first of the two lemmata above.

Turning to the lower bound we look at another decomposition. Suppose $\mathbf{w} \in \mathcal{C}_A^{(\infty)}$ satisfies $\mathrm{den}(\mathbf{w}) > (A+1)^2 X$. Then we write $\mathbf{w} = (\mathbf{v}, k, \mathbf{u})$ with $\mathbf{v}, \mathbf{u} \in \mathcal{C}_A^{(\infty)}$ and $k \in \mathbb{N}$ such that

$$\frac{X}{A+1} < \mathrm{den}(\mathbf{v}) \leq X < \mathrm{den}((\mathbf{v}, k)).$$

This decomposition shows

$$\sum_{\substack{\mathbf{w} \in \mathcal{C}_A^{(\infty)}, \\ (A+1)^2 X < \mathrm{den}(\mathbf{w})}} \mathrm{den}(\mathbf{w})^{-s} \ll \sum_{\substack{\mathbf{v} \in \mathcal{C}_A^{(\infty)}, \\ \frac{X}{A+1} < \mathrm{den}(\mathbf{v}) \leq X}} \mathrm{den}(\mathbf{v})^{-s} \sum_{\frac{X}{\mathrm{den}(\mathbf{v})} \leq k \leq A} k^{-s} \underbrace{\sum_{\mathbf{u} \in \mathcal{C}_A^{(\infty)}} \mathrm{den}(\mathbf{u})^{-s}}_{=\zeta(s,A)}.$$

Dividing both sides by $\zeta(s, A)$ and taking $s \to D(A)$ we find that

$$1 \ll \sum_{\substack{\mathbf{v} \in \mathcal{C}_A^{(\infty)}, \\ \frac{X}{A+1} < \mathrm{den}(\mathbf{v}) \leq X}} \mathrm{den}(\mathbf{v})^{-D(A)} \sum_{\frac{X}{\mathrm{den}(\mathbf{v})} \leq k \leq A} k^{-D(A)} \ll \sum_{\substack{\mathbf{v} \in \mathcal{C}_A^{(\infty)}, \\ \frac{X}{A+1} < \mathrm{den}(\mathbf{v}) \leq X}} \mathrm{den}(\mathbf{v})^{-D(A)} \left(\frac{X}{\mathrm{den}(\mathbf{v})}\right)^{1-D(A)}.$$

This can be written as

$$X^{D(A)} \ll \int_{X/(A+1)}^{X} \frac{X}{t} dF_A(t).$$

Partial integration (in the Riemann-Stieltjes sense) gives

$$X^{D(A)} \ll F_A(X) - (A+1) F_A(X/(A+1)) + X \int_{X/(A+1)}^{X} t^{-2} F_A(t) dt.$$

Now suppose that for every $\epsilon > 0$ we have $F_A(X) < (\epsilon X)^{D(A)}$ for sufficiently large X. Then we get

$$X^{D(A)-1} \ll \int_{X/(A+1)}^{\epsilon X} t^{-2+D(A)} dt + \int_{\epsilon X}^{X} t^{-2} (\epsilon X)^{D(A)} dt \ll (\epsilon X)^{D(A)-1}.$$

This gives a contradiction for sufficiently small $\epsilon > 0$ and we obtain $F_A(X) \gg X^{D(A)}$. This completes the proof. □

2. Fuchsian Groups

The hyperbolic plane (also upper half plane) is given by $\mathbb{H} = \{z \in \mathbb{C} \colon \mathrm{Im}(z) > 0\}$. The metric is given by $ds^2 = \frac{dx^2+dy^2}{y^2}$ and the measure is $d\mu(z) = \frac{dxdy}{y^2}$. The group $\mathrm{PSL}_2(\mathbb{R})$ acts on $\mathbb{H}$ via Möbius transformations. More precisely

$$g.z = \frac{az+b}{cz+d} \text{ for } g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{R}).$$

It turns out that $\mathrm{PSL}_2(\mathbb{R})$ is the group of orientation preserving isometries of $\mathbb{H}$.

The geodesics of $\mathbb{H}$ are precisely the arcs of circles intersecting $\partial\mathbb{H}$ orthogonally. The basic example is the (degenerate circle) $i\mathbb{R}_+$. Given two points $z, w \in \overline{\mathbb{H}} =$

$\mathbb{H} \cup \partial\mathbb{H}$ there is a unique geodesic segment, denoted by $[z, w]$ for now, connecting z and w. We define the hyperbolic distance by

$$d(z, w) = l([z, w]), \tag{4}$$

where $l(\cdot)$ denotes the length of a curve with respect to the line element ds. A short computation shows that

$$\cosh(d(z, w)) = 1 + \frac{|z - w|^2}{2\,\mathrm{Im}(z)\,\mathrm{Im}(w)}.$$

The geodesic distance gives rise to so called geodesic polar coordinates, which we will use frequently.

Elements $T \in \mathrm{PSL}_2(\mathbb{R})$ can be classified by their fixed points in $\mathbb{H}$ (i.e. $z \in \mathbb{H}$ with $Tz = z$). According to this we make the following definition:

- T is called *elliptic* if $\mathrm{Tr}(T) < 2$. This implies that T has one fixed point in $\mathbb{H}$;
- T is called *parabolic* if $\mathrm{Tr}(T) = 2$ and $T \neq 1$. In this case T has a single degenerate fixed point in $\partial\mathbb{H}$;
- T is *hyperbolic* if $\mathrm{Tr}(T) > 2$. In this situation T has two distinct fixed points in $\partial\mathbb{H}$.

The (complex) Möbius transform

$$z \mapsto \frac{z - i}{z + i}$$

maps the upper half to the unit disc. This gives rise to the disc-model $\mathbb{B}$ for the hyperbolic plane.

A *Fuchsian group* is a discrete subgroup of $\mathrm{PSL}_2(\mathbb{R})$. We say a subgroup $\Gamma \subset \mathrm{PSL}_2(\mathbb{R})$ acts *properly discontinuously* if any compact subset of $\mathbb{H}$ contains only finitely many orbit points. In this case the quotient $\Gamma\backslash\mathbb{H}$ is well defined as a metric space.

Lemma 2.1. *A subgroup $\Gamma \subset \mathrm{PSL}_2(\mathbb{R})$ acts properly discontinuously on $\mathbb{H}$ if and only if it is Fuchsian.*

Proof. Exercise. □

A *fundamental domain* $\mathcal{F} \subset \mathbb{H}$ for a Fuchsian group Γ is a closed region such that $\Gamma\mathcal{F} = \mathbb{H}$ and $\mathcal{F}^\circ$ does contain at most one point of each Γ-orbit. The *limit set* $\Lambda(\Gamma) \subset \partial\mathbb{H}$ of a Fuchsian group Γ is the set of limit points of all orbits Γz for $z \in \mathbb{H}$.

Theorem 2.2 (Poincaré, Fricke-Klein)**.** *The possibilities for the limit set of a Fuchsian group Γ are:*

- $\sharp\Lambda(\Gamma) \in \{0, 1, 2\}$ *(in this case Γ is called elementary);*
- $\Lambda(\Gamma)$ *is a perfect nowhere dense subset of $\partial\mathbb{H}$ (in this case Γ is called of the second kind);*

- $\Lambda(\Gamma) = \partial\mathbb{H}$ *(in this case* Γ *is called of the first kind).*

Proof. [1, Theorem 2.9]. □

We call Γ *cofinite* (*cocompact*) if the quotient $\Gamma\backslash\mathbb{H}$ has finite volume (is compact).

Theorem 2.3. *A Fuchsian group is cofinite if and only if it is of the first kind.*

Proof. [9, Section 4.5] □

The absolute Poincaré series of a Fuchsian group is given by

$$\mathcal{P}_\Gamma(z, w; s) = \sum_{\gamma\in\Gamma} e^{-sd(z\gamma.w)},$$

which converges for $s \in \mathbb{C}$ with sufficiently large real part. The *exponent of convergence* of Γ is given by

$$\delta = \inf\{s \geq 0 \colon \mathcal{P}_\Gamma(z, w; s) < \infty\} \text{ for some } w, z.$$

Exercise 2.1. *Show that the Poincare series defined in* (1) *converges absolutely for* $\mathrm{Re}(s) > \delta$*, where* δ *is the exponent of convergence of* Γ*.*

Lemma 2.4 (Selberg's Lemma)**.** *A finitely generated group of matrices over a field of characteristic* 0 *has a torsion free subgroup of finite index.*

Proof. We call a group G *residually finite*, if for each $f \in G \setminus \{1\}$ there is a finite homomorphic image H_g so that the image of g in H_g is not the identity. (Roughly speaking: there are many finite quotients.)

We actually proof the following more general statement:

Let A be a finitely generated integral domain of characteristic 0 and consider the group $G = \mathrm{GL}_n(A)$. Then G is residually finite and G contains a normal subgroup of finite index which is torsion free.[4]

To prove this we consider the quotient field F of A. It is a finite algebraic extension, lets call the degree k, over the purely transcendental field $K = \mathbb{Q}(x_1, \ldots, x_m)$. Next we express the finite set of generators of A in terms a basis for F over K. The coefficients will feature denominators which are contained in a finitely generated ring B. There is an integer s and a polynomial f such that

$$B = \mathbb{Z}[\frac{1}{s}][x_1, \ldots, x_m, \frac{1}{f}].$$

Given an n-dimensional vector space V over F we have the natural representation $\mathrm{End}_F(V) \to \mathrm{End}_K(V)$. By considering $V = F^n$ we obtain an injective homomorphis

$$\rho \colon \mathrm{GL}_n(F) \to \mathrm{GL}_{nk}(K).$$

[4]If the characteristic of A is positive, then one can still show that G contains a normal subgroup of finite index in which every element of finite order is unipotent. The statement that G is residually finite also remains true.

The homomorphism ρ represents the group $G = \mathrm{GL}_n(A)$ as a subgroup off $\mathrm{GL}_{nk}(B)$. Therefore it suffices to prove the statement for the group $G' = \mathrm{GL}_{nk}(B)$.

Take $g \in G' \setminus \{1\}$ and let $w(x_1, \dots, x_m, \frac{1}{f})$ be a non-zero entry in the matrix $g - 1$. Fix a prime not dividing s, so that modulo p not all coefficients of w are 0. Further take v sufficiently large, so that $u = f^v w$ is a polynomial in $x_1, \dots, x_m$. Finally choose $a_1, \dots, a_m$ in the algebraic closure of $\mathbb{F}_p$ so that $u(a_1, \dots, a_m) \neq 0$. With these choices made $w(a_1, \dots, a_m, f(a_1, \dots, a_m)^{-1}) \neq 0$ and the kernel $\mathfrak{b}$ of the homomorphism

$$\pi \colon B \to \mathbb{F}_p(a_1, \dots, a_m)$$

is a maximal ideal of finite index. We conclude that the induced homomorphism $\Pi \colon \mathrm{GL}_{nk}(B) \to \mathrm{GL}_{nk}(B/\mathfrak{b})$ has finite image. By construction we have $\Pi(g) \neq 1$. This shows that G' (and thus G) is residually finite.

Let g be an element of order $1 < a < \infty$ in G. Of course g satisfies the polynomial $X^a - 1$. The minimal polynomial of g has distinct roots and the eigenvalues are roots of unity. Since the coefficients of the characteristic polynomial of g are symmetric functions in its roots these coefficients are algebraic integers in $K = \mathbb{Q}(x_1, \dots, x_m)$. We conclude that the trace of an element of finite order in G' is an integer. Of course its absolute value is $\leq nk$. Thus there are only a finite number of traces of these elements of finite order. We call the set of these traces T. Consider a prime p which does not divide s, the coefficients of f, and the non-zero integers $t - nk$ for $t \in T$. (These are finitely many conditions so that there are infinitely many such primes.) Write Ω_p for the algebraic closure of $\mathbb{F}_p$. We can now find an homomorphsim $\sigma \colon A \to \Omega_p$. (For example by extending the reduction modulo p naturally to $\mathbb{Z}[1/s]$ and by sending x_i to $a_i \in \Omega_p$ with $f(a_1, \dots, a_m) = 0$.) We find that $\sigma(A) = \mathbb{F}_p(a_1, \dots, a_m)$ is a finite field, so that the kernel $\mathfrak{a} = \ker(\sigma)$ is a maximal ideal of finite index in A. Let $\Sigma \colon \mathrm{GL}_{nk}(A) \to \mathrm{GL}_{nk}(A/\mathfrak{a})$ be the corresponding natural homomorphism. Its kernel $G(\mathfrak{a})$, called the (principal) congruence subgroup of level $\mathfrak{a}$, has finite index and is normal. Consider the subgroup $G_0 = G' \cap G(\mathfrak{a})$, which is of finite index in G' and normal. Now consider an element of finite order $g \in G_0$. Obviously $\mathrm{Tr}(g) \in T$ and $\mathrm{Tr}(g) \equiv nk \bmod \mathfrak{a}$. We conclude that $\mathrm{Tr}(g) - nk$ is an integer, which reduces to 0 modulo $\mathfrak{a}$. Therefore p divides $\mathrm{Tr}(g) - nk$ and our choice of p implies that $\mathrm{Tr}(g) = nk$. But this already implies that $g = 1$. □

A Fuchsian group is called *geometrically finite* if there exists a fundamental domain which is a finite sided convex polygon. It can be shown (see [1, Theorem 2.10]) that Γ is geometrically finite if and only if Γ is finitely generated.

We end this section by looking at the classification of hyperbolic ends, which will be of importance later on. Before stating the classification we carefully introduce the main players.

Given a hyperbolic transformation $T \in \mathrm{PSL}_2(\mathbb{R})$ we obtain the cyclic hyperbolic group $\langle T \rangle$ generated by T. Write

$$l = l(T) = \min_{z \in \mathbb{H}} d(z, Tz).$$

The quotient $C_l = \langle T \rangle \backslash \mathbb{H}$ will be called a *hyperbolic cylinder* of diameter l. After conjugation if necessary we can identify the generator T with the map $z \mapsto e^l z$. Write Γ_l for the corresponding cyclic group. It is easy to verify that

$$\mathcal{F}_l = \{1 \leq |z| \leq e^l\}$$

is a fundamental domain for Γ_l. The y-axis is the lift of the only simple closed geodesic on C_l with length l. A *funnel* F_l is half of a hyperbolic cylinder of diameter l with boundary given the central geodesic. Note that $\mathrm{Vol}(F_l, \mu) = \infty$.

Similarly we can define *parabolic cylinders*: take a parabolic element T, form the parabolic cyclic group $\langle T \rangle$ and consider the quotient $\langle T \rangle \backslash \mathbb{H}$. We can assume that T is given by the map $z \mapsto z+1$ and we write Γ_∞ for the corresponding cyclic group. A fundamental domain for Γ_∞ is

$$\mathcal{F}_\infty = \{0 \leq \mathrm{Re}(z) \leq 1\}.$$

A circle lying in $\mathbb{H}$ which is tangent to $\partial\mathbb{H}$ is called a *horocycle*. (These are exactly curves stabilized by parabolic transformations.) A *cusp* is the small end of a parabolic cylinder with boundary the unique closed horocycle of length one. The volume of a cusp (normalized as in our definition) is 1.

Now let Γ be nonelementary. Recall that in this case $\Lambda(\Gamma)$ is either a perfect nowhere dense set or equal to $\partial\mathbb{H}$. Suppose we are in the first case (i.e. Γ is of the second kind). Then $\partial\mathbb{H} \setminus \Lambda(\Gamma) = \bigcup_{j \in \mathbb{N}} I_j$ is the countable union of open intervals I_j. Suppose that γ_j is the geodesic connecting the endpoints of I_j. Let H_j be the half plane bounded by γ_j and I_j. The *Nielsen region* if a Fuchsian group Γ_j is the set $\widetilde{N} = \mathbb{H} \setminus (\bigcup_j H_j)$. The quotient $N = \Gamma \backslash \widetilde{N}$ is called the *convex core* of $\Gamma \backslash \mathbb{H}$. (Note that if Γ is of the first kind, then the Nielsen region of Γ is simply $\mathbb{H}$.)

By passing from $\Gamma \backslash \mathbb{H}$ to its convex core we remove a finite set of funnels. However, there still may be some cusps which we would like to isolate as well. Given a parabolic fixed point $p \in \partial\mathbb{H}$ let Γ_p be the parabolic cyclic subgroup of Γ fixing p. Let σ_p be the unique horocycle tangent to p such that $\Gamma_p \backslash \sigma_p$ has length one. Take O_p to be the open region bounded by σ_p, so that $\Gamma_p \backslash O_p$ is precisely a cusp. The *truncated Nielsen region* is

$$\widetilde{K} = \widetilde{N} \setminus \Big(\bigcup_{p \text{ parabolic f.p.}} O_p \Big).$$

We put $K = \Gamma \backslash \widetilde{K}$ and call it the *compact core* of $\Gamma \backslash \mathbb{H}$. (Note that it requires some work to see that the quotient is actually well defined, see [1, Lemma 2.12].)

Theorem 2.5. *Let Γ be nonelementary and geometrically finite. Then the compact core K of $\Gamma\backslash\mathbb{H}$ is compact and $(\Gamma\backslash\mathbb{H}) \setminus K$ is a finite disjoint union of cusps and funnels.*

Proof. [1, Theorem 2.13] □

3. Strong Approximation

Strong approximation can be formulated in many ways. The easiest way is probably in terms of congruences. Suppose we have a family of polynomials

$$f_\alpha(x_1, \dots, x_d) \in \mathbb{Z}(x_1, \dots, x_d),\ \alpha \in I.$$

For any $\mathbb{Z}$-algebra R we have the set of points

$$X(R) = \{(\alpha_1, \dots, \alpha_d) \in R^d \colon f(\alpha_1, \dots, \alpha_d) = 0 \text{ for all } \alpha \in I\}.$$

($X \subset \mathbb{A}^d_{\mathbb{Z}}$ is the closed affine subscheme defined by these polynomials.) For each $m \in \mathbb{N}$ we have the natural reduction map

$$\rho \colon X(\mathbb{Z}) \to X(\mathbb{Z}/m\mathbb{Z}).$$

The question is for which m these maps are surjective. Ideally one would want the map to be surjective for all m but in reality one needs to restrict to m that are coprime to some fixed $N = N(X)$.

For algebraic groups the theory of strong approximation becomes particularly useful. We state a general theorem:

Theorem 3.1 (Mathews, Vaserstein, Weisfeiler)**.** *Let G be a connected simply connected absolutely almost simple algebraic group defined over $\mathbb{Q}$. Let Γ be a finitely generated subgroup of $G(\mathbb{Q})$ that is Zariski dense in G. Then the reduction Γ_p of Γ is equal to $G(\mathbb{F}_p)$ for sufficiently large p.*

The proof is very deep and requires a lot of representation theory as well as the classification of finite simple groups. We will only discuss an simpler version, where the computations are more hands on.

For the reminder of this section let Γ be a finitely generated non-elementary subgroup of $\mathrm{SL}_2(\mathbb{Z})$. We write $\Gamma(q)$ to be the kernel of the reduction map $R_q \colon \mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ restricted to Γ. Our goal is to prove that for $p \gg_\Gamma 1$ (i.e. large enough) one has $\Gamma/\Gamma(p) \cong \mathrm{SL}_2(\mathbb{F}_p)$. According to Selberg's lemma we can assume that Γ is torsion free.

Note that the principal congruence subgroup $K(2) = \ker(R_2)$ modulo 2 is free. Since subgroups of free groups are free we conclude that $\Gamma(2) \subset K(2)$ is free. We will now apply the following theorem.

Theorem 3.2 (Stallings)**.** *A finitely generated torsion free group which contains a free subgroup of finite index is free.*

Note that $\Gamma(2) \subset \Gamma$ is obviously of finite index, so that we can conclude that Γ is free.

Take a symmetric set of generators $S = \{A_1, \dots, A_k\}$ of Γ and consider the Cayley graphs

$$\mathcal{G}_p = \mathcal{G}(\Gamma/\Gamma(p); S).$$

Let $c(\mathcal{G}_p)$ be the length of the shortest cycle also called the *girth*. Also define $d(\mathcal{G}_p)$ be the largest integer such that any two walks in $\mathcal{G}_p$ starting at the identity E with length at most $d(\mathcal{G}_p)$ end at different vertices. (This can be thought of as the injectivity radius.) Obviously we have

$$c(\mathcal{G}_p) \geq 2d(\mathcal{G}_p) - 1.$$

Let $Y_p = R_p(\Gamma)$ denote the image of Γ under the reduction map R_p. In particular we have $Y_p \cong \Gamma/\Gamma(p)$ and Y_p is a subgroup of $\mathrm{SL}_2(\mathbb{F}_p)$. The images of the generators will be denoted by $A_{i,p} = R_p(A_i)$ for $i = 1, \dots, k$ and we write $S_p = R_p(S)$. In particular we have

$$\mathcal{G}_p = \mathcal{G}(Y_p, S_p).$$

Take two walks $p = (p_0, \dots, p_r)$ and $s = (s_0, \dots, s_t)$ with starting point $E = s_0 = p_0$ and common end $s_t = p_r$. We obtain the corresponding words $V = (v_1, \dots, v_r)$ and $W = (w_1, \dots, w_t)$ over S_p and we have $p_i = v_1 \cdots v_i$ as well as $s_j = w_1 \cdots w_j$. In particular, since $s_t = p_r$ we have

$$v_1 \cdots v_r = w_1 \cdots w_t.$$

We lift the words V and W to words $\tilde{V}$ and $\tilde{W}$ over S. (This is done by taking the appropriate preimage A_i of $A_{i,p}$ or A_i^{-1} of $A_{i,p}^{-1}$.) Note that V, W are reduced and different so that $\tilde{V}$ and $\tilde{W}$ are reduced and different. We obtain

$$\tilde{v_1} \cdots \tilde{v_r} \neq \tilde{w_1} \cdots \tilde{w_t}$$

since Γ is free and generated by $A_1, \dots, A_k$. Thus we can look at the matrix

$$M = M(V, W) = \tilde{v_1} \cdots \tilde{v_r} - \tilde{w_1} \cdots \tilde{w_t} \in \mathrm{Mat}_{2\times 2}(\mathbb{Z}).$$

Obviously $M \equiv 0 \bmod p$ but $M \neq 0$. Therefore

$$\|M\| = \sup_{x \neq 0} \frac{\|Mx\|}{\|x\|} \geq p,$$

where the norm $\|\cdot\|$ of $x \in \mathbb{R}^2$ is the usual 2-norm. We obtain

$$\max(\|\tilde{v}_1 \cdots \tilde{v}_r\|, \|\tilde{w}_1 \cdots \tilde{w}_t\|) \geq \frac{p}{2}.$$

We put $\alpha = \max_k \|A_k\|$ and obtain

$$\alpha^{\max(r,t)} \geq \frac{p}{2}.$$

We directly obtain

$$d(\mathcal{G}_p) \geq \log_\alpha(\frac{p}{2}) \text{ and } c(\mathcal{G}_p) \geq 2\log_\alpha(\frac{p}{2}) - 1.$$

We recall the following (standard) result classifying subgroups of $\mathrm{SL}_2(\mathbb{F}_p)$.

Theorem 3.3. *Let $p \geq 5$ be prime. Then any subgroup of $\mathrm{SL}_2(\mathbb{F}_p)$ is isomorphic to one of the following subgroups:*

- *The dihedral groups of order $2\left(\frac{p\pm 1}{2}\right)$ and their subgroups;*
- *A group H of order $p\left(\frac{p-1}{2}\right)$ and its subgroups. If $H_1 \equiv N_1$ is a subgroup of H, then its factor group H/H_1 is cyclic;*
- *A_4, S_4 or A_5.*

Suppose $Y_p \neq \mathrm{SL}_2(\mathbb{F}_p)$ for p large enough. Then Y_p is one of the groups listed in the Theorem above. The idea is that certain groups are excluded immediately, because they violate the girth bound derived above. They key input is that the remaining subgroups have trivial second commutator. More precisely

$$(x_1y_1x_1^{-1}y_1^{-1})(x_2y_2x_2^{-1}y_2^{-1})(y_1x_1y_1^{-1}x_1^{-1})(y_2x_2y_2^{-1}x_2^{-1}) = 1 \tag{5}$$

for all $x_1, x_2, y_1, y_2 \in Y_p$. Therefore, taking x_1, x_2, y_1, y_2 to be generators we find a closed cycle of length 16. We find $2\log_\alpha(\frac{p}{2}) \leq 17$, which is a contradiction for $p > 2\alpha^{\frac{17}{2}}$. We have established the following theorem.

Theorem 3.4. *Let Γ be a non-elementary finitely generated subgroup of $\mathrm{SL}_2(\mathbb{Z})$, then $\Gamma/\Gamma(p) \cong \mathrm{SL}_2(\mathbb{F}_p)$ for all $p \gg_\Gamma 1$.*

4. Spectral Theory

A key tool for us is the spectral theory of orbifolds $\Gamma\backslash\mathbb{H}$. The operator in question is the Laplace-Beltrami operator

$$\Delta = -y^2(\partial_x^2 + \partial_y^2).$$

Note that $\Delta \circ g = g \circ \Delta$ for all $g \in \mathrm{SL}_2(\mathbb{R})$. We quickly check

$$\langle \Delta F, G\rangle_{\mathbb{H}} = \int_{\mathbb{H}} \nabla F \overline{\nabla G} d\mu = \langle F, \Delta G\rangle_{\mathbb{H}}.$$

Thus, given a Fuchsian group Γ we can view Δ as an unbounded non-negative self-adjoint operator on $L^2(\Gamma\backslash\mathbb{H}, \mu)$. (More precisely we can start by defining Δ on $\mathcal{D} = \{f \in \mathcal{C}_0^\infty(\Gamma\backslash\mathbb{H})\colon f, \Delta f \in L^2(\Gamma\backslash\mathbb{H}, \mu)\}$ and then consider the Friedrich's extension.)

The spectrum of this operator depends heavily on Γ. Particularly strong are the differences between the situations cocompact, cofinite but not cocompact and not cofinite. We are most interested in the final case since this is the situation arising from thin groups.

We say λ is an *eigenvalue* of Δ on $\Gamma\backslash\mathbb{H}$, with *eigenfunction* ϕ, if $\phi \in L^2(\Gamma\backslash\mathbb{H}, \mu)$ and

$$\Delta\phi = \lambda\phi.$$

It will be convenient to identify $\Gamma\backslash\mathbb{H}$ with a suitable fundamental domain $\mathcal{F}$ for Γ. We write $\Omega(\mathcal{F})$ for the spectrum of the Laplace-Beltrami operator Δ on

$L^2(\mathcal{F},\mu) = L^2(\Gamma,\backslash\mathbb{H},\mu)$. Let $\lambda_0(\mathcal{F})$ denote the bottom of the spectrum and write $\lambda_1(\mathcal{F})$ for the next eigenvalue.

4.1. **The Full Hyperbolic Plane.** We start by looking at the full hyperbolic plane (i.e. $\Gamma = \{\pm 1\}$). In this case we can guess certain (pseudo)-eigenfunctions. (Note that we have not yet specified the domain of Δ!) Given $s \in \mathbb{C}$ we check

$$\Delta y^s = s(1-s)y^s.$$

The *resolvent* (of Δ for $\mathbb{H}$) is defined by

$$R_{\mathbb{H}}(s) = (\Delta - s(1-s))^{-1} \text{ for } \operatorname{Re}(s) > \frac{1}{2} \text{ and } s \notin [\frac{1}{2}, 1].$$

Note that obviously we have

$$(\Delta - s(1-s))R_{\mathbb{H}}(s)f(z) = f(z) = \int_{\mathbb{H}} y^2\delta(z-z')f(z')d\mu(z').$$

Put $R_{\mathbb{H}}(s)f(z) = \int_{\mathbb{H}} R_{\mathbb{H}}(s;z,z')f(z')d\mu(z')$. Then we obtain

$$(\Delta - s(1-s))R_{\mathbb{H}}(s;z,z') = y^2\delta(z-z'), \tag{6}$$

where Δ acts on the z-coordinate.[5] Since Δ is $\mathrm{SL}_2(\mathbb{R})$-invariant, the resolvent kernel depends only on the hyperbolic distance between z and z'. Thus there is a function f_s so that

$$R_{\mathbb{H}}(s;z,z') = f_s(d(z,z')).$$

Putting $r = d(z,z')$ we can swap to polar coordinates (r,θ). The Laplacian then reads

$$\Delta = -\frac{1}{\sinh(r)}\partial_r(\sinh(r)\partial_r) - \frac{1}{\sinh(r)^2}\partial_\theta^2.$$

Before solving (6) we look at the corresponding homogeneous equation in polar coordinates:

$$\left[-\frac{1}{\sinh(r)}\partial_r(\sinh(r)\partial_r) - s(1-s)\right] f_s(r) = 0.$$

Changing coordinates by setting $g_s(\sigma) = f_s(r)$ for $\sigma = \cosh(r/2)^2$ yields

$$\sigma(1-\sigma)g_s'' + (1-2\sigma)g_s' - s(1-s)g_s = 0.$$

This is a special case of the classical hypergeometric equation.[6] We get the solution

$$g_s(\sigma) = c_s\sigma^{-s}{}_2F_1(s,s;2s;\sigma^{-1}) = c_s\frac{\Gamma(2s)}{\Gamma(s)^2}\int_0^1 \frac{(t(1-t))^{s-1}}{(\sigma-t)^s}dt.$$

[5] We are essentially dealing with the classical Green's function.

[6] The equation reads

$$z(1-z)h''(z) + (c-(a+b+1)z)h'(z) - abh(z) = 0.$$

The typical solution is the Gauß hypergeometric function $h(z) = {}_2F_1(a,b;c;z)$ which is regular at $z=0$. We took a different solution since we require regularity at ∞.

Where we used Euler's integral representation for the hypergeometric function (valid for $\mathrm{Re}(s) > 0$). We want to choose c_s so that $R_{\mathbb{H}}(s; z, z') = g_s(\sigma(z, z'))$ and confirm our choice of g_s. To do so we integrate (6) in polar coordinates over a small disc of radius ϵ (i.e. $B_0(\epsilon)$). We get

$$\begin{aligned} 1 &= -2\pi \int_0^\epsilon \left[(\sinh(r) f_s'(r))' + s(1-s)\sinh(r) f_s(r)\right] dr \\ &= -2\pi \sinh(\epsilon) f_s'(\epsilon) - 2\pi s(1-s) \int_0^\epsilon \sinh(r) f_s(r) dr. \end{aligned}$$

Supposing that f_s is locally L^2 we see that the integral vanishes as $\epsilon \to 0$. Thus we obtain

$$f_s'(r) = -\frac{1}{2\pi r} + O(1) \text{ as } r \to 0.$$

This implies that the appropriate boundary condition is $f_s(r) \sim -\frac{1}{2\pi}\log(r)$. Recall that $\sigma \sim 1 + r^2/4$ so that this translates into

$$g_s(\sigma) \sim -\frac{1}{4\pi}\log(\sigma - 1) \text{ as } \sigma \to 1.$$

A quick analysis of Euler's integral reveals that g_s has the correct asymptotic for $c_s = \frac{1}{4\pi}\frac{\Gamma(s)^2}{\Gamma(2s)} = 2^{-1-2s}\frac{\Gamma(s)}{\sqrt{\pi}\Gamma(s+\frac{1}{2})}$. We have obtained the following theorem:

Proposition 4.1. *The resolvent kernel is given by*

$$R_{\mathbb{H}}(s; z, z') = g_s(\cosh(d(z, z')/2)^2),$$

for

$$g_s(\sigma) = 2^{-1-2s}\frac{\Gamma(s)}{\sqrt{\pi}\Gamma(s+\frac{1}{2})}\sigma^{-s}{}_2F_1(s, s; 2s; \sigma^{-1}) = \frac{1}{4\pi}\int_0^1 \frac{t^{s-1}(1-t)^{s-1}}{(\sigma - t)^2} dt.$$

Note that the integral representation is only valid for $\mathrm{Re}(s) > 0$.

We define

$$E_{\mathbb{H}}(s; z, x') = \lim_{y' \to 0} (y')^{-s} R_{\mathbb{H}}(s; z, z') \text{ for } z' = x' + iy'$$

and call the resulting function *generalized eigenfunctions* of Δ. (The adjective generalized is added to indicate that they are not in L^2.) One can compute

$$E_{\mathbb{H}}(s; z, x') = \frac{1}{2s-1}\frac{\Gamma(s)}{\sqrt{\pi}\Gamma(s-\frac{1}{2})}\left[\frac{y}{(x-x')^2 + y^2}\right]^s$$

using our explicit formula for the resolvent kernel. One quickly verifies

$$(\Delta - s(1-s))E_{\mathbb{H}}(s; \cdot, x') = 0$$

justifying the term eigenfunction.

Proposition 4.2. *Meromorphically for $s \in \mathbb{C}$ we have*

$$R_{\mathbb{H}}(s;z,w) - R_{\mathbb{H}}(1-s;z,w) = -(2s-1)\int_{-\infty}^{\infty} E_{\mathbb{H}}(s;z,x')E_{\mathbb{H}}(1-s;w,x')dx'.$$

Proof. We consider the region $\Sigma_T = [-T,T]\times[T^{-1},T]$. One computes

$$\begin{aligned} &R_{\mathbb{H}}(s;z,w) - R_{\mathbb{H}}(1-s;z,w) \\ &= \lim_{T\to\infty}\int_{\Sigma_T} \left[R_{\mathbb{H}}(s;z,z')\Delta_{z'}R_{\mathbb{H}}(1-s;z',w) - R_{\mathbb{H}}(1-s;z',w)\Delta_{z'}R_{\mathbb{H}}(s;z,z')\right] d\mu(z'). \end{aligned}$$

By Greens formula we obtain

$$\begin{aligned} &R_{\mathbb{H}}(s;z,w) - R_{\mathbb{H}}(1-s;z,w) \\ &= \lim_{T\to\infty}\int_{\partial\Sigma_T} \left[\partial_{v'}R_{\mathbb{H}}(s;z,z')R_{\mathbb{H}}(1-s;z',w) - R_{\mathbb{H}}(1-s;z',w)\partial_{v'}R_{\mathbb{H}}(s;z,z')\right] ds(z'). \end{aligned}$$

Note that since $z\in\mathbb{H}$ is fixed we have $\sigma(z,z')\to\infty$ as $T\to\infty$ and $z'\in\partial\Sigma_T$. The explicit formula for the resolvent kernel shows that

$$R_{\mathbb{H}}(s;z,z') \sim c_s\sigma(z,z')^s \text{ and } R_{\mathbb{H}}(1-s;w,z') \sim c_{1-s}\sigma(w,z')^{s-1}.$$

The normal derivatives have the same asymptotics. With this at hand it becomes easy bound the contributions of everything but the bottom edge. (One can show $O(T^{-1})$ for this contribution.) To treat the bottom edge we put $y' = T^{-1}\to 0$. Recalling the definition of the generalized eigenfunction shows that

$$R_{\mathbb{H}}(s;z,z') = (y')^s E_{\mathbb{H}}(s;z,x') + O((y')^{s+1}).$$

The normal derivative is simply $\delta_{v'} = -y'\partial_{y'}$. We compute

$$y'\partial_{y'}R_{\mathbb{H}}(s;z,z') = s(y')^s E_{\mathbb{H}}(s;z,x') + O((y')^{s+1}).$$

The length element simplifies to $ds(z') = (y')^{-1}dx'$. So taking $T\to\infty$ yields

$$\lim_{T\to\infty}\int_{\partial\Sigma_T}\partial_{v'}R_{\mathbb{H}}(s;z,z')R_{\mathbb{H}}(1-s;z',w)ds(z') = -s\int_{-\infty}^{\infty} E_{\mathbb{H}}(s;z,x')E_{\mathbb{H}}(1-s;w,x')dx'.$$

The second term is computed similarly. □

Theorem 4.3. *The spectrum of Δ on $L^2(\mathbb{H},\mu)$ is absolutely continuous and equal to $[\frac{1}{4},\infty)$.*

Proof. Let P_I denote the spectral projector of Δ onto $I\subset\mathbb{R}$. This spectral projector can be computed using Stone's formula (which is a direct consequence of the resolvent functional calculus). One needs to be slightly careful with the parametrization $s(1-s) = \lambda \pm i\epsilon$ in the definition of the resolvent. We have

$$\frac{1}{2}(P_{[a,b]} + P_{(a,b)}) = \lim_{\epsilon\to 0}\frac{1}{2\pi i}\int_a^b [R_{\mathbb{H}}(s_+(z,\epsilon)) - R_{\mathbb{H}}(s_-(z,\epsilon))]dz.$$

Here $s = s_\pm(z,\epsilon)$ with $\mathrm{Re}(s) > \frac{1}{2}$ is given by

$$s(1-s) = z \pm i\epsilon.$$

This is $s_\pm(z,\epsilon) = \frac{1}{2} \pm \sqrt{\frac{1}{4} - (z \pm i\epsilon)}$. For $0 \leq z \leq \frac{1}{4}$ the ϵ-limits coincide, so that the two terms cancel each other. We conclude that $P_{[0,\frac{1}{4})} = 0$. Therefore there can be no spectrum below $\frac{1}{4}$.

On the other hand suppose $z \geq \frac{1}{4}$ and write $z = \frac{1}{4} + \xi$. Then we can rewrite Stone's formula as

$$\frac{1}{2}(P_{[a,b]} + P_{(a,b)}) = \frac{1}{2\pi i} \int_{\sqrt{a-\frac{1}{4}}}^{\sqrt{b-\frac{1}{4}}} [R_{\mathbb{H}}(\frac{1}{2} - i\xi) - R_{\mathbb{H}}(\frac{1}{2} + i\xi)] 2\xi d\xi,$$

for $b > a \geq \frac{1}{4}$. The kernel of the spectral measure is given in the proposition above and one reads of that the spectrum is absolutely continuous. □

A continuous compactly supported function $k(\cdot,\cdot)\colon \mathbb{H} \times \mathbb{H} \to \mathbb{C}$ is called a *point pair invariant*[7] if

$$k(gz, gw) = k(z,w) \text{ for all } z, w \in \mathbb{H} \text{ and all } g \in \mathrm{SL}_2(\mathbb{R}).$$

Such a function depends only on the hyperbolic distance and we abuse notation to write

$$k(z,w) = k(u(z,w)).$$

Thus we can view k as a function from $\mathbb{R}_{\geq 0} \to \mathbb{C}$.

Theorem 4.4. *Let k be a point pair invariant and suppose that $\phi\colon \mathbb{H} \to \mathbb{C}$ is a function with $\Delta\phi = \lambda\phi$. Write $\lambda = s(1-s)$ with $s = \frac{1}{2} + it$. Then*

$$\int_{\mathbb{H}} k(z,w)\phi(w)dw = h(t)\phi(z),$$

where h is the Selberg/Harish-Chandra transform of k given by

$$\begin{aligned} q(v) &= \int_v^\infty k(u)(u-v)^{-\frac{1}{2}} du, \\ g(r) &= 2q(\sinh(r/2)^2), \\ h(t) &= \int_{-\infty}^\infty e^{irt} g(r) dr. \end{aligned}$$

Proof. This is proved in several steps. First it is easy to check using polar coordinates that

$$\Delta_z k(z,w) = \Delta_w k(z,w).$$

[7]Note that we include strong regularity conditions in the definition of a point pair invariant. This is not standard but makes our live easier. Note that these conditions can be slightly relaxed.

From this one deduces that the invariant integral operators commute with Δ. Indeed put $T_k f(z) = \int_{\mathbb{H}} k(z,w) f(w) d\mu(w)$, then

$$\Delta T_k f(z) = \int_{\mathbb{H}} \Delta_z k(z,w) f(w) d\mu(w) = \int_{\mathbb{H}} \Delta_w k(z,w) f(w) d\mu(w)$$
$$= \int_{\mathbb{H}} k(z,w) [\Delta f](w) d\mu(w) = T_k \Delta f(z).$$

Given $f\colon \mathbb{H} \to \mathbb{C}$ define the mean value operator (at $w \in \mathbb{H}$) by

$$f_w(z) = \int_{G_w} f(gz) dg \text{ for } G_w = \{g\colon gw = w\}.$$

The averaged function f_w is radial at w, meaning that $f_w(z)$ only depends on the distance between z and w. Obviously $f_z(z) = f(z)$. One also easily checks that $(T_k f)(z) = (T_K f_z)(z)$.

The upshot is that $\omega(z,w) = {}_2F_1(s, 1-s; 1; u(z,w))$ is the unique function in z, which is radial at w and satisfies $\omega(w,w) = 1$ as well as $[\Delta_z - s(1-s)]\omega(z,w) = 0$. (To see this one looks at the corresponding differential equation obtained by considering the eigenvalue equation in geodesic polar coordinates. This is similar to earlier arguments.)

Uniqueness of $\omega(z,w)$ implies directly that a function ϕ with $\Delta\phi = s(1-s)\phi$ (i.e. $\lambda = s(1-s)$) satisfies

$$\phi_w(z) = \omega(z,w)\phi(w).$$

From this we obtain that ϕ is an eigenfunction of all invariant integral operators T_k. Furthermore, the eigenvalue only depends on k and λ. More precisely, there is $\Lambda = \Lambda(\lambda, k)$ with

$$T_k \phi(z) = \int_{\mathbb{H}} k(z,w)\phi(w) d\mu(w) = \Lambda\phi(z).$$

To compute Λ we can now choose $\phi(z) = \mathrm{Im}(z)^s$ and specialize to $z = i$. We get

$$\Lambda = \int_{\mathbb{H}} k(i,w) \,\mathrm{Im}(w)^s d\mu(w)$$
$$= 2\int_0^\infty \int_0^\infty k(\frac{x^2 + (y-1)^2}{4y}) y^{s-2} dy dx.$$

One concludes by changing variables to $x = 2\sqrt{uy}$ and $y = e^r$. □

4.2. **Compact Quotients.** In the compact case any function $\phi\colon \Gamma\backslash\mathbb{H} \to \mathbb{C}$ satisfying $\Delta\phi = \lambda\phi$ is automatically in $L^2(\Gamma\backslash\mathbb{H}, \mu)$. Indeed, by elliptic regularity such a function is smooth and thus square integrable when restricted to compact subsets (such as fundamental domains for Γ) of $\mathbb{H}$. It turns out that the full spectrum is exhausted by eigenvalues.

Theorem 4.5. *If Γ is co-compact, then there is a complete orthonormal basis $(\phi_i)_{i\in\mathbb{N}_0}$ for $L^2(\Gamma\backslash\mathbb{H},\mu)$ such that $\Delta\phi_i = \lambda_i\phi_i$ and*

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

Proof. One can check that the domain of Δ (constructed as above using the Friedrichs extension) is the Sobolev space

$$H^2(\Gamma\backslash\mathbb{H},\mu) = \{u \in L^2(\Gamma\backslash\mathbb{H},\mu)\colon \Delta u \in L^2(\Gamma\backslash\mathbb{H},\mu)\}.$$

The inclusion $H^2(\Gamma\backslash\mathbb{H},\mu) \to L^2(\Gamma\backslash\mathbb{H},\mu)$ is compact, so that $(\Delta+1)^{-1}$ is a compact self-adjoint operator on $L^2(\Gamma\backslash\mathbb{H},\mu)$. The theorem follows from the spectral theorem for compact operators.[8] □

Given a point-pair invariant k we define the automorphic kernel

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

Note that this converges due to the regularity of k. Furthermore, the result is Γ-invariant in both variables. We observe that

$$\int_{\Gamma\backslash\mathbb{H}} K_\Gamma(z,w)f(w)d\mu(w) = \int_{\mathbb{H}} k(z,w)f(w)d\mu(w)$$

for all $f \in L^2(\Gamma\backslash\mathbb{H},\mu)$. Spectrally expanding K_Γ yields the following pre-trace formula:

$$K_\Gamma(z,w) = \sum_{i\geq 0} h(t_i)\phi_i(z)\overline{\phi_i(w)}.$$

It can be shown that the right hand side converges absolutely and uniformly on compacta. (Recall that we write $\lambda_i = \frac{1}{4} + t_j^2$.)

4.3. **Non-Compact Finite Volume Quotients.** The basic spectral theorem in this case reads as follows

Theorem 4.6 (Lax-Phillips)**.** *Let Γ be cofinite but non-cocompact. The spectrum of Δ on $L^2(\Gamma\backslash\mathbb{H},\mu)$ has absolutely condinuous spectrum $[\frac{1}{4},\infty)$. The discrete spectrum consists of finitely many eigenvalues in $[0,\frac{1}{4})$. Furthermore, there are examples with infinitely many embedded eigenvalues in $[\frac{1}{4},\infty)$.*

While the discrete spectrum will remain mysterious one can give a more precise description of the absolutely continuous part.

Let Γ be a Fuchsian group of the first kind. (In particular Γ has finite co-volume and finitely many generators.) We assume that Γ is not co-compact. Recall that

[8] Let A be a compact self-adjoint operator on a Hilbert space $\mathcal{H}$, then there exists an orthonormal basis $\{\phi_j\}$ for $\mathcal{H}$ such that $A\phi_j = \lambda_j\phi_j$. The eigenvalues λ_j are real and accumulate only at 0.

the cusps of Γ are fixed points of parabolic elements in Γ. Due to the classification of ends these are the only hyperbolic ends that can appear. We denote (equivalences of) cusps by gothic letters $\mathfrak{a}, \mathfrak{b}, \ldots$. Note that

$$\Gamma_{\mathfrak{a}} = \{\gamma \in \Gamma \colon \gamma\mathfrak{a} = \mathfrak{a}\} = \langle \gamma_{\mathfrak{a}} \rangle.$$

There is a matrix $\sigma_{\mathfrak{a}} \in \mathrm{PSL}_2(\mathbb{R})$, called a scaling matrix, such that

$$\sigma_{\mathfrak{a}}\infty = \mathfrak{a} \text{ and } \sigma_{\mathfrak{a}}^{-1}\gamma_{\mathfrak{a}}\sigma_{\mathfrak{a}} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

We define the *Eisenstein series*

$$E_{\mathfrak{a}}(z,s) = \sum_{\gamma \in \Gamma_{\mathfrak{a}}\backslash\Gamma} \mathrm{Im}(\sigma_{\mathfrak{a}}^{-1}\gamma z)^s$$

for $z \in \mathbb{H}$ and $\mathrm{Re}(s) > 1$. It is a deep result that these Eisenstein series have an analytic continuation and satisfy a functional equation. Given $\psi \in \mathcal{C}_c^\infty(\mathbb{R}_+)$ we define the *incomplete Eisenstein series* by

$$E_{\mathfrak{a}}(z|\psi) = \frac{1}{2\pi i}\int_{(c)} E_{\mathfrak{a}}(z,s)[\mathfrak{M}\psi](s)ds$$

where

$$[\mathfrak{M}\psi](s) = \int_0^\infty \psi(y)y^{-s-1}dy$$

is the Mellin transform.

We define the space of *cusp forms* to be the set of all smooth and bounded functions $\phi \colon \Gamma\backslash\mathbb{H} \to \mathbb{C}$ with

$$\langle \phi, E_{\mathfrak{a}}(\circ|\psi)\rangle = 0 \text{ for all } \psi \in \mathcal{C}_c^\infty(\mathbb{R}_+) \text{ and all cusps } \mathfrak{a}.$$

Write $L_{\mathrm{cusp}}(\Gamma\backslash\mathbb{H},\mu)$ for the closure of this space in $L^2(\Gamma\backslash\mathbb{H},\mu)$. Roughly speaking this is the complement of the space of incomplete Eisenstein series. It is very educational computation to show that the orthogonality condition precisely translates into a vanishing condition at all cusps of Γ.

Proposition 4.7. *The Laplace-Beltrami operator Δ has pure point spectrum in $L^2_{\mathrm{cusp}}(\Gamma\backslash\mathbb{H},\mu)$ and the eigenspaces have finite dimensions. There is a complete orthogonal system $\{\phi_j\}_{j\in\mathbb{N}}$ of cusp forms with $\Delta\phi_j = \lambda_j\phi_j$ so that*

$$f(z) = \sum_j \langle f, \phi_j\rangle\phi_j(z).$$

The spectral expansion converges in the norm-topology but can be upgraded to absolute and uniform convergence on compacta with f is assumed to be more regular.

The cuspidal part of the spectrum is very mysterious and there is a common believe that (infinitely many) cusp forms only exist if there is a reason for this. The remaining part of the spectrum turns out to be much more explicit (in a sense).

Proposition 4.8. *We have the Δ-invariant decomposition*

$$L^2_{\mathrm{cusp}}(\Gamma\backslash\mathbb{H},\mu)^\top = L^2_{\mathrm{res}}(\Gamma\backslash\mathbb{H},\mu) \oplus \bigoplus_{\mathfrak{a}} L^2_{\mathfrak{a}}(\Gamma\backslash\mathbb{H},\mu).$$

The spectrum of Δ on $L^2_{\mathrm{res}}(\Gamma\backslash\mathbb{H},\mu)$ is discrete and it consists of finitely many points $0 \le \lambda_j < \frac{1}{4}$. The spectrum of Δ on $L^2_{\mathfrak{a}}(\Gamma\backslash\mathbb{H},\mu)$ is absolutely continuous and covers the segment $[\frac{1}{4},\infty)$ uniformly with multiplicity 1.

For $f \in L^2(\Gamma\backslash\mathbb{H},\mu)$ we have

$$f(z) = \sum_i \langle f,\phi_j\rangle\phi_j(z) + \sum_{\mathfrak{a}} \frac{1}{4\pi}\int_{-\infty}^{\infty} \langle f, E_{\mathfrak{a}}(\circ,\frac{1}{2}+ir)\rangle E_{\mathfrak{a}}(z,\frac{1}{2}+ir)dr,$$

which converges in the norm topology. (An upgrade of convergence is again possible when f is more regular.) Note that the j-sum combines an eigenbasis of cusp forms (possibly infinite) with a finite eigenbasis for the residual part.

This spectral expansion can be applied to the automorphic kernel K_Γ associated to a point pair invariant k as in the co-compact case. Indeed one obtains the pretrace formula

$$K_\Gamma(z,w) = \sum_j h(t_j)\phi_j(z)\overline{\phi_j(w)} + \sum_{\mathfrak{a}} \frac{1}{4\pi}\int_{-\infty}^{\infty} h(r)E_{\mathfrak{a}}(z,\frac{1}{2}+ir)\overline{E_{\mathfrak{a}}(w,\frac{1}{2}+ir)}dr,$$

which converges absolutely and uniformly on compacta. (Recall that the ϕ_j satisfy $\Delta\phi_j = \lambda_j\phi_j$ for $\lambda_j = \frac{1}{4}+t_j^2$. Note that not all of the ϕ_j are cusp forms and that we might encounter the situation $0 \le \lambda_j < \frac{1}{4}$.)

We have omitted most proofs in this section, because they are partly similar to what we will do in the next subsection.

4.4. **Infinite Volume Quotients.** Let Γ be a geometrically finite Fuchsian group of the second kind. Our goal is to sketch a proof the following theorem

Theorem 4.9 (Patterson, Lax-Phillips)**.** *Assume that the exponent of convergence δ of Γ satisfies $\delta > \frac{1}{2}$. Then the bottom of the spectrum $\lambda(\mathcal{F}) = \delta(1-\delta)$ is an isolated eigenvalue of multiplicity one. Furthermore there are finitely many discrete eigenvalues in the interval $[0,\frac{1}{4})$ and the spectrum is continuous in $[\frac{1}{4},\infty)$.*

Furthermore we will put some effort in deriving a technical tool resembling the pretrace equality that we stated for Γ with finite co-volume. We start by looking at the model resolvents of funnels and cusps.

4.4.1. *The Model Resolvent of a Hyperbolic Cylinder.* The basic model for an hyperbolic cylinder is $C_l = \Gamma_l\backslash\mathbb{H}$ for $\Gamma_l = \langle z \mapsto e^l z\rangle$. We have the fundamental domain

$$\mathcal{F}_l = \{z \in \mathbb{H}\colon 1 \le |z| \le e^l\}.$$

There is a very convenient set of coordinates, called Fermi coordinates, given as follows:

$$z = e^t \frac{e^r + i}{e^r - i}$$

for $t \in \mathbb{R}/l\mathbb{Z}$ and $r \in \mathbb{R}$. One checks that in the coordinates (r, t) we have

$$ds^2 = dr^2 + \cosh(r)^2 dt^2.$$

Recall that earlier we computed the resolvent kernel $R_{\mathbb{H}}(s; z, z')$ for the full hyperbolic plane. The analysis also shows that

$$R_{\mathbb{H}}(s, z, e^{kl} z') = O(e^{-s|k|l}).$$

Hence we can define

$$R_{C_l}(s; z, z') = \sum_{k \in \mathbb{Z}} R_{\mathbb{H}}(s; z, e^{kl} z').$$

This defines an analytic function of s as long as $\mathrm{Re}(s) > 0$. Note thatthis already establishes the analytic continuation of $R_{C_l}(s; z, z')$ across the critical line $\mathrm{Re}(s) = \frac{1}{2}$ supporting the (expected) continuous spectrum. With a more detailed analysis one can show the following result

Proposition 4.10. *The resolvent $R_{C_l}(s)$ has a meromorphic continuation to $s \in \mathbb{C}$ with poles at $s \in \mathbb{Z}_{<0} + \frac{2\pi i}{l}\mathbb{Z}$.*

One can go further and compute an explicit Fourier decomposition for the resolvent kernel. We will omit the details.

For the spectral theory of infinite volume hyperbolic surfaces the resolvent for funnel ends F_l will be of great importance. A Funnel end F_l is exactly half of a hyperbolic cylinder. We can model it on $\{z \in \mathcal{F}_l \colon \mathrm{Re}(z) > 0\}$. This corresponds to $r > 0$ in the Fermi coordinates. This implies that the funnel resolvent is given by

$$R_{F_l}(s; z, w) = R_{C_l}(s; z, w) - R_{C_l}(s; z, -\overline{w}).$$

This yields a meromorphic continuation of the the funnel resolvent. Analysing the pole structure shows that the poles of R_{F_l} are precisely those of R_{C_l} with odd real part.

4.4.2. *The Model Resolvent of a Parabolic Cylinder.* A parabolic cylinder has the standard model $C_\infty = \Gamma_\infty \backslash \mathbb{H}$ where $\Gamma_\infty = \langle z \mapsto z + 1 \rangle$. The resolvent kernel can be written as

$$R_{C_\infty}(s; z, z') = \sum_{k \in \mathbb{Z}} R_{\mathbb{H}}(s; z, z' - k).$$

Recall that we can write $R_{\mathbb{H}}(s; z, z') = g_s(\sigma(z, z'))$. Furthermore, analysing the description of g_s as hypergeometric function yields

$$g_s(\sigma) = \sum_{n=0}^{N-1} \frac{1}{4\pi} \frac{\Gamma(s+n)^2}{\Gamma(2s+n)} \sigma - s - n + O(\sigma^{-s-N}).$$

Define

$$J(a,b;s)=\sum_{k\in\mathbb{Z}}[(a+k)^2+b^2]^{-s}.$$

It can be shown that $J(a,b;s)$ has a meromorphic continuation to $s\in\mathbb{C}$. These observations give us

$$R_{C_\infty}(s;z,z')=\sum_{n=0}^{N-1}\frac{1}{4\pi}\frac{\Gamma(s+n)^2}{\Gamma(2s+n)}J(a,b;s+n)(4yy')^{s+n}+O\left(\sum_{k\in\mathbb{Z}}\sigma(z,z'-k)^{-s-N}\right).$$

From this it is easy to deduce

Proposition 4.11. *The resolvent $R_{C_{infty}}(s)$ for the parabolic cylinder has a meromorphic continuation to $s\in\mathbb{C}$. The only pole of is at $s=\frac{1}{2}$.*[9]

4.4.3. *The Spectral Theorem.* Let $X=\Gamma\backslash\mathbb{H}$ be a geometrically finite Fuchsian group of the second kind. In particular X has infinite volume, is non-elementary and features at least one funnel end.

Key to the spectral theory is the analytic continuation of the resolvent R_X originally define by

$$R_X(s)=(\Delta_X-s(1-s))^{-1}\text{ for }\operatorname{Re}(s)>\frac{1}{2}\text{ and }s\notin[\frac{1}{2},1].$$

Recall that by using the truncated Nielsen region we obtained a decomposition

$$X=K\cup F\cup C,$$

where K is the compact core, F is the disjoint union of funnels $F_1,\dots,F_{n_f}$ and C consists of the cusps $C_1,\dots,C_{n_c}$.

A key ingredient for the analytic continuation of the rsolvent is an appropriate compactification of X. Put $\Omega(\Gamma)=\partial\mathbb{H}\setminus\lambda(\Gamma)$. Any (Dirichlet) fundamental domain $\mathcal{F}$ for Γ will meat $\Omega(\Gamma)$ in a finite collection of discjoint arcs each corresponding to a funnel end. let $\mathcal{P}(\Gamma)$ denote the parabolic fixed points. We have

$$\overline{X}=\Gamma\backslash(\mathbb{H}\cup\Omega(\Gamma)\cup\mathcal{P}(\Gamma)).$$

We derive the smooth structure from the compactification of a Dirichlet fundamental domain in the Riemann-sphere topology. Functions $f\in\mathcal{C}^\infty(\overline{X})$ are simply functions $f\in\mathcal{C}^\infty(X)$ that behave *nicely at infinity*. We define the boundary defining function

$$\rho(r)=\begin{cases}2e^{-r} & \text{in } F,\\ e^{-r} & \text{in } C.\end{cases}\tag{7}$$

We extend ρ to a smooth non-vanishing function inside K.

[9]To show this final statement concerning the pole one has to work a bit harder. This can for example be deduced by computing an appropriate Fourier decomposition.

Theorem 4.12. *Let X be a geometrically finite hyperbolic surface. For $N > 0$ the resolvent $R_X(s)$ extends for $\mathrm{Re}(s) > \frac{1}{2} - N$ fo a finitely meromorphic family of operators*[10]

$$R_X(s)\colon \rho^N L^2(X) \to \rho^{-N} L^2(X).$$

Proof. We briefly sketch the proof. At each funnel $F_j \subset F$ we let $R_{F_j}(s)$ denote the resolvent for $\Delta|_{F_j}$ with Dirichlet boundary condition at the boundary geodesic. Further, we define $R_{C_j}(s)$ as the pull back of the model resolvent $R_{C_\infty}(s)$. We group these resolvents together and write

$$R_F(s) = \oplus_j R_{F_j}(s) \text{ and } R_C(s) = \oplus_i R_{C_i}(s).$$

We view these as operators for the full surface acting by 0 outside the funnels or cusps.

Define cutoffs χ_a which is a smoothed version of $\mathbb{1}_{r\leq a}$, where r is the geodesic distance form the compact core, with support in $r \leq a + 1$. We define

$$\begin{aligned}
&M_i = \chi_2 R_X(s_0)\chi_1 \text{ for some fixed } \mathrm{Re}(s_0) > 1,\\
&M_f(s) = (1-\chi_0)R_F(s)(1-\chi_1),\\
&M_c(s) = (1-\chi_0)R_C(s)(1-\chi_1) \text{ and}\\
&M(s) = M_i + M_f(s) + M_c(s).
\end{aligned}$$

One computes

$$(\Delta_X - s(1-s))M(s) = I - L_i(s) - L_f(s) - L_c(s)$$

for some error terms L_i, L_f and L_c. The goal is to show that these errors are finitely meromorphic and compact on the weighted Hilbert space. Once this is established, the result follows from standard resolvent estimates (establishing the combined error is invertible at some s) and the analytic Fredholm theorm.

That L_f and L_c are nice can be deduced by studying the corresponding model resolvent. Let us only sketch the idea for L_i. One checks that

$$L_i(s) = -[\Delta, \chi_2]R_X(s_0)\chi_1 + (s(1-s) - s_0(1-s_0))M_i.$$

This is polynomial in s, so that we only need to show compactness. To this end we observe that $[\Delta, \chi_2]R_X(s_0)\chi_1$ is a smoothing operator. This is due to the disjoint supports of $[\Delta, \chi_2]$ and χ_1. On the other hand M_i is compact because Δ_X is a second-order elliptic differential operator in the interior (standard elliptic parametrix construction). □

[10] A family of bounded operators $A(s)$ on a Hilbert space H, parametrized by $s \in U \subset \mathbb{C}$ is finitely meromorphic if for each point $a \in U$ we have a Laurent series representation $A(s) = \sum_{k=-m}^{\infty}(s-a)^k A_k$ converging in operator topology in some neighbourhood of a, where the coefficients A_k are finite-rank operators for $k < 0$.

Due to the explicit understanding of the spectral theory for funnel and cusp ends one can describe the structure of $R_X(s)$ in much more detail. Pursuing this would go beyond the purpose of these notes. We will now turn towards establishing important properties of the spectrum itself.

Weyl's criterion says that λ is in the essential spectrum of Δ if and only if there is a sequence $\phi_n \in L^2(\Gamma\backslash\mathbb{H}, \mu)$ with

$$\|(\Delta - \lambda)\phi_n\| \to 0.$$

Proposition 4.13. *For geometrically finite Γ of infinite co-volume the interval $[\frac{1}{4}, \infty)$ is contained in the essential spectrum of Δ.*

Proof. Note that $X = \Gamma\backslash\mathbb{H}$ must contain at least one funnel. After possibly conjugating the group we can assume that

$$\{z \in \mathbb{H} \colon |\mathrm{Re}(z)| < 1,\ \mathrm{Im}(z) < 1\} \subset \mathcal{F}.$$

We define pick $\psi_n \,\mathrm{n}\, \mathcal{C}^\infty(\mathbb{R}^2)$ such that

$$\psi_n(x,t) = \begin{cases} 0 & \text{if } |x| \geq 1 \text{ and } t \notin [0,n], \\ 1 & \text{if } |x| \leq \frac{1}{2} \text{ and } t \in [1, n-1]. \end{cases}$$

We can do this in such a way that the derivatives of second order are bounded independently of n. For $\mathrm{Re}(s) = \frac{1}{2}$ define

$$u_n(z) = y^s \psi_n(x, -\log(y)).$$

One checks that $\|u_n\|^2 \geq n - 2$ and

$$\|(\Delta_X - s(1-s))u_n\| = O(1)$$

. Taking $\phi_n = u_n/\|u_n\|$ now satisfies

$$\|(\Delta_X - s(1-s))\phi_n\| \ll n^{-1}.$$

A slight modification making this sequence orthogonal (for example by making the supports disjoint) allows us to apply Weyl's criterion and conclude the proof. □

Proposition 4.14. *For geometrically finite Γ of infinite co-volume the discrete spectrum consists of finitely many eigenvalues in the interval $(0, \frac{1}{4})$.*

Proof. This follows by Stone's formula for the spectral projectors. The argument being similar to the one in the proof of Theorem 4.3. One sees that the spectral projectors $\frac{1}{2}(P[a,b] - P(a,b))$ are zero away from the points $\lambda = s(1-s) \leq \frac{1}{4}$, where s is a pole of $R_X(s)$. Since $R_X(s)$ is finitely meromorphic there can be only finitely many poles in the relevant region. □

Proposition 4.15. *For a non-elementary geometricall finite Γ with infinite co-volume Δ has no L^2-eigenvalues in $[\frac{1}{4}, \infty)$.*

Proof. The proof relies on a unique continuation principle and is quite technical. We omit the details. Let us just note that this phenomenon holds due to the existence of funnels in the infinite volume case. □

4.5. **Patterson-Sullivan Theory.** Given a point $x \in \mathbb{B}$ we let ν_x denote the corresponding point measure at x. Now define the probability measures

$$\mu^{(s)} = \left(\sum_{\gamma\in\Gamma} e^{-sd(x,\gamma x)}\right)^{-1} \sum_{\gamma\in\Gamma} e^{-sd(x,\gamma x)}\nu_{\gamma x}.$$

We can take $x = 0$ to be the origin in the disc model $\mathbb{B}$. By Alaoglu's theorem we find a sequence $s_j \to \delta$ such that $\mu^{(s_j)}$ converges weakly to some limiting measure μ. The support of this measure is on the unit circle $\partial\mathbb{B}$.

Lemma 4.16. *For $\xi \in \Gamma$ we have*

$$\xi^\star\mu = |\xi'|^\delta \cdot \mu.$$

Here $\xi^\star\mu(E) = \mu(\xi.E)$.

Proof. Let E be a Borel subset of $\mathbb{B}$, $\xi \in \Gamma$ and $s > \delta$. By definition we have

$$\begin{aligned}\mu^{(s)}(\xi.E) &= \left(\sum_{\gamma\in\Gamma} e^{-sd(x,\gamma x)}\right)^{-1} \sum_{\substack{\gamma\in\Gamma,\\ \gamma.x\in\xi.E}} e^{-sd(x,\gamma x)} \\ &= \left(\sum_{\gamma\in\Gamma} e^{-sd(x,\gamma x)}\right)^{-1} \sum_{\substack{\gamma\in\Gamma,\\ \gamma.x\in E}} e^{-sd(\xi^{-1}x,\gamma x)}.\end{aligned}$$

Now we need the following property of the Poisson kernel $P(z,q) = \frac{1-|z|^2}{|z-q|^2}$:

$$\lim_{w\to q} e^{d(z,w)-d(z',w)} = \frac{P(z',q)}{P(z,q)} \text{ for } q \in \partial\mathbb{B}.$$

With this at hand we observe that

$$e^{-sd(\xi^{-1}x,w)} \sim e^{-sd(x,w)}P(\xi^{-1}x,q)^s \text{ for } w \to q.$$

(Note that since $q \in \partial\mathbb{B}$ and $x = 0$ is the origin we have $P(0,q) = 1$.) Now the Poisson kernel satisfies $P(\gamma z,\gamma q) = |\gamma'(q)| = P(z,q)$ for $z \in \mathbb{B}$, $q \in \partial\mathbb{B}$ and $\gamma \in \mathrm{PSU}(1,1)$. We obtain $P(\xi^{-1}x,q) = |\xi'(q)|$. For a sequence $\gamma_j x \to q \in \Lambda(\Gamma)$ we have

$$e^{-sd(\xi^{-1}x,\gamma_j x)} \sim e^{-sd(x,\gamma_j x)}|\xi'(q)|^s.$$

Since μ is supported on the boundary the result follows. □

We now define

$$F(z) = \int_{\partial B} P(z,q)^\delta d\mu(q), \tag{8}$$

which we may call *Patterson function*. Note that F is Γ-invariant and thus descends to a function on $\Gamma\backslash\mathbb{B}$. One checks that

$$(\Delta - \delta(1-\delta))F = 0.$$

Theorem 4.17 (Beardon)**.** *For any nonelementary Fuchsian group* Γ *we have* $\delta > 0$*. Furthermore, if* Γ *contains parabolic elements then* $\delta > \frac{1}{2}$*.*

Proof. Suppose $\delta = 0$. Then μ is Γ-invariant. If $T \in \Gamma$ is hyperbolic with fixed points $q_\pm \in \partial\mathbb{B}$, then we can partition $\partial\mathbb{B} \setminus \{q_\pm\}$ into a countable finite collection of disjoint intervals that are mapped to each other by powers of T. By invariance of μ and because the total mass is finite we conclude that $\mu(\partial\mathbb{B} \setminus \{q_\pm\}) = 0$. We further conclude that $\{q_\pm\}$ is a finite orbit of Γ. This implies that Γ is elementary. (A similar argument works if we assume that T is parabolic.)

Now assume that Γ is non-elementary and contains a parabolic element T. Conjugating Γ if necessary allows us to assume that T fixes 1 and maps i to -1. Therefor

$$T^n = \begin{pmatrix} 1 + in/2 & -in/2 \\ in/2 & 1 - in/2 \end{pmatrix}.$$

One checks that

$$|(T^n)'(z)| = |1 + in(z-1)|^{-2}.$$

Let $E = \{e^{i\theta} \colon \pi/2 < \theta \leq \pi\}$, so that $\{T^n E \colon n \in \mathbb{Z}\}$ forms a disjoint cover of $\partial\mathbb{B} \setminus \{1\}$. Since Γ is non-elementary μ can not concentrate entirely on $\{1\}$. We conclude that $\mu(E) > 0$. We obtain

$$1 > \mu(\partial\mathbb{B} \setminus \{1\}) = \sum_n \mu(T^n E) = \sum_n \int_E |(T^n)'(z)|^\delta d\mu(z) \geq \mu(E) \sum_n (1 + 4n^2)^{-\delta}.$$

However, the right hand side converges only for $\delta > \frac{1}{2}$. □

Proposition 4.18. *For* Γ *non-elementary and geometrically finite, the measure* μ *has no atoms.*

Proof. The proof relies on the following characterization of limit points. We call $q \in \Lambda(\Gamma)$ radial limit point if there exists a geodesic ray η_q in $\mathbb{B}$ with endpoint q and an orbit Γw such that $\{z \in \Gamma w \colon d(z, \eta_q) < r\}$ is infinite for some $r > 0$. (Example: If q is an attracting hyperbolic fixed point of $T \in \Gamma$, then q is radial.) It can be shown that if Γ is geometrically finite, then all points in $\Lambda(\Gamma)$ are either parabolic fixed points or radial limit points.

It is now easy to see that there can be no atoms at radial limit points. Indeed, without loss of generality we can assume that $q = 1$ and take a geodesic ray η ending at 1 such that $d(\gamma_n^{-1} 0, \eta) < C$ for some sequence $\gamma_n \in \Gamma$. This implies that γ_n^{-1} approaches 1 within a sector of the form

$$\{|\mathrm{Im}(z)| \leq c\,\mathrm{Re}(1 - z)\}.$$

(This can be best seen geometrically by drawing a picture.) We obtain

$$|\gamma_n'| = \frac{1 - |\gamma_n^{-1} 0|^2}{|\gamma_n^{-1} 0 - 1|^2},$$

so that $|\gamma_n'(1)| \to \infty$. Furthermore we know that

$$\mu(\{\gamma_n 1\}) = \gamma_n^* \mu\{1\} = |\gamma_n'(1)|^\delta \mu\{1\}.$$

This implies that $\mu\{1\} = 0$, so that there can not be an atom at 1.

The argument that there are no atoms at parabolic fixed points is more involved. We omit the details. (Note that in our applications later we assume that there are no parabolic fixed points anyway!) □

It is well known that the geodesic flow (on $\mathcal{S}X$) is ergodic with respect to the Liouville measure. This, together with the fact that μ has no atoms implies the following result

Proposition 4.19. *The action of Γ on $\partial\mathbb{B}$ is ergodic with respect to μ. Furthermore, the product action $T\colon (q, q') \mapsto (T_q, Tq')$ of Γ on $(\partial\mathbb{B} \times \partial\mathbb{B})_- = \{(q, q')\colon q \neq q'\}$ is ergodic with respect to $\tilde{\mu}$. Here $d\tilde{\mu}(q, q') = \frac{d\mu(q)d\mu(q')}{|q-q'|^{2\delta}}$.*

Theorem 4.20 (Patterson, Sullivan)**.** *Let Γ be a geometrically finite Fuchsian group of second type. Then $\dim_H(\Lambda(\Gamma)) = \delta$. Even more, if Γ has no parabolic fixed points then the Patterson-Sullivan measure μ is a constant multiple of the Hausdorff measure $H^\delta|_{\Lambda(\Gamma)}$.*

Proof. We sketch the proof only for the case of Γ with no parabolic elements. (I.e. if the convex core $\Gamma\backslash\tilde{N}$ is compact. Sometimes this property is refereed to as Γ being convex cocompact.)

Given $w \in \mathbb{B}$ and $r > 0$ so that $0 \notin B(w, r)$. Then we define the shadow of $B(w, r)$on $\partial\mathbb{B}$ by

$$I(w, r) = \{q \in \partial\mathbb{B}\colon d([0, q], w) < r\}.$$

If r is constant and $|w|$ is bounded away from zero we have the estimate

$$|I(w, r)| \asymp 1 - |w|.$$

for the euclidean arc length. A key input for the proof is Sullivan's shadow lemma, which states that for fixed (sufficiently large) r and all but finitely many $\gamma \in \Gamma$ we have

$$\mu(I(\gamma 0; r)) \asymp |I(\gamma 0, r)|^\delta.$$

To use this one shows that all small intervals in $\partial\mathbb{B}$ can be approximated by shadows of the form $I(\gamma 0, r)$. It is here were our simplifying assumption on Γ comes in, since it can be seen that neighborhoods of parabolic fixed points can not be approximated this way. More precisely, (if there are no parabolic fixed points) one can show the following. For $q \in \Lambda(\gamma)$ let I_q denote an interval in $\partial\mathbb{B}$ centered at q. There exists $\epsilon > 0$ such that for any $|I_q| < \epsilon$ we have $\mu(I_q) \asymp |I_q|^\delta$, uniformly in q.

With these technical pre-requisites taken for granted we can conclude the proof. Indeed we easily see from the definition of the Hausdorff measure that $\mu(A) \asymp H^\delta(A)$ for any Borel set $A \subset \Lambda(\Gamma)$. This means that μ is absolutely continuous

with respect to H^δ. More precisely $d\mu = f dH^\delta$ for some f on $\Lambda(\Gamma)$. Note that the function f is Γ-invariant, so that by ergodicity it must be constant.

Thus we have seen that $\mu = c \cdot H^\delta|_{\Lambda(\Gamma)}$. Since $\mu(\Lambda(\Gamma)) = 1$, the statement about the Hausdorff measure follows directly. □

4.5.1. *Resonances.* Recall that we are working with geometrically finite Fuchsian groups Γ of the second kind. The key to the proof of the spectral theorem was the meromorphic continuation of the resolvent $R_X(s)$, where $X = \Gamma\backslash\mathbb{H}$.

Definition 4.1. The poles of $R_X(s)$ are called **resonances**. We write $\mathcal{R}_X$ for the set of all resonances.

We will mainly be interested in resonances for $\delta \leq \frac{1}{2}$, because in this case Δ_X has no discrete eigenvalues and the resonances are somehow the natural replacement. Note that if $\delta < \frac{1}{2}$, then Γ can not contain parabolic elements.

Theorem 4.21 (Patterson)**.** *Assume that $\Gamma\backslash\mathbb{H}$ is geometrically finite, non-elementary and of infinite volume. There is a resonance of multiplicity one at the point $s = \delta$ such that*

$$\mathrm{res}_{s=\delta} R_X(s; z, w) = c(\Gamma) F(z) F(w),$$

where F is the Patterson function defined in (8). Furthermore, there are no other resonances in the half-plane $\mathrm{Re}(s) \geq \delta$.

Proof. Again we only give a rough sketch of the proof. The key technical input is the following result concerning the resolvent kernel. Recall the Poincaré series

$$\mathcal{P}_\Gamma(z, w; s) = \sum_{\gamma \in \Gamma} e^{-sd(z, \gamma w)}$$

of Γ. One can show that

$$R_X(s; z, w) = \frac{4^{s-1}}{\pi} \frac{\Gamma(s)^2}{\Gamma(2s)} \mathcal{P}_\Gamma(z, w; s) + H(s; z, w),$$

for $\mathrm{Re}(s) > \delta$ and a function $H(s; z, w)$ which is holomorphic for $\mathrm{Re}(s) > \delta - 1$. This follows directly from our explicit knowledge of $R_{\mathbb{H}}(s; z, w)$ and the relation $R_X(s; z, w) = \sum_{\gamma \in \Gamma} R_{\mathbb{H}}(s; z, \gamma w)$, which is valid for $\mathrm{Re}(s) > \delta$.

The next step is to show that $R_X(s; z, w)$ as well as $\mathcal{P}_\Gamma(z, w; s)$ both have a pole at $\delta = s$. This can be seen by alluding to Landau's theorem, which says that a Dirichlet series has a pole at its abscissa of convergence.

To identify the residue we have to recall the construction of the Patterson-Sullivan measure. Note that we can do so using an arbitrary base point:

$$\mu_w^{(s)} = \left(\sum_{\gamma \in \Gamma} e^{-sd(0, \gamma w)} \right)^{-1} \sum_{\gamma \in \Gamma} e^{-sd(0, \gamma w)} \delta_{\gamma w}$$

and $\mu_w = \lim_{s_j \to \delta} \mu_w^{(s_j)}$. It turns out that μ_w is absolutely continuous with respect to the original Patterson-Sullivan measure $\mu = \mu_0$. One computes that

$$\int_{\partial \mathbb{B}} P(z,q)^\delta d\mu_w(q) = \lim_{s \to \delta} \frac{\mathcal{P}_\Gamma(z,w;s)}{\mathcal{P}_\Gamma(0,w;s)}.$$

A quick computation shows that $\mathcal{P}_\Gamma(z,w;s)$ has a pole of order 1 at $s = \delta$. This implies that the corresponding pole of the resolvent also has order one. It is a general result that then we can write

$$R_X(s;z,w) = \sum_{k=1}^{q} \frac{\phi_k(z)\phi_k(w)}{s-\delta} + (\text{holomorphic}),$$

for linearly independent, real valued generalized eigenfunctions. We obtain that

$$\int_{\partial \mathbb{B}} P(z,q)^\delta d\mu_w(q) = a(w) \sum_{k=1}^{q} \phi_k(z)\phi_k(w).$$

By choosing points $w_1, \ldots, w_q$ suitably we find a matrix $(A_{k,j})_{k,j}$ so that

$$\phi_k(z) = \int_{\partial \mathbb{B}} P(z,q)^\delta h_k(q) d\mu_0(q) \text{ for } h_k d\mu_0 = \sum_{j=1}^{q} A_{kj} d\mu_{w_j}.$$

One checks that h_k is (almost everywhere) Γ-invariant. By ergodicity we conclude that h_k is constant for all k. This implies that $q = 1$ and $\phi_k = c \cdot F$ as desired.

Finally we need to show that there are no other resonances on the line $\mathrm{Re}(s) = \delta$. We argue by contradiction assuming there is such a resonance ζ. The idea is that we can run the same construction that led to the Patterson-Sullivan measure with ζ replaced by δ. We get a measure

$$\sigma = \lim_{s_j \to \zeta} (s_j - \zeta) \sum_{\gamma \in \Gamma} e^{-s_j d(0,\gamma 0)} \delta(\gamma 0).$$

One goes on to show that σ has the same properties as μ and is absolutely continuous with respect to μ:

$$d\sigma(q) = \psi(q) d\mu(q).$$

Furthermore

$$\psi(\gamma q) = |\gamma'(q)|^{\zeta - \delta} \psi(q).$$

Similarly one constructs $\tilde{\sigma}$, which is absolutely continuous with respect to $\tilde{\mu}$. Using ergodicity one finds that $\tilde{\sigma} = c\tilde{\mu}$, so that

$$\psi(q)\psi(q') = c|q - q'|^{2(\zeta - \delta)},$$

for almost all q, q'. With a bit more work this identity can be extended to all $(q,q') \in (\partial \mathbb{B} \times \partial \mathbb{B})_-$. Taking the limit $q_j \to q$ in $\Lambda(\Gamma)$ and using continuity of ψ we get $\psi(q)^2 = 0$. But this implies that σ is 0, contradicting the existence of ζ. $\square$

4.6. **Pre-Trace Inequality and the Convolution of Kernels.** Recall that a point pair invariant was a compactly supported smooth function $k\colon \mathbb{H}\times\mathbb{H}\to\mathbb{C}$ satisfying $k(gz,gw)=k(z,w)$ for all $z,w\in\mathbb{H}$ and all $g\in \mathrm{SL}_2(\mathbb{R})$. We abuse notation and write $k(z,w)=k(u(z,w))$, which is possible since k depends only on the hyperbolic distance. The associated automorphic kernel (with respect to a geometrically finite Fuchsian group) is defined by

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

The convolution of two kernels is defined by

$$[k_1\star k_2]=\int_{\mathbb{H}} k_1(z,x)k_2(x,w)d\mu(x).$$

We are interested in a very specific point-pair invariant:

$$k_1(z,w)=\mathbb{1}_{u(z,w)\leq (X-2)/4}. \tag{9}$$

This is of interest since we have

$$K_1(z,w)=\sum_{\gamma\in\Gamma} k_1(z,\gamma w)=\sharp\{\gamma\in\Gamma\colon 4u(z,\gamma w)+2\leq X\}.$$

In particular, taking $z=w=i$ this simplifies to

$$N_1(\sqrt{X},\Gamma)=K_1(i,i)=\sharp\{\gamma\in\Gamma\colon \|\gamma\|^2\leq X\}. \tag{10}$$

We will need to establish a good estimate for the self-convolution of this kernel.

Proposition 4.22. *Let $k=k_1\star\overline{k_1}$. Then we have*

$$k(z,w)\ll e^{T-\frac{\rho(z,w)}{2}} \text{ if } \rho(z,w)\leq 2T$$

and 0 for $\rho(z,w)\geq 2T$, where $e^T+e^{-T}=X$.

Proof. To prove this it is more convenient to work in the disc model $\mathbb{B}$.

By definition of the convolution we find that

$$k(z,w)=\mathrm{Area}_h(B(z,T)\cap B(w,T)),$$

here $B(z,T)$ is the hyperbolic disc with center z and radius T. Denote $B(z,T)\cap B(w,T)=E$. Obviously $E=\emptyset$ if $\rho(z,w)>2T$. Thus we can assume the contrary and without loss of generality we take $z=0$.

Note that $\rho(0,\zeta)=T$ is an Euclidean circle centered at 0 and with radius R determined by

$$\sinh(T/2)=\frac{R}{\sqrt{1-R^2}} \text{ or equivalently } \tanh(T/2)=R.$$

Thus, if $\rho(0,w)=s$, then we can assume that w has euclidean coordinates $(\tanh(s/2),0)$. Put $d=\tanh(s/2)$.

Next we observe that the set of points $\rho(w,\zeta)$ is a Euclidean circle of radius R_0 centered at $(a_0,0)$, where

$$a_0 = d\frac{1-R^2}{1-R^2d^2} \text{ and } R_0 = R\frac{1-d^2}{1-R^2d^2}.$$

We consider the hyperbolic triangle $(0,w,\zeta)$, where $\rho(0,\zeta)=\rho(w,\zeta)=T$. By the hyperbolic cosine law we get

$$\cosh(T) = \cosh(s)\cosh(T) - \sinh(s)\sinh(T)\cos(\alpha) \text{ and } \cos(\alpha) = \frac{\tanh(s/2)}{\tanh(T)}.$$

Let $v=(s/2,0)$. Then our goal is to show that $E \subset B(v,r)$ for $\cosh(r) = \frac{\cosh(T)}{\cosh(s/2)}$. Once this is established we are done since, by the hyperbolic area formula for discs, we get

$$\text{Area}_h(B(v,r)) = 4\pi\sinh(r/2)^2 = 2\pi\left(\frac{\cosh(T)}{\cosh(s/2)} - 1\right) \ll e^(T - s/2).$$

To see the inclusion we let A,B be the points where $B(0,T)$ intersects $B(\zeta,T)$. Let C is any point in $B(w,T)$ with $\rho(C,0)=T$ and denote the angle $\widehat{Czv}$ by ϕ. We must have $\phi<\alpha$. Using the hyperbolic cosine law once again we find that

$$\cosh(\rho(0,C)) \leq \frac{\cosh(T)}{\cosh(s/2)}.$$

A similar estimate holds for $C \in B(0,T)$ with $\rho(C,w)=T$. This shows the claim. □

Spectrally expanding the automorphic kernel K_Γ associated to $k = k_1 \star \overline{k_1}$ we obtain the following very important pretrace inequality.

Proposition 4.23. *Suppose Γ is a geometrically finite Fuchsian group and take $k = k_1 \star \overline{k_1}$ be as above. Let $\lambda_0,\dots,\lambda_j$ be the discrete eigenvalues in $[0,\frac{1}{4})$ with corresponding eigenfunctions $\phi_0,\dots,\phi_j$. Then*

$$K(z,z) \geq \sum_{\lambda_i<\frac{1}{4}} |h_1(t_i)|^2|\phi_i(z)|^2,$$

where h_1 is the Selberg/Harish-Chandra transform of k_1.

Proof. We first note that the Selberg/Harish-Chandra transform translates convolutions into products. In particular we have

$$\int_{\mathbb{H}} k(x,y)\phi(y)d\mu(y) = |h_1(t)|^2\phi(x).$$

By the spectral theorem for Δ on $L^2(\Gamma\backslash\mathbb{H},\mu)$ we know that there are finitely many eigenvalues in $[0,\frac{1}{4})$ which make up the discrete spectrum:

$$L^2_{\text{disc}}(\Gamma\backslash\mathbb{H},\mu) = V_1 = \bigoplus_{\lambda_i<\frac{1}{4}} \mathbb{C}\phi_i.$$

The rest of the spectrum is absolutely continuous and covers $[\frac{1}{4},\infty)$ (there are no embedded eigenvalues). Put $V_2 = L^2_{\text{cont}}(\Gamma\backslash\mathbb{H},\mu)$, so that $L^2(\Gamma\backslash\mathbb{H},\mu) = V_1 \oplus V_2$.

By construction as a convolution the integral operator $T_k\colon L^2(\Gamma\backslash\mathbb{H},\mu) \to L^2(\Gamma\backslash\mathbb{H},\mu)$ is non-negative:

$$\langle T_k f, f\rangle \geq 0 \text{ for all } f \in L^2(\Gamma\backslash\mathbb{H},\mu).$$

Spectrally expanding the automorphic kernel yields

$$K(z,z) = \sum_{\lambda_i<\frac{1}{4}} |h_1(t_i)|^2|\phi_i(z)|^2 + B(z,z) \tag{11}$$

where $B(z,w) = K(z,w) - \sum_{\lambda_i<\frac{1}{4}} h(\lambda_i)\phi_i(z)\overline{\phi_i(w)}$. It turns out that B is the kernel for the operator $T_B = \text{Pr}_{V_2}\circ T_k$, which is obviously non-negative. For fixed z we define

$$f_n(w) = \delta_{d(z,w)\leq 1/n} \tag{12}$$

and observe that

$$B(z,z) = \lim_{n\geq\infty} \langle Bf_n, f_n\rangle \geq 0. \tag{13}$$

Thus we can drop $B(z,z)$ from the pre-trace formula and the result follows. □

Finally we remark that the Selberg/Harish-Chandra transform of k_1 can be computed explicitly (**Exercise**) and one finds

$$h_1(t) = 2\sqrt{\pi}\frac{\Gamma(s-\frac{1}{2})}{\Gamma(s+1)}X^s + O(\sqrt{X}) \text{ for } \frac{1}{2} < s \leq 1.$$

Here $\lambda = s(1-s)$ and $s = \frac{1}{2} + it$.

4.7. **Spectral Theory on the Group Level.** So far we have looked at the spectral theory of the Laplace-Beltrami operator Δ on quotients $\Gamma\backslash\mathbb{H}$. Put $G = \text{SL}_2(\mathbb{R})$. Since we can identify $\mathbb{H} = G/K$ for $K = \text{SO}_2(\mathbb{R})$ and Δ is G-invariant, the spectral theory of Δ is equivalent to the decomposition of $L^2(\Gamma\backslash G)$ into irreducible G-modules. Here G acts on $L^2(\Gamma\backslash G)$ by right translation.

The Cartan decomposition of G reads $G = KA^+K$, for

$$A^+ = \{a_t = \text{diag}(e^{-\frac{t}{2}}, e^{\frac{t}{2}})\colon t \geq 0\}.$$

Note that K is abelian and can be parametrized as

$$K = \{k_\theta = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta)\end{pmatrix} : \theta \in [0,2\pi)\}.$$

Let Γ be a Fuchsian group of second type with critical exponent $\delta = \delta_\Gamma > \frac{1}{2}$. We order the eigenvalues of the Laplacian on $\Gamma\backslash\mathbb{H}$ below $\frac{1}{4}$ by

$$0 < \delta(1-\delta) = \lambda_0 < \lambda_1 \leq \ldots \leq \lambda_N < \frac{1}{4}.$$

Another parametrisation of the eigenvalues is $\lambda_j = s_j(1-s_j)$ with $s_j > \frac{1}{2}$. The corresponding Laplace eigenfunctions are denoted by φ_j.

As mentioned above the group G acts on the Hilbert space $V = L^2(\Gamma\backslash G)$ by right translation. The G-span of φ_j in V will be denoted by V_{φ_j}. As a G-representation it is isomorphic to the complementary series representation with parameter s_j. (We are normalizing everything so that the principal series lie on the critical line $\mathrm{Re}(s) = \frac{1}{2}$.) This gives us the decomposition

$$V = V_{\varphi_0} \oplus V_{\varphi_1} \oplus \ldots \oplus V_{\varphi_N} \oplus V_{temp}.$$

The spaces V_{φ_j} will be called the automorphic model of the corresponding G-representation. We will now also recall the line model. Given a function $f\colon \mathbb{R} \to \mathbb{C}$ we define

$$[\mathcal{I}f](y) = \int_{\mathbb{R}} \frac{f(x)}{|x-y|^{2(1-s)}} dx.$$

Further, we introduce the pairing

$$\langle f_1, f_2 \rangle = \int_{\mathbb{R}} f_1(x)\overline{[\mathcal{I}f_2](x)} dx.$$

Let V_s denote the space of $f\colon \mathbb{R} \to \mathbb{C}$ with $\langle f, f\rangle < \infty$. The G-action on this space is given by

$$\pi\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) f(x) = |-bx+d|^{-2s} f\left(\frac{ax-c}{-bx+d}\right).$$

We have constructed a G-representation (π, V_s). This is the line model for the complementary series with parameters s. (As mentioned above we have $V_{\varphi_j} \cong V_{s_j}$ as G-module.)

Write $\mathcal{H}$ for one of the irreducible spaces V_{φ_j}. The (dense) subspace of smooth vectors in $\mathcal{H}$ will be denoted by $\mathcal{H}^\infty$. We have the following decomposition on K-isotypic components:

$$\mathcal{H}^\infty = \bigoplus_{k\in\mathbb{Z}} \mathcal{H}^{(2k)},$$

where $\mathcal{H}^{(2k)} = \mathbb{C} \cdot v_{2k}$. This simple structure of the K-isotypic parts is due to the fact that K is abelian and thus its representation theory is very easy. Indeed v_{2k} is a function of weight $2k$, that means it satisfies

$$v_{2k}(gk_\theta) = e^{2ik\theta} v_{2k}(g),$$

for all $g \in G$.

Recall that the Lie algebra $\mathfrak{g}$ of G acts on $\mathcal{H}^\infty$ by differential operators. A basis of $\mathfrak{g} = \mathfrak{sl}_2(\mathbb{R})$ is given by

$$h = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, e = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \text{ and } f = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

In the complexified Lie algebra $\mathfrak{g}_{\mathbb{C}}$ we define the two important operators

$$\mathcal{R} = h + i(e + f) \text{ and } \mathcal{L} = h - i(e + f).$$

These operators go by the names raising and lowering operator. We will need the following expressions for them in KA^+K (i.e. $g = k_{\theta_1} a_t k_{\theta_2} \leftrightsquigarrow (\theta_1, t, \theta_2)$) coordinates:

$$\mathcal{R} = e^{2i\theta_2} \left(-i\operatorname{csch}(t) \frac{\partial}{\partial \theta_1} + 2\frac{\partial}{\partial t} + i \coth(t) \frac{\partial}{\partial \theta_2} \right) \text{ and}$$
$$\mathcal{L} = e^{-2i\theta_2} \left(i\operatorname{csch}(t) \frac{\partial}{\partial \theta_1} + 2\frac{\partial}{\partial t} - i \coth(t) \frac{\partial}{\partial \theta_2} \right).$$

The proof is a straight forward but cumbersome computation that can be found in [7, Lemma 2.7]. Another important operator is the Casimir operator

$$\mathcal{C} = \frac{1}{2}h^2 + ef + fe \in \mathcal{U}(\mathfrak{g}_{\mathbb{C}}).$$

It acts on $\mathcal{H}^\infty$ as scalar multiplication by $-2\lambda = -2s(1-s)$. The following expression in Cartan coordinates is well known:

$$\frac{1}{2}\mathcal{C} = \frac{\partial^2}{\partial t^2} + (\coth(t))\frac{\partial}{\partial t} + \operatorname{csch}(t)^2 \left(\frac{\partial^2}{\partial \theta_1^2} + \frac{\partial^2}{\partial \theta_2^2} \right) - \frac{2\cosh(t)}{\sinh(t)^2} \cdot \frac{\partial^2}{\partial \theta_1 \partial \theta_2}.$$

Finally we can change variables $(\theta_1, t, \theta_2) \leftrightsquigarrow (\theta_1, r, \theta_2)$ for $r = \tanh(t/2)$ or (almost) inversely $e^t = \frac{1+r}{1-r}$. One has

$$\frac{\partial}{\partial t} = \frac{1}{2}(1 - r^2)\frac{\partial}{\partial r}.$$

Note that $\operatorname{csch}(t) = \frac{1-r^2}{2r}$ and $\cosh(t) = \frac{1+r^2}{2r}$. Thus we have

$$\frac{1}{2}\mathcal{C} = \frac{(1-r^2)^2}{4} \cdot \frac{\partial^2}{\partial r^2} + \frac{(1-r^2)^2}{4r} \cdot \frac{\partial}{\partial r} + \frac{(1-r^2)^2}{16r^2} \cdot \left(\frac{\partial^2}{\partial \theta_1^2} + \frac{\partial^2}{\partial \theta_2^2} \right) - \frac{1-r^4}{8r^2} \cdot \frac{\partial^2}{\partial \theta_1 \partial \theta_2},$$
$$\mathcal{R} = e^{2i\theta_2} \left(-i\frac{1-r^2}{2r} \cdot \frac{\partial}{\partial \theta_1} + (1-r^2)\frac{\partial}{\partial r} + i\frac{1+r^2}{2r} \cdot \frac{\partial}{\partial \theta_2} \right) \text{ and}$$
$$\mathcal{L} = e^{-2i\theta_2} \left(i\frac{1-r^2}{2r} \cdot \frac{\partial}{\partial \theta_1} + (1-r^2)\frac{\partial}{\partial r} - i\frac{1+r^2}{2r} \cdot \frac{\partial}{\partial \theta_2} \right).$$

We turn our attention towards the line model $\mathcal{H} = V_s$. Here we can describe the weight $2k$ element $f_{2k,s} \in V_s^{(2k)}$ element explicitly by

$$f_{2k,s}(x) = c(x - i)^{k-s}(x + i)^{-k-s}$$

up to some constant $c \in \mathbb{C}$. To see this we consider the action of $Y = e - f \in \mathfrak{g}$ on $f \in V_s^{(2k)}$. It is given by $Y.f = \frac{\partial}{\partial \theta_2} f = 2ikf$. On the other hand, in the line model Y acts on V_s by

$$Y.f(x) = 2sxf(x) + (1+x^2)f'(x).$$

Thus we get the ordinary differential equation $2sxf(x) + (1+x^2)f'(x) = 2ikf(x)$. One checks that the proposed function $f_{2k,s}$ is the up to scaling unique solution.

We will now derive a suitable basis for $\mathcal{H}$ using the weight $2k$ elements v_{2k}. Starting point is a weight 0 (i.e. K-fixed) vector v_0, which we normalize by $\langle v_0, v_0 \rangle = 1$. The raising (resp. lowering) operator takes $\mathcal{H}^{(2k)}$ to $\mathcal{H}^{2k+2}$ (resp. $\mathcal{H}^{2-2}$). However, it does not respect the norm. Thus we need the following computation.

Lemma 4.24. *let $\mathcal{X} \in \{\mathcal{R}, \mathcal{L}\}$. Then, for any $k \geq 0$ we have*

$$\langle \mathcal{X}^k v_0, \mathcal{X}^k v_0 \rangle = 2^{2k} \frac{\Gamma(s+k)\Gamma(1-s+k)}{\Gamma(s)\Gamma(1-s)} = b_{k,s}.$$

Proof. We do the case $\mathcal{X} = \mathcal{R}$. The remaining case is left as an exercise. Recall that

$$\mathcal{L}\mathcal{R} = 2\mathcal{C} + Y^2 + 2iY, \text{ for } Y = e - f.$$

Therefore this operator acts on $\mathcal{H}^{(2k)}$ by scalar multiplication with

$$-4\lambda + (2ik)^2 + 2i(2ik) = -4(s+k)(1-s+k).$$

Note that $\langle \mathcal{R}v, w \rangle = -\langle v, \mathcal{L}w \rangle$. With these facts gathered we can compute

$$\begin{aligned}
&\langle \mathcal{X}^k v_0, \mathcal{X}^k v_0 \rangle = (-1)^k \langle \mathcal{L}^k \mathcal{R}^k v_0, v_0 \rangle \\
&= (-1)^k (-4(s)(1-s)(-4(s+1)(1-s+1)) \cdots (-4(s+k-1)(1-s+k-1))) \langle v_0, v_0 \rangle \\
&= (-1)^k (-1)^k 4^k \frac{\Gamma(s+k)\Gamma(1-s+k)}{\Gamma(s)\Gamma(1-s)}.
\end{aligned}$$

□

For v_0 fixed as above we now define the convenient basis

$$v_{2k} = \frac{1}{\sqrt{b_{|k|,s}}} \cdot \begin{cases} \mathcal{R}^k v_0 & \text{if } k > 0, \\ \mathcal{L}^{|k|} v_0 & \text{if } k < 0. \end{cases}$$

The next goal is to connect v_{2k} in the automorphic model to the functions $f_{2k,s}$ in the line model. To do this we need the following result:

Lemma 4.25. *We have*

$$[\mathcal{I} f_{2k,s}] = \frac{4^{1-s}\pi(-1)^k \Gamma(2s-1)}{\Gamma(s-k)\Gamma(s+k)} f_{2k,1-s}.$$

Proof. Since the intertwiner $\mathcal{I}$ preserves the group action we must have $\mathcal{I}f_{2k,s} \in V_{1-s}^{(2k)}$. Therefore it must be a multiple of $f_{2k,1-s}$ and it suffices to compute $\mathcal{I}f_{2k,s}(0)$ to determine this multiple. This boils down to evaluating the integral

$$\int_{\mathbb{R}} \frac{(y-i)^{k-s}(y+i)^{-k-s}}{|y|^{2(1-s)}} dy.$$

Computing this and recalling that $f_{sk,1-s}(0) = (-i)^{k-(1-s)} \cdot i^{-k-(1-s)}$ concludes the proof. □

It is now easy to compute the norms of $f_{2k,s}$:

Lemma 4.26. *For $f_{2k,s}$ as above we have*

$$\langle f_{2k,s}, f_{2k,s} \rangle = \frac{4^{1-s}\pi^2(-1)^k\Gamma(2s-1)}{\Gamma(s-k)\Gamma(s+k)} = \tilde{b}_{k,s}.$$

Proof. This follows directly from the definition of the inner product and our computation of $\mathcal{I}f_{2k,s}$ above. □

We have established how to generate the full G-module V_j from the original eigenfunction φ_j. This was done by using the ladder operators $\mathcal{R}$ and $\mathcal{L}$ go generate a nice basis in the automorphic model. Furthermore we understand the image of this basis in the line model. The latter is very useful for explicit computations as we will see below.

Given $v_{2k} \in \mathcal{H}^{(2k)}$ we write $v_{2k}(\theta_1, t, \theta_2) = v_{2k}(k_{\theta_1} a_t k_{\theta_2})$ and we can also apply the change of variables t to r described earlier. One obtains the Fourier expansion

$$v_{2k}(\theta_1, r, \theta_2) = e^{2ik\theta_2} \sum_{n\in\mathbb{Z}} v_{2n,2k}(r) e^{2in\theta_1}.$$

(Note that we only pick up even frequencies because the center of G acts trivially.) Applying the Casimir operator to the Fourier expansion (term wise) yields the equations

$$\frac{(1-r^2)^2}{4} \cdot \frac{\partial^2}{\partial r^2} v_{2n,2k}(r) + \frac{(1-r^2)^2}{4r} \cdot \frac{\partial}{\partial r} v_{2n,2k}(r) + \left(-\frac{(1-r^2)^2}{4r^2}\left(n^2+k^2\right) + \frac{1-r^4}{2r^2} nk + s(1-s) \right) v_{2n,2k}(r) = 0.$$

We want to solve this equation, but to do so we need to have some regularity. Since v_{2k} is regular at the origin (actually everywhere) also $v_{2n,2k}$ is regular at $r = 0$. Approximately we have

$$(1+O(r))\frac{\partial^2}{\partial r^2} + \frac{1+O(r)}{r}\frac{\partial}{\partial r} - \frac{1+O(r)}{r^2}(n-k)^2 = 0.$$

This gives us an asymptotic solution of the form

$$\begin{cases}(c_1 r^{n-k} + c_2 r^{k-n})(1+O(r)) & \text{if } n \neq k \\ (c_1 + c_2 \log(r))(1+O(r)) & \text{if } n = k.\end{cases}$$

We find that $c_2 = 0$ for $n \geq k$ and $c_1 = 0$ for $n < k$. Either way there is a multiplicity one principle so that $v_{2n,2k}$ is a constant multiple of the unique solution $\Phi_{2n,2k}$ to the ODE above. Explicitly one computes

$$\Phi_{2n,2r}(r) = (1-r^2)^s r^{|n-k|}{}_2F_1(s-\epsilon_{n,k}k, s+\epsilon_{n,k}n; 1+|n-k|; r^2)$$

for

$$\epsilon_{n,r} = \begin{cases}1 & \text{if } n \geq k \\ -1 & \text{else.}\end{cases}$$

For convenience we write

$$\Phi_{2n,2k}(k_{\theta_1} a_t k_{\theta_2}) = e^{2in\theta_1}\Phi_{2n,2k}(r)e^{2ik\theta_2}.$$

We have obtained the Fourier expansion

$$v_{2k}(g) = \sum_{n\in\mathbb{Z}} c_{2n,2k}\Phi_{2n,2k}(g).$$

Given the coefficients $c_{2n} = c_{2n,0}$ of v_0 we can determine $c_{2n,2k}$ for all k using the ladder operators. To do this one computes, see [7, Lemma 2.27], that

$$\mathcal{R}\Phi_{2n,2k} = -2\Phi_{2n,2k+2} \times \begin{cases}-(n-k) & \text{if } n > k, \\ \frac{(s+k)(1-s+k)}{1-n+k} & \text{if } n \leq k\end{cases} \text{ and}$$

$$\mathcal{L}\Phi_{2n,2k} = -2\Phi_{2n,2k-2} \times \begin{cases}\frac{(s-k)(1-s-k)}{1+n-k} & \text{if } n \geq k, \\ -(k-n) & \text{if } n < k.\end{cases}$$

One deduces that

$$\mathcal{R}^k v_0(g) = (-1)^k 2^k \sum_{n\in\mathbb{Z}} d(n,k)\cdot c_{2n}\Phi_{2n,2k}(g),$$

for $k \geq 0$ and with

$$d(n,k) = \begin{cases}(-1)^k \frac{\Gamma(n+1)}{\Gamma(n-k+1)} & \text{if } n \geq k, \\ (-1)^n \frac{\Gamma(n+1)\Gamma(s+k)\Gamma(1-s+k)}{\Gamma(k-n+1)\Gamma(s+n)\Gamma(1-s+n)} & \text{if } 1 \leq n \leq k-1, \\ \frac{\Gamma(|n|+1)\Gamma(s+k)\Gamma(1-s+k)}{\Gamma(k+|n|+1)\Gamma(s)\Gamma(1-s)} & \text{if } n \leq 0.\end{cases}$$

Similarly we can work out the action of $\mathcal{L}^k$ on the Fourier coefficients. This shows

Proposition 4.27. *For $k \geq 0$, the value at the origin of v_0 acted on by ladder operators is related to its Fourier coefficients by*

$$\mathcal{R}^k v_0(e) = c_{2k} 2^k \Gamma(k+1)$$

and

$$\mathcal{L}^k v_0(e) = c_{-2k} 2^k \Gamma(k+1)$$

We conclude this discussion on Fourier expansions by remarking that in general the coefficients c_{2n} can be rather complicated. However, since the first eigenfunction φ_0 can be described using the Patterson-Sullivan-measure μ we can be more precise. Indeed we simply have

$$\varphi_0(\theta_1, r, \theta_2) = \int_0^\pi \left(\frac{1-r^2}{|re^{2i\theta_1} - e^{2i\alpha}|^2} \right)^\delta d\mu(\alpha).$$

The Fourier development discussed above reads

$$\varphi_0(\theta_1, r, \theta_2) = \sum_{n \in \mathbb{Z}} c_{2n} \Phi_{2n,0}(r) e^{2in\theta_1}.$$

Define the $2n$-th Fourier coefficient of μ by

$$\hat{\mu}(2n) = \int_0^\pi e^{2in\alpha} d\mu(\alpha).$$

Proposition 4.28. *The relationship between the coefficients c_{2n} and $\hat{\mu}$ is given by*

$$c_{2n} = \frac{1}{\Gamma(\delta)} \cdot \frac{\Gamma(\delta + |n|)}{\Gamma(1+|n|)} \hat{\mu}(-2n).$$

Proof. Recalling the shape of $\Phi_{2n,0}$ in terms of the Gauß hypergeometric function yields

$$\sum_n c_{2n} r^{|n|} {}_2F_1(\delta, \delta+|n|; 1+|n|; r^2) e^{2in\theta_1} = \int_0^\pi (r^2 - r(e^{i(2\theta_1 - 2\alpha)} + e^{i(2\alpha - 2\theta_1)}) + 1)^{-\delta} d\mu(\alpha).$$

The result follows by writing down the appropriate series expansions of both sides and comparing coefficients. □

For the record we state the following asymptotic of $\Phi_{2n,2k}$ as $t \to \infty$:

$$\Phi_{2n,2k}(a_t) = 4^{1-s} e^{-t(1-s)} \frac{\Gamma(1+|n-k|)\Gamma(2s-1)}{\Gamma(s-\epsilon_{n,k}k)\Gamma(s+\epsilon_{n,k}n)} (1 + O(nke^{-1})).$$

This is [7, Lemma 2.30].

We turn towards matrix coefficients. If π denotes the tight-regular representation on the irreducible space $\mathcal{H}$, then we write

$$M_{2n,2k}(g) = \langle \pi(g) v_{2k}, v_{2n} \rangle.$$

Here $v_{2k} \in \mathcal{H}^{(2k)}$ is the normalized basis element constructed from v_0 by applying the ladder operators. We observe straight away that

$$M_{2n,2k}(k_{\theta_1} g k_{\theta_2}) = e^{2in\theta_1} M_{2n,2k}(g) e^{2ik\theta_2}.$$

Since $M_{2n,2k}$ are also eigenfunctions of the Casimir operator one sees that $M_{2n,2k}$ is a multiple of $\Phi_{2n,2k}$. For example, if $n = k$ one directly obtains $M_{2n,2n}(g) = \Phi_{2n,2n}$. The general case is treated in the following lemma.

Lemma 4.29. *For integers $n, k \in \mathbb{Z}$ we have*

$$M_{2k,2n}(g) = \frac{(-1)^k 4^{1-s}\pi^2\Gamma(2s-1)}{\Gamma(1+|n-k|)\Gamma(s-\epsilon_{n,k}n)\Gamma(s+\epsilon_{n,k}k)} \cdot \frac{\Phi_{2n,2k}(g)}{\sqrt{\tilde{b}_{k,s}\tilde{b}_{n,s}}}.$$

Proof. To prove this result we switch to the line model, where we can work explicitly. Of course it is enough to look at $g = a_t$, for $t = t(r)$ as usual. We have

$$M_{2k,2n}(a_t) = \langle \pi(a_t)v_{2k}, v_{2n}\rangle = \frac{\langle \pi(a_t)f_{2k,s}, f_{2n,s}\rangle}{\sqrt{\tilde{b}_{k,s}\tilde{b}_{n,s}}}.$$

The inner product is now given by

$$\langle \pi(a_t)f_{2k,s}, f_{2n,s}\rangle = \frac{4^{1-s}\pi(-1)^n\Gamma(2s-1)}{\Gamma(s-n)\Gamma(s+n)} \cdot \int_{\mathbb{R}} \left(\frac{1+r}{1-r}\right)^s \left(\left(\frac{1+r}{1-r}\right)x - i\right)^{k-s} \left(\left(\frac{1+r}{1-r}\right)x + i\right)^{-k-s} \cdot \overline{(x-i)^{n-(1-s)}(x+i)^{-n-(1-s)}}dx.$$

We know already that in r this is a constant multiple of $\Phi_{2k,2n}$. Thus it suffices to asymptotically evaluate the integral as $r \to 1$. This can be done for example using Laplace's method and one obtains

$$\int_{\mathbb{R}} \left(\frac{1+r}{1-r}\right)^s \left(\left(\frac{1+r}{1-r}\right)x - i\right)^{k-s} \left(\left(\frac{1+r}{1-r}\right)x + i\right)^{-k-s} \cdot \overline{(x-i)^{n-(1-s)}(x+i)^{-n-(1-s)}}dx = \frac{(-1)^{k+n}\pi\Gamma(s+\epsilon_{n,k}n)}{\Gamma(1+|n-k|)\Gamma(s+\epsilon_{n,k}k)}\Phi_{2k,2n}(r).$$

Combining all the constants completes the proof. □

Finally let us cite some results concerning the matrix coefficients of tempered representations.

Lemma 4.30. *Let (π, V) be a tempered unitary representation of G. Then, for any vectors $v, w \in V$ whose K-span is one-dimensional, we have*

$$|\langle \pi(k_{\theta_1}a_t k_{\theta_2})v, w\rangle| \ll te^{-\frac{t}{2}}\|v\|_2\|w\|_2$$

with absolute implied constant when $t \to \infty$.

Another estimate can be given in terms of the Sobolev norm

$$\mathcal{S}v = \|v\|_2 + \|d\pi(h).v\|_2 + \|d\pi(e).v\|_2 + \|d\pi(f).v\|_2.$$

(Recall that h, e, f are an orthonormal basis of $\mathfrak{g}$ acting on $\mathcal{H}$ infinitesimal by $d\pi$.)

Lemma 4.31. *Let $\Theta > \frac{1}{2}$ and (π, V) be a unitary representation of G which does not weakly contain any complementary series representation with parameter $s > \Theta$. Then for any smooth vectors $v, w \in V^\infty$ we have*

$$|\langle \pi(k_{\theta_1} a_t k_{\theta_2}) v, w\rangle| \ll e^{-\Theta t} (\|v\|\|w\|)^{\frac{1}{2}} (\mathcal{S}v\mathcal{S}w)^{\frac{1}{2}}$$

as $t \to \infty$ with absolute implied constant.

5. Expansion, Lattice Point Counting and Spectral Gaps

We will first introduce the three players in this section and then dedicate a subsection too each. Note that since these concepts are highly intertwined the structure of this section is not linear.

First let us briefly talk about lattice point counting. This typically means counting certain elements of a discrete group (i.e. lattice) contained in an archimedean ball. More precisely let $\|\cdot\|$ denote the Frobenius norm on 2×2 matrices. Given a subgroup (or more generally a subset) Γ of $\mathrm{SL}_(\mathbb{Z})$ we define the congruence subgroups $\Gamma(q)$ as the kernel of the natural projection $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ intersected with Γ. Then we are interested in the size of

$$N_1(T, \Gamma(q)) = \sharp(\Gamma(q) \cap B_T) = \sharp\{\gamma \in \Gamma(q) \colon \|\gamma\| \leq T\}.$$

Of course there exist many variations of this counting problem.

Recall that given $X = \Gamma\backslash\mathbb{H}$ we have seen in the previous section that parts of the discrete spectrum of Δ may be contained in the interval $[0, \frac{1}{4})$. We call such eigenvalues *exceptional eigenvalues*. Of course there are at most finitely many such eigenvalues and we number them by $0 \leq \lambda_0(\Gamma) < \lambda_1(\Gamma) \leq \cdots \leq \lambda_{\max}(\Gamma) < \frac{1}{4}$. A deep result due to Selberg states that for congruence subgroups Γ of $\mathrm{SL}_2(\mathbb{Z})$ we have the lower bound $\frac{3}{16} \leq \lambda_1(\Gamma)$. This is what we call a (quantitative) *spectral gap*. It is uniform in the sense that it holds for all congruence subgroups.[11] Here we are mostly interested in the following situation. Let Γ be a finitely generated non-elementary subgroup of $\mathrm{SL}_2(\mathbb{Z})$. In particular, X has infinite volume and we write $\delta = \delta(\Gamma)$ for the Hausdorff dimension of the limiting set of Γ. We have two very different situations. If $\frac{1}{2} < \delta < 1$, then we have seen that $\lambda_0(\Gamma) = \delta(1-\delta)$. On the other hand, if $\delta \leq \frac{1}{2}$, then Δ has no discrete spectrum. In this case one needs to consider so called scattering resonances to make sense of the spectral gap.

Finally we turn towards *expansion*. Given an undirected k-regular graph $\mathcal{G}$ with vertices V and a subset X of V, the expansion of X, called $c(X)$, dis defined by $\frac{\sharp N(X)}{\sharp X}$, where $N(X)$ is the set of neighbors of X. The expansion coefficient of a graph is

$$c(\mathcal{G}) = \inf_{\sharp X < \frac{1}{2}\sharp\mathcal{G}} c(X).$$

[11] The assumption that Γ is a congruence subgroup can not be lifted completely, since one can construct Γ with arbitrary small eigenvalues. However, in the congruence case the value $\frac{3}{16}$ can be slightly improved.

We say that a family of k-regular graphs $\mathcal{G}_n$ forms a family of expanders if there is a fixed constant $C > 0$ so that

$$\liminf_{n\to\infty} c(\mathcal{G}_n) \geq C.$$

The connection to the topics above lies in considering Cayley graphs of $\Gamma/\Gamma(p)$ for p large.

5.1. **Lattice Point Counting.** We start with the following counting result.

Lemma 5.1. *For a prime p we have*

$$N_1(T, K(p)) = \sum_{\gamma\in K(p), \|\gamma\|\leq T} 1 \ll \frac{T^{2+\epsilon}}{p^3} + \frac{T^{1+\epsilon}}{p} + 1.$$

The implicit constant is independent of p.

Proof. We follow the method of Sarnak and Xue. The problem obviously transforms into counting $(a, b, c, d) \in \mathbb{Z}^4$ with

$$\begin{aligned} |a|, |b|, |c|, |d| &\leq T, \\ ad - bc &= 1, \\ a \equiv d \equiv 1 \bmod p &\text{ and} \\ b \equiv c \equiv 0 \text{ mid } p. & \end{aligned}$$

We claim that we get the congruence conditions together with the determinant restriction yields the strong condition

$$a + d \equiv 2 \bmod p^2. \tag{14}$$

To see this we write

$$a = 1 + \lambda_1 p,\ b = \lambda_3 p,\ c = \lambda_4 p \text{ and } d = 1 + \lambda_2 p.$$

We get the equation

$$(1 + \lambda_1 p)(1 + \lambda_2 p) - \lambda_3\lambda_4 p^2 = 1.$$

This can be rewritten as

$$(\lambda_1 + \lambda_2)p = (\lambda_3\lambda_4 - \lambda_1\lambda_2)p^2.$$

With this at hand we get

$$a + d = 2 + (\lambda_1 + \lambda_2)p = 2 + (\lambda_3\lambda_4 - \lambda_1\lambda_2)p^2 \equiv 2 \bmod p^2.$$

The case $a = d = 1$ contributes $O(\frac{T}{p} + 1)$ possibilities. This is since b or c must vanish and $|b|, |c| \ll T$ as well as $p \mid (b, c)$.

The case $|a| > 1, d = 1$ or $a = 1, |d| > 1$ contributes at most $O(\frac{T^{1+\epsilon}}{p^2})$. This is since there are $O(\frac{T}{p^2})$ choices for a or d due to (14) and since $bc = 1 - ad$ we have $O(T^\epsilon)$ possibilities for (b, c).

Finally, if $|a|, |d| > 1$, then we choose a in $O(\frac{T}{p})$ ways and $a+d$ in $O(\frac{T}{p^2})$ ways. This determines (a,d) and we conclude by observing that there are $O(T^{2\epsilon})$ choices for (b,c) due to the divisor bound. □

Using spectral theory different types of lattice point counting results can be obtained.

Theorem 5.2. *Let Γ be a finitely generated subgroup of $\mathrm{SL}_2(\mathbb{Z})$ with $\delta_\Gamma > \frac{1}{2}$. Let $q_0 \in \mathbb{N}$ be given so that $\Gamma/\Gamma(q) \cong \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ for all square-free q with $(q, q_0) = 1$. (Such a q_0 exists by the strong approximation combined with Gorusat's Lemma.) There is $\epsilon > 0$ depending on Γ, so that for any square-free q with $(q, q_0) = 1$ and any $g \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ we have*

$$\sharp\{\gamma \in \Gamma \colon \|\gamma\| \leq T \text{ and } \gamma \equiv g \bmod q\} = \frac{c_\Gamma T^{2\delta}}{\sharp\,\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} + O(q^3 T^{2\delta-\epsilon_1}).$$

Proof. The starting point is the observation

$$\|g\|^2 = a^2 + b^2 + c^2 + d^2 = 4u(g.i, i) + 2 \tag{15}$$

Recall that $\cosh(d(z,w)) = 1 + 2u(z,w)$. Thus we define the more general counting problem

$$N_\Gamma(X; z, w) = \sharp\{\gamma \in \Gamma \colon d(z, \gamma w) \leq X\}.$$

A spectral argument due to Lax and Phillips yields

$$|N_{\Gamma(q)}(X; z, w) - \sum_j c_j \varphi_{j,q}(z)\overline{\varphi_{j,q}(w)} e^{s_{j,q}X}| = O(q^3 X^{\frac{5}{6}} e^{X/2}).$$

Here we are summing over the exceptional part of the spectrum $\lambda_j(\Gamma(q)) = s_{j,q}(1 - s_{j,q})$ with associated eigenfunctions $\varphi_{j,q}$. Note that $s_{0,q} = \delta$.

Note that the L^2-normalization of the ground state implies

$$\varphi_{0,q} = \frac{1}{\sharp\,\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}\varphi_{0,1}.$$

Inserting the spectral gap and translating X to T completes this proof (sketch). □

The method of Lax-Phillips to prove lattice point counting results using spectral theorem fails for Γ with $\delta_\Gamma \leq \frac{1}{2}$. In these cases one has to use a different approach.

Theorem 5.3 (Bourgain-Gamburd-Sarnak 2011)**.** *Let Γ be a finitely generated subgroup of $\mathrm{SL}_2(\mathbb{Z})$ with $0 < \delta_\Gamma \leq \frac{1}{2}$. Let q_0 be the ramified modulus coming from strong approximation. There are $\epsilon_1 > 0$ and $C > 0$ depending on Γ so that for square-free q with $(q, q_0) = 1$ and any $g \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ we have*

$$\sharp\{\gamma \in \Gamma \cap B_T \colon \gamma \equiv g \bmod q\} = \frac{c_\Gamma T^{2\delta}}{\sharp\,\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}\left(1 + O(e^{-c\sqrt{\log(T)}})\right) + O(q^C T^{2\delta-\epsilon_1}).$$

To give a proof here would go beyond the scope of these notes. As mentioned above spectral theory can not be used directly here. Instead one rewrites the counting problem as a renewal equation and uses dynamical methods. More precisely one encounters the (dynamical) Ruelle zeta function, which is closely linked to the transfer operator. The resonances appear as zeros of this zeta function. Note that expansion is still crucial for the proof to work.

We will later need a version of this theorem for the semigroup Γ_A (or even $\Gamma_{\mathfrak{A}}$). However, since $\delta < \frac{1}{2}$, the group Γ can not contain parabolic elements. This is crucial for the proof. It turns out, that, since Γ_A satisfies this as well, the proof can be easily adapted to that situation.

Later we will need another basic counting result for $\Gamma = \mathrm{SL}_2(\mathbb{Z})$. Since this seems to be the appropriate place we state them now.

Lemma 5.4. *Let $X \gg 1$. There exists a function*

$$\varphi_X \colon \mathrm{SL}_2(\mathbb{R}) \to \mathbb{R}_{\geq 0}$$

which approximates the indicator function of an archimedean ball. More precisely

$$\varphi_X(g) \geq 1 \text{ for } \|g\| \leq X$$

and

$$\sum_{\xi \in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\xi) \ll X^2.$$

Furthermore, for any $y \geq 1$, and any $\gamma_0 \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ we have

$$\sum_{\substack{\xi \in \mathrm{SL}_2(\mathbb{Z}),\\ \xi \equiv \gamma_0 \ mod\ q}} \varphi_X(\xi) = \frac{1}{\sharp\, \mathrm{SL}_2(Z/q\mathbb{Z})} \sum_{\xi \in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\xi) + O(X^{\frac{3}{2}}).$$

Proof. This can be deduced easily using spectral theory using Selberg's 3/16-theorem to control the exceptional eigenvalues. □

5.2. **Spectral Gap.** We now proof Gamburd's $\frac{5}{6}$-Theorem, which features a uniform spectral gap for a family of principal congruence subgroups $\Gamma(p) \subset \Gamma$ for p sufficiently large under the assumption $\delta > \frac{5}{6}$. If Γ is given this spectral gap is in some sense quantitative.

Theorem 5.5 (Gamburd)**.** *Let $\Gamma = \langle A_1, \ldots, A_k \rangle$ be a finitely generated subgroup of $\mathrm{SL}_2(\mathbb{Z})$ with $\delta > \frac{5}{6}$. Let $\mathcal{F}(p) = \Gamma(p)\backslash\mathbb{H}$. For p large enough we have*

$$\Omega(\mathcal{F}(p)) \cap [\delta(1-\delta), \frac{5}{36}) = \Omega(\mathcal{F}(1)) \cap [\delta(1-\delta), \frac{5}{36}).$$

In particular $\Omega(\mathcal{F}(p))$ has a spectral gap, that is for p large

$$\lambda_1(\mathcal{F}(p)) \geq \min(\lambda_1(\mathcal{F}(1)), \frac{5}{36}).$$

Proof. Take p large enough so that $\Gamma/\Gamma(p) \cong \mathrm{SL}_2(\mathbb{F}_p)$. Further assume that

$$\Omega(\mathcal{F}(p)) \cap [\delta(1-\delta), \frac{5}{36}) \neq \Omega(\mathcal{F}(1)) \cap [\delta(1-\delta), \frac{5}{36}).$$

If this is the case we must have a new discrete eigenvalue λ in $(\delta(1-\delta), \frac{5}{36})$. Let V_λ be the corresponding eigenspace. Since the Laplacian commutes with deck transformations (i.e. Γ) V_λ must contain a non-trivial irreducible representation of $\mathrm{SL}_2(\mathbb{F}_p)$. (This is a theorem due to Frobenius.) Since any non-trivial representation of $\mathrm{SL}_2(\mathbb{F}_p)$ has dimension at least $\frac{p-1}{2}$ we find

$$m(\lambda, \mathcal{F}(p)) \geq \frac{p-1}{2}.$$

We will now use spectral theory to give an upper bound for the multiplicity and thus obtain a contradiction.

Recall the point pair invariant (9):

$$k_1(z,w) = \mathbb{1}_{u(z,w) \leq \frac{X-2}{4}}$$

and $k = k_1 \star \overline{k_1}$. The corresponding automorphic kernel is denoted by $K(z,w) = K_X(z,w)$. Let $\mathcal{K}_p$ be the compact part of $\mathcal{F}(p)$. Then we have

$$\begin{aligned}
\int_{\mathcal{K}_p} K(z,z)d\mu(z) &= \sum_{\gamma \in \Gamma(p)} \int_{\mathcal{K}_p} k(z, \gamma z) d\mu(z) \\
&= \sum_{\gamma \in \Gamma(p)} \sum_{\delta \in \Gamma/\Gamma(p)} \int_{\mathcal{K}_1} k(\delta^{-1}z, \gamma\delta^{-1}z) d\mu(z) \\
&\ll p^3 \sum_{\gamma \in \Gamma(p)} \int_{\mathcal{K}_1} k(z, \gamma z) d\mu(z) \\
&\ll p^3 \sum_{\gamma \in K(p)} \int_{\mathcal{K}_1} k(z, \gamma z) d\mu(z).
\end{aligned}$$

As noted in (10) the kernel K_1 evaluated at (i,i) counts exactly $N_1(\sqrt{X}, \Gamma(p))$. This generalizes to

$$K_1(z,z) = \sharp \left\{ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(p) \colon f_z(\gamma) = f_z(a,b,c,d) \leq \frac{X-1}{2} \right\},$$

for some homogeneous positive-definite quadratic form in a, b, c, d. The dependence on z is continuous. Thus, since $\mathcal{K}_1$ is compact there is σ, so that

$$\frac{\|\gamma\|^2}{\sigma} \leq f_z(\gamma) \leq \sigma \|\gamma\|^2,$$

for all $z \in \mathcal{K}_1$. We get

$$K_{1,\frac{X}{\sigma}}(i,i) \leq K_{1,X}(z,z) \leq K_{1,\sigma X}.$$

Recall that $\cosh(T) = 2+4X$. Thus by Proposition 4.22 together with our counting result we get

$$\begin{aligned}\sum_{\gamma\in K(p)}\int_{\mathcal{K}_1} k(z,\gamma z)d\mu(z) &\ll e^T\int_{\mathcal{K}_1}\sum_{\gamma\in K(p)} e^{-\frac{\rho(z,\gamma z)}{2}}\delta_{\rho(z,\gamma z)\leq 2T}d\mu(z)\\ &\ll X\sup_{z\in\mathcal{K}_1}\sum_{\gamma\in K(p)}\frac{1}{\sqrt{u(z,\gamma z)}}\delta_{u(z,\gamma z)\leq X^2}\\ &\ll X\sum_{\substack{l\ll X^2,\\ \text{dyadic}}}\frac{N_1(\sqrt{l},K(p))}{\sqrt{l}} \ll X\sum_{\substack{l\ll X^2,\\ \text{dyadic}}}[\frac{\sqrt{l}}{p^3}+1]\\ &\ll \frac{X^{2+\epsilon}}{p^3}+X^{1+\epsilon}.\end{aligned}$$

This gives

$$\int_{\mathcal{K}_p} K(z,z)d\mu(z) \ll X^{2+\epsilon}+p^3X^{1+\epsilon}.$$

On the other hand we can lower bound the automorphic kernel using the pre-trace inequality:

$$\int_{\mathcal{K}_p} K(z,z)d\mu(z) \geq \int_{\mathcal{K}_p}\sum_{\lambda_{j,p}<\frac{1}{4}} X^{2s_{j,p}}|\phi_{j,p}(z)|^2d\mu(z).$$

To further estimate this expression we need to take care of the cusps and funnels of $\mathcal{F}(p)$. Indeed by concretely estimating the integrals of $\phi_{j,p}$ over the regions around cusps and funnels one can show[12]

$$\int_{\mathcal{K}_p}|\phi_{j,p}(z)|^2d\mu(z) \gg_\Gamma \int_{\mathcal{F}(p)}|\phi_{j,p}(z)|^2d\mu(z).$$

We get

$$\begin{aligned}\int_{\mathcal{K}_p}\sum_{\lambda_{j,p}<\frac{1}{4}} X^{2s_{j,p}}|\phi_{j,p}(z)|^2d\mu(z) &\geq C\sum_{\lambda_{j,p}<\frac{1}{4}} X^{2s_{j,p}}\int_{\mathcal{F}(p)}|\phi_{j,p}(z)|^2d\mu(z)\\ &\ll \sum_{\lambda_{j,p}<\frac{1}{4}} X^{2s_{j,p}}\cdot m(\lambda_{j,p}).\end{aligned}$$

Dropping everything but the contribution of one eigenvalue $0\leq\lambda<\frac{1}{4}$ we get

$$m(\lambda,p) \ll X^{2(1-s)+\epsilon}+p^3X^{1-2s+\epsilon}.$$

Choosing $X\asymp p^3$ yields

$$m(\lambda,p) \ll p^{6(1-s)}.$$

[12]This is technically the most difficult part of the proof, which we omit!

Combining this with the lower bound for the multiplicity we get (for p large enough)

$$6(1-s) > 1.$$

This translates into $s < \frac{5}{6}$ or $\lambda > \frac{5}{36}$. □

Remark 5.6. The proof can be generalized to give the following stronger result. Let Γ be a Fuchsian group of the second kind with $\delta > \frac{5}{6}$. Then there is $N = N(\Gamma)$ so that

$$\Omega(\mathcal{F}(p)) \cap [\delta(1-\delta), \frac{5}{36}) \subset \Omega(\mathcal{F}(q')),$$

for all $q \in \mathbb{N}$ with $q' = (q, N)$.

Next we give a generalization of this result to all $\frac{1}{2} < \delta$. However, the result is slightly weaker.

Theorem 5.7 (Bourgain-Gamburd-Sarnak 2011)**.** *Let Γ be a finitely generated subgroup of $\mathrm{SL}_2(\mathbb{Z})$ with $\delta > \frac{1}{2}$. There is an $\epsilon = \epsilon(\Gamma)$ so that*

$$\lambda_1(\Gamma(q)) \geq \lambda_0(\Gamma(q)) + \epsilon$$

for all square-free $q \geq 1$.

Proof. The proof can be found in [3, Section 2]. In contrast to what we have seen above, to prove this theorem the expansion properties of the associated Cayley graphs are established using combinatorial techniques. The spectral gap then follows from Theorem 5.10 below. □

We turn towards the case $\delta \leq \frac{1}{2}$. Note that for this to be the case Γ can not contain parabolic elements. This is a key property in what follows. Indeed, the same methods work for the semigroup Γ_A, precisely since the latter does not contain any parabolic elements.

Recall that the resolvent was the meromorphic continuation of $R_X(s) = (\Delta_X - s(1-s))$ for $X = \Gamma\backslash\mathbb{H}$. In the absence of exceptional eigenvalues we have the spectral gap is replaced by a **resonance free region**. Recall that $R_X(s)$ has a simple pole at $s = \delta$ and no further poles on the line $\mathrm{Re}(s) = \delta$. This resonance replaces the base eigenvalue. One has the following result

Theorem 5.8 (Bourgain-Gamburd-Sarnak 2011)**.** *Let Γ be a finitely generated subgroup of $\mathrm{SL}_2(\mathbb{Z})$ with $\delta_\Gamma \leq \frac{1}{2}$. For $q \geq 1$ square-free we write $X(q) = \Gamma(q)\backslash\mathbb{H}$. There is $\epsilon = \epsilon(\Gamma) > 0$ such that $R_{X(q)}(s)$ is holomorphic, with exception of a simple pole at $\delta = s$, for*

$$\mathrm{Re}(s) > \delta - \epsilon \min(1, \frac{1}{\log(1 + \mathrm{Im}(s))}).$$

This is also proved using dynmaical methods and we will not give any details. Let us just mention, that the resonances can be interpreted as zeros of the Selberg zeta function. The latter turns out to be the dynamical zeta function that we

already mentioned in the lattice point counting section. The result above is thus equivalent to a zero-free region of this zeta function. Again expansion plays a huge role in establishing this zero-free region.

5.3. **Expansion.**

Theorem 5.9. *Let $S = \{A_1, \dots, A_k\}$ be a symmetric set of generators in $\mathrm{SL}_2(\mathbb{Z})$ and let $\Gamma = \langle A_1, \dots, A_k \rangle$. If the Hausdorff dimension of the limit set $\delta(\Gamma) > \frac{5}{6}$, then $\mathcal{G}_p = G(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), S)$ is a family of expanders.*

Proof. We start by reformulating our spectral gap theorem on the group level. Write

$$L^2(\Gamma(p)\backslash \mathrm{SL}_2(\mathbb{R})) = V_{\varphi_0} \oplus L^2_0(\Gamma(p)\backslash \mathrm{SL}_2(\mathbb{R}))$$

as (unitary) $\mathrm{SL}_2(\mathbb{R})$ representations. (Here φ_0 is the ground state with eigenvalue $\lambda_0(\Gamma)$.) Then we define

$$R = \bigcup_{p \gg 1} \{\pi \in \widehat{\mathrm{SL}_2(\mathbb{R})} \colon \mathrm{Hom}(\pi, L^2_0(\Gamma(p)\backslash \mathrm{SL}_2(\mathbb{R}))) \neq \{0\}\}.$$

When $\widehat{\mathrm{SL}_2(\mathbb{R})}$ is equipped with the Fell topology, then the spectral gap tells us that V_{φ_0} is not contained in $\overline{R}$.

We will now use Fell's continuity of induction to deduce the trivial representation ρ_0 of Γ is isolated from certain other Γ-modules. We define

$$H(p) = L^2(\mathrm{SL}_2(\mathbb{F}_p)) = 1 \oplus H_0(p). \tag{16}$$

In particular,

$$H_0(p) = \{f \in H(p) \colon \langle f, 1 \rangle = \sum_x f(x) = 0\}.$$

Note that Γ acts on $H(p)$ and for p sufficiently large the decomposition in (16) is true in the sense of Γ-modules. (Where $\gamma f(x) = f(x\gamma)$.)

By induction in stages we have

$$\mathrm{Ind}_{\Gamma(p)}^{\mathrm{SL}_2(\mathbb{R})} 1 = \mathrm{Ind}_{\Gamma}^{\mathrm{SL}_2(\mathbb{R})} \mathrm{Ind}_{\Gamma(p)}^{\Gamma} 1 = \mathrm{Ind}_{\Gamma}^{\mathrm{SL}_2(\mathbb{R})} \rho_0 \oplus \mathrm{Ind}_{\Gamma}^{\mathrm{SL}_2(\mathbb{R})} H_0(p).$$

Note that $\pi_{\lambda_0(\Gamma)} \subseteq \mathrm{Ind}_{\Gamma}^{\mathrm{SL}_2(\mathbb{R})} \rho_0$. Let

$$T = \bigcup_{p \gg 1} H_0(p) \subseteq \widehat{\Gamma}.$$

Then ρ_0 is isolated with respect to T. (Otherwise one gets a contradiction to our spectral gap for $\Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ using that $\tau_j \to \rho_0$ implies that $\mathrm{Ind}_{\Gamma}^{\mathrm{SL}_2(\mathbb{R})} \tau_j \to \mathrm{Ind}_{\Gamma}^{\mathrm{SL}_2(\mathbb{R})}(\rho_0) \supseteq \pi_{\lambda_0(\Gamma)}$.)

Recall that from the definition of the Fell topology it follows that there is $\epsilon > 0$ depending only on Γ and S such that for all $f \in H_0(p)$ we have

$$\|\gamma f - f\| > \epsilon \|f\|$$

for some $\gamma \in S$.

From here it is easy to conclude. Let A be a subset of $V_p = \mathrm{SL}_2(\mathbb{F}_p)$ of size a. Denote the compliment of A by B and write $b = n - a$, where $n = \sharp V_p$. Define

$$f(x) = \begin{cases} b & \text{if } x \in A, \\ -a & \text{if } x \in B. \end{cases}$$

Obviously $f \in H_0(p)$. We have

$$\|f\|^2 = nab.$$

Let

$$E_\gamma(A, B) = \{x \in V | x \in A \text{ and } x\gamma \in B \text{ or } x \in B \text{ and } x\gamma \in A\}.$$

Then

$$\|\gamma f - f\|^2 = n^2 \cdot \sharp E_\gamma(A, B).$$

Now there is $\gamma \in S$ so that

$$\sharp N(A) \geq \frac{1}{2}\sharp E_\gamma(A, B) = \frac{\|\gamma f - f\|^2}{2n^2} \geq \frac{\epsilon^2 \|F\|^2}{2n^2} = \epsilon^2 \frac{ab}{2n} = \frac{\epsilon^2}{2}\left(1 - \frac{\sharp A}{n}\right) \cdot \sharp A.$$

Since ϵ does not depend on p we are done. □

This theorem has a direct generalization.

Theorem 5.10. *Let $\Gamma = \langle S \rangle$ be a finitely generated subgroup of $\mathrm{SL}_2(\mathbb{R})$ with $\delta(\Lambda) > \frac{1}{2}$. Further let N_j be a family of finite-index normal subgroups of Γ. Then the following are equivalent:*

1. *The Cayley graphs $G(\Gamma/N_j, S)$ form a family of expanders;*
2. *There is $\epsilon = \epsilon(\Gamma) > 0$ such that $\lambda_1(N_j) \geq \lambda_0(N_j) + \epsilon$.*

Proof. The implication (2) $\implies$ (1) is a direct extension of the argument featured above.

We turn towards (1) $\implies$ (2). Let us set some notation. Write $V_j = L^2(\Gamma/N_j)$ equipped with the right regular representation R_j. Then we consider

$$H_j = \{F \colon \mathbb{H} \to V_j \colon F(\gamma z) = R_q(\gamma)F(z) \text{ for all } \gamma \in \Gamma\}.$$

Let φ_0 be the ground state (i.e. the eigenfunction corresponding to the bottom of the spectrum) with eigenvalue $\lambda_0(\Gamma)$. Let $H_{0,j}$ be the subspace of H_j orthogonal to $\varphi_0 \otimes \mathrm{Id}$. We need to show that

$$\int_{\mathcal{F}_1} \|\nabla F\|^2 d\mu \geq (\lambda_0 + c) \int_{\mathcal{F}_1} \|F\|^2 d\mu,$$

where $\mathcal{F}_1$ is a fundamental domain for Γ.

Recall that by the expansion property we have $\epsilon > 0$ depending only on S such that for all $F \in H_0(q)$ we have

$$\|F(\gamma z) - F(z)\| \geq \epsilon \|F(z)\| \tag{17}$$

for some $\gamma \in S$.

Write $f = \|F\|$ and decompose it as

$$f(z) = a\varphi_0(z) + b(z).$$

Here

$$\int_{\mathcal{F}_1} \varphi_0(z)\overline{b(z)}d\mu = 0 \text{ and } \int_{\mathcal{F}_1} |f(z)|^2 d\mu = a^2 + \int_{\mathcal{F}_1} |b(z)|^2 d\mu = 1.$$

Further write $F(z) = (F_1(z), \dots, F_{k_j}(z))$, where $k_j = \sharp\Gamma/N_j$.

One computes that

$$\|\nabla F\|^2(z) \geq |\nabla f|^2(z).$$

We obtain

$$\begin{aligned}\frac{\int_{\mathcal{F}_1} \|\nabla F\|^2 d\mu}{\int_{\mathcal{F}_1} \|F\|^2 d\mu} &\geq \frac{\int_{\mathcal{F}_1} \langle \Delta f, f\rangle d\mu}{\int_{\mathcal{F}_1} |f(z)|^2 d\mu} \\ &= \int_{\mathcal{F}_1} \langle a\lambda_0\varphi_0 + \Delta b, a\varphi_0 + b\rangle = a^2\lambda_0 + \langle \Delta b, b\rangle \\ &\geq \lambda_0 + (\lambda_1 - \lambda_0) \int_{\mathcal{F}_1} |b(z)| d\mu.\end{aligned}$$

Since there are only finitely many discrete eigenvalues we have $\lambda_1 - \lambda_0 = c_1(\Gamma) > 0$. If we have $\int_{\mathcal{F}_1} |b(z)|^2 d\mu > \epsilon_1 > 0$, then we are done.

Let us consider the critical case when $\int_{\mathcal{F}_1} |b(z)|^2 d\mu = 0$. in this case we can assume that $a = 1$ and we write

$$F(z) = u(z)\varphi_0(z) \text{ with } \|u(z)\|^2 = \sum_{i=1}^{k_j} |u_i(z)|^2 = 1.$$

Computing with derivatives yields

$$\|\nabla\varphi_0 u\|^2 = |\nabla\varphi_0|^2 + \varphi_0^2 \|\nabla u\|^2.$$

This implies

$$\frac{\int_{\mathcal{F}_1} \|\nabla F\|^2 d\mu}{\int_{\mathcal{F}_1} \|F\|^2 d\mu} \geq \lambda_0 + \frac{\int_{\mathcal{F}_1} \varphi_0(z)^2 \|\nabla u\|^2 d\mu}{\int_{\mathcal{F}_1} |\varphi_0(z)| d\mu}.$$

Our goal is to show that the remaining quotient is bounded from below.

Assume that

$$\frac{\int_{\mathcal{F}_1} \varphi_0(z)^2 \|\nabla u\|^2 d\mu}{\int_{\mathcal{F}_1} |\varphi_0(z)| d\mu} < \kappa.$$

Our goal is to reach a contradiction for small κ. By some foliation argument (applied after savely removing regions around funnels and cusps) one obtains

$$\int_{B(z,\delta)} \varphi_0(z)^2 \|u(\gamma z) - u(z)\| d\mu(z) < \kappa \int_{B(z,\delta)} \varphi_0(z)^2 d\mu,$$

for all $\gamma \in S$. But by (17) we reach a contradiction for some $\gamma \in S$. This completes this sketch of a proof. □

In general one has the following result

Theorem 5.11 (Bourgain-Gamburd-Sarnak 2010)**.** *Let $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$ be finitely generated and Zariski dense SL_2. Let S be a finite symmetric set of generators of Γ. Then, for squarefree q the graphs $G(\Gamma/\Gamma(q), S)$ form an expander family.*

Proof. A key technical input is a sum-product estimate for subsets $A \subset \mathbb{Z}/q\mathbb{Z}$. The proof can be found in [2]. □

6. Counting results for the semi-group Γ_A

For thin groups Γ with $\delta \leq \frac{1}{2}$ (i.p. there are no parabolic elements in Γ) the counting result Theorem 5.3 is a key ingredient for the execution of the affine sieve. This has been discussed in [3, Theorem 1.6 and 1.7]. For the applications to Zaremba's conjecture we will need similar counting results for the semi-group Γ_A. In this section we state the relevant results and make a couple of remarks concerning the proofs.

Theorem 6.1. *Recall the definition of the semigroup Γ_A and take $A = 2$. There is an absolute square-free integer $\mathfrak{B} \geq 1$, absolute constants $c, C > 0$ and an absolute spectral gap $\Theta > 0$ so that for any square free $q \equiv 0 \bmod \mathfrak{B}$ and any $\omega \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ we have*

$$\sharp\{\gamma \in \Gamma_2 \cap B_Y : \gamma \equiv \omega \bmod q\} = \frac{\sharp \mathrm{SL}_2(\mathbb{Z}/\mathfrak{B}\mathbb{Z})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \cdot \sharp\{\gamma \in \Gamma_2 \cap B_Y : \gamma \equiv \omega \bmod \mathfrak{B}\} + O(\sharp\{\gamma \in \Gamma_2 : \|\gamma\| < Y\} \cdot \mathfrak{G}(Y; q)),$$

where

$$\mathfrak{G}(Y; q) = \begin{cases} e^{-c\sqrt{\log(Y)}} & \text{if } q \leq C \log(Y), \\ q^C Y^{-\Theta} & \text{if } q > C \log(Y). \end{cases}$$

Proof. Note that Γ_2 is free and every non-identity element is hyperbolic. For this reason the proof of Theorem 5.3 directly generalizes to this setting. (The error term is a consequence of a zero-free region for the congruence transfer operator coupled with a Tauberian theorem. One should note the similarity to the typical prime number theorem error term.) □

This result will be used in the proof of Theorem 1.12 to construct a set Π in which it taylored for the execution of a sieving argument. In this context the restriction to square-free moduli q is no obstacle. However, in the application to Zaremba's conjecture (Theorem 1.4) we will need a similar result to construct a set which is suitable for the application of the (orbital) circle method. In this case it is crucial that all moduli q are allowed.

Theorem 6.2 (Theorem 8.1, [5]). *Let $\Gamma = \Gamma_{\mathfrak{A}}$ be the usual semigroup. There exists an integer $\mathfrak{B} = \mathfrak{B}(\mathfrak{A}) \geq 1$ and a constant $\mathfrak{c} = \mathfrak{c}(\mathfrak{A}) > 0$ so that the following holds. For any $(q, \mathfrak{B}) = 1$, any $w \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$, any $\gamma_0 \in \Gamma$ and parameters $T, H \to \infty$ with $H < e^{c\sqrt{\log(T)}}$, there is a constant $C(\gamma_0) > 0$ so that*

$$\sharp\{\gamma \in \Gamma \colon \gamma \equiv w \bmod q,\ |v_+(\gamma) - \mathfrak{v}| < H^{-1} \text{ and } \frac{\|\gamma\gamma_0\|}{\|\gamma\|} \leq T\}$$
$$= C(\gamma_0) T^{2\delta} \frac{\mu(\mathcal{I})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} + O(T^{2\delta} e^{-\mathfrak{c}\sqrt{\log(T)}}).$$

Here $\mathcal{I}$ is the interval of length H^{-1} about $\mathfrak{v}$ and the implied constant does not depend on T, H, q, w or γ_0.

Similarly, if $\mathfrak{B} \mid q$ we have

$$\sharp\{\gamma \in \Gamma \colon \gamma \equiv w \bmod q,\ v_+(\gamma) \in \mathcal{I} \text{ and } \frac{\|\gamma\gamma_0\|}{\|\gamma\|} \leq T\}$$
$$= \frac{\sharp \mathrm{SL}_2(\mathfrak{O})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \sharp\{\gamma \in \Gamma \colon \gamma \equiv w \bmod \mathfrak{B}, v_+(\gamma) \in \mathcal{I} \text{ and } \frac{\|\gamma\gamma_0\|}{\|\gamma\|} \leq T\}$$
$$+ O(T^{2\delta} e^{-\mathfrak{c}\sqrt{\log(T)}}).$$

Proof. Again the methods from the proof of Theorem 5.3 carry over to this setting with not much modification. (This is because $\Gamma_{\mathfrak{A}}$ is a free semigroup with only hyperbolic elements.) The bottleneck is that the expansion property from Theorem 5.11 only holds for square-free q. Luckily the relevant results to lift the square-free assumption are known by now. This allows one to prove the theorem as stated (and also Theorem 5.3 holds for arbitrary q).

Finally, note that the condition $\frac{\|\gamma\gamma_0\|}{\|\gamma_0\|} \leq T$ can (essentially) be phrased as

$$d(\gamma\gamma_0 i, i) - d(\gamma_0 i, i) < C \log(T).$$

The latter expression appears naturally in the dynamical approach to Theorem 5.3. □

What we will actually apply later is a direct corollary of this result. Unfortunately we need some notation that is only introduced in Section 9 below to properly state the result:

Corollary 6.3. *With the notation as in Theorem 6.2 we have for any $T, H, H_1 \to \infty$ with $H_1 = o(H)$ and $H < e^{c\sqrt{\log(T)}}$ and any $(q, \mathfrak{B}) = 1$, $w \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ that*

$$\sharp\{\gamma \in \Gamma \colon \gamma \equiv w \bmod q,\ |v_+(\gamma) - \mathfrak{v}| < H^{-1} \text{ and } |\lambda(\gamma) - T| < \frac{T}{H_1}\}$$
$$= C(\mathfrak{v}) T^{2\delta} \frac{\mu(\mathcal{I})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} (1 + O(H_1^{-1} + \frac{H_1}{H})) + O(T^{2\delta} e^{-\mathfrak{c}\sqrt{\log(T)}}).$$

Here $\mathcal{I}$ is the interval of length H^{-1} about $\mathfrak{v}$ and the implied constant does not depend on T, H, H_1 q or w.

Similarly, if $\mathfrak{B} \mid q$ we have

$$\sharp\{\gamma \in \Gamma \colon \gamma \equiv w \bmod q,\, v_+(\gamma) \in \mathcal{I} \text{ and } |\lambda(\gamma) - T| < \frac{T}{H_1}\}$$
$$= \frac{\sharp \mathrm{SL}_2(\mathfrak{Q})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \sharp\{\gamma \in \Gamma \colon \gamma \equiv w \bmod \mathfrak{B}, v_+(\gamma) \in \mathcal{I} \text{ and } |\lambda(\gamma) - T| < \frac{T}{H_1}\}$$
$$+ O(T^{2\delta} e^{-\mathfrak{c}\sqrt{\log(T)}}).$$

Proof. Later we will see that

$$\|\gamma\| = \frac{\lambda(\gamma)}{|\langle v_+(\gamma), v_-^{\perp}(\gamma)\rangle|}(1 + O(\|\gamma\|^{-2})).$$

Suppose γ, γ_0 both lie in $\mathcal{I}$ (an interval of length H^{-1} about $\mathfrak{v}$) and that $\|\gamma\|, \|\gamma_0\| > H$. Then the above yields

$$\|\gamma_0\| = \frac{\lambda(\gamma_0)}{|\langle \mathfrak{v}, v_-^{\perp}(\gamma_0)\rangle|}(1 + O(H^{-1}))$$

and

$$\|\gamma\gamma_0\| = \frac{\lambda(\gamma)\lambda(\gamma_0)}{|\langle \mathfrak{v}, v_-^{\perp}(\gamma_0)\rangle|}(1 + O(H)).$$

This leads to

$$\frac{\|\gamma\gamma_0\|}{\|\gamma_0\|} = \lambda(\gamma)(1 + O(H^{-1})).$$

Now we observe that $C(\gamma_0)$ approaches a constant $C(\mathfrak{v})$ as $v_+(\gamma_0) \to \mathfrak{v}$ and $\|\gamma_0\| \to \infty$. □

To implement this corollary we have to combine it with a (standard) randomness extraction argument. In absence of a better place we record the relevant lemma here.

Lemma 6.4. *Let $\mu = \mu_S$ be a probability measure of finite subset $S \subset \mathrm{SL}_2(\mathbb{Z})$:*

$$\mu(\gamma) = \frac{1}{\sharp S} \sum_{s \in S} \mathbb{1} s = \gamma.$$

Fix $\eta > 0$, let $q_0 < Q$ be a fixed modulus, fix $w_0 \in \mathrm{SL}_2(\mathbb{Z}/q_0\mathbb{Z})$ and let $\mathfrak{Q} = \mathfrak{Q}_{q_0} \subset [1, Q]$ be the set of moduli $q < Q$ with $q_0 \mid q$. Assume that for all $q \in \mathfrak{Q}$ and all $w \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ with $w \equiv w_0 \bmod q_0$ the projection

$$\pi_q[\mu](w) = \sum_{\gamma \equiv w \bmod q} \mu(\gamma)$$

is near the uniform measure on $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ *conditioned on being* $\equiv w_0 \ mod\ q_0$,

$$\left\| \pi_q[\mu] - \frac{\sharp \mathrm{SL}_2(\mathbb{Z}/q_0\mathbb{Z})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \right\|_{L^\infty|_{\equiv w_0 \ mod\ q_0}} = \max_{\substack{w\in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z}),\\ w\equiv w_0 \ mod\ q_0}} |\pi_q[\mu](w) - \frac{\mathrm{SL}_2(\mathbb{Z}/q_0\mathbb{Z})}{\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}| < \eta.$$

Then for any T *with*

$$\eta - 2\log(Q) < T = o(\sqrt{\sharp S})$$

there exist T *distinct points* $\gamma_1, \ldots, \gamma_T \in S$ *such that the probability measure* $\nu = \nu_{T,\gamma_1,\ldots,\gamma_T}$ *defined by*

$$\nu = \frac{1}{T}(\delta_{\gamma_1} + \ldots + \delta_{\gamma_T})$$

has the same property. That is, for all $q \in \mathfrak{Q}$ *the projection* $\pi_q\nu$ *is also nearly uniform:*

$$\max_{q\in\mathfrak{Q}} \left(\left\| \pi_q[\nu] - \frac{\sharp \mathrm{SL}_2(\mathbb{Z}/q_0\mathbb{Z})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \right\|_{L^\infty|_{\equiv w_0 \ mod\ q_0}} \right) \ll \eta.$$

Proof. We define

$$\mathcal{D} = \sum_{\gamma\in \mathrm{SL}_2(\mathbb{Z})^T} \max_{q\in\mathfrak{Q}} \max_{\substack{w\in \mathrm{SL}_2(q),\\ w\equiv w_0 \bmod q_0}} |\frac{1}{T}\sum_{j=1}^{T} \mathbb{1}_{\gamma_j\equiv w \bmod q} - \frac{\sharp \mathrm{SL}_2(\mathbb{Z}/q_0\mathbb{Z})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}|\mu^{(T)}(\gamma),$$

where $\mu^{(T)}$ is the product measure on $\mathrm{SL}_2(\mathbb{Z})^T$ and $\gamma = (\gamma_1, \ldots, \gamma_T)$. This is the expectation with respect to μ of quantity we are aiming to estimate.

Using our assumptions gives the bound

$$\mathcal{D} < \eta + \sum_{\gamma\in \mathrm{SL}_2(\mathbb{Z})^T} \sum_{\xi\in \mathrm{SL}_2(\mathbb{Z})^T} \max_{q\in\mathfrak{Q}} \max_{w\in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} |\frac{1}{T}\sum_{j=1}^{T} f_w(\gamma_j,\xi_j)|\mu^{(T)}(\gamma)\mu(T)(\xi),$$

for

$$f_w(\gamma_j,\xi_j) = \mathbb{1}_{\gamma_j\equiv w \bmod q} - \mathbb{1}_{\xi_j\equiv w \bmod q}.$$

Note that for fixed w, $f_w(\gamma_j,\xi_j)$ are independent mean zero random variables bounded by 1. The contraction principle gives

$$\mathcal{D} < \eta + \sum_{\gamma}\sum_{\xi} \mathcal{D}(\xi,\gamma)\mu^{(T)}(\gamma)\mu^{(T)}(\xi).$$

Here

$$\mathcal{D}(\gamma,\xi) = 2^{-T} \sum_{\epsilon\in\{\pm1\}^T} \max_{q\in\mathfrak{Q}} \max_{w\in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} |\frac{1}{T}\sum_{j=1}^{T} \epsilon_j f_w(\gamma_j,\xi_j)| \tag{18}$$

For a p chosen later we can apply Khintchine's inequality to get

$$\begin{aligned}\mathcal{D}(\gamma,\xi) &\leq 2^{-T}\sum_{\epsilon\in\{\pm 1\}^T}\left(\sum_{q\in\mathfrak{Q}}\sum_{w\in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}|\frac{1}{T}\sum_{j=1}^{T}\epsilon_j f_w(\gamma_j,\xi_j)|^p\right)^{\frac{1}{p}}\\ &\ll\left(\sum_{q\in\mathfrak{Q}}\sum_{w\in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}2^{-T}\sum_{\epsilon\in\{\pm 1\}^T}|\frac{1}{T}\sum_{j=1}^{T}\epsilon_j f_w(\gamma_j,\xi_j)|^p\right)^{\frac{1}{p}}\\ &\ll\left(\sum_{q\in\mathfrak{Q}}\sum_{w\in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}p^{\frac{p}{2}}\left(\sum_{j=1}^{T}\frac{|f_w(\gamma_j,\xi_j)|^2}{T^2}\right)^{\frac{p}{2}}\right)^{\frac{1}{p}}\\ &\ll Q^{\frac{4}{p}}p^{\frac{1}{2}}T^{-\frac{1}{2}}.\end{aligned}$$

We choose $p=\log(Q)$ and get

$$\mathcal{D}(\gamma,\xi)\ll\log(Q)^{\frac{1}{2}}T^{-\frac{1}{2}}.$$

For $T>\eta^{-2}\log(Q)$ we get

$$\mathcal{D}\ll\eta.$$

Note that the number of T-tuples γ with distinct entries is $\frac{(\sharp S)!}{(\sharp S-T)!}\asymp(\sharp S)^T$ for $T=o(\sqrt{\sharp S})$. This concludes the proof. □

7. Hyperbolic Sector Counting

In this section we supply several technical counting results that will be needed later on. We will mostly focus on the infinite volume case. In the classical case of finite co-volume such estimates were given for example by Good.

Define the Sobolev type norm

$$\mathcal{S}_{\infty,T}f=\max_{X\in\{0,X_1,X_2,X_3\}}\sup_{g\in G,\|g\|<T}|d\pi(X).f(g)|.$$

We start with some technical discussion. We start by working with a Fuchsian group $\Gamma\subset G=\mathrm{SL}_2(\mathbb{R})$ of the second type with $\delta>\frac{1}{2}$. As before we write $K=\mathrm{SO}_2(\mathbb{R})$. The point spectrum of the Laplacian acting on $L^2(\Gamma\backslash\mathbb{H})$ is

$$0<\delta(1-\delta)=\lambda_0<\lambda_1\leq\ldots\leq\lambda_N<\frac{1}{4}.$$

We write $\lambda_j=s_j(1-s_j)$ with $s_j>\frac{1}{2}$.

First we will (asymptotically) evaluate

$$\mathcal{N}(T)=\sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<T}}e^{2in\theta_1(\gamma)}e^{2ik\theta_2(\gamma)}.$$

This will be done spectrally as follows. For $g \in G$ define

$$f_T(g) = e^{2in\theta_1(g)} e^{2ik\theta_2(g)} \mathbb{1}_{\|g\|<T}.$$

The associated automorphic kernel $\mathcal{F}_T \colon \Gamma\backslash G \times \Gamma\backslash G \to \mathbb{C}$ given by

$$\mathcal{F}_T(g,h) = \sum_{\gamma\in\Gamma} f_T(g^{-1}\gamma h)$$

satisfies $\mathcal{F}_T(e,e) = \mathcal{N}(T)$.

Fix $\eta > 0$ (in terms of T) and choose a smooth function $\psi \colon \Gamma\backslash G \to \mathbb{R}$ with $\int_{\Gamma\backslash G} \psi = 1$ and compact support in a ball of radius η around $e \in G$. Consider the integral

$$\mathcal{H}(T) = \langle \mathcal{F}_T, \psi\otimes\psi\rangle = \int_{\Gamma\backslash G}\int_{\Gamma\backslash G} \mathcal{F}_T(g,h)\psi(g)\psi(h)dgdh.$$

Lemma 7.1. *We have*

$$\mathcal{H}(T) = \mathcal{N}(T) + O(\eta(1+|n|+|k|)T^{2\delta}).$$

Proof. We write $\mathcal{F}_T(g,h) = \mathcal{N}(T) + (\mathcal{F}_T(g,h) - \mathcal{F}_T(e,e))$. Inserting this in the definition of $\mathcal{H}(T)$ using that ψ has mass one we obtain $\mathcal{H}(T) = \mathcal{N}(T) + \mathcal{E}(T)$ with

$$\begin{aligned}\mathcal{E}(T) &= \int_{\Gamma\backslash G}\int_{\Gamma\backslash G} (\mathcal{F}_T(g,h) - \mathcal{F}_T(e,e))\psi(g)\psi(h)dgdh \\ &= \sum_{\gamma\in\Gamma}\int_{\mathfrak{G}...mma\backslash G}\int_{\Gamma\backslash G} (f_T(g^{-1}\gamma h) - f_T(\gamma))\psi(g)\psi(h)dgdh.\end{aligned}$$

We consider several cases exploiting the support properties of f_T and ψ:

- If $\|\gamma\| \geq \frac{T}{1-\eta}$, then $f_T(g^{-1}\gamma h)$ and $f_T(\gamma)$ vanish.
- If $\frac{T}{1+\eta} < \|\gamma\| \leq \frac{T}{1-\eta}$, then we have the trivial estimate
$$|f_T(g^{-1}\gamma h) - f_T(\gamma)| \leq 2.$$
- If $\|\gamma\| \leq \frac{T}{1+\eta}$, then $\|g^{-1}\gamma h\| < T$ and $\|\gamma\| < T$. Since $|e^{2in\theta_1(g^{-1}\gamma h)} - e^{2in\theta_1(\gamma)}| \ll |n|\eta$ we deduce that
$$|f_T(g^{-1}\gamma h) - f_T(\gamma)| \ll (|n|+|k|)\eta.$$

Combining these observations we get

$$\begin{aligned}\mathcal{E}(T) &\ll \sum_{\substack{\gamma\in\Gamma,\\ \frac{T}{1+\eta}<\|\gamma\|\leq\frac{T}{1-\eta}}} 1 + (|n|+|k|)\eta \sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|\leq\frac{T}{1+\eta}}} 1 \\ &\ll \eta T^{2\delta} + (|n|+|k|)\eta T^{2\delta}.\end{aligned}$$

Here we have used a standard counting argument due to Lax-Phillips. □

Lemma 7.2. *The inner product $\mathcal{H}(T)$ can be expressed as follows*

$$\mathcal{H}(T) = \int_G f_T(g)\langle \pi(g)\psi, \psi\rangle dg.$$

Proof. We insert the definition of $\mathcal{F}(T)$ into the definition of $\mathcal{H}(T)$. By unfolding we get

$$\mathcal{H}(T) = \int_{\Gamma\backslash G} \left(\int_G f_T(g)\psi(xy)dg\right)\psi(x)dx.$$

We conclude by interchanging integrals and recognizing the desired inner product. □

Next we decompose ψ into its Fourier series:

$$\psi(gk_\theta) = \sum_{m\in\mathbb{Z}} \psi_{2m}(g)e^{2im\theta}.$$

Inserting this above yields

$$\mathcal{H}(T) = \sum_{m,l}\int_G f_T(g)\langle \pi(g)\psi_{2m}, \psi_{2l}\rangle dg = \int_G f_T(g)\langle \pi(g)\psi_{-2k}, \psi_{-2n}\rangle dg.$$

To obtain the second equality we observed that the G-integral vanishes unless $m = -k$ and $l = -n$ by orthogonality.

At this point we expand the matrix coefficient $\langle \pi(g)\psi_{-2k}, \psi_{-2n}\rangle$. spectrally. To do so we recall that the first frequency is $\lambda_0 = \delta(1-\delta)$ and it comes with the corresponding eigenfunction φ_0. Let V be the closure of the G-span of ϕ_0. We obtain

$$\langle \pi(g)\psi_{-2k}, \psi_{-2n}\rangle = \underbrace{\langle \psi_{-2k}, v_{-2k}\rangle}_{=\langle \psi, v_{-2k}\rangle}\cdot\underbrace{\langle v_{-2n}, \psi_{-n}\rangle}_{=\langle v_{-2n},\psi\rangle}\cdot\langle \pi(g)v_{-2k}, v_{-2n}\rangle + \langle \pi(g)\psi^\perp_{-2k}, \psi^\perp_{-2n}\rangle.$$

We can now prove the following estimate

Lemma 7.3. *As $T\to\infty$ we have*

$$\mathcal{H}(T) = \langle \psi, v_{-2k}\rangle\langle v_{-2n}, \psi\rangle\int_0^{2\log(T)}\langle \pi(a_t)v_{-2k}, v_{-2n}\rangle \sinh(t)dt + O(\|\psi\|_2^2 T\log(T)).$$

Proof. The main term comes directly from our observations above, so that we only need to estimate

$$\int_G f_T(g)\langle \pi(g)\psi^\perp_{-2k}, \psi^\perp_{-2n}\rangle dg.$$

By the triangle inequality and mixing (i.e. Lemma 4.30) we get the bound

$$\|\psi\|^2\int_{A^+}\mathbb{1}_{\|a_t\|<T}te^{-\frac{t}{2}}\sinh(t)dt \ll \|\psi\|_2^2 T\log(T).$$

Here we used the explicit description of the Haar measure da_t as $\sinh(t)dt$. □

We are now ready to proof the following master theorem

Theorem 7.4. *Let Γ be a Fuchsian group of the second kind with critical exponent $\delta > \frac{1}{2}$. Let*

$$0 < \delta(1-\delta) = \lambda_0 < \lambda_1 \leq \ldots \leq \lambda_N < \frac{1}{4}$$

be the exceptional eigenvalues of Δ on $\Gamma\backslash\mathbb{H}$. Then, for integers n and k there are constants $c_1, \ldots, c_N \in \mathbb{C}$ depending on n and k such that

$$\sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<T}} e^{2in\theta_1(\gamma)}e^{2ik\theta_2(\gamma)} = \widehat{\mu}(2n)\widehat{\mu}(2k)\sqrt{\pi}\frac{\Gamma(\delta-\frac{1}{2})}{\Gamma(\delta+1)}T^{2\delta} + \sum_{j=1}^{N} c_j(n,k)T^{2s_j} + O(T^{\frac{1}{4}+2\delta\cdot\frac{3}{4}}\log(T)^{\frac{1}{4}}(1+|n|+|k|)^{\frac{3}{4}})$$

as $T \to \infty$. Here $|c_j(n,k)| \ll |c_j(0,0)|$,as n and k vary and the implied constants depend only on Γ.

Proof. Recall that we want to estimate precisely $\mathcal{N}(T)$ and we already related this to $\mathcal{H}(T)$ up to an negligible error. We start directly from the last lemma.

Recall that ψ had unit mass. This implies

$$\langle v_{-2n}, \psi\rangle = v_{-2n}(e) + O(\eta) \text{ and } \|\psi\|^2 \ll \eta{-}3.$$

(The last exponent is 3, since this is the dimension of G.)

We end up with

$$\mathcal{N}(T) = \overline{v_{-2k}(e)}v_{-2n}(e)\int_0^{2\log(T)} \langle \pi(a_t)v_{-2k}, v_{-2n}\rangle \sinh(t)dt + O(\eta(1+|n|+|k|)T^{2\delta} + \eta - 3T\log(T)).$$

Choosing η optimally gives the desired error.

It remains to evaluate the main term. This follows from the considerations in Section 4.7 and is left as an **Exercise**. □

From this theorem we can derive several important consequences that will be used later.

Theorem 7.5. *Let Γ be a Fuchsian group of the second kind with $\delta_\Gamma > \frac{5}{6}$. Fix $\gamma_0 \in \Gamma$ and a congruence subgroup $\Gamma_1(q) \subset \Gamma$ of level $q \geq 1$. Let $f\colon G \to \mathbb{C}$ be a smooth function with $|f| \leq 1$. There is a fixed integer $\mathfrak{B}$ depending only on Γ such that for $q = q_1q_2$, $q_1 \mid \mathfrak{B}$,*

$$\sum_{\substack{\gamma\in\gamma_0\Gamma_1(q),\\ \|\gamma\|<T}} f(\gamma) = \frac{1}{[\Gamma\colon\Gamma_1(q)]}\left(\sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<T}} f(\gamma) + \mathcal{E}_{q_1}\right) + O\left(T^{\frac{12}{7}\delta+\frac{5}{21}}(1+\mathcal{S}_{\infty,T}f)^{\frac{6}{7}}\right).$$

Here $\mathcal{E}_{q_1} \ll T^{2\delta-\alpha_0}$, with $\alpha_0 > 0$ and all implied constants are independent of q_2 and γ_0.

Proof. The ramification number $\mathfrak{B}$ is constructed using strong approximation. Recall Remark 5.6 for the corresponding spectral gap. We write

$$0 < \delta(1-\delta) = \lambda_0^q < \lambda_1^q \leq \ldots \leq \lambda_{N_0(q)}^q < \frac{1}{4}.$$

We call an eigenvalue coming from smaller levels old-forms. In particular, due to Gamburd's $\frac{5}{6}$ theorem all eigenvalues below $\frac{5}{6}(1-\frac{5}{6})$ are old, except for possible finitely many that come from level q_1.

We view f in KA^+K-coordinates: $f(g) = f(\theta_1(g), t(g), \theta_2(g))$. We define $f_T(g) = f(g)\mathbb{1}_{\|g\|<T}$ and

$$\mathcal{F}_{T,q}(g,h) = \sum_{\gamma \in \Gamma_0(q)} f_T(g^{-1}\gamma h).$$

Clearly we have

$$\mathcal{F}_{T,q}(\gamma_0^{-1}, 1) = \mathcal{N}_q(T) = \sum_{\substack{\gamma \in \gamma_0\Gamma_1(q), \\ \|\gamma\|<T}} f(\gamma).$$

Fix $\eta > 0$ and a smooth tes function $\psi\colon G \to \mathbb{R}$ with unit mass (i.e. $\int_G \psi(g)dg = 1$) and compact support in a ball of radius η around 1. We write

$$\Psi(g) = \sum_{\gamma \in \Gamma(q)} \psi(\gamma_g)$$

and $\Psi_{q,\gamma_0}(g) = \Psi_q(\gamma_0 g)$. As before we consider the integral

$$\mathcal{H}_q(T) = \langle \mathcal{F}_T, \Psi_{q,\gamma_0} \otimes \Psi_q \rangle = \int_{\Gamma(q)\backslash G} \int_{\Gamma(q)\backslash G} \mathcal{F}_T(\gamma_0^{-1} g, h)\Psi_q(g)\Psi_q(h)dgdh.$$

This is a good approximation to $\mathcal{N}_q$:

$$\mathcal{H}_q(T) = \mathcal{N}_q(T) + O(\eta(1+\mathcal{S}_{\infty,T})T^{2\delta}).$$

One further computes

$$\mathcal{H}_q(T) = \int_G f_T(g)\langle \pi(g)\Psi_q, \Psi_{q,\gamma_0}\rangle_{\Gamma(q)\backslash G} dg.$$

For notational purposes we make the following assumption on the exceptional spectrum. The spectrum below $\frac{5}{36}$ consists only of

- The base eigenvalue $\lambda_0 = \delta(1-\delta)$ with ground state
$$\varphi^{(q)} = [\Gamma\colon \Gamma(q)]^{-\frac{1}{2}}\varphi^{(1)};$$
- One eigenform from
$$\tilde{\varphi}^{(q)} = \frac{1}{\sqrt{\Gamma\colon \Gamma(q)}}\sqrt{[\Gamma\colon \Gamma(q_1)]}\tilde{\varphi}^{(q_1)}$$
from the bad level q_1.

Here $\varphi^{(1)} \in L^2(\Gamma\backslash G)$ and $L^2(\Gamma(q_1)\backslash G)$ are normalized newforms. We write V (resp. $\tilde{V}$) for the irreducible subspace of $L^2(\Gamma(q)\backslash G)$ generated by $\varphi(q)$ (resp. $\tilde{\varphi}^{(q)}$). We can decompose

$$\Psi_q = \Psi_q|_V + \Psi_q|_{\tilde{V}} + \Psi_q^{\perp}$$

accordingly. A similar decomposition holds for Ψ_{q,γ_0}. The projections are given by

$$\Psi_q|_V = \sum_{k\in\mathbb{Z}} \langle \Psi_q, \varphi_{2k}^{(q)}\rangle \varphi_{2k}^{(q)}$$

and so on. We use this decomposition to write

$$\mathcal{H}_q(T) = W_q(T) + \tilde{W}_q(T) + W_q^{\perp}(T).$$

The pieces are given by the obvious expressions, for example

$$W_q(T) = \int_G f_T(g)\langle \pi(g)\Psi_q|_V, \Psi_{q,\gamma_0}|_V\rangle_{\Gamma(q)\backslash G} dg.$$

A simple computation shows that

$$spec_matest\langle \Psi_q, \varphi_{2k}^{(q)}\rangle_{\Gamma(q)\backslash G} = \frac{1}{\sqrt{\Gamma\colon \Gamma(q)}}\langle \Psi_1, \varphi_{2k}^{(1)}\rangle_{\Gamma\backslash G}$$

$$\text{and } \langle \pi(g)\varphi_{2k}^{(q)}, \varphi_{2k'}^{(q)}\rangle_{\Gamma(q)\backslash G} = \langle \pi(g)\varphi_{2k}^{(1)}, \varphi_{2k'}^{(1)}\rangle_{\Gamma\backslash G}.$$

These calculations are straight forward from the definitions, but it is important to get the volumes right. This leads to

$$W_q(T) = \frac{1}{\Gamma\colon \Gamma(q)} W_1(T) \text{ for } W_1(T) = \int_G f_T(g)\langle \pi(g)\Psi_1|_V, \Psi_1|_V\rangle_{\Gamma\backslash G}.$$

The same argument yields

$$\tilde{W}_q(T) = \frac{1}{\Gamma\colon \Gamma(q)} \mathcal{E}_{q_1}(T),$$

for

$$\mathcal{E}_{q_1}(T) = [\Gamma\colon \Gamma(q_1)] \int_G f_T(g)\langle \pi(g)\Psi_{q_1}|_{\tilde{V}}, \Psi_{q_1,\tilde{\gamma}_0}|_{\tilde{V}}\rangle_{\Gamma(q_1)\backslash G}.$$

This integral can be estimated independent of q_2 and by taking the supremum over $\tilde{\gamma} \in \Gamma(q_1)\backslash\Gamma$, making the estimate independent of γ_0.

Finally we can estimate $W_q^{\perp}$ using the spectral gap. Indeed we have $\|\Psi_q\| \ll \eta^{-\frac{3}{2}}$ and $\mathcal{S}\Psi_q \ll \eta^{-\frac{9}{2}}$, so that

$$W_q^{\perp}(T) \ll T^{2\cdot\frac{5}{6}} \|\Psi_q\|_2 \mathcal{S}\Psi_q \ll T^{\frac{5}{3}}\eta^{-6}$$

by Lemma 4.31.

Putting everything together yields

$$\mathcal{N}_q(T) = \frac{1}{[\Gamma\colon \Gamma(q)]}(\mathcal{N}_1(q) + \mathcal{E}_{q_1}(T)) + O(T^{\frac{5}{3}}\eta - 6 + \eta(1 + \mathcal{S}_{\infty,T} f)T^{2\delta}).$$

We conclude by choosing $\eta = (1+\mathcal{S}_{\infty,T}f)^{-\frac{1}{7}}T^{-2(\delta-\frac{5}{6})/7}$. □

Theorem 7.6. *Let $\mathbf{v}_0, \mathbf{w} \in \mathbb{Z}^2$ and assume that $n \in \mathbb{Z}$, $\frac{N}{K_0} < |n| < N$, $|\mathbf{w}| < N^{1-\sigma}$, $|\mathbf{v}_0| \leq 1$, and $|n| < |\mathbf{v}_0||\mathbf{w}|N^\sigma$. Then*

$$\sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<N^\sigma}} \mathbb{1}_{\{|\langle \mathbf{v}_0\cdot\gamma,\mathbf{w}\rangle - n|<\frac{N}{2K_0}\}} \gg \frac{N^{2\delta\sigma}}{K_0} + O(N^{\sigma(\frac{3}{4}+\frac{1}{2}\delta)}\log(N)^{\frac{1}{4}}).$$

Proof. We decompose $\gamma = k_u a_\rho k_v \in KA^+K$. We have $1 < \rho \approx \|\gamma\|$, so that $\rho < T$. Write $\mathbf{v} = (a,b)$ and $\mathbf{w} = (c,d)$ and compute

$$\begin{aligned} |\langle \mathbf{v}\gamma, \mathbf{w}\rangle - n| &= |(a,b)k_u a_\rho k_v (c,d)^\top - n| \\ &\approx |\rho(a\cos(u) - b\sin(u))(-c\sin(v) + d\cos(v)) - n| \\ &< \frac{N}{2K_0}. \end{aligned}$$

This follows by direct computations expanding the product $k_u a_\rho k_v$ and we omit the details.

We can rewrite this as

$$|\rho|\mathbf{v}||\mathbf{w}|\cos(\mathfrak{u})\cos(\mathfrak{v}) - n| < \frac{N}{2K_0}$$

or

$$|\frac{\rho}{N^\sigma}\cos(\mathfrak{u})\cos(\mathfrak{v}) - \frac{n}{|\mathbf{v}||\mathbf{w}|N^\sigma}| < \frac{N^{1-\sigma}}{2|\mathbf{v}||\mathbf{w}|K_0}$$

where $\mathfrak{u}$ is the angle between (a,b) and $(\cos(u), -\sin(u)$ while $\mathfrak{v}$ is the angle between (c,d) and $(\cos(v),\sin(v))$. We definte $\mathcal{A} = \frac{n}{N^\sigma|\mathbf{v}||\mathbf{w}|}$ and $\mathcal{B} = \frac{N^{1-\sigma}}{|\mathbf{v}||\mathbf{w}|}$.

Note that $\cos(\mathfrak{u})$ and $\cos(\mathfrak{v})$ range in intervals independent of K_0. Dividing these intervals in sectors: $u \in \Psi_\alpha$ and $v \in \Phi_\beta$ yields

$$\sum_{\substack{\gamma\in\Gamma\\ N^\sigma(\mathcal{A}-\frac{\mathcal{B}}{2K_0})<\|\gamma\|<N^\sigma(\mathcal{A}+\frac{\mathcal{B}}{2K_0})}} \mathbb{1}_{u\in\Psi_\alpha}\mathbb{1}_{v\in\Psi_\beta} \gg \frac{1}{K_0}\left(\mu(\Psi_\alpha)\mu(\Phi_\beta)c_0N^{2\delta\sigma} + \sum_j c_j N^{2s_j\sigma}\right) + O(T^{\frac{3}{4}+\frac{\delta}{2}}\log(T)^{\frac{1}{4}}).$$

by (smoothing and) applying Theorem 7.4. Note that $\Psi = \cup_\alpha\Psi_\alpha$ and $\Phi = \cup_\beta\Phi_\beta$ are K_0-independent intervals we have $\mu(\Psi), \mu(\Phi) \gg 1$. This completes the proof. □

Theorem 7.7. *Fix $(c,d) \in \mathbb{Z}^2$ and $y = (y_1, y_2) \in \mathbb{Z}^2$ with $|y| < \sqrt{N}$, $|(c,d)| < N^\sigma$ and $|y| < N^{\frac{1}{2}-\sigma}|(c,d)|$. Then*

$$\sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<N^{\frac{1}{2}-\sigma}}} \mathbb{1}_{\{|(c,d)\gamma - y|<\frac{\sqrt{N}}{K}\}}\mathbb{1}_{\{(c,d)\gamma\equiv y \bmod q\}} \ll \frac{N^{\delta(1-2\sigma)}}{K^{1+\delta}q^2} + N^{(\frac{1}{2}-\sigma)(\frac{12}{7}\delta+\frac{5}{21})}.$$

Proof. Writ $\Gamma_0(q)$ for the subgroup of Γ which stabilizes (c,d) modulo q. This subgroup obviously has level q (i.e. it contains $\Gamma(q)$). We decompose $\gamma = \gamma_0\gamma_1 \in \Gamma$ with $\gamma_0 \in \Gamma_0(q)$ and $\gamma_1 8\Gamma_0(q)\backslash\Gamma$. Using Theorem 7.5 we get

$$\begin{aligned}
&\sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<N^{\frac12-\sigma}}} \mathbb{1}_{\{|(c,d)\gamma-y|<\frac{\sqrt{N}}{K}\}}\mathbb{1}_{\{(c,d)\gamma\equiv y \bmod q\}}\\
&= \sum_{\gamma_1\in\Gamma_0(q)\backslash\Gamma} \mathbb{1}_{\{(c,d)\gamma_1\equiv y \bmod q\}} \sum_{\substack{\gamma\in\Gamma_0(q)\cdot\gamma_1,\\ \|\gamma\|<N^{\frac12-\sigma}}} \mathbb{1}_{\{|(c,d)\gamma-y|<\frac{\sqrt{N}}{K}\}}\\
&\ll \frac{1}{[\Gamma\colon\Gamma_0(q)]} \sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<N^{\frac12-\sigma}}} \mathbb{1}_{\{|(c,d)\gamma-y|<\frac{\sqrt{N}}{K}\}} + O(N^{(\frac12-\sigma)(\frac{12}{7}\delta+\frac{5}{21})}).
\end{aligned}$$

It remains to estimate

$$\mathcal{N} = \sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<N^{\frac12-\sigma}}} \mathbb{1}_{\{|(c,d)\gamma-y|<\frac{\sqrt{N}}{K}\}}.$$

This is again done by writing $\gamma = K_u a_\rho k_v$ in KA^+K coordinates with $1 < \rho \approx \|\gamma\| < T$. To analyse this we let $\mathfrak{u}$ bet the angle between (c,d) and $(\cos(u), -\sin(u)$. Similarly let $\mathfrak{v}$ be the angle between (y_1,y_2) and $(\cos(v),\sin(v))$. A direct computation shows that

$$\frac{N}{K^2} > (\rho\cdot|(c,d)|\cos(\mathfrak{u}) - |\mathbf{y}|\cos(\mathfrak{v}))^2 + |\mathbf{y}|^2(1-\cos(\mathfrak{v})^2).$$

We break this inequality in two pieces.

First consider

$$\frac{N}{K^2} > |\mathbf{y}|^2(1-\cos(\mathfrak{v})^2).$$

This requires $\mathfrak{v} \ll \frac{\sqrt{N}}{|\mathbf{y}|K}$. This forces v to be contained in an interval Φ of length $\ll \frac{\sqrt{N}}{|\mathbf{y}|K}$.

The second piece can be written as

$$\left|\frac{\rho}{N^{\frac12-\sigma}}\cos(\mathfrak{u}) - \mathcal{A}\right| \ll \frac{1}{K},$$

for $\mathcal{A} = \frac{|\mathbf{y}|}{N^{\frac12-\sigma}|(c,d)|}\cos(\mathfrak{v})$. Since $\rho < T$, u ranges over a constant interval Ψ. Breaking the sum into sectors and using Theorem 7.4 yields

$$\mathcal{N} \ll \sum_{\substack{\gamma\in\Gamma,\\ \mathcal{A}N^{\frac12-\sigma}(1-\frac{c_1}{K})<\|\gamma\|<\mathcal{A}N^{\frac12+\sigma}(1-\frac{c_1}{K})}} \mathbb{1}_{u\in\Psi}\mathbb{1}_{v\in\Phi} \ll \frac{1}{K}\mu(\Psi)\mu(\Phi)N^{(\frac12-\sigma)2\delta}.$$

Note that $[\Gamma\colon \Gamma_0(q)] \gg q^2$. One completes the proof by observing that since $|\Phi| \ll K^{-1}$ we have $\mu(\Phi) \ll K^{-\delta}$. □

The following lemma is of similar flavor as the results above.

Lemma 7.8. *For $(qK)^{\frac{13}{5}} < Y < X$, and vectors $\eta, \eta' \in \mathbb{Z}^2$ having co-prime coordinates with $|\eta| \asymp \frac{X}{Y}$ and $|\eta'| \asymp Y$, we have*

$$\sharp\{\gamma \in \mathrm{SL}_2(\mathbb{Z})\colon \|\gamma\| \asymp Y,\, |\gamma\eta - \eta'| < \frac{X}{YK} \text{ and } \gamma\eta \equiv \gamma' \bmod q\} \ll \left(\frac{Y}{qK}\right)^2.$$

The implied constant is absolute, but it depends on the implied constants in the assumption.

Proof. Write $G(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})$ and let

$$G_\eta(q) = \{\gamma \in G(\mathbb{Z})\colon \gamma\eta \equiv \eta \bmod q\}$$

the stabilizer of η modulo q. We write $G(\mathbb{Z}) \cong G(\mathbb{Z})/G_\eta(q) \times G_{\eta(q)}$.

We define the region

$$R_{Y,K} = R = \{g \in \mathrm{SL}_2(\mathbb{Z})\colon \|g\|Y,\, |g\eta - \eta'| < \frac{X}{YK}\}.$$

The methods from this section apply in this setting and give estimates of the form

$$\begin{aligned}\sum_{\gamma\in G(\mathbb{Z})} \mathbb{1}_{\gamma\in R}\mathbb{1}_{\gamma\eta\equiv\eta' \bmod q} &= \sum_{w\in G(\mathbb{Z})/\Gamma_\eta(q)} \mathbb{1}_{w\eta\equiv\eta' \bmod q} \sum_{\gamma'\in G_\eta(\mathbb{Z})} \mathbb{1}_{w\gamma'\in R} \\ &\ll \sum_{w\in G(\mathbb{Z})/\Gamma_\eta(q)} \mathbb{1}_{w\eta\equiv\eta' \bmod q} \left(\left(\frac{Y}{qK}\right)^2 + Y^{2\Theta+\epsilon}\right) \\ &\ll \left(\frac{Y}{qK}\right)^2 + Y^{2\Theta+\epsilon}. \end{aligned} \tag{19}$$

Here we take $\Theta = \frac{1}{2} + \frac{7}{64}$ to be the best known bound towards the Ramanujan conjecture. The first term dominates as long as $q^2K^2 < Y^{\frac{25}{32}}$. Note that $\frac{64}{25} + \epsilon < \frac{13}{5}$ for ϵ small enough. This concludes the proof sketch. □

Later we will also need a sector estimate for the semi group Γ_A (actually the result also holds for the more general version $\Gamma_{\mathfrak{A}}$ introduced later).

Proposition 7.9. *Let $v_+(\gamma)$ be the expanding eigenvector of a matrix $\gamma \in \Gamma_A$. Further take a density point $x \in \mathfrak{C}_A$ (the limiting set of Γ_A) and put $\mathfrak{v} = \frac{(x,1)}{\sqrt{1+x^2}}$. There is a constant $c = c(A) > 0$ so that as long as $H < e^{c\sqrt{\log(T)}}$ we have*

$$\sharp\{\gamma \in \Gamma_A\colon \|\gamma\| < T \text{ and } |v_+(\gamma) - \mathfrak{v}| < H^{-1}\} \gg \frac{T^{2\delta}}{H}$$

as $T \to \infty$.

Proof. Let $\mathcal{I}$ be an interval of length H^{-1} around $\mathfrak{v}$. Asymptotic estimates like

$$\sharp\{\gamma \in \Gamma \colon \|\gamma\| < T \text{ and } |v_+(\gamma) - \mathfrak{v}| < H^{-1}\} \sim C \cdot T^{2\delta}\mu(\mathcal{I}) \tag{20}$$

are well known for non-elementary convex-cocompact (i.e. no parabolic elements) subgroups of $\mathrm{SL}_2(\mathbb{Z})$. Here μ is the δ-dimensional Hausdorff measure lifted from the limiting set $\Lambda(\Gamma)$ to $\mathbb{P}^1$ by setting $d\mu(x, y) = d\mu(x/y)$. Note that the proof of these results uses symbolic dynamics and the renewal method. It is here were it becomes essential that there are no parabolic elements.

The same method can be set up for Γ replaced by the free semigroup Γ_A. (In this case the transition matrix turns out to be trivial, since all words are allowed.) One obtains the asymptotic (20) with an error term bounded by

$$T^{2\delta}e^{-c\sqrt{\log(T)}}.$$

The result follows since $\mathfrak{v}$ corresponds to a density point of $\mathfrak{C}_A$, so that $\mu(\mathcal{I}) \gg H^{-\delta-\epsilon}$. The constraint on H ensures that the error term is nicely dominated.

This concludes this quick sketch. □

8. Proof of Theorem 1.2

We are now ready to put things together. The proofs of the Theorems stated in the introduction all rely on a version of the *Hardy-Littlewood circle method*. However each one has its own intricacies. We start by proving Theorem 1.2 in detail.

To complete this proof we closely follow the original analysis from [4]. The basic idea is to apply standard circle method analysis to the coefficients of the trigonometric polynomial

$$\sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<N}} e(\langle \mathbf{v}_0\gamma, \mathbf{w}_0\rangle\theta).$$

But for technical reasons we need to consider a modified version of this sum.

Recall that $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$ is finitely generated, free and contains no parabolic elements. Suppose $\alpha_1, \dots, \alpha_l$ are the free generators of Γ. Thus each element in Γ can be uniquely written as a reduced word in the letters $\alpha_1, \dots, \alpha_l$. We write $l(\gamma)$ for the word length. Let δ denote the Hausdorff dimension of the limit set of Γ and assume $\delta > \frac{1}{2}$.

We define

$$\Xi_k = \{\xi \in \Gamma \colon \|\xi\| < N^{\frac{1}{2}},\ l(\xi) = k \text{ and the reduced word giving } \xi \text{ does end on } \alpha_l\}.$$

For some $\sigma < \frac{1}{4}$ we set

$$\Pi = \{\varpi \in \Gamma \colon \|\varpi\| < N^{\frac{1}{2}-\sigma} \text{ and the reduced word giving } \varpi \text{ does start with } \alpha_1\}.$$

Without loss of generality we can assume that $\alpha_1 \neq \alpha_l^{-1}$.

Exercise 8.1. *Show that there is $k_0 \in \mathbb{N}$ so that $\sharp\Xi_{k_0} \gg N^{\delta}\log(N)^{-1}$ and $\sharp\Pi \gg N^{\delta(1-2\sigma)}$ with an explicit constant only depending on Γ. (Hint: Pigeonhole principle and $\log(\|\gamma\|) \ll_\Gamma l(\gamma) \ll_\Gamma \log(\|\gamma\|)$.)*

We put $\Xi = \Xi_{k_0}$ for $k_0 \in \mathbb{N}$ as in the exercise. For fixed N and $\theta \in [0,1]$ we set

$$S_N(\theta) = \sum_{\xi\in\Xi}\sum_{\varpi\in\Pi}\sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<N^\sigma}} e(\langle \mathbf{v}_0\gamma\xi\varpi, \mathbf{w}_0\rangle\theta).$$

The nth Fourier coefficients of this trigonometric polynomials are given by

$$\begin{aligned} R_N(n) = \widehat{S_N}(n) &= \int_0^1 S_N(\theta)e(-n\theta)d\theta \\ &= \sum_{\xi\in\Xi}\sum_{\varpi\in\Pi}\sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<N^\sigma}} \mathbb{1}_{\{\langle \mathbf{v}_0\gamma\xi\varpi,\mathbf{w}_0\rangle = n\}}. \end{aligned}$$

The point is that the products $\xi\varpi$ are unique, since the last letter of ξ can not cancel with the last letter of ϖ. Thus, since Γ has no parabolic elements the vectors $\mathbf{v}\xi\varpi$ are all distinct. The additional element γ needs to be included in order to make it possible to apply spectral methods, for which it is impossible to detect restrictions on the letters.

We will split the integral defining $R_N(n)$ into major and minor arcs. This is done by setting $Q_0 = N^{\alpha_0}$ and $K_0 = N^{\kappa_0}$. The major arcs are given by

$$\mathfrak{M} = \{\theta = \frac{a}{q} + \beta \colon q < Q_0 \text{ and} |\beta| < K_0/N\}.$$

Of course the minor arcs are simply the complement: $\mathfrak{m} = [0,1] \setminus \mathfrak{M}$.

Exercise 8.2. *Almost all (with respect to Lebesgue measure) $\theta \in [0,1]$ are of the form $\theta = \frac{a}{q} + \beta$ with $q < N^{\frac{1}{2}}$, $(a,q) = 1$ and $|\beta| < \frac{1}{qN^{\frac{1}{2}}}$. (Hint: Consider irrational numbers.)*

For analytic reasons we will also introduce the triangle function

$$\psi(x) = \begin{cases} 0 & \text{if } |x| \geq 1, \\ 1-x & \text{if } 0 < x < 1, \\ 1+x & \text{if } -1 < x \leq 0. \end{cases} \tag{21}$$

Exercise 8.3. *Show that ψ is a self-convolution and compute its Fourier transform to be $\widehat{\psi}(y) = \left(\frac{\sin(\pi y)}{\pi y}\right)^2$. In particular it is positive.*

We need some more definitions:

$$\begin{aligned}\Psi_{N,K_0}(\beta) &= \sum_{m\in\mathbb{Z}} \psi\left((\beta+m)N/K_0\right),\\ \mathfrak{M}(\theta) &= \sum_{1\le q\le Q_0}\sum_{(a,q)=1} \Psi_{N,K_0}(\theta-\frac{a}{q}),\\ \mathcal{M}_N(n) &= \int_0^1 \mathfrak{M}(\theta)S_N(\theta)e(-n\theta)d\theta,\\ \mathfrak{m}(\theta) &= 1-\mathfrak{M}(\theta) \text{ and}\\ \mathcal{E}_N(n) &= \int_0^1 \mathfrak{m}(\theta)S_N(\theta)e(-n\theta)d\theta.\end{aligned}$$

We will proceed by giving a lower bound for $\mathcal{M}_N(n)$ the so called major arc contribution (or the main term). Later we will establish a mean square estimate for $\mathcal{E}_N(n)$ the minor arc contribution (or the error).

Theorem 8.1. *There is a set $\mathfrak{E}(N)\subset[-N,N]$ of size $\sharp\mathfrak{E}(N)\ll N^{1-\epsilon_0}$ such that the following holds. Suppose $K_0=N^{\kappa_0}$ and $Q_0=N^{\alpha_0}$ with*

$$\kappa_0<\frac{3}{2}\sigma(\delta-\frac{1}{2}) \text{ and } 21\alpha_0+13\kappa_0<(2\delta-\frac{5}{3})\sigma.$$

Then, for $|n|<N$ and $n\notin\mathfrak{E}(N)$, the main term is

$$\mathcal{M}_N(n)=\begin{cases}\gg \log\log(10+|n|)^{-1}\log(N)^{-1}N^{2\delta-1} & \text{if } n\in\mathcal{A},\\ 0 & \text{else.}\end{cases}$$

Proof. We fix $\varpi\in\Pi$ and $\xi\in\Xi$ and write

$$\mathbf{w}=\mathbf{w}_0\varpi^\top\xi^\top.$$

Furthermore we define the congruence subgroup of Γ of level q by

$$\Gamma(q)\subset\Gamma_1(q)=\{\gamma\in\Gamma\colon \mathbf{v}_0\gamma\equiv\mathbf{v}_0 \bmod q\}.$$

Every $\gamma\in\Gamma$ can be written as $\gamma=\gamma_1\gamma_2$ with $\gamma_1\in\Gamma_1(q)$ and $[\gamma_2]\in\Gamma_1(q)\backslash\Gamma$. This composition then gives

$$\langle\mathbf{v}_0\cdot\gamma_1\gamma_2,\mathbf{w}\rangle\equiv\langle\mathbf{v}_0\cdot\gamma_2,\mathbf{w}\rangle \bmod q.$$

This allows us to split the sum

$$\begin{aligned}&\sum_{\substack{\gamma\in\Gamma,\\ \|\gamma\|<N^\sigma}} e\left(\langle\mathbf{v}_0\cdot\gamma,\mathbf{w}\rangle(\frac{a}{q}+\beta)\right)\\ &\qquad=\sum_{\gamma_2\in\Gamma(q)\backslash\Gamma} e\left(\langle\mathbf{v}_0\cdot\gamma_2,\mathbf{w}\rangle\frac{a}{q}\right)\sum_{\substack{\gamma_1\in\Gamma(q),\\ \|\gamma_1\gamma_2\|<N^\sigma}} e\left(\langle\mathbf{v}_0\cdot\gamma_1\gamma_2,\mathbf{w}\rangle\beta\right).\end{aligned}$$

An application of Theorem 7.5 (assuming a spectral gap $(\Theta, \mathfrak{B}) = (\frac{5}{6}, 1)$ for simplicity) yields

$$\sum_{\substack{\gamma_1 \in \Gamma(q), \\ \|\gamma_1\gamma_2\| < N^\sigma}} e(\langle \mathbf{v}_0 \cdot \gamma_1\gamma_2, \mathbf{w}\rangle \beta) = \frac{1}{[\Gamma\colon \Gamma(q)]} \sum_{\substack{\gamma \in \Gamma, \\ \|\gamma\| < N^\sigma}} e(\langle \mathbf{v}_0 \cdot \gamma, \mathbf{w}\rangle \beta) + O(K_0^{\frac{6}{7}} N^{\sigma(\frac{12}{7}\delta + \frac{5}{21})}),$$

for $|\beta| < \frac{K_0}{N}$. Inserting this in the main term yields

$$\begin{aligned}\mathcal{M}_N(n) = &\sum_{\xi, \varpi} \sum_{q \le Q_0} \sum_{(a,q)=1} \sum_{\gamma_2 \in \Gamma(q)\backslash\Gamma} \frac{e((\langle \mathbf{v}_0 \cdot \gamma_2, \mathbf{w}\rangle - n)\frac{a}{q})}{[\Gamma\colon \Gamma_1(q)]} \\ &\sum_{\substack{\gamma \in \Gamma, \\ \|\gamma\| < N^\sigma}} \int_0^1 \Psi_{N,K_0}(\beta) e((\langle \mathbf{v}_0 \cdot \gamma, \mathbf{w}\rangle - n)\beta) d\beta \\ &+ O(N^{2\delta(1-\sigma)} Q_0^{3+\epsilon} \frac{K_0}{N} K_0^{\frac{6}{7}} N^{\sigma(\frac{12}{7}\delta + \frac{5}{21})}).\end{aligned}$$

Realizing that the a-sum yields a Ramanujan sum brings us to the following singular series:

$$\mathfrak{S}_{N,\xi,\varpi}(n) = \sum_{q < Q_0} \frac{1}{[\Gamma\colon \Gamma(q)]} \sum_{\gamma_2 \in \Gamma(q)\backslash\Gamma} c_q(\langle \mathbf{v}_0 \cdot \gamma_2, \mathbf{w}\rangle - n).$$

The singular integral reads

$$\tau_{N,\xi,\varpi}(n) = \sum_{\substack{\gamma \in \Gamma, \\ \|\gamma\| < N^\sigma}} \int_0^1 \Psi_{N,K_0}(\beta) e((\langle \mathbf{v}_0 \cdot \gamma, \mathbf{w}\rangle - n)\beta) d\beta.$$

Thus the main term reads

$$\mathcal{M}_N(n) = \sum_{\xi, \varpi} \mathfrak{S}_{N,\xi,\varpi} \tau_{N,\xi,\varpi}(n) + O(N^{2\delta(1-\sigma)} Q_0^{3+\epsilon} \frac{K_0}{N} K_0^{\frac{6}{7}} N^{\sigma(\frac{12}{7}\delta + \frac{5}{21})}).$$

We begin by looking at the singular integral. By definition of Ψ_{N,K_0} we obtain

$$\int_0^1 \Psi_{N,K_0}(\beta) e(x\beta) d\beta = \int_{\mathbb{R}} \psi(\beta \frac{N}{K_0}) e(x\beta) d\beta = \frac{K_0}{N} \widehat{\psi}(x \frac{K_0}{N}) \ge \frac{2K_0}{5N} \mathbb{1}_{|x| < \frac{N}{2K_0}}.$$

The final lower bound follows since $\widehat{\psi}$ is positive and one checks that $\widehat{\psi}(y) > 0.4$ for $|y| < \frac{1}{2}$ using the exact formula. We obtain

$$\tau_{N,\xi,\varpi}(n) \ge \frac{2K_0}{5N} \sum_{\substack{\gamma \in \Gamma, \\ \|\gamma\| < N^\sigma}} \mathbb{1}_{|\langle \mathbf{v}_0 \cdot \gamma, \mathbf{w}\rangle - n| < \frac{N}{2K_0}}.$$

We want to apply Theorem 7.6 to give a lower bound for the γ-sum. This can be done for $|\mathbf{w}| \asymp N^{1-\sigma}$ and we get

$$\begin{aligned}\tau_{N,\xi,\varpi}(n) &\gg \frac{2K_0}{5N} \cdot \frac{N^2\delta\sigma}{K_0} + O(\frac{K_0}{N} N^{(2\delta+3)\sigma/4} \log(N)^{\frac{1}{4}}) \\ &\gg N^{2\delta\sigma-1} + O(\frac{K_0}{N} N^{(2\delta+3)\sigma/4} \log(N)^{\frac{1}{4}}).\end{aligned}$$

On the other hand, if $|\mathbf{w}| \ll N^{1-\sigma-\epsilon}$, then $|\langle \mathbf{v}_0 \cdot \gamma, \mathbf{w}\rangle| \ll N^{1-\epsilon}$ and the corresponding values for n can be absorbed into the exceptional set $\mathfrak{E}(N)$. Recall that $K_0 = N^{\kappa_0}$. The error term is fine since

$$\kappa_0 < \frac{3}{2}\sigma(\delta - \frac{1}{2}).$$

We define the *Ramanujan sum*

$$c_q(x) = \sum_{(a,q)=1} e(\frac{ax}{q}).$$

Exercise 8.4. *Show that $c_p(x)$ is $p-1$ if $x \equiv 0$ mod p and -1 otherwise.*

We turn towards the singular series. Let us pretend that $\Gamma(q)\backslash\Gamma \cong \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ for all q. (The modification needed for the general case are minor.) The singular series reads

$$\mathfrak{S}_{N,\xi,\varpi}(n) = \sum_{q<Q_0} \frac{1}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \sum_{\gamma \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} c_q(\langle \mathbf{v}_0\gamma, \mathbf{w}\rangle - n) \tag{22}$$

This can be extended to all $q \in \mathbb{N}$. One sees that the full q-sum is eulerian and that the main contribution comes from the primes:

$$\prod_p \left(1 + \frac{1}{\sharp \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})} \sum_{\gamma \in \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})} c_p(\langle \mathbf{v}_0\gamma, \mathbf{w}\rangle - n)\right)$$

After changing representatives we can assume without loss of generality that $\mathbf{v}_0 = \mathbf{w} = (0,1)$. Then $\langle \mathbf{v}_0\gamma, \mathbf{w}\rangle = d_\gamma$, where d_γ is the lower right entry of γ. We first treat the case $p \mid n$. An easy counting argument shows that there are $p^2 - p$ choices for γ with $d_\gamma \equiv 0 \bmod p$. Since $\sharp \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}) = p(p^2-1)$ we find that there are remaining matrices (i.e. $p \nmid d_\gamma$) $p^3 - p^2$. Taking the evaluations of the Ramanujan sums into account yields

$$\frac{1}{\sharp \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})} \sum_{\gamma \in \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})} c_p(d_\gamma - n) = \frac{1}{p^3-p}\left[(p-1)(p^2-p) + (-1)(p^3-p^2)\right] = -\frac{1}{p+1}.$$

On the other hand if $p \nmid n$, then we first count matrices γ with $d_\gamma \equiv n \not\equiv 0 \bmod p$. One directly sees that there are p^2 such matrices. Accounting for the remaining

matrices as before yields

$$\frac{1}{\sharp \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})} \sum_{\gamma \in \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})} c_p(d_\gamma - n) = \frac{1}{p^3 - p}\left[(p-1)p^2 + (-1)(p^3 - p^2 - p)\right] = \frac{1}{p^2-1}.$$

Combining both cases yields

$$\mathfrak{S}_{N,\xi,\varpi}(n) \gg \prod_{p \nmid n}\left(1 + \frac{1}{p^2+1}\right)\prod_{p|n}\left(1 - \frac{1}{p+1}\right) \gg \log\log(n)^{-1}.$$

We end up with

$$\mathcal{M}_N(n) \gg N^{2\delta-1}\log(N)^{-1}\log\log(n)^{-1} + O(N^{2\delta(1-\sigma)}Q_0^{3+\epsilon}\frac{K_0}{N}K_0^{\frac{6}{7}}N^{\sigma(\frac{12}{7}\delta + \frac{5}{21})}).$$

Recalling that $Q_0 = N^{\alpha_0}$ and looking at the error term we are done since

$$3\alpha_0 + \frac{13}{7}\kappa_0 < (2\delta - \frac{5}{3})\sigma/7$$

by assumption. □

We will now deal with the minor arcs. To do so we will look at different ranges

$$W_{Q,K} = \{\theta = \frac{a}{q} + \beta \colon (a,q) = 1,\, q \sim Q,\, |\beta| \sim \frac{K}{N}\}.$$

It will take us three (long) theorems to gather all the information we need.

Theorem 8.2. *Write $\theta = \frac{a}{q} + \beta$ with $q \leq N^{\frac{1}{2}}$, $|\beta| < \frac{1}{qN^{\frac{1}{2}}}$ and $|\beta| \sim \frac{K}{N}$. Then*

$$|S_N(\theta)| \ll N^{\frac{3\delta+1}{2}}\left(\frac{1}{K^{(1+\delta)/2}q} + N^{-\frac{1}{84}(6\delta-5)(1-2\sigma)}\right).$$

Proof. We set,

$$\mu(x) = \sum_{\xi \in \Xi} \mathbb{1}_{x = \mathbf{w}_0\xi^\top} \text{ and } \nu(y) = \sum_{\varpi \in \Pi}\sum_{\substack{\gamma \in \Gamma, \\ \|\gamma\| < N^\sigma}} \mathbb{1}_{y = \mathbf{v}_0\gamma\varpi} \tag{23}$$

so that

$$\begin{aligned} S_N(\theta) &= \sum_{\xi \in \Xi}\sum_{\varpi \in \Pi}\sum_{\substack{\gamma \in \Gamma, \\ \|\gamma\| < N^\sigma}} e(\langle \mathbf{v}_0\gamma\varpi, \mathbf{w}_0\xi^\top\rangle\theta) \\ &= \sum_{x,y \in \mathbb{Z}^2} \mu(x)\nu(y)e(\langle x, y\rangle\theta). \end{aligned}$$

Note that because γ contains no parabolic elements we must have $\mu(x) \leq 1$ for all x. On the other hand, due to possible cancellation between γ and ϖ we possibly have $\nu(y) > 1$. However, for fixed γ the value of $\mathbf{v}_0\gamma\varpi$ is unique, and there are $N^{2\delta\sigma}$ choices for γ. Hence we roughly have

$$\nu(y) \ll N^{\delta\sigma}.$$

Thus we have

$$\sum_x \mu(x) \asymp \sum_y \nu(y) \asymp N^\delta.$$

Let us also note that $\mu(x)$ and $\nu(y)$ vanish for $\|x\|, \|y\| \geq N^{\frac{1}{2}}$.

We break $\nu = \sum_\alpha \nu_\alpha$ in 64 pieces each of which is supported in a square of side lengths $\frac{1}{4}N^{\frac{1}{2}}$ and satisfies $|\nu_\alpha| \leq |\nu|$. We then have

$$|S_N(\theta)| \leq \sum_\alpha |S_\alpha(\theta)|,$$

for

$$S_\alpha(\theta) = \sum_{x,y\in\mathbb{Z}^2} \mu(x)\nu_\alpha(y)e(\langle x, y\rangle\theta). \tag{24}$$

Fix a smooth non-negative function $\Upsilon\colon \mathbb{R}^2 \to \mathbb{R}$, which is bounded below by 1 on the unit square $[-1,1]^2$ and with $\mathrm{supp}(\widehat{\Upsilon}) \subset B_{\frac{1}{10}}(0)$.

Inserting Υ and applying Cauchy-Schwarz yields

$$\begin{aligned}|S_\alpha(\theta)| &\ll \left(\sum_x |\mu(x)|^2\right)^{\frac{1}{2}} \left(\sum_x |\sum_y \nu_\alpha(y)e(\langle x, y\rangle\theta)|^2 \Upsilon(\frac{x}{\sqrt{N}})\right)^{\frac{1}{2}} \\ &\ll N^{\delta/2} \left(\sum_x |\sum_y \nu_\alpha(y)e(\langle x, y\rangle\theta)|^2 \Upsilon(\frac{x}{\sqrt{N}})\right)^{\frac{1}{2}}.\end{aligned}$$

Opening the square and interchanging sums yields

$$|S_\alpha(\theta)|^2 \ll N^\delta \sum_{y,y'} \nu_\alpha(y)\nu_\alpha(y') \sum_x e(\langle x, y-y'\rangle\theta)\Upsilon(\frac{x}{N^{\frac{1}{2}}}).$$

Note that

$$\int_{\mathbb{R}^2} e(\langle x, y-y'\rangle\theta)\Upsilon(\frac{x}{N^{\frac{1}{2}}})e(-\langle x, k\rangle)dx = N\widehat{\Upsilon}(N^{\frac{1}{2}}(\theta(y-y') - k)).$$

Thus, applying Poisson summation in the x-sum yields

$$|S_\alpha(\theta)|^2 \ll N^{\delta+1} \sum_{y,y'} \nu_\alpha(y)\nu_\alpha(y') \sum_k \widehat{\Upsilon}(N^{\frac{1}{2}}(\theta(y-y') - k)).$$

Due to the support of $\widehat{\Upsilon}$ there is at most one contribution in the k-sum. It is $\ll |\widehat{\Upsilon}(0)| \ll 1$. More precisely this contribution appears when

$$\{\theta(y-y')\} \leq \frac{1}{10\sqrt{N}}, \tag{25}$$

where $\{\cdot\}$ denotes the distance to the closest $\mathbb{Z}^2$-lattice point. Let us have a closer look at this distance. To do so we use that $\theta = \frac{a}{q} + \beta$. We get

$$|\frac{a}{q}(y-y')| \leq |\theta(y-y')| + |\beta(y-y')|.$$

Due to the support of ν_α we can bound $|y-y'| \leq \frac{\sqrt{2}}{4}N^{\frac{1}{2}}$. Since $|\beta| \leq \frac{1}{q\sqrt{N}}$ we have

$$|\beta(y-y')| \leq \frac{1}{q\sqrt{N}} \cdot \frac{\sqrt{2}}{4}N^{\frac{1}{2}} \leq \frac{\sqrt{2}}{4q}.$$

By recalling (25) we get

$$\{\frac{a}{q}(y-y')\} \leq \frac{1}{10\sqrt{N}} + \frac{\sqrt{2}}{4q}. < 0.5 \cdot q^{-1}.$$

We conclude that $y \equiv y' \bmod q$. This implies that

$$|\beta(y-y')| = \{\theta(y-y')\} \leq \frac{1}{10\sqrt{N}}.$$

We get

$$|S_N(\theta)|^2 \ll N^{\delta+1} \sum_y \nu(y) \sum_{\substack{y' \equiv y \bmod q, \\ |y-y'| \leq \frac{1}{10|\beta|\sqrt{N}}}} \nu(y').$$

Let $|\beta| = \frac{K}{N}$ and assume $K \gg 1$. (The opposite case is handled similarly and left as an exercise.) Considering the inner sum for fixed y yields

$$\begin{aligned}\sum_{\substack{y' \equiv y \bmod q, \\ |y-y'| \leq \frac{1}{10|\beta|\sqrt{N}}}} \nu(y') &= \sum_{\varpi \in \Pi} \sum_{\substack{\gamma \in \Gamma, \\ \|\gamma\| < N^\sigma}} \mathbb{1}_{\substack{(c,d)\varpi \equiv y \bmod q, \\ |(c,d)\varpi - y| \leq \frac{1}{10|\beta|N^{\frac{1}{2}}}}} \\ &\leq \sum_{\substack{\gamma \in \Gamma, \\ \|\gamma\| < N^\sigma}} \sum_{\substack{\varpi \in \Gamma, \\ \|\varpi\| < N^{\frac{1}{2}-\sigma}}} \mathbb{1}_{\substack{(c,d)\varpi \equiv y \bmod q, \\ |(c,d)\varpi - y| \leq \frac{1}{10|\beta|N^{\frac{1}{2}}}}},\end{aligned}$$

where we have written $\mathbf{v}_0\gamma = (c,d)$. By making the exceptional set larger (if necessary) we can assume tat $|y| \asymp N^{\frac{1}{2}}$ and $|(c,d)| \asymp N^\sigma$. This assumptions allow us to apply Theorem 7.7 to obtain

$$\sum_{\substack{\varpi \in \Gamma, \\ \|\varpi\| < N^{\frac{1}{2}-\sigma}}} \mathbb{1}_{\substack{(c,d)\varpi \equiv y \bmod q, \\ |(c,d)\varpi - y| \leq \frac{1}{10|\beta|N^{\frac{1}{2}}}}} \ll \frac{N^{\delta-2\delta\sigma}}{K^{1+\delta}q^2} + N^{(\frac{1}{2}-\sigma)(\frac{12}{7}\delta+\frac{5}{21})}.$$

Going back to $S_N(\theta)$ we observe that the γ-sum contributes $N^{2\delta\sigma}$ and the y-sum gives a contribution of at most N^δ. Therefore we obtain

$$\begin{aligned}S_N(\theta) &\ll N^{\frac{\delta+1}{2}} \left(N^\delta N^{2\delta\sigma} \left(\frac{N^{\delta-2\delta\sigma}}{K^{1+\delta}q^2} + N^{(\frac{1}{2}-\sigma)(\frac{12}{7}\delta+\frac{5}{21})} \right) \right)^{\frac{1}{2}} \\ &\ll \frac{N^{(3\delta+1)/2}}{K^{(1+\delta)/2}q} + N^{\delta+\frac{1}{2}+\delta\sigma+(\frac{1}{2}-\sigma)(\frac{6}{7}\sigma+\frac{5}{42})}.\end{aligned}$$

This completes the proof. □

Theorem 8.3. *Recall that $\theta = \frac{a}{b} + \beta$, with $\frac{Q}{2} \leq q < Q < \sqrt{N}$ and $|\beta| < \frac{1}{qN^{\frac{1}{2}}}$. Fix β with $|\beta| < \frac{2}{Q\sqrt{N}}$ and let*

$$P_{Q,\beta} = \{\theta = \frac{a}{q} + \beta \colon q \sim Q,\ (a,q) = 1\}.$$

In particular we have $\sharp P_{Q,\beta} \asymp Q^2$. We have

$$\sum_{\theta \in P_{Q,\beta}} |S_N(\theta)| \ll N^{1+\delta+\epsilon} Q \left(Q^{-\frac{1}{2}} + N^{-\sigma} + N^{-\frac{\sigma+1}{2}} Q \right).$$

Proof. We write

$$\begin{aligned} S_N(\theta) &= \sum_{\xi \in \Xi} \sum_{\varpi \in \Pi} \sum_{\substack{\gamma \in \Gamma \\ \|\gamma\| < N^\sigma}} e(\langle \mathbf{v}_0 \gamma, \mathbf{w}_0 (\xi\varpi)^\top \rangle \theta) \\ &= \sum_{\mathbf{x} \in B_{N^{1-\sigma}}} \sum_{\mathbf{y} \in B_{N^\sigma}} \mu(\mathbf{x}) \nu(\mathbf{y}) e(\langle \mathbf{x}, \mathbf{y} \rangle \theta), \end{aligned}$$

where μ and ν are certain measures supported in $B_{N^{1-\sigma}}$ and B_{N^σ} respectively. We can write them explicitly as

$$\mu(\mathbf{x}) = \sum_{\substack{\xi \in \Xi, \\ \varpi \in \Pi}} \mathbb{1}_{\mathbf{x} = \mathbf{w}_0 (\xi\varpi)^\top} \text{ and } \nu(\mathbf{y}) = \sum_{\substack{\gamma \in \Gamma, \\ \|\gamma\| < N^\sigma}} \mathbb{1}_{\mathbf{y} = \mathbf{v}_0 \gamma}.$$

By constructions products of the form $\xi\varpi$ are unique, so that the sums have at most one term (i.e. $\mu, \nu \leq 1$). We choose $\zeta(\theta) \in S^1$ so that $|S_N(\theta)| = \zeta(\theta) S_N(\theta)$. For any $\Omega \subset [0,1]$ we have

$$\int_\Omega |S_N(\theta)| d\theta = \sum_{\mathbf{x}} \mu(\mathbf{x}) \sum_{\mathbf{y}} \nu(\mathbf{y}) \int_\Omega \zeta(\theta) e(\langle \mathbf{x}, \mathbf{y} \rangle \theta) d\theta.$$

We now want to apply Cauchy-Schwarz. To do so we recall the properties of the function Υ defined the previous minor arc estimate. We get

$$\int_\Omega |S_N(\theta)| d\theta \leq \left(\sum_{\mathbf{x}} \mu(\mathbf{x})^2 \right)^{\frac{1}{2}} \left(\sum_{\mathbf{x}} |\sum_{\mathbf{y}} \nu(\mathbf{y}) \int_\Omega \zeta(\theta) e(\langle \mathbf{x}, \mathbf{y} \rangle \theta) d\theta| \Upsilon(\frac{\mathbf{x}}{N^{1-\sigma}}) \right)^{\frac{1}{2}}.$$

Recall the bound

$$\sum_{\mathbf{x}} \mu(\mathbf{x})^2 \leq \sum_{\mathbf{x}} \mu(\mathbf{x}) \ll N^{2\delta(1-\sigma)}.$$

Using this estimate and opening the square yields

$$\int_\Omega |S_N(\theta)| d\theta \ll N^{\delta(1-\sigma)} \left(\sum_{\mathbf{y}, \mathbf{y}'} \nu(\mathbf{y}) \nu(\mathbf{y}') \int_\Omega \int_\Omega \zeta(\theta) \overline{\zeta(\theta')} \sum_{\mathbf{x}} e(\langle \mathbf{x}, \mathbf{y}\theta - \mathbf{y}'\theta' \rangle) \Upsilon(\frac{\mathbf{x}}{N^{1-\sigma}}) d\theta d\theta' \right)^{\frac{1}{2}}.$$

We apply Poisson summation in the $\mathbf{x}$-sum. Since $\widehat{\Upsilon}$ has support in $B_{1/10}$ there is only a contribution if $\mathbf{y}\theta - \mathbf{y}'\theta'$ is in a small neighborhood of an integer lattice point. We get

$$\int_\Omega |S_N(\theta)|d\theta \ll N^{(\delta+1)(1-\sigma)} \left(\sum_{\mathbf{y},\mathbf{y}'} \nu(\mathbf{y})\nu(\mathbf{y}') \int_\Omega \int_\Omega \mathbb{1}_{\|\mathbf{y}\theta-\mathbf{y}'\theta'\|<\frac{1}{10N^{1-\sigma}}} d\theta d\theta' \right)^{\frac{1}{2}}. \quad (26)$$

Of course the same argument works with the integral over Ω replaced by a sum over a discrete set of θ's. More precisely we replace $\int_\Omega$ by $\sum_{\theta\in P_{Q,\beta}}$. The result is

$$\sum_{\theta\in P_{Q,\beta}} |S_N(\theta)| \ll N^{(\delta+1)(1-\sigma)} \left(\sum_{\mathbf{y},\mathbf{y}'} \nu(\mathbf{y})\nu(\mathbf{y}') \sum_{\theta\in P_{Q,\beta}} \sum_{\theta'\in P_{Q,\beta}} \mathbb{1}_{\|\mathbf{y}\theta-\mathbf{y}'\theta'\|<\frac{1}{10N^{1-\sigma}}} \right)^{\frac{1}{2}}.$$

This is essentially a point counting problem which can be described as follows. We are interested in the number of elements of a set $A \subset B^2_{N^\sigma} \times P^2_{Q,\beta}$. An element $(\mathbf{y},\mathbf{y}',\theta,\theta') \in A$ is determined by the numbers $\mathbf{y} = (y_1, y_2)$, $\mathbf{y}' = (y_1', y_2')$, $\theta = \frac{a}{b} + \beta$ and $\theta' = \frac{a'}{b'} + \beta$. Exploiting the support properties of our sums we get the following conditions:

(1) $|\mathbf{y}|, |\mathbf{y}'| < N^\sigma$ primitive;
(2) $\mathbf{y}, \mathbf{y}' \neq (0,0)$;
(3) $\|y_1\theta - y_1'\theta'\| < \frac{1}{10N^{1-\sigma}}$ and $\|y_2\theta - y_2'\theta'\| < \frac{1}{10N^{1-\sigma}}$, where $\|\cdot\|$ is the distance to the nearest integer.

A first computation shows that

$$\begin{aligned}\|(y_2'y_1 - y_1'y_2)\frac{a}{q}\| &= \|y_2'(y_1\frac{a}{q} - y_1'\frac{a'}{q'}) - y_1'(y_2\frac{a}{q} - y_2'\frac{a'}{q'})\| \\ &\leq \|y_2'(y_1\frac{a}{q} - y_1'\frac{a'}{q'})\| + \|y_1'(y_2\frac{a}{q} - y_2'\frac{a'}{q'})\|.\end{aligned}$$

Now we use that θ and θ' come with the same β, which satisfies $|\beta| < \frac{2}{QN^{\frac{1}{2}}}$:

$$\|y_1\frac{a}{q} - y_1'\frac{a'}{q'}\| \leq \|y_1\theta - y_1'\theta'\| + |\beta(y_1 - y_1')| \leq \frac{1}{10N^{1-\sigma}} + \frac{2}{QN^{\frac{1}{2}}} \cdot 2N^\sigma.$$

Recall that y_2' is an integer bounded by N^σ. Thus we get

$$\|y_2'(y_1\frac{a}{q} - y_1'\frac{a'}{q'})\| \leq \frac{1}{10N^{1-2\sigma}} + \frac{4N^{2\sigma}}{QN^{\frac{1}{2}}}.$$

The same agument works for $y_2' = y_1'$, $y_1 = y_2$ and $y_1' = y_2'$. We get

$$\|(y_2'y_1 - y_1'y_2)\frac{a}{q}\| \leq \frac{1}{5N^{1-2\sigma}} + \frac{8N^{2\sigma}}{QN^{\frac{1}{2}}}.$$

We conclude that

$$\|(y_2'y_1 - y_1'y_2)a\| \leq \frac{q}{5N} + \frac{8qN^{2\sigma}}{QN^{\frac{1}{2}}} \leq \frac{1}{5N^{\frac{1}{2}}} + \frac{8N^{2\sigma}}{N^{\frac{1}{2}}}.$$

Choosing σ so that $N^\sigma < \frac{1}{4}N^{\frac{1}{4}}$ (i.e. $\sigma < \frac{1}{4} - \frac{\log(4)}{\log(N)}$) we find that

$$\|(y_2'y_1 - y_1'y_2)a\| < 1.$$

Since $(a, q) = 1$ and we are dealing with integers we conclude that

$$y_2'y_1 - y_1'y_2 \equiv 0 \bmod q.$$

Similarly we see that

$$y_2'y_1 - y_1'y_2 \equiv 0 \bmod q'.$$

Let $\tilde{q}$ be the least common multiple of q and q' (i.e. $\tilde{q} = [q, q']$). Then we have $\frac{1}{2}Q \leq \tilde{q} \leq Q^2$ and

$$y_2'y_1 - y_1'y_2 \equiv 0 \bmod \tilde{q}.$$

We will distinguish three cases:

- $y_1y_2' - y_2y_1' \neq 0$ (call this region A_1);
- $y_1y_2' - y_2y_1' = 0$ but $y_1y_2y_1y_2' \neq 0$ (call this region A_2);
- $y_1y_2' - y_2y_1' = 0$ and $y_1y_2y_1y_2' = 0$ (call this region A_3).

Before we estimate this contributions individually let us recall that

$$\sum_{\theta \in P_{Q,\beta}} |S_N(\theta)| \ll N^{(\delta+1)(1-\sigma)}\sqrt{\sharp A} = N^{(\delta+1)(1-\sigma)}\sqrt{\sharp A_1 + \sharp A_2 + \sharp A_3}.$$

Claim 1: The contribution of points with $y_1y_2' - y_2y_1' \neq 0$ (i.e. Case 1) to A is bounded by $\sharp A_1 \ll N^{(1+\delta)2\sigma}Q$.

To see this we argue as follows. First observe that since

$$\tilde{q} \mid (y_2'y_1 - y_1'y_2) \text{ and } 0 \neq |y_2'y_1 - y_1'y_2| \leq 2N^{2\sigma},$$

we must have $\tilde{q} \leq 2N^{2\sigma}$. We obtain the bound

$$\tilde{q} \leq \min(Q^2, 2N^{2\sigma}) \leq \sqrt{2}QN^\sigma.$$

We can now deduce that $\|y_1\frac{a}{q} - y_1'\frac{a'}{q'}\| = 0$. Indeed, by assuming the contrary we obtain the inequality

$$\begin{aligned}\frac{1}{\sqrt{2}QN^\sigma} \leq \frac{1}{\tilde{q}} \leq \|y_1\frac{a}{q} - y_1'\frac{a'}{q'}\| &\leq \|y_1\theta - y_1'\theta'\| + |\beta(y_1 - y_1')| \\ &\leq \frac{1}{10N^{1-\sigma}} + \frac{4N^\sigma}{QN^{\frac{1}{2}}}.\end{aligned}$$

Using $Q < \sqrt{N}$ and $N^{2\sigma} < \frac{1}{16}\sqrt{N}$ we can rewrite this as

$$\frac{1}{\sqrt{2}} \leq \frac{QN^{2\sigma}}{10N} + \frac{4N^{2\sigma}}{\sqrt{N}} \leq \frac{1}{160} + \frac{1}{4},$$

which is a contradiction.

We conclude that

$$y_1\frac{a}{q} \equiv y_1'\frac{a'}{q'} \bmod 1.$$

A similar argument yields

$$y_2\frac{a}{q} \equiv y_2'\frac{a'}{q'} \bmod 1.$$

To put these observations to use we need some more notation. Put $d = (q, q')$ and write $q = dq_1$ as well as $q' = dq_1'$. Of course we have $(q_1, q_1') = 1$. We obtain

$$y_1 a q_1' \equiv y_1' a' q_1 \bmod dq_1q_1'.$$

We can read off the divisibility properties:

$$q_1 \mid y_1,\ q_1 \mid y_2,\ q_1' \mid y_1' \text{ and } q_1' \mid y_2'.$$

At this point we recall that $\mathbf{y} = (y_1, y_2)$ is a primitive vector. Thus $(y_1, y_2) = 1$, forcing $q_1 = 1$ and $d = q$. Similarly we obtain $q_1' = 1$ and $d = q'$ so that $q = q'$. So we can write our congruences as

$$y_1 a \equiv y_1' a' \bmod q \text{ and } y_2 a \equiv y_2' a' \bmod q.$$

Now we are ready to count this contribution to A. Observe that there are $\ll Q$ choices for q and $\ll q^2$ choices for (a, a'). Furthermore there are $\ll N^{2\sigma\delta}$ choices for primitive pairs (y_1, y_2). At this point y_1' and y_2' are determined modulo q. Therefore we have $\ll N^{2\sigma}q^{-2}$ choices for them. Combing this we get

$$\sharp A_1 \ll \sum_{q\sim Q} q^2 N^{2\sigma\delta}\frac{N^{2\sigma}}{q^2} \ll N^{2\sigma(1+\delta)}Q. \tag{27}$$

This proves Claim 1.

Case 2: The contribution of points with $y_1y_2' - y_2y_1' = 0$ but $y_1y_2y_1'y_2' \neq 0$ (i.e. Case 2) to A is bounded by $\sharp A_2 \ll N^{2\sigma\delta+\epsilon}Q^2 + N^{2\sigma\delta+\sigma-1+\epsilon}Q^4$.

In this situation the divisibility condition is vacuous, but we know that y_1, y_2, y_1' and y_2' are all nonzero. Using that $\mathbf{y} = (y_1, y_2)$ and $\mathbf{y}' = (y_1', y_2')$ are primitive and $y_2'y_1 = y_1'y_2$ we find that

$$y_1 = \pm y_1' \text{ and } y_2 = \pm y_2'.$$

Thus there are $\ll N^{2\sigma\delta}$ choices for $\mathbf{y}$ and $\mathbf{y}'$.

Set $q_1 = (y_1, q)$ and $q_1' = (y_1', q') = (y_1, q')$. Since q_1, q_1' both divide y_1 there are $\ll N^\epsilon$ choices for q_1 and q_1'. Without loss of generality we can assume $q_1 \leq q_1'$. We fix a' and q'. There are $\ll \frac{Q^2}{q_1'}$ possibilities to do so.

Write $y_1 = q_1z_1$ and $q = q_1q_2$. Then

$$\|y_1\frac{a}{q} - y_1'\frac{a'}{q'} + \beta(y_1 - y_1')\| < \frac{1}{10N^{1-\sigma}}$$

can be rewritten as

$$\|z_1\frac{a}{q_2}-\psi\|<\frac{1}{10N^{1-\sigma}},$$

where $\psi=y_1'\frac{a'}{q'}-\beta(y_1-y_1')$ is already completely determined. The grid in the unit interval of possible values of $z_1\frac{a}{q_2}$ as a and q_2 vary has mesh of size at least $4q_1^2Q^{-2}$. Hence the set of values of $z_1\frac{a}{q_2}$, which are as close as required to ψ bounded by

$$\ll\frac{Q^2}{q_1^2N^{1-\sigma}}+1.$$

Fix a point $\tilde{\psi}$ in the grid with $z_1\frac{a}{q_2}\equiv\tilde{\psi}$ mod 1. Since $(q_2,a,z_1)=1$ this determines q_2 uniquely. Now also a is determined modulo q_2. Thus we have $q_1=q/q_2$ possible values for a.

Combining everything we can count

$$\begin{aligned}\sharp A_1&\ll\sum_{\mathbf{y}}\nu(y)\sum_{\mathbf{y}'=\mathbf{y}}\sum_{\substack{q_1|y_1\\ q_1'|y_1,\\ q_1<q_1'\leq\min(Q,N^\sigma)}}Q\cdot\frac{Q}{q_1'}\left(1+\frac{Q^2}{q_1^2N^{1-\sigma}}\right)q_1\\&\ll N^{2\sigma\delta+\epsilon}Q^2+N^{2\sigma\delta+\sigma-1+\epsilon}Q^4.\end{aligned}$$

This is exactly Claim 2.

Case 3: The contribution of points with $y_1y_2'-y_2y_1'=0$ and $y_1y_2y_1'y_2'=0$ (i.e. Case 3) to A is bounded by $\sharp A_3\ll N^\sigma Q^2$.

This case is particularly easy. Without loss of generality we assume $y_1=0$. Since $\mathbf{y}$ is primitive we get $y_2=\pm1$. The equation $y_1y_2'=y_2y_1'$, then yields (without loss of generality) $y_1'=0$ and $y_2'=\pm1$. From here we can argue as in the proof of Claim 2 (with $q_1=q_1'=1$) and we find

$$\sharp A_3\ll Q\cdot Q\cdot\left(1+\frac{Q^2}{N^{1-\sigma}}\right)\ll Q^2+Q^4N^{\sigma-1}\ll Q^2+Q^2N^\sigma\ll Q^2N^\sigma.$$

This establishes Claim 3.

Combing the three claims we get

$$\begin{aligned}\sum_{\theta\in P_{Q,\beta}}|S_N((\theta))|&\ll N^{(\delta+1)(1-\sigma)}\sqrt{\sharp A_1+\sharp A_2+\sharp A_3}\\&\ll N^{(\delta+1)(1-\sigma)}\left(N^{2\sigma(1+\delta)}Q+N^{2\sigma\delta+\epsilon}Q^2+N^{2\sigma\delta+\sigma-1+\epsilon}Q^4+N^\sigma Q^2\right)^{\frac{1}{2}}\\&\ll N^{(\delta+1)(1-\sigma)}QN^{\sigma(1+\delta)+\epsilon}\left(Q^{-\frac{1}{2}}+N^{-\sigma}Q^4+N^{-\frac{\sigma}{2}-\frac{1}{2}}Q\right).\end{aligned}$$

This completes the proof. □

Theorem 8.4. *We have*

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \log(N) \left(N^{\delta(1+\sigma)} \|S_N(\theta)|_{W_{Q,K}}\|_\infty + N^{2\delta+1-\sigma} \right).$$

Proof. We first prove the following. Let $\Omega \subset [0,1]$ be a finite union of open intervals. Then

$$\int_\Omega |S_N(\theta)| d\theta \ll \max(N^{(\sigma+1)\delta}, N^{\delta+(1-\sigma)/2} |\Omega|^{\frac{1}{2}}). \tag{28}$$

To see this we rewrite (26) as

$$\int_\Omega |S_N(\theta)| d\theta \ll N^{(\delta+1)(1-\sigma)} \left(\sum_{\mathbf{y},\mathbf{y}'} \nu(\mathbf{y})\nu(\mathbf{y}') \operatorname{Vol}(\{(\theta,\theta') \colon \|\theta y_1 - \theta' y_1'\|, \|\theta y_2 - \theta' y_2'\| < \frac{1}{N^{1-\sigma}}\}) \right)^{\frac{1}{2}}.$$

Put $Y = \begin{pmatrix} y_1 & y_2 \\ -y_1' & -y_2' \end{pmatrix}$. We treat two cases. First, if $\det(Y) \neq 0$, then the map $(\theta,\theta') \mapsto (\theta,\theta')Y$ is a measure preserving map modulo $(1,1)$. We get

$$\operatorname{Vol}(\{(\theta,\theta') \colon \|\theta y_1 - \theta' y_1'\|, \|\theta y_2 - \theta' y_2'\| < \frac{1}{N^{1-\sigma}}\}) \ll N^{2\sigma-2}.$$

Furthermore there are up to $N^{4\sigma\delta}$ choices for $\mathbf{y}$ and $\mathbf{y}'$.

If $\det(Y) = 0$, then we must have $\mathbf{y} = \pm\mathbf{y}'$. In this case there are only $N^{2\sigma\delta}$ choices. Without loss of generality we can assume $y_1 \neq 0$. Fixing $\theta' \in \Omega$ contributes at most $|\Omega|$. Then θ must satisfy $\|y_1\theta - \theta_0\| < \frac{1}{N^{1-\sigma}}$ for some fixed θ_0. Hence we get a contribution of $N^{2\sigma\delta}|\Omega|N^{\sigma-1}$.

Putting both cases together yields

$$\int_\Omega |S_N(\theta)| d\theta \ll N^{(\delta+1)(1-\sigma)}(N^{4\sigma\delta+2\sigma-2} + |\Omega| N^{2\sigma\delta+\sigma-1})^{\frac{1}{2}} \ll N^{\delta(1+\sigma)} + N^{\delta+(1-\sigma)/2}|\Omega|^{\frac{1}{2}}$$

which is exactly (28).

We put $W_{Q,K} = \Omega$. The trivial bound is $S_N(\theta) \ll N^{2\delta}$ and we can take a dyadic subdivision $M \ll N^{2\delta}$ of $\ll \log(N)$ terms and decompose Ω into level sets

$$\Omega = \bigsqcup_\alpha \Omega_\alpha \tag{29}$$

according to the size of $|S_N(\theta)|$. More precisely, if $\theta \in \Omega_\alpha$, then $\frac{M}{2} \leq |S_N(\theta)| < M$ with $M = M_\alpha \ll N^{2\delta}$.

We have the trivial estimate

$$\frac{1}{|\Omega_\alpha|} \int_{\Omega_\alpha} |S_N(\theta)| d\theta \asymp \sup_{\theta \in \Omega_\alpha} |S_N(\theta)|.$$

We will now use this observation and (28). This yields

$$\begin{aligned}\int_\Omega |S_N(\theta)|^2 d\theta &\ll \log(N) \sup_\alpha \int_{\Omega_\alpha} |S_N(\theta)|^2 d\theta \\ &\ll \log(N) \sup_\alpha \sup_{\theta\in\Omega_\alpha} |S_N(\theta)| \int_{\Omega_\alpha} |S_N(\theta)| d\theta \\ &\ll \log(N) \sup_\alpha \sup_{\theta\in\Omega_\alpha} |S_N(\theta)| \max\left(N^{(\sigma+1)\delta}, N^{\delta+(1-\sigma)/2}|\Omega_\alpha|^{\frac{1}{2}}\right) \\ &\ll \log(N) \max\left(N^{(\sigma+1)\delta} \sup_{\theta\in\Omega}|S_N(\theta)|, \sup_\alpha N^{\delta+(1-\sigma)/2}|\Omega_\alpha|^{-\frac{1}{2}} \int_{\Omega_\alpha} |S_N(\theta)| d\theta\right) \\ &\ll \log(N) \max\left(N^{(\sigma+1)\delta} \sup_{\theta\in\Omega}|S_N(\theta)|, \sup_\alpha N^{2\delta+(1-\sigma)}\right) \\ &\ll \log(N) \left(N^{(\sigma+1)\delta}\|S_N(\theta)|_\Omega\|_\infty + N^{2\delta+1-\sigma}\right)\end{aligned}$$

and completes the proof. □

The next step is to gather the estimates above and deduce that the minor arcs are small. We do this for the choices

$$Q = N^\alpha \text{ and } K = N^\kappa.$$

We assume $\alpha, \kappa \in [0, \frac{1}{2}]$. We first deal with those θ for which $\mathfrak{m}(\theta) = 1$ (these are most). We will now derive 3 lemmata each dealing with different situations. The first one works only if K or Q is large. The second lemma deals with both K and Q that are small but still to large to be covered by the major arcs.

Lemma 8.5. *As $N \to \infty$ we have*

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll N^{4\delta-1-\eta},$$

if

$$\begin{aligned}\alpha + \frac{1+\delta}{2}\kappa &> \frac{3}{2}(1-\delta) + \delta\sigma \\ \sigma &> 2(1-\delta) \text{ and} \\ \sigma &< \frac{132\cdot\delta - 131}{96\cdot\delta - 10}.\end{aligned}$$

Proof. Using Theorem 8.4 and then Theorem 8.2 to estimate $\|S_N(\theta)|_{W_{Q,K}}\|_\infty$ we get

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll N^{\delta(\sigma+1)+(3\delta+1)/2}\left(\frac{1}{K^{(1+\delta)/2}} + N^{-\frac{1}{84}(1-2\sigma)(6\delta-5)}\right) + N^{2\delta+1-\sigma}.$$

The assumption $\sigma > 2(1-\delta)$ is needed to control the last term on the right hand side. For the middle term we need $\sigma < \frac{132\cdot\delta-131}{96\cdot-10}$. Finally, $\alpha + \frac{1+\delta}{2}\kappa > \frac{3}{2}(1-\delta) + \delta\sigma$

is needed to control the first term of this bound. (Observe that we require δ to be close to one. This implies that we essentially need $\alpha+\kappa > \sigma$ for this estimate.) □

Lemma 8.6. *As $N \to \infty$ we have*

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll N^{4\delta-1-\eta},$$

for

$$\kappa > \frac{1-\delta}{\delta} \text{ and}$$
$$1-\delta+\kappa+2\alpha < \frac{1}{42}(6\delta-5)(1-2\sigma).$$

Proof. Note that $\mathrm{Vol}(W_{Q,K}) \asymp Q^2\frac{K}{N}$. We apply Theorem 8.2 and get

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll Q^2\frac{K}{N}\left(\frac{N^{3\delta+1}}{K^{1+\delta}Q^2} + N^{-\frac{1}{42}(6\delta-5)(1-2\sigma)}\right)$$
$$= \frac{N^{3\delta}}{K^{\delta}} + Q^2KN^{3\delta-\frac{1}{42}(6\delta-5)(1-2\sigma)}.$$

The conditions follow directly when fitting this bound to our claim. □

Lemma 8.7. *As $N \to \infty$ we have*

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll N^{4\delta-1-\eta},$$

for

$$(1-\delta)\kappa + 3(\delta-1) < \alpha,$$
$$(1-\delta)\kappa/2 + 3(1-\delta)/2 < \sigma,$$
$$(1-\delta)\kappa/2 + 3(1-\delta)/2 + \alpha < (1+\sigma)/2,$$
$$3(1-\delta)/2 + \kappa + \alpha/2 < \frac{1}{84}(6\delta-5)(1-2\sigma),$$
$$3(1-\delta)/2 + \kappa + \alpha < \frac{1}{84}(6\delta-5)(1-2\sigma) + \sigma \text{ and}$$
$$3(1-\delta)/2 + \kappa + 2\alpha < \frac{1}{84}(6\delta-5)(1-2\sigma) + \sigma/2 + \frac{1}{2}.$$

Proof. By applying Theorem 8.2 and Theorem 8.3 we get

$$\int_{W_{Q,k}} |S_N(\theta)|^2 d\theta \ll \sup_{\theta\in W_{K,Q}} |S_N(\theta)| \frac{K}{N} \sum_{\theta\in P_{Q,\beta}} |S_N(\theta)|$$
$$\ll N^{(3\delta+1)/2}\left(\frac{1}{K^{(1+\delta)/2}Q} + N^{-\frac{1}{84}(6\delta-5)(1-2\sigma)}\right)\cdot\frac{K}{N}\cdot N^{\delta+1+\epsilon}Q\left(Q^{-\frac{1}{2}} + N^{-\sigma} + N^{-\sigma/2-\frac{1}{2}}Q\right).$$

Each condition comes from one of the 6 terms that one gets after multiplying out the bound above. □

Lastly we need to estimate the contribution of those θ, where θ is non-trivial. We obtain

$$\int_{\mathfrak{m}(\theta)\neq 1} |\mathfrak{m}(\theta)|^2 |S_N(\theta)|^2 d\theta \ll \sum_{q\leq Q} \sum_{(a,q)=1} \int_{|\beta|<K/N} (\beta N/K)^2 \frac{N^{3\delta+1}}{(N\beta)^{1+\delta}Q^2} d\beta \ll N^{3\delta}K^{-\delta}.$$

Here we used the bound from Theorem 8.2, but only the first term of the estimate is required.

To finish of the minor arc estimate we need to make sense of all the constraints on σ, δ, α and κ. More precisely we need to find values for δ and σ, so that each (α, κ) either lies in the major arcs or satisfies the conditions in one of the lemmatas above.

Theorem 8.8. *Assume that $\delta > 0,9999493550$. Then there exists a choice for σ such that as $N \to \infty$ there is some $\eta > 0$ with*

$$\sum_{|n|<N} |\mathcal{E}_N(n)|^2 = \int_0^1 |\mathfrak{m}(\theta)|^2 |S_N(\theta)|^2 d\theta \ll N^{4\delta-1-\eta}.$$

Proof. This is basic but cumbersome. First, one can plot the region in the (σ, δ) plane in which the following conditions hold:

$$\sigma > 2(1-\delta),$$

$$21((1-\delta)(1-\delta)/\delta + 3(1-\delta)) + 13(1-\delta)/\delta < (2\delta - \frac{5}{3})\sigma \text{ and}$$

$$\sigma < \frac{132\delta - 131}{96\delta - 10}.$$

Solving for the minimal δ in this region one finds that it is the largest root of

$$1020 - 8897x - 5010x^2 - 12888x^3.$$

(The lower bound on δ given in the statement is an upper approximation to this root.) The proof is concluded by looking at all the other conditions to see that together with the major arcs the full (α, κ) plane is covered. □

We can now proof Theorem 1.2:

Theorem 8.9 (Bourgain-Kontorovich 2010)**.** *Let Γ be thin, free, finitely generated with no parabolic elements and assume that $\delta > 1 - 5 \times 10^{-5}$. Then there is $\eta_0 > 0$ so that*

$$\frac{\sharp(S \cap [1, N])}{\sharp(\mathcal{A} \cap [1, N])} = 1 + O(N^{-\eta_0}).$$

This can be read as (a quantitative version of) for almost every n we have that n is represented (by $(\Gamma, \mathbf{v}_0, \mathbf{w}_0)$) if and only if n is admissible.

Proof. Let $\mathfrak{E}(N)$ be the set of exceptions up to N. Let $\mathcal{Z}$ denote those integers passing local obstructions. (For simplicity in our treatment of the major arcs we assumed that $\mathcal{Z} = \mathbb{Z}$.) We now apply Theorem 8.1 and Theorem 8.8 to get

$$\begin{aligned}\sharp\mathfrak{E}(N) &= \sum_{\substack{n\in\mathcal{Z}\cap B_N,\\ |\mathcal{E}_N(n)|>\mathcal{M}_N(n)}} 1\\ &\ll \sum_{\substack{|n|<N,\\ |\mathcal{E}_N(n)|\gg\frac{N^{2\delta-1}}{\log\log(n)\log(N)}}} 1.\\ &\ll \sum_{|n|<N} |\mathcal{E}_N(n)|^2 \log\log(n)\log(N)N^{2-4\delta}\\ &\ll N^{1-\delta}\log(N)\log\log(N).\end{aligned}$$

This finishes the proof. □

9. Proof of Theorem 1.4

In this section we follow [5] to present the proofs of the theorems concerning Zaremba's conjecture. Let us recall (and slightly generalize) the statements.

Fix a finite set $\mathfrak{A} \subset \mathbb{N}$, which we call an alphabet. We write $\mathfrak{C}_{\mathfrak{A}}$ to denote the collection of all $x \in (0,1)$ with continued fraction expansion $x = [a_1, a_2, \ldots]$ with $a_i \in \mathfrak{A}$ for all $i \in \mathbb{N}$. We further write

$$\mathfrak{R}_{\mathfrak{A}} = \{\frac{b}{d} = [a_1, \ldots, a_k] \colon 0 < b < d,\ (b,d) = 1 \text{ and } a_j \in \mathfrak{A} \text{ for } j = 1, \ldots, k\}.$$

Finally let $\mathfrak{D}_{\mathfrak{A}} \subset \mathbb{N}$ denote the set of denominators that appear in $\mathfrak{R}_{\mathfrak{A}}$. Of course we have seen all this if $\mathfrak{A} = \mathbb{N} \cap [1, A]$. In this case the subscript was simply A.

Let $\delta_{\mathfrak{A}}$ denote the Hausdorff dimension of $\mathfrak{C}_{\mathfrak{A}}$. We state the following fixed version of Hensley's generalization of Zaremba's conjecture.[13]

Conjecture 9.1. *If $\delta_{\mathfrak{A}} > \frac{1}{2}$, then the set of denominators $\mathfrak{D}_{\mathfrak{A}}$ contains every sufficiently large admissible integer.*

We will prove the following theorem

Theorem 9.1. *There exists $\delta_0 < 1$ so that, if the dimension $\delta_{\mathfrak{A}}$ exceeds δ_0, then the set of denominators $\mathfrak{D}_{\mathfrak{A}}$ contains almost every admissible integer. More precisely, there is a constant $c = c(\mathfrak{A})$ so that*

$$\frac{\sharp(\mathfrak{D}_{\mathfrak{A}} \cap [N/2, N])}{\sharp(\mathfrak{A}_{\mathfrak{A}} \cap [N/2, N])} = 1 + O(e^{-c\sqrt{\log(N)}}),$$

[13] Hensley formulated this conjecture without the extra admissibility condition. However, Bourgain and Kontorovich observed that $\mathfrak{A} = \{2, 4, 6, 8, 10\}$ satisfies the Hausdorff dimension condition but $\mathfrak{D}_{\mathfrak{A}}$ modulo 4 contains only $0, 1, 2$. Thus integers congruent to 3 modulo 4 can not be represented.

where $\mathfrak{A}_{\mathfrak{A}} = \{d \in \mathbb{Z} \colon d \in \mathfrak{D}_{\mathfrak{A}} \text{ mod } q \text{ for all } q > 1\}$ *is the set of admissible integers. Furthermore, each* d *produced above appears with multiplicity*

$$\gg N^{2\delta_{\mathfrak{A}} - \frac{1001}{1000}}.$$

Remark 9.2. This implies Theorem 1.4 by taking $\mathfrak{A} = \{1, \dots, 50\}$. As noted above there are no local obstructions in this case. Of course one needs to verify that in this case $\delta_{\mathfrak{A}} > \delta_0$. But it can be computed that $\delta_{\mathfrak{A}} \asymp 0.986$, while the proof yields $\delta_0 \asymp 0.984$. Later δ_0 was improved to $5/6$ by Huang allowing $A = 5$. It is unlikely that the methods can be pushed all the way down to $\delta_0 = \frac{1}{2}$.

Recall that $\Gamma_{\mathfrak{A}}$ is the semigroup (freely and finitely) generated by the matrix products

$$\begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & a' \end{pmatrix} \text{ for } a, a' \in \mathfrak{A}.$$

We write $\Gamma = \Gamma_{\mathfrak{A}}$. If $\widetilde{\Gamma}_{\mathfrak{A}}$ is the semigroup in $\mathrm{GL}_2(\mathbb{Z})$ generated by the matrices $\begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix}$ with $a \in \mathfrak{A}$, then we can consider the orbit $\mathcal{O}_{\mathfrak{A}} = \widetilde{\Gamma}_{\mathfrak{A}} \cdot e_2$. We find that

$$\mathfrak{D}_{\mathfrak{A}} = \langle \mathcal{O}_{\mathfrak{A}}, e_2 \rangle.$$

Since we want to work with a semigroup in $\mathrm{SL}_2(\mathbb{Z})$ we observe that

$$\mathcal{O}_{\mathfrak{A}} = \Gamma_{\mathfrak{A}} \cdot e_2 \cup \bigcup_{a \in \mathfrak{A}} \Gamma_{\mathfrak{A}} \cdot \begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix} e_2.$$

This allows us to work without loss of generality with Γ in place of $\widetilde{\Gamma}_{\mathfrak{A}}$.

Remark 9.3. The reduction of Γ modulo q is all of $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ for all q co-prime too a certain bad modulus $\mathfrak{B}$. This follows from Goursat's Lemma and strong approximation as discussed earlier. For the alphabet $\mathfrak{A} = \{1, 2\}$ (and all alphabets containing it), it is easy to see that $\mathfrak{B} = 1$.

For simplicity we assume that the reduction of Γ is full from now on.[14] Given $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma \setminus \{1\}$ one obtains (using induction) that

$$1 \leq a \leq \min(b, c) \leq \max(b, c) < d.$$

Recall that we have the inequalities

$$\frac{1}{2}\|\gamma\| \leq \mathrm{Tr}(\gamma) \leq 2\|\gamma\|$$

for the Frobenius norm. Furthermore we have

$$\|\gamma\|_\infty = d \leq |\gamma e_2| = \sqrt{b^2 + d^2} < \|\gamma\| < 2|\gamma e_2| < 4\|\gamma\|_\infty,$$

[14]This is anyway true for all alphabets containing 1 and 2. The modifications needed for the general case are minor.

for $\gamma \in \Gamma$.

Given $\gamma \in \Gamma$ we denote the expanding eigenvalue by $\lambda(\gamma) = \lambda_+(\gamma)$ and the contracting eigenvalue by $\lambda_-(\gamma) = \frac{1}{\lambda_+(\gamma)}$. The normalized eigenvectors are denoted by $v_+(\gamma)$ and $v_-(\gamma)$ respectively. Of course we have

$$\lambda(\gamma) = \frac{\mathrm{Tr}(\gamma) + \sqrt{\mathrm{Tr}(\gamma)^2 - 4}}{2} > 1.$$

(In particular Γ contains no parabolic elements, since all the eigenvalues are real.)

Proposition 9.4. *Let $\gamma, \gamma' \in \Gamma$. Then we have*

$$\lambda(\gamma\gamma') = \lambda(\gamma)\lambda(\gamma') \left[1 + O\left(|v_+(\gamma) - v_+(\gamma')| + \|\gamma\|^{-2} + \|\gamma'\|^{-2}\right)\right].$$

Furthermore,

$$|v_+(\gamma\gamma') - v_+(\gamma)| \ll \|\gamma\|^{-2} \text{ and } |v_-(\gamma\gamma') - v_-(\gamma')| \ll \|\gamma'\|^{-2}.$$

Proof. First, for large γ we have

$$\lambda(\gamma) = \frac{\mathrm{Tr}(\gamma) + \sqrt{\mathrm{Tr}(\gamma)^2 - 4}}{2} = \mathrm{Tr}(\gamma) + O(\|\gamma\|^{-1}).$$

We can also approximate the eigenvectors by

$$v_+(\gamma) = \frac{(b, \lambda_+(\gamma) - a)}{\sqrt{b^2 + (\lambda_+(\gamma) - a)^2}} = \frac{(b, d)}{\sqrt{b^2 + d^2}} + O(\|\gamma\|^{-2}) \text{ and}$$

$$v_-(\gamma) = \frac{(d - \lambda_-(\gamma), c)}{\sqrt{(d - \lambda_-(\gamma))^2 + c^2}} = \frac{(-d, c)}{\sqrt{c^2 + d^2}} + O(\|\gamma\|^{-2}) \text{ and}$$

Taking the inner product yields

$$|\langle v_+(\gamma), v_-(\gamma)^\top \rangle| = \frac{bc + d^2}{\sqrt{(b^2 + d^2)(c^2 + d^2)}} + O(\|\gamma\|^2) \geq \frac{1}{2},$$

for large γ. Thus the angle between expanding and contracting eigenvector does not degenerate. We will only proof the almost multiplicativity of the (expanding) eigenvalues, since the eigenvector estimates are similar. Note that by our first observation it suffices to look at the traces:

$$\begin{aligned}
|\mathrm{Tr}(\gamma\gamma') - \mathrm{Tr}(\gamma)\,\mathrm{Tr}(\gamma')| &= |(aa' + dc' + cb' + dd') - (a + d)(a' + d')| \\
&\leq \frac{d}{d'}|\frac{bc'd'}{d} - a'd'| + \frac{d'}{d}|\frac{cb'd}{d'} - ad| \\
&\leq \frac{d}{d'}(1 + c'|\frac{bd'}{d} - b'|) + \frac{d'}{d}(1 + c|\frac{b'd}{d'} - b|) \\
&= \frac{d}{d'} + \frac{d'}{d} + (cd' + c'd)|\frac{b}{d} - \frac{b'}{d'}|.
\end{aligned}$$

We can clearly estimate

$$|\frac{b}{d} - \frac{b'}{d'}| \ll |v_+(\gamma) - v_+(\gamma')| + \|\gamma\|^{-2} + \|\gamma'\|^{-2}.$$

We conclude that

$$|\mathrm{Tr}(\gamma\gamma') - \mathrm{Tr}(\gamma)\,\mathrm{Tr}(\gamma')| \ll dd'(|v_+(\gamma) - v_+(\gamma')| + \|\gamma\|^{-2} + \|\gamma'\|^{-2}).$$

which implies the first part of the statement. □

We now set up the (orbital) circle method. We can not simply sum over $\{\gamma \in \Gamma\colon \|\gamma\| \le X\}$, since for this our treatment of the minor arcs does not work. Instead we have to replace this nice set by a more complicated one which we construct now. Given a density point $x \in \mathfrak{C} = \mathfrak{C}_{\mathfrak{A}}$ we let

$$\mathfrak{v} = \mathfrak{v}_x = \frac{(x,1)}{\sqrt{1+x^2}}$$

be the corresponding unit vector. We let N be large enough and let $\delta > \delta_0 = \frac{307}{312}$. Define

$$\mathfrak{b} = \frac{1}{1000}(\delta - \delta_0) > 0.$$

Let $\alpha_0 > 0$ be a parameter chosen later. Put

$$B = N^{\mathfrak{b}} \text{ and } \mathfrak{Q} = e^{\alpha_0\sqrt{\log(N)}}$$

and define

$$\mathcal{U} = \{u_0 = \frac{B}{100} < u \le \frac{99}{100}B\colon u - u_0 \in \frac{2B}{\mathfrak{Q}^5}\cdot\mathbb{N}\} \subset [\frac{1}{100}B, \frac{99}{100}B].$$

The cardinality of $\mathcal{U}$ is roughly $\mathfrak{Q}^4$:

$$\sharp\mathcal{U} \asymp \mathfrak{Q}^5.$$

Proposition 9.5. *For each $u \in \mathcal{U}$, there is a non-empty set $\aleph_u \in \Gamma$, all of the same cardinality (i.e. $\sharp\aleph_u = \sharp\aleph'_u$ for all $u, u' \in \mathcal{U}$), so that the following holds. For every $\mathfrak{a} \in \aleph_u$, its expanding eigenvector is restricted by*

$$|v_+(\mathfrak{a}) - \mathfrak{v}| < \mathfrak{Q}^{-5}$$

and its expanding eigenvalue is restricted by

$$|\lambda(\mathfrak{a}) - u| < \frac{B}{\mathfrak{Q}^5}.$$

In particular $\lambda(\mathfrak{a}) \in (\frac{B}{200}, B)$ for large N. Moreover, for any $q < \mathfrak{Q}$, any $\omega \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ and any $u \in \mathcal{U}$ we have

$$\sharp\{\mathfrak{a} \in \aleph_u\colon \mathfrak{a} \equiv \omega \bmod q\} = \frac{\sharp\aleph_u}{\sharp\,\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}(1 + O(\mathfrak{Q}^{-4})),$$

where the implied constant does not depend on q, ω or u.

We define

$$\aleph = \bigsqcup_{u\in\mathcal{U}} \aleph_u.$$

Proof. We set $R = \sharp \mathrm{SL}_2(\mathbb{Z}/\mathfrak{B}\mathbb{Z})$. Furthermore, define

$$\alpha_0 = \frac{\beta \mathfrak{c}}{40R} \text{ and } T = N^{c_1},$$

for $\mathfrak{c}$ as in Theorem 6.2 and a small parameter c_1 to be determined later. Finally put

$$H = \mathfrak{Q}^{12} \text{ and } H_1 = \mathcal{Q}^6$$

and define the set

$$\mathcal{S}(T) = \{\gamma \in \Gamma \colon |v_+(\gamma) - \mathfrak{v}| < H^{-1}, \, |\lambda(\gamma) - T| < \frac{T}{H_1}\}.$$

We have the crude estimate

$$\sharp \mathcal{S}(T) \gg T^{2\delta} \mathfrak{Q}^{-18} + O(T^{2\delta} e^{-\mathfrak{c}\sqrt{\log(T)}}).$$

As long as

$$c_1 > \left(\frac{3\mathfrak{b}}{4R}\right)^2 \tag{30}$$

we have

$$e^{-\mathfrak{c}\sqrt{\log(T)}} \ll \mathfrak{Q}^{-30}.$$

Thus, applying the pigeon hole principle. we find $\mathfrak{s}_T \in \mathcal{S}(T)$ so that

$$\mathcal{S}'(T) = \{s \in \mathcal{S}(T) \colon s \equiv \mathfrak{s}_T \bmod \mathfrak{B}\} \tag{31}$$

satisfying

$$\sharp \mathcal{S}'(T) \geq \frac{\sharp \mathcal{S}(T)}{\sharp \mathrm{SL}_2(\mathbb{Z}/\mathfrak{B}\mathbb{Z})} \gg T^{2\delta} \mathcal{Q}^{-18}.$$

For this set the counting statement from Corollary 6.3 remains significant. Indeed, for $q < \mathfrak{Q}$ with $\mathfrak{B} \mid q$ and any $w \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ with $w \equiv \mathfrak{s}_T \bmod \mathfrak{B}$, we get

$$\begin{aligned}
&\sharp\{s \in \mathcal{S}'(T) \colon s \equiv w \bmod q\} = \sharp\{s \in \mathcal{S}(T) \colon s \equiv w \bmod q\} \\
&= \frac{\sharp \mathrm{SL}_2(\mathbb{Z}/\mathfrak{B}\mathbb{Z})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \sharp\{s \in \mathcal{S}(T) \colon s \equiv w \equiv \mathfrak{s}_T \bmod \mathfrak{B}\}(1 + O(\mathfrak{Q}^{-6})) + O(T^{2\delta}\mathfrak{Q}^{-30}) \\
&= \frac{\sharp \mathrm{SL}_2(\mathbb{Z}/\mathfrak{B}\mathbb{Z})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \sharp\mathcal{S}'(T)(1 + O(\mathfrak{Q}^{-6})) + O(T^{2\delta}\mathfrak{Q}^{-30}).
\end{aligned}$$

By our observations above the main term is $\gg T^{2\delta}\mathfrak{Q}^{-21}$ and therefor dominates the error.

Since R is the order of $\mathrm{SL}_2(\mathbb{Z}/\mathfrak{B}\mathbb{Z})$ we find that each

$$\gamma \in \mathcal{S}'(T)\mathfrak{s}_T^{R-1}$$

satisfies $\gamma \equiv 1 \bmod \mathfrak{B}$. We pick representatives $\{\gamma_1, \dots, \gamma_R\} = \mathrm{SL}_2(\mathbb{Z}/\mathfrak{B}\mathbb{Z})$ and take $x_1, \dots, x_R \in \Gamma$ so that

$$x_r \equiv y_r \bmod \mathfrak{B} \text{ for } r = 1, \dots, R$$

To ensure the existence of these elements we use the assumption that $\Gamma/\Gamma(q) \cong \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ for all q. We can assume that the x_r are of size $\asymp_{\mathfrak{A}} 1$.

We still need to fix a single auxilliary element $\mathfrak{f}_0 \in \Gamma$ with

$$\lambda(\mathfrak{f}_0) \asymp B^{\frac{1}{100}} \text{ and } |v_+(\mathfrak{f}_0) - \mathfrak{v}| < \mathfrak{Q}^{-6}.$$

We obtain

$$v_+(s \cdot \mathfrak{s}_T^{R-1}\mathfrak{f}_0 x_r) = \mathfrak{v}(1 + O(\mathfrak{Q}^{-6})),$$

for any $s \in \mathcal{S}'(T)$. Furthermore

$$\lambda(s \cdot \mathfrak{s}_T^{R-1}\mathfrak{f}_0 x_r) = T^R \lambda(\mathfrak{f}_0 x_r)(1 + O(\mathfrak{Q}^{-6})).$$

For each $u \in \mathcal{U}$, $u \asymp B$ and each $r = 1, \dots, R$ we take $T = T_{u,r}$ so that

$$T^R \lambda(\mathfrak{f}_0 x_r) = u.$$

This boils down to $T_{u,r} \asymp B^{\frac{99}{100R}} = N^{\frac{99\mathfrak{b}}{100R}}$. This determines c_1 and (30) is easily satisfied.

In summary we have defined sets

$$\mathcal{B}_{u,r} = \mathcal{S}'(T_{ur}) \cdot (\mathfrak{s}_{T_{u,r}})^{R-1}\mathfrak{f}_0 x_r \subset \Gamma$$

for each u and r.For each $\mathfrak{a} \in \mathfrak{B}_{u,r}$ we control the expanding vector

$$|v_+(\mathfrak{a}) - \mathfrak{v}| \ll \mathfrak{Q}^{-6}$$

and the eigenvalue

$$\lambda(\mathfrak{a}) = u(1 + O(\mathfrak{Q}^{-6})).$$

By construction we have, for all $q < \mathfrak{Q}$ with $\mathfrak{B} \mid q$ and all $w \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ with $w \equiv \mathfrak{f}_0 x_r \bmod \mathfrak{B}$, that

$$\sharp\{\mathfrak{a} \in \mathfrak{B}_{u,r} \colon \mathfrak{a} \equiv w \bmod q\} = \frac{\sharp \mathrm{SL}_2(\mathbb{Z}/\mathfrak{B}\mathbb{Z})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \sharp\mathfrak{B}_{u,r}(1 + O(\mathfrak{Q}^{-5})). \tag{32}$$

But we can bound the cardinality of $\mathfrak{B}_{u,r}$ from below by $\gg N^c$.

For fixed u we apply Lemma 6.4 to $\mathfrak{B}_{u,r}$ with $\eta = \mathfrak{Q}^{-5}$ and $q_0 = \mathfrak{B}$. This way we obtain sets $\mathfrak{B}'_{u,r} \subset \mathfrak{B}_{u,r}$ of size $\gg N^c$ for which (32) still holds. Without loss of generality we can assume that the cardinality of $\mathfrak{B}'_{u,r}$ is independent of r. We define

$$\tilde{\aleph}_u = \bigsqcup_{r=1}^{R} \mathcal{B}'_{u,r}.$$

For $\mathfrak{B} \mid q < \mathfrak{Q}$ and $w \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ we record that

$$\sharp\{\mathfrak{a} \in \tilde{\aleph}_u \colon \mathfrak{a} \equiv w \bmod q\} = \frac{\sharp \mathrm{SL}_2(\mathbb{Z}/\mathfrak{B}\mathbb{Z})}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \sharp\mathfrak{B}_{u,r}(1 + O(\mathfrak{Q}^{-5})).$$

Here r is given by $w \equiv \mathfrak{f}_0 x_r \bmod \mathfrak{B}$. Recall that $\sharp\mathfrak{B}'_{u,r} = \frac{\sharp\tilde{\aleph}_u}{R}$ so that we have

$$\sharp\{\mathfrak{a} \in \tilde{\aleph}_u \colon \mathfrak{a} \equiv w \bmod q\} = \frac{\sharp\tilde{\aleph}_u}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}(1 + O(\mathfrak{Q}^{-5})).$$

Finally we drop the condition $\mathfrak{B} \mid q$ by summing over certain arithmetic progressions. This is possible since $\mathfrak{B} \ll_{\mathfrak{A}} 1$. This allows us to apply Lemma 6.4 again to $\tilde{\aleph}_0$ with $\eta = \mathfrak{Q}^{-5}$ and $q_0 = 1$. We obtain the desired sets

$$\aleph_u \subset \tilde{\aleph}_0.$$

Obviously we can ensure that they all have the same cardinality. □

Proposition 9.6. *Given $M \gg 1$ and $H = e^{c\sqrt{\log(M)}}$, there exists $\frac{1}{4}M \leq L \leq 4M$, an integer $k \asymp \log(M)$ and a set $\Xi = \Xi(M, H; L, k) \subset \Gamma$ with the following properties. For all $\gamma \in \Xi$, the expanding eigenvalue satisfies*

$$L(1 - \frac{1}{\log(L)}) < \lambda(\gamma) < L,$$

the expanding eigenvector is controlled by

$$|v_+(\gamma) - \mathfrak{v}| < \frac{1}{H},$$

and the wordlength metric l (in our usual generators of Γ) is exactly

$$l(\gamma) = k.$$

Furthermore, we have

$$\frac{L^{2\delta}}{H \log(L)^2} \ll \sharp\Xi \ll L^{2\delta}.$$

Proof. The set is constructed using the following algorithm:

(1) Let $S_1 \subset \Gamma$ be defined by

$$S_1 = \{\gamma \in \Gamma \colon \frac{M}{2} < \|\gamma\| < M,\ |v_+(\gamma) - \mathfrak{v}| < H^{-1}\}.$$

By Proposition 7.9 we have $\sharp S_1 \gg M^{2\delta} H^{-1}$.

(2) Note that the expanding eigenvalue of $\gamma \in S_1$ can be bounded by $\frac{1}{4}M \leq \lambda(\gamma) \leq 4M$. By the pigeon hole principle we find an L in this range so that

$$\underbrace{\sharp\{\gamma \in S_1 \colon L(1 - \frac{1}{\log(L)}) < \lambda(\gamma) < L\}}_{=S_2(L)=S_2} \gg \frac{L^{2\delta}}{H \log(L)}.$$

(3) Finally we note that $l(\gamma) \asymp \log(\|\gamma\|)$, with implied constants only depending on $\mathfrak{A}$. (In other words the metrics l is commensurable with the archimedean one.) We can use the pigeon hole principle again to find some k with

$$\underbrace{\sharp\{\gamma \in S_2 \colon l(\gamma) = k\}}_{=\Xi(M,H;L,k)=\Xi} \gg \frac{L^{2\delta}}{H \log(L)^2}.$$

□

We will now use the set $\aleph$ and the sets $\Xi(M,H;L,k)$ to construct an ensemble Ω_N that is used to set up the circle method. We start by taking

$$M=\sqrt{N}/B=N^{\frac{1}{2}-\mathfrak{b}} \text{ and } H=\mathfrak{Q}^5.$$

Note that the parameter α_0 was chosen sufficiently small so that $H<e^{c\sqrt{\log(M)}}$. Using the previous result we find sets $\widetilde{\Xi}_1=\Xi(M,H;L,k)$ with $L=\alpha_1 M$ for $\alpha_1\in(\frac{1}{4},4)$. Set

$$\widetilde{N}_1=L=\alpha_1 N^{\frac{1}{2}-\mathfrak{b}},\ N_1=B\widetilde{N}_1=\alpha_1 N^{\frac{1}{2}} \text{ and } \Xi_1=\aleph\widetilde{\Xi}_1.$$

Note that the representation of an element in Ξ_1 as a product of elements in $\aleph$ and $\widetilde{\Xi}_1$ is unique since the wordlength of elements in $\widetilde{\Xi}_1$ is fixed. We have the easy estimate

$$\sharp\Xi_1\geq\sharp\widetilde{\Xi}_1\gg\widetilde{N}_1^{2\delta-\epsilon}\gg N^{\delta-2\delta\mathfrak{b}-\epsilon}.$$

The first step is now to (re)-set

$$M=\frac{N_1^{\frac{1}{2}}}{\alpha_1}=\frac{N^{\frac{1}{4}}}{\sqrt{\alpha_1}} \text{ and } H=\log(M).$$

Generate the set $\Xi_2=\Xi(M,H;L,k)$. Further set $N_2=L=\alpha_2 M$, where L is the newly obtained parameter while constructing $\Xi(M,H;L,k)$ and $\alpha_2\in(\frac{1}{4},4)$. We have

$$\sharp\Xi_2\gg\frac{N_2^{2\delta}}{\log(N_2)^3}.$$

We now iterate the first step as follows. We start with $j=3$ and iterate up to $j=J-1$, where $2^{J-1}=c\log(N)$. (The constant c will be determined in a bit but is independent of N.) For each j we set

$$M=\frac{N_{j-1}^{\frac{1}{2}}}{\alpha_{j-1}}=\frac{N^{2^{-j}}}{\alpha_{j-1}^{\frac{1}{2}}\alpha_{j-2}^{\frac{1}{4}}\cdots\alpha_1^{2^{-j+1}}} \text{ and } H=\log(M).$$

We then generate the set $\Xi_j=\Xi(M,H;L,k)$ and set $N_j=L=\alpha_j M$ for $\alpha_j\in(\frac{1}{4},4)$. Note that

$$\sharp\Xi_j\gg\frac{N_j^{2\delta}}{\log(N_j)^3} \text{ and } \frac{1}{16}N^{2^{-j}}<N_j<16N^{2^{-j}}.$$

The final set (i.e. $j=J$) is constructed with

$$M=\frac{N_{J-1}}{\alpha_{J-1}^2} \text{ and } H=\log(M).$$

As before we take $\Xi_J=\Xi(M,H;L,k)$ and define

$$N_J=L=\frac{\alpha_J N^{2^{-J+1}}}{\alpha_{J-1}\cdots\alpha_1^{2^{-J+1}}}\asymp N^{2^{-J+1}}=e^{\frac{1}{c}}\ll 1.$$

Finally note that $\frac{1}{4} < N_J/M = \alpha_J < 4$, so that

$$\frac{1}{4} < \frac{N_1 \cdots N_J}{N} = \frac{B\widetilde{N}_1 N_2 \cdots N_J}{N} < 4.$$

Set

$$\Omega_N = \Xi_1 \cdots \Xi_J = \aleph \cdot \widetilde{\Xi}_1 \cdot \Xi_2 \cdots \Xi_J.$$

We will now record some properties of Ω_N that follow directly from its construction. For $\gamma \in \Omega_N$ writhe

$$\Gamma = \mathfrak{a}\widetilde{\xi}_2 \xi_2 \cdots \xi_J \text{ with } \mathfrak{a} \in \aleph,\ \widetilde{\xi}_1 \in \widetilde{\Xi}_1 \text{ and } \xi_j \in \Xi_j \text{ for } 2 \le j \le J.$$

Due to the wordlength restrictions this decomposition is unique.

Lemma 9.7. *For any $2 \le j_1 \le j_2 \le J$, any $\xi_j \in \Xi_j$ for $j_1 \le j \le j_2$ and any $\mathfrak{a} \in \aleph, \widetilde{\xi}_1 \in \widetilde{\Xi}_1$ we have the following inequalities*

$$\begin{aligned}
&|v_+(\widetilde{\xi}_1 \cdot \xi_2 \cdots \xi_J) - \mathfrak{v}| \ll \mathfrak{Q}^{-5},\\
&\frac{1}{2} < \frac{\lambda(\xi_{j_1} \cdots \xi_{j_2})}{N_{j_1} \cdots N_{j_2}} < 2,\\
&\frac{1}{2} < \frac{\lambda(\widetilde{\xi}_1 \xi_2 \cdots \xi_{j_2})}{\widetilde{N}_1 N_2 \cdots N_{j_2}} < 2 \text{ and}\\
&\frac{1}{2} < \frac{\lambda(\mathfrak{a}\widetilde{\xi}_1 \xi_2 \cdots \xi_{j_2})}{\lambda(\mathfrak{a})\widetilde{N}_1 N_2 \cdots N_{j_2}} < 2.
\end{aligned}$$

Proof. We first look at the eigenvector estimate. Observe that

$$|v_+(\widetilde{\xi}_1 \cdot \xi_2 \cdots \xi_J) - \mathfrak{v}| \le |v_+(\widetilde{\xi}_1 \cdot \xi_2 \cdots \xi_J) - v_+(\widetilde{\xi}_1)| + |v_+(\widetilde{\xi}_1 - \mathfrak{v}| \ll \|\widetilde{\xi}_1\|^{-2} + \mathfrak{Q}^{-5}.$$

This establishes the first estimate. We will only prove the first (second in total) of the three eigenvalue estimates, since the remaining two are very similar to establish. First observe that as above we get

$$|v_+(\xi_{j_1} \cdots \xi_{j_2}) - \mathfrak{v}| \ll \frac{1}{\log(N_j)}.$$

We claim that

$$\lambda(\xi_{j_1} \cdots \xi_{j_2}) = N_{j_1} \cdots N_{j_2} \cdot \left(1 + O(\frac{1}{\log(N_{j_1})} + \ldots + \frac{1}{\log(N_{j_2})})\right).$$

This is obvious for $j_1 = j_2$ and if $j_1 = j_2 - 1$ we have

$$\begin{aligned}
\lambda(\xi_{j_2-1}\xi_{j_2}) &= \lambda(\xi_{j_2-1})\lambda(\xi_{j_2}) \cdot \left(1 + O(|v_+(\xi_{j_2-1}) - v_+(\xi_{j_2})| + \|\xi_{j_2-1}\|^{-2} + \|\xi_{j_2}\|^{-2})\right)\\
&= N_{j_2-1} \cdot N_{j_2} \cdot \left(1 + O(\frac{1}{\log(N_{j_2-1})} + \frac{1}{\log(N_{j_2})})\right).
\end{aligned}$$

The general case follows by induction using the estimate

$$\begin{aligned}&\lambda(\xi_{j_1}\cdots\xi_{j_2})\\&=\lambda(\xi_{j_1})\lambda(\xi_{j_1+1}\cdots\xi_{j_2})\cdot\big(1+O(|v_+(\xi_{j_1})-v_+(\xi_{j_1+1}\cdots\xi_{j_2})|+\|\xi_{j_1}\|^{-2}+\|\xi_{j_1+1}\cdots\xi_{j_2}\|^{-2})\big)\\&=N_{j_1}\cdot)\lambda(\xi_{j_1+1}\cdots\xi_{j_2})\cdot\left(1+O(\frac{1}{\log(N_{j_1})}+\frac{1}{\log(N_{j_1+1})})\right).\end{aligned}$$

We can rewrite this as

$$\lambda(\xi_{j_1}\cdots\xi_{j_2})=N_{j_1}\cdots N_{j_2}\cdot\left(1+O\left(\frac{2^J}{\log(N)}\right)\right).$$

The claim follows when taking the constant c (remember earlier ...) sufficiently small. □

A direct corollary of this is the bound on the (archimedean) norms:

$$\|\gamma\|\le 2\lambda(\gamma)\le 16N.$$

We can also estimate the size of Ω_N as follows:

$$\sharp\Omega_N=\sharp\Xi_1\cdots\sharp\Xi_J\gg\widetilde{N}_1^{2\delta-\epsilon}\cdot\frac{N_2^{2\delta}}{\log(N_2)^3}\cdots\frac{N_J^{2\delta}}{\log(N_J)^3}\gg N^{2\delta-2\delta\mathfrak{b}-\epsilon}.$$

This ensures that the set Ω_N is not to small. Of course we also have

$$\sharp\Xi_j\cdots\sharp\Xi_J\gg(N_j\cdots N_J)^{2\delta}e^{-c(J-j)\log\log(N_j)}.$$

We define the all important trigonometric polynomial

$$S_N(\theta)=\sum_{\gamma\in\Omega_N}e(\theta\langle\gamma e_2,e_2\rangle).$$

The major arcs (of level $\mathfrak{Q}$) are defined by

$$\mathfrak{M}_{\mathfrak{Q}}=\bigsqcup_{q<\mathfrak{Q}}\bigsqcup_{(a,q)=1}\left[\frac{a}{q}-\frac{\mathfrak{Q}}{N},\frac{a}{q}+\frac{\mathfrak{Q}}{N}\right].$$

We define

$$\nu_q(a)=\frac{1}{\sharp\,\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}\sum_{\omega\in\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}e\left(\frac{a}{q}\langle\omega e_2,e_2\rangle\right).$$

Theorem 9.8. *There exists a function $\varpi_N\colon\mathbb{R}/\mathbb{Z}\to\mathbb{C}$ (given explicitly below) such that*

(1) *The Fourier transform $\widehat{\varpi}_N\colon\mathbb{Z}\to\mathbb{C}$ given by*

$$\widehat{\varpi}_N(n)=\int_0^1\varpi_N(\theta)e(-n\theta)d\theta$$

is real-valued and nonnegative, with

$$\varpi_N(0)=\sum_n\widehat{\varpi_N}(n)\ll\sharp\Omega_N.$$

(2) *For* $\frac{1}{15}N < n < \frac{1}{5}N$ *we have*

$$\widehat{\varpi}_N(n) \gg \frac{\sharp\Omega_N}{N}.$$

(3) *On the major arcs* $\theta = \frac{a}{q} + \beta \in \mathfrak{M}_{\mathfrak{Q}}$ *we have*

$$S_N(\frac{a}{q} + \beta) = \nu_q(a)\varpi_N(\beta)(1 + O(\mathfrak{Q}^{-4})).$$

Proof. We write $\Omega_N = \aleph\Omega'$ with $\Omega' = \widetilde{\Xi}_1\Xi_2 \cdots \Xi_J$. This lets us decompose

$$S_N(\theta) = \sum_{\mathfrak{a}\in\aleph}\sum_{\gamma\in\Omega'} e(\theta\langle\mathfrak{a}\gamma e_2, e_2\rangle).$$

Recall the properties

$$|v_+(\mathfrak{a}) - \mathfrak{v}| < \mathfrak{Q}^{-5}, |v_+(\gamma) - \mathfrak{v}| \ll \mathfrak{Q}^{-5}$$

and $\lambda(\mathfrak{a}) \asymp B$.

We have $(x, y)^\perp = (-y, x)$. Furthermore one has the elementary decomposition

$$w = \frac{\langle w, v_-^\perp\rangle}{\langle v_+, v_-^\perp\rangle}v_+ + \frac{\langle w, v_+^\perp\rangle}{\langle v_-, v_+^\perp\rangle}v_-$$

for any linearly independent $v_+, v_- \in \mathbb{R}^2$ and $w \in \mathbb{R}^2$. We deduce that for any unit vector $w \in \mathbb{R}^2$ and any large $\xi \in \Gamma$ we have

$$\xi w = \lambda(\xi)\frac{\langle w, v_-^\perp(\xi)\rangle}{\langle v_+(\xi), v_-^\perp(\xi)\rangle}v_+(\xi)\left(1 + O(\frac{1}{\|\xi\|^2})\right).$$

We obtain

$$\langle\xi e_2, e_2\rangle = \lambda(\xi)\frac{\langle w, v_-^\perp(\xi)\rangle}{\langle v_+(\xi), v_-^\perp(\xi)\rangle}\langle v_+(\xi), e_2\rangle\left(1 + O(\frac{1}{\|\xi\|^2})\right).$$

Applying this our present situation gives

$$\langle\gamma e_2, e_2\rangle = \lambda(\gamma)\frac{\langle e_2, v_-^\perp(\gamma)\rangle}{\langle v_+(\gamma), v_-^\perp(\gamma)\rangle}\langle\mathfrak{v}, e_2\rangle\left(1 + O(\frac{1}{\mathfrak{Q}^5})\right).$$

Similarly we get

$$\begin{aligned}\langle\mathfrak{a}\gamma e_2, e_2\rangle &= \lambda(\mathfrak{a}\gamma)\frac{\langle e_2, v_-^\perp(\mathfrak{a}\gamma)\rangle}{\langle v_+(\mathfrak{a}\gamma), v_-^\perp(\mathfrak{a}\gamma)\rangle}\langle v_+(\mathfrak{a}\gamma), e_2\rangle\left(1 + O(\frac{1}{N^2})\right)\\ &= \lambda(\mathfrak{a})\lambda(\gamma)\frac{\langle e_2, v_-^\perp(\gamma)\rangle}{\langle v_+(\gamma), v_-^\perp(\gamma)\rangle}\langle\mathfrak{v}, e_2\rangle\left(1 + O(\frac{1}{\mathfrak{Q}^5})\right).\end{aligned}$$

Combining the last two equations yields

$$\langle\mathfrak{a}\gamma e_2, e_2\rangle = \lambda(\mathfrak{a})\langle\gamma e_2, e_2\rangle + O(N\mathfrak{Q}^{-5}).$$

We turn towards computing $S_N(\theta)$ for $\theta = \frac{a}{q} + \beta \in \mathfrak{M}_{\mathfrak{Q}}$ with $|\beta| < \frac{\mathfrak{Q}}{N}$. From the definition we get

$$\begin{aligned} S_N(\frac{a}{q}+\beta) &= \sum_{\mathfrak{a}\in\aleph}\sum_{\gamma\in\Omega'} e(\frac{a}{q}\langle \mathfrak{a}\gamma e_2, e_2\rangle) e(\beta\langle \mathfrak{a}\gamma e_2, e_2\rangle) \\ &= \sum_{\mathfrak{a}\in\aleph}\sum_{\gamma\in\Omega'} e(\frac{a}{q}\langle \mathfrak{a}\gamma e_2, e_2\rangle) e(\beta\lambda(\mathfrak{a})\langle \gamma e_2, e_2\rangle) + O(\mathfrak{Q}^{-4}\sharp\Omega_N) \\ &= \sum_{\gamma\in\Omega'}\sum_{\omega\in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} e(\frac{a}{q}\langle \omega\gamma e_2, e_2\rangle) \sum_{\substack{\mathfrak{a}\in\aleph,\\ \mathfrak{a}\equiv\omega \bmod q}} e(\beta\lambda(\mathfrak{a})\langle \gamma e_2, e_2\rangle) + O(\mathfrak{Q}^{-4}\sharp\Omega_N) \end{aligned}$$

We take a closer look at the inner most sum, using the construction of $\aleph$ in terms ot the $\aleph_u$'s:

$$\begin{aligned} \sum_{\substack{\mathfrak{a}\in\aleph,\\ \mathfrak{a}\equiv\omega \bmod q}} e(\beta\lambda(\mathfrak{a})\langle \gamma e_2, e_2\rangle) &= \sum_{u\in\mathcal{U}} \sum_{\substack{\mathfrak{a}\in\aleph_u,\\ \mathfrak{a}\equiv\omega \bmod q}} e(\beta\lambda(\mathfrak{a})\langle \gamma e_2, e_2\rangle) \\ &= \sum_{u\in\mathcal{U}} e(\beta u\langle \gamma e_2, e_2\rangle)\cdot \sharp\{\mathfrak{a}\in\aleph_u\colon \mathfrak{a}\equiv\omega \bmod q\}\cdot\left(1+O(\mathfrak{Q}^{-4})\right). \end{aligned}$$

The cardinality of all $\aleph_u$ is the same and by construction we have

$$\sharp\{\mathfrak{a}\in\aleph_u\colon \mathfrak{a}\equiv\omega \bmod q\} = \frac{\sharp\aleph}{\sharp\mathcal{U}\cdot\sharp\,\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}(1+O(\mathfrak{Q}^{-4})).$$

The implied constant is absolute.

Inserting this above yields

$$S_N(\frac{a}{q}+\beta) = \underbrace{\frac{1}{\sharp\,\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}\sum_{\omega\in\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} e(\frac{a}{q}\langle \omega e_2, e_2\rangle)}_{=\nu_q(a)} \frac{\sharp\aleph}{\sharp\mathcal{U}} \sum_{\gamma\in\Omega'}\sum_{u\in\mathcal{U}} e(\beta u\langle \gamma e_2, e_2\rangle)\cdot(1+O(\mathfrak{Q}^{-4})).$$

It remains to polish the archimedean contribution. For now consider γ and u fixed. If $m\in\mathbb{Z}$ satisfies

$$|m - u\langle \gamma e_2, e_2\rangle| \le B\langle \gamma e_2, e_2\rangle\cdot\mathfrak{Q}^{-5},$$

we have

$$e(\beta u\langle \gamma e_2, e_2\rangle) = e(\beta m)(1+O(\mathfrak{Q}^{-4})).$$

There are $2B\langle \gamma e_2, e_2\rangle\mathfrak{Q}^{-5} + O(1)$ integers m in this range. We obtain

$$e(\beta u\langle \gamma e_2, e_2\rangle) = \frac{\mathfrak{Q}^5}{2B\langle \gamma e_2, e_2\rangle} \sum_{\substack{m\in\mathbb{Z}\\ |\frac{m}{\langle \gamma e_2, e_2\rangle}-u|\le B\mathfrak{Q}^{-5}}} e(\beta m)(1+O(\mathfrak{Q}^{-4})).$$

With this in mind we set

$$\varpi_N(\beta) = \frac{\sharp\aleph}{\sharp\mathcal{U}}\sum_{\gamma\in\Omega'}\frac{\mathfrak{Q}^5}{2B\langle \gamma e_2, e_2\rangle}\sum_{m\in\mathbb{Z}} e(\beta m)\sum_{u\in\mathcal{U}} \mathbb{1}_{|\frac{m}{\langle \gamma e_2, e_2\rangle}-u|\le B\mathfrak{Q}^{-5}}.$$

This gives us

$$S_N(\frac{a}{q}+\beta) = \nu_q(a)\varpi_N(\beta)(1+O(\mathfrak{Q}^{-4}))$$

as required. It remains to verify some properties of $\varpi_N(\beta)$.

The Fourier transform of ϖ_N is given by

$$\widehat{\varpi}_N(n) = \frac{\sharp\aleph}{\sharp\mathcal{U}} \sum_{\gamma\in\Omega'} \frac{\mathfrak{Q}^5}{2B\langle\gamma e_2, e_2\rangle} \sum_{u\in\mathcal{U}} \mathbb{1}_{|\frac{n}{\langle\gamma e_2,e_2\rangle}-u|\leq B\mathfrak{Q}^{-5}}$$

and clearly has the desired properties.

We can estimate

$$\frac{N}{4B} < \langle\gamma e_2, e_2\rangle < \frac{4N}{B}.$$

Thus for $\frac{N}{25} < n < \frac{N}{5}$ we have

$$\frac{B}{100} < \frac{n}{\langle\gamma e_2, e_2\rangle} < \frac{99}{100}B.$$

By the spacing in $\mathcal{U}$ the innermost sum has at least one contribution. We obtain the lower bound

$$\widehat{\varpi}_N(n) \gg \frac{\sharp\aleph}{\sharp\mathcal{U}} \sum_{\gamma\in\Omega'} \frac{\mathfrak{Q}^5}{2B\langle\gamma e_2, e_2\rangle} \gg \frac{\sharp\aleph\cdot\sharp\Omega'}{N} = \frac{\sharp\Omega_N}{N}.$$

This completes the proof. □

This is all we need to treat the major arc contribution. Recall the triangle function defined in (21). We rescale this function to

$$\psi_N(x) = \psi(\frac{N}{\mathfrak{Q}}x)$$

and make it periodic by averaging:

$$\Psi_N(\theta) = \sum_{m\in\mathbb{Z}} \psi_N(\theta+m).$$

The desired function, which has support on the major arcs, is

$$\Psi_{\mathfrak{Q},N}(\theta) = \sum_{q<\mathfrak{Q}} \sum_{(a,q)=1} \Psi_N(\theta-\frac{a}{q}).$$

The set up is now very analogous to the one in the proof of Theorem 1.2. The representation number is

$$R_N(n) = \widehat{S}_N(n) = \int_0^1 S_N(\theta)e(-n\theta),$$

We split $R_N(n) = \mathfrak{M}_N(n) + \mathfrak{E}_N(n)$ into major arcs and minor arcs (or error) by setting

$$\mathfrak{M}_N(n) = \int_0^1 \Psi_{\mathfrak{Q},N}(\theta)S_N(\theta)e(-n\theta)d\theta.$$

Of course we then must have

$$\mathfrak{E}_N(n) = \int_0^1 (1 - \Psi_{\mathfrak{Q},N}(\theta))S_N(\theta)e(-n\theta)d\theta.$$

Theorem 9.9. *For $\frac{1}{20}N \leq n \leq \frac{1}{10}N$ we have*

$$\mathfrak{M}_N(n) \gg \frac{\sharp\Omega_N}{N \cdot \log\log(N)}.$$

Proof. Fix $n \in \frac{1}{20}[N, 2N]$. We first rewrite the major arcs as

$$\begin{aligned}\mathfrak{M}_N(n) &= \sum_{q<\mathfrak{Q}} \sum_{(a,q)=1} \int_0^1 \Psi_N(\theta - \frac{a}{q})S_N(\theta)e(-n\theta)d\theta \\ &= \sum_{q<\mathfrak{Q}} \sum_{(a,q)=1} e(-n\frac{a}{q}) \int_0^1 \Psi_N(\beta)S_N(\frac{a}{q} + \beta)e(-n\beta)d\beta. \qquad (33)\end{aligned}$$

Inserting the approximation for $S_N(\frac{a}{q} + \beta)$ on major arcs yields the following:

$$\mathfrak{M}_N(n) = \sum_{q<\mathfrak{Q}} \sum_{(a,q)=1} \nu_q(a)e(-n\frac{a}{q}) \int_0^1 \Psi_N(\beta)\varpi_N(\beta)e(-n\beta)d\beta + O(\mathfrak{Q}\mathfrak{Q}\frac{\mathfrak{Q}}{N}(\sharp\Omega_N)\mathfrak{Q}^{-4}).$$

Note that this expression is already split in archimedean and modular component:

$$\mathfrak{M}_N(n) = \mathfrak{S}_{\mathfrak{Q}}(n) \cdot \Pi_N(n) + O(\frac{\sharp\Omega_N}{N\mathfrak{Q}}),$$

for

$$\mathfrak{S}_{\mathfrak{Q}}(n) = \sum_{q<\mathfrak{Q}} \sum_{(a,q)=1} \nu_q(a)e(-n\frac{a}{q})$$

and

$$\Pi_N(n) = \int_0^1 \Psi_N(\beta)\varpi_N(\beta)e(-n\beta)d\beta.$$

The singular series already appeared in (22). The treatment there showed that

$$\mathfrak{S}_{\mathfrak{Q}}(n) \gg \log\log(n)^{-1}.$$

Turning towards the singular integral we first observe that

$$\Pi_N(n) = \sum_{m\in\mathbb{Z}} \widehat{\Psi}_N(n-m)\widehat{\varpi}_N(m) = \frac{\mathfrak{Q}}{N} \sum_{m\in\mathbb{Z}} \widehat{\psi}(\frac{\mathfrak{Q}}{N}(n-m))\widehat{\varpi}_N(m).$$

Recall that $\widehat{\psi}$ is positive and satisfies $\widehat{\psi}(y) > \frac{2}{5}$ for $|y| \leq \frac{1}{2}$. We get

$$\Pi_N(n) \geq \frac{2\mathfrak{Q}}{5N} \sum_{|m-n|<\frac{N}{2\mathfrak{Q}}} \widehat{\varpi}_N(m).$$

For N sufficiently large the summation condition forces $\mathfrak{m}N \in [\frac{1}{25}, \frac{1}{5}]$. Using the properties of $\widehat{\varpi}_N$ in this range gives

$$\Pi_N(n) \gg \frac{\mathfrak{Q}}{N} \cdot \frac{N}{2\mathfrak{Q}} \cdot \frac{\sharp\Omega_N}{N} \gg \frac{\sharp\Omega_N}{N}.$$

This completes the proof. □

We turn now to the treatment of the minor arcs. As usual this will be quite involved. The goal is to estimate

$$\sum_{n\in\mathbb{Z}} |\mathfrak{E}_N(n)|^2 = \int_0^1 |1 - \Psi_{\mathfrak{Q},N}(\theta)|^2 |S_N(\theta)|^2 d\theta.$$

We consider the regions

$$W_{Q,K} = \left\{ \theta = \frac{a}{q} + \beta \colon \frac{Q}{2} \leq q < Q,\ (a,q) = 1,\ \frac{K}{2N} \leq |\beta| < \frac{K}{N} \right\}.$$

The parameters K and Q will vary dyadically in the ranges

$$Q < N^{\frac{1}{2}} \text{ and } K < \frac{N^{\frac{1}{2}}}{Q}.$$

Note that if $K \ll 1$, then we interpret the condition on β as $|\beta| \ll N^{-1}$.

Proposition 9.10. *Let N, Q, K be as above and write $\theta = \frac{a}{b} + \beta \in W_{Q,K}$. Then*

$$|S_N(\theta)| \ll N^{2\delta} \left(\frac{N^{1-\delta}}{KQ} \right)$$

as $N \to \infty$.

Proof. We write

$$\Omega_N = \Xi_1 \cdot \Omega' \text{ for } \Omega' = \Xi_2 \cdots \Xi_J.$$

Recall that for $\gamma \in \Xi_1$ and $\omega \in \Omega'$ we have

$$|\gamma^\top e_2|, |\omega e_2| < 50 N^{\frac{1}{2}}.$$

Furthermore, $\sharp\Xi_1, \sharp\Omega' \ll N^\delta$.

We rewrite $S_N(\theta)$ as follows

$$S_N(\theta) = \sum_{\mathbf{x}\in\mathbb{Z}^2} \sum_{\mathbf{y}\in\mathbb{Z}^2} \mu(\mathbf{x})\nu(\mathbf{y}) e(\theta\langle \mathbf{x}, \mathbf{y}\rangle),$$

where μ and ν are (counting) measures defined by

$$\mu(\mathbf{x}) = \sum_{\gamma\in\Xi_1} \mathbb{1}_{\mathbf{x}=\gamma^\top e_2} \text{ and } \mu(\mathbf{x}) = \sum_{\omega\in\Omega'} \mathbb{1}_{\mathbf{y}=\omega e_2}.$$

The projection $\omega \to \omega \cdot e_2$ is one-to-one. (This is so because the continued fraction of a rational number, if restricted to have even length, is unique.) Similarly the map $\gamma \mapsto \gamma^\top \cdot e_2$ is one-to-one. Thus we have

$$\|\mu\|_\infty, \|\nu\|_\infty \leq 1.$$

At this point we decompose ν into (say 100000) blocks

$$\nu = \sum_\alpha \nu^{(\alpha)}$$

so that if $\mathbf{y}, \mathbf{y}' \in \mathrm{supp}(\nu^\alpha)$, then we have

$$|\mathbf{y} - \mathbf{y}'| < \frac{1}{2} N^{\frac{1}{2}}.$$

We bound $|S_N(\theta)| \leq \sum_\alpha |S_N^{(\alpha)}(\theta)|$ with

$$S_N^{(\alpha)}(\theta) = \sum_{\mathbf{x} \in \mathbb{Z}^2} \sum_{\mathbf{y} \in \mathbb{Z}^2} \mu(\mathbf{x}) \nu^{(\alpha)}(\mathbf{y}) e(\theta \langle \mathbf{x}, \mathbf{y} \rangle).$$

We will estimate each $S_N^{(\alpha)}$ independent of α obtaining the desired bound for S_N.

Take a smooth test function $\Upsilon \colon \mathbb{R}^2 \to \mathbb{R}_+$ as follows. On $[-1,1]^2$ the function is bigger than 1 and the Fouriertransfrom $\widehat{\Upsilon}$ is supported in $B_1(0)$.

Applying Cuachy-Schwarz, artificially inserting Υ and opening the squares leads to

$$|S_N^{(\alpha)}(\theta)| \ll \underbrace{\left(\sum_{\mathbf{x}} \mu(\mathbf{x})^2 \right)^{\frac{1}{2}}}_{\ll N^{\delta/2}} \left(\sum_{\mathbf{x}} \Upsilon(\frac{\mathbf{x}}{50 N^{\frac{1}{2}}}) \sum_{\mathbf{y}} \nu^{(\alpha)}(\mathbf{y}) \sum_{\mathbf{y}'} \nu^{(\alpha)}(\mathbf{y}') e(\langle \mathbf{x}, \mathbf{y} - \mathbf{y}' \rangle \theta) \right)^{\frac{1}{2}}.$$

Using Poisson summation in the remaining $\mathbf{x}$-sum and exploiting the support of $\widehat{\Upsilon}$ we obatin

$$|S_N^{(\alpha)}(\theta)| \ll N^{\frac{1}{2}+\frac{\delta}{2}} \left(\sum_{\mathbf{y}, \mathbf{y}'} \nu^{(\alpha)}(\mathbf{y}) \nu^{(\alpha)}(\mathbf{y}') \mathbb{1}_{\|(\mathbf{y}-\mathbf{y}')\theta\| < \frac{1}{50\sqrt{N}}} \right)^{\frac{1}{2}}.$$

Here $\|\cdot\|$ is the distance to the nearest lattice point in $\mathbb{Z}^2$. We have

$$\|(\mathbf{y} - \mathbf{y}')\frac{a}{q}\| \leq \|(\mathbf{y} - \mathbf{y}')\theta\| + |y - y'||\beta| < \frac{1}{50\sqrt{N}} + \frac{K}{2\sqrt{N}} < \frac{1}{Q}.$$

Since $q < Q$ we find that $\|(\mathbf{y} - \mathbf{y}')\frac{a}{q}\| = 0$, which implies

$$\mathbf{y} \equiv \mathbf{y}' \bmod q.$$

We obtain

$$|\mathbf{y} - \mathbf{y}'| \ll \frac{\sqrt{N}}{K}.$$

So far we found

$$|S_N^{(\alpha)}(\theta)| \ll N^{\frac{1}{2}+\frac{\delta}{2}} \left(\sum_{\mathbf{y}} \nu^{(\alpha)}(\mathbf{y}) \sum_{\substack{\mathbf{y}' \\ |\mathbf{y}-\mathbf{y}'| \ll \frac{\sqrt{N}}{K}}} \mathbb{1}_{\mathbf{y} \equiv \mathbf{y}' \bmod q,} \right)^{\frac{1}{2}}.$$

Using $Q < \frac{N^{\frac{1}{2}}}{K}$ and estimating the $\mathbf{y}'$-sum trivially yields

$$|S_N^{(\alpha)}(\theta)| \ll N^{\frac{1}{2}+\frac{\delta}{2}} \left(\sum_{\mathbf{y}} \nu^{(\alpha)}(\mathbf{y}) \frac{N}{Q^2K^2} \right)^{\frac{1}{2}} \ll \frac{N^{1+\delta}}{QK}.$$

□

This bound already suffices to treat certain parameter-ranges.

Theorem 9.11. *Assume $Q < \sqrt{N}$ and $K < \frac{\sqrt{N}}{Q}$. Then*

$$\int_{Q_{Q,K}} |S_N(\theta)|^2 d\theta \ll \frac{(\sharp\Omega_N)^2}{N} \cdot \frac{N^{2(1-\delta)+4\mathfrak{b}}}{K}.$$

Proof. Estimating trivially using the proposition above yields

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \frac{K}{N} Q^2 \left(\frac{N^{1+\delta}}{QK} \right)^2 \ll N^{4\delta-1} \cdot \frac{N^{2(1-\delta)}}{K}.$$

We conclude by using $\sharp\Omega_N \gg N^{2\delta-2\delta\mathfrak{b}-\epsilon}$. □

For the next (bilinear form) estimate we introduce

$$P_{Q,\beta} = \{\theta = \frac{a}{q} + \beta \colon \frac{Q}{2} \leq Q < Q,\ (a,q) = 1\}.$$

Proposition 9.12. *For all $\epsilon > 0$ we have*

$$\sum_{\theta \in P_{Q,\beta}} |S_N(\theta)| \ll_\epsilon N^{2\delta} Q^2 N^{1-\delta+\epsilon} \left(\frac{1}{Q^{\frac{3}{2}}} + \frac{1}{QN^{\frac{1}{8}}} \right).$$

Proof. Set $\Omega' = \Xi_1\Xi_2$ and $\Omega'' = \Xi_2 \cdots \Xi_J$. Of course we have $\Omega_N = \Omega'\Omega''$. Furthermore,

$$|\gamma^\top e_2| < 300N^{\frac{3}{4}} \text{ and } |\omega e_2| < 2000N^{\frac{1}{4}},$$

for $\gamma \in \Omega'$ and $\omega \in \Omega''$. Recall $\sharp\Omega' \ll N^{\frac{3\delta}{2}}$ and $\sharp\Omega'' \ll N^{\delta 2}$.

Proceeding as in the proof of the previous proposition (also using analogous notation) we need to bound

$$S_N^{(\alpha)} = \sum_{\mathbf{x}} \sum_{\mathbf{y}} \mu(\mathbf{y}) \nu^{(\alpha)}(\mathbf{y}) e(\theta \langle \mathbf{x}, \mathbf{y} \rangle).$$

The difference being that we arrange $\nu^{(\alpha)}$ so that for $\mathbf{y}, \mathbf{y}' \in \mathrm{supp}(\nu^{(\alpha)})$ we have

$$|\mathbf{y} - \mathbf{y}'| < \frac{1}{10000} N^{\frac{1}{4}}.$$

We proceed by writing

$$\sum_{\theta\in P_{Q,\beta}}|S_N^{(\alpha)}(\theta)| = \sum_{q\asymp Q}\sum_{(a,q)=1}\zeta(\theta)S_N^{(\alpha)}(\theta)$$
$$= \sum_{q\asymp Q}\sum_{(a,q)=1}\zeta(\theta)\sum_{\mathbf{x}}\sum_{\mathbf{y}}\mu(\mathbf{y})\nu^{(\alpha)}(\mathbf{y})e(\theta\langle\mathbf{x},\mathbf{y}\rangle).$$

Choose the bump function Υ essentially as earlier but assuming that the Fourier transform is supported in a ball of radius $1/40$ of the origin. Applying Cauchy-Schwarz, inserting Υ, exchanging order of summation and applying Poisson summation yields

$$\sum_{\theta\in P_{Q,\beta}}|S_N(\theta)| \ll N^{3\delta/4}\left(\sum_{\mathbf{x}}\Upsilon(\frac{\mathbf{x}}{300N^{\frac{3}{4}}})|\sum_{q\asymp Q}\sum_{(a,q)=1}\zeta(\theta)\sum_{\mathbf{y}}\nu^{(\alpha)}(\mathbf{y})e(\theta\langle\mathbf{x},\mathbf{y}\rangle)|^2\right)^{\frac{1}{2}}$$
$$\ll N^{\frac{3(\delta+1)}{4}}\mathcal{X}^{\frac{1}{2}},$$

for

$$\mathcal{X} = \mathcal{X}_{Q,\beta} = \sum_{q,q'}\sum_{a,a'}\sum_{\mathbf{y},\mathbf{y}'}\nu^{(\alpha)}(\mathbf{y})\nu^{(\alpha)}(\mathbf{y}')\mathbb{1}_{\|\mathbf{y}\theta-\mathbf{y}'\theta'\|<\frac{1}{12000N^{\frac{3}{4}}}}.$$

Coordinate wise the condition in the sum reads

$$\|y_1\theta - y_1'\theta'\|, \|y_2\theta - y_2'\theta'\| < \frac{1}{12000N^{\frac{3}{4}}}.$$

Note that $y_1y_2y_1'y_2' \neq 0$. Recall that $|\beta| < \frac{K}{N} < \frac{1}{N^{\frac{1}{2}}Q}$. We get

$$\|y_1\frac{a}{q} - y_1'\frac{a'}{q'}\| \leq \|y_1\theta - y_1'\theta'\| + |(y_1-y_1')\beta| < \frac{1}{12000N^{\frac{3}{4}}} + \frac{1}{1000N^{\frac{1}{4}}Q}.$$

A similar estimate holds for y_2, y_2'.

We put $Y = \begin{pmatrix} y_1 & y_1' \\ y_2 & y_2' \end{pmatrix}$. The determinant is given by

$$\mathcal{Y} = \det(Y) = y_1y_2' - y_1'y_2.$$

We can estimate

$$\|\mathcal{Y}\frac{a}{q}\| \leq \|y_2'\left(y_1\frac{a}{q} - y_1'\frac{a'}{q'}\right)\| + \|y_1'\left(y_2'\frac{a'}{q'} - y_2\frac{a}{q}\right)\| < \frac{1}{Q}.$$

This forces $\mathcal{Y} \equiv 0 \bmod q$. Running the same argument with the roles of a, q and a', q' interchanged we get $\mathcal{Y} \equiv 0 \bmod q'$. Thus, if $\mathfrak{q} = [q, q']$, we have

$$\mathcal{Y} \equiv 0 \bmod \mathfrak{q}.$$

Note that $\frac{1}{2}Q \leq \mathfrak{q} < Q^2$.

We decompose $\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2$ according to $\mathcal{Y} \neq 0$ or $\mathcal{Y}00$. The desired bound will follow directly from the estimates for $\mathcal{X}_1$ and $\mathcal{X}_2$, which we will prove now.

We start by considering $\mathcal{X}_1$. Here we need to make use of the condition $\mathcal{Y} \neq 0$. Observe

$$|\mathcal{Y}| \leq |y_1(y_2' - y_2)| + |(y_1 - y_1')y_2| < N^{\frac{1}{2}}.$$

Since $\mathfrak{q} \mid \mathcal{Y}$ and $\mathcal{Y} \neq 0$ we have

$$\mathfrak{q} \leq \min(Q^2, N^{\frac{1}{2}}) \leq QN^{\frac{1}{4}}.$$

This forces

$$y_1\frac{a}{q} - y_1'\frac{a'}{q'} \equiv 0 \bmod 1 \text{ and } y_2\frac{a}{q} - y_2'\frac{a'}{q'} \equiv 0 \bmod 1.$$

Introducing $\tilde{q} = (q, q')$, $q = q_1\tilde{q}$ and $q' = q_1'\tilde{q}$ so that $\mathfrak{q} = q_1q_1'\tilde{q}$ and

$$y_1aq_1' \equiv y_1'a'q_1 \bmod \mathfrak{q} \text{ and } y_2aq_1' \equiv y_2'a'q_1 \bmod \mathfrak{q}.$$

We deduce that $q_1 \mid y_1$ and $q_1 \mid y_2$ since $(a, q) = (a', q') = 1$. But $(y_1, y_2) = 1$, so that $q_1 = 1$. Similar we deduce that $q_1' = 1$. This implies that $q = q' = \mathfrak{q}$.

Fixing $\mathbf{y}, \mathbf{y}' \in \Omega'' e_2$ determines $\mathcal{Y}$. Since $q \mid \mathcal{Q}$ this leaves N^{ϵ} choices for q and $\ll Q$ choices for a. But this determines a' since $y_1a \equiv y_1'a' \bmod q'$ (recall $q = q'$) and $1 \leq a \leq q'$. We arrive at

$$\mathcal{X}_1 \ll \sum_{\substack{\mathbf{y},\mathbf{y}',\\ \mathcal{Y}\neq 0}} \nu^{(\alpha)}(\mathbf{y})\nu^{(\alpha)}(\mathbf{y}') \sum_{\substack{q|\mathcal{Y},\\ \frac{Q}{2}\leq q<Q}} \sum_{a \bmod q} 1 \ll N^{\delta+\epsilon}Q.$$

We turn towards the estimate for $\mathcal{X}_2$. Note that $\mathcal{Y} = 0$ implies $\frac{y_1}{y_2} = \frac{y_1'}{y_2'}$. By the uniqueness of the continued fraction expansion we obtain $\mathbf{y} = \mathbf{y}'$. There are $N^{\delta/2}$ choices for $\mathbf{y}$ and we fix one of these. To save space we set $N' = \frac{1}{12000}N^{\frac{3}{4}}$. Then we have the condition

$$\|y_1(\frac{a}{q} - \frac{a'}{q'})\| < \frac{1}{N'}.$$

Put $v = (y_1, q)$, $v' = (y_1, q')$ and write $q = vr$ as well as $y_1 = vz$ with $(z, r) = 1$. Without loss of generality we have $v \leq v'$. Since $v, v' \mid y_1$, there are at most N^{ϵ} choices for these numbers. Now there are $\ll Q/v'$ choices for $q' \equiv 0 \bmod v'$ and $\ll Q$ choices for $(a', q') = 1$. Having made these choices we fixed $y_1\frac{a'}{q'} \bmod 1$, which we denote by ψ. Our condition above reads

$$\|z\frac{a}{r} - \psi\| < \frac{1}{N'}.$$

Define the set

$$\mathcal{U}_z = \{\frac{za}{r} \bmod 1 \colon Q/(2v) \leq r < Q/v,\ a \leq Q \text{ with } (a, vr) = 1\}.$$

Points $u \in \mathcal{U}_z$ are separated by a distance of at least v^2/Q^2. We conclude that the intersection of $\mathcal{U}_z$ with the interval $[\psi - \frac{1}{N'}, \psi + \frac{1}{N'}]$ contains at most $\frac{Q^2}{v^2N'} + 1$ points. Once $u = \frac{f}{r} \in \mathcal{U}_z$ is determined we have v possible values for $a \bmod q$, as $q = rv$.

In summary we get

$$\begin{aligned}\mathcal{X}_1 &\ll \sum_{\mathbf{y}} \nu^\alpha(\mathbf{y}) \sum_{\substack{v,v'|y_1,\\ v\le v'}} \sum_{q'\equiv 0 \bmod v'} \sum_{(a',q')=1} \sum_{\frac{f}{r}\in \mathcal{U}_z\cap[\psi-\frac{1}{N'},\psi+\frac{1}{N'}]} \sum_{\substack{a<q,\\ a\equiv f \bmod r}} 1\\ &\ll \sum_{\mathbf{y}} \nu^\alpha(\mathbf{y}) \sum_{\substack{v,v'|y_1,\\ v\le v'}} \frac{Q}{v'}\cdot Q\left(1+\frac{Q^2}{v^2N'}\right) v\\ &\ll N^{\delta/2+\epsilon}Q^2\left(\frac{Q^2}{N^{\frac{3}{4}}}+1\right).\end{aligned}$$

This completes the estimate of $\mathcal{X}_2$ and therefore the proof. □

This leads to a minor arc bound, which is useful for large Q.

Theorem 9.13. *Assume $Q < N^{\frac{1}{2}}$ and $KQ < N^{\frac{1}{2}}$. Then*

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \frac{(\sharp\Omega_N)^2}{N}\cdot N^{2(1-\delta)+4\mathfrak{b}}\left(\frac{1}{Q^{\frac{1}{2}}}+\frac{1}{N^{\frac{1}{8}}}\right).$$

Proof. We estimate

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \sup_{\theta\in W_{Q,K}} |S_N(\theta)|\cdot\frac{K}{N}\sup_{\beta\asymp\frac{K}{N}}\sum_{\theta\in P_{Q,\beta}} |S_N(\theta)|.$$

Inserting the estimates from the propositions above yields

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll N^{2\delta+1+\epsilon}\left(\frac{1}{Q^{\frac{1}{2}}}+\frac{1}{N^{\frac{1}{8}}}\right).$$

The claim follows directly by recalling how large Ω_N is at least. □

We still need suitable estimates for the situation when K is small.

Proposition 9.14. *Assume $\theta\in W_{Q,K}$ with $1\ll KQ < N^{\frac{5}{52}}$. Then*

$$|S_N(\theta)| \ll \sharp\Omega_N\left(\frac{e^{c\log\log(KQ)^2}}{(KQ)^{1-(1-\delta)\frac{52}{5}}}\right).$$

Proof. Since $N_i \asymp N^{2^{-i}}$ there is $1\le j\le J$ so that

$$\frac{1}{100}(KQ)^{\frac{13}{5}} < N_j < (KQ)^{\frac{26}{5}} < N^{\frac{1}{2}}.$$

We define the three set

$$\Omega^{(1)} = \Xi_1\cdots\Xi_{j-1},\ \Omega^{(2)} = \Xi_j \text{ and } \Omega^{(3)} = \Xi_{j+1}\cdots\Xi_J.$$

For $g_i \in \Omega^{(i)}$ and $M = N_{j+1} \cdots N_J$ we have the estimates

$$\begin{aligned}\lambda(g_1) &\asymp \frac{N}{MN_j},\\ \lambda(g_2) &\sim N_j \text{ and}\\ \lambda(g_3) &\sim M.\end{aligned}$$

Note that $\frac{N_j}{\log(N_j)} \ll M \ll N_j \log(N_j)$ and $J - j \asymp \log\log(M)$. Finally recall that

$$\sharp\omega^{(2)} \gg \frac{N_j^{2\delta}}{\log(N_j)^3} \text{ and } \sharp\Omega^{(3)} \gg \frac{M^{2\delta}}{e^{c\log\log(M)^2}}.$$

We are ready to start the estimation of $S_N(\theta)$. First write

$$|S_N(\theta)| \ll \sum_{g_1\in\Omega^{(1)}} \sum_{g_3\in\Omega^{(3)}} \left| \sum_{g_2\in\Omega^{(2)}} e(\langle g_3 e_2, g_2^\top g_1^\top e_2\rangle\theta)\right|.$$

For fixed g_1 put $\eta = g_1^\top e_2$ and note that $|\eta| \asymp \frac{N}{MN_j}$. Continuing as previously we obtain the estimate

$$\sum_{g_3\in\Omega^{(3)}} \left| \sum_{g_2\in\Omega^{(2)}} e(\langle g_3 e_2, g_2^\top g_1^\top e_2\rangle\theta)\right| \ll (\sharp\Omega^{(3)})^{\frac{1}{2}} \cdot M \cdot (\sharp\mathcal{S})^{\frac{1}{2}},$$

for

$$\mathcal{S} = \{(g, g') \in [\Omega^{(2)}]^\top \times [\Omega^{(2)}]^\top : \|\langle (g - g')\eta, e_i\rangle\| \ll \frac{1}{M} \text{ for } i = 1, 2\}.$$

Note that we have extended the g_3-sum to a sum over $g_3 e_2 \in \{z \in \mathbb{Z}^2 : |z| \ll M\}$. Since $\theta \in W_{K,Q}$ we can write

$$\|\langle (g - g')\eta, e_i\rangle \frac{a}{q}\| = \|\langle (g - g')\eta, e_i\rangle\theta\| + \|\langle (g - g')\eta, e_i\rangle\beta\|$$

and estimate

$$\|\langle (g - g')\eta, e_i\rangle\beta\| \ll N_j \cdot \frac{N}{MN_j} \cdot \frac{K}{N} = \frac{K}{M}.$$

For fixed g' we can enlarge the count $\sharp\mathcal{S}$ by allowing $g \in \mathrm{SL}_2(\mathbb{Z})$ with $\|g\| \ll N_j$. An application of Lemma 7.8 with $\eta' = g'\eta$, $X = N/M$ and $Y = N_j$ yields

$$\sharp\mathcal{S} \ll \sharp\Omega^{(2)} \cdot \frac{N_j^2}{K^2Q^2}.$$

In summary we have

$$\sum_{g_3\in\Omega^{(3)}}\left|\sum_{g_2\in\Omega^{(2)}} e(\langle g_3e_2, g_2^\top g_1^\top e_2\rangle\theta)\right| \ll (\sharp\Omega^{(2)}\cdot\sharp\Omega^{(3)})^{\frac{1}{2}}\cdot M\cdot\frac{N_j}{KQ}$$
$$\ll \sharp\Omega^{(2)}\cdot\sharp\Omega^{(3)}\frac{(MN_j)^{1-\delta}e^{c\log\log(M)^2}\log(N_j)^3}{KQ}.$$

The result follows directly after executing the g_1-sum trivially and recalling bounds for N_j and M. □

Corollary 9.15. *Suppose that* $1\ll KQ<N^{\frac{5}{52}}$*, then we have*

$$\int_{W_{K,Q}}|S_N(\theta)|^2d\theta\ll\frac{(\sharp\Omega_N)^2}{N}\cdot\frac{Q^{(1-\delta)\frac{104}{5}}e^{c\log\log(KQ)^2}}{K^{1-(1-\delta)\frac{104}{5}}}.$$

Proof. This follows directly after inserting the L^∞ bound above in the integral. □

This is still not sufficient for all K.

Proposition 9.16. *Suppose* $1\ll KQ<N^{\frac{5}{52}}$*, then*

$$\sum_{\theta\in P_{Q,\beta}}|S_N(\theta)|\ll\sharp\Omega_N\cdot Q^2\left(\frac{(KQ)^{(1-\delta)\frac{52}{5}}e^{c\log\log(KQ)^2}}{Q^{\frac{3}{2}}}\right).$$

Proof. Using the same decomposition $\Omega_N=\Omega^{(1)}\Omega^{(2)}\Omega^{(3)}$ as in the previous proof we follow the proof of Proposition 9.12. It is relatively straight forward to arrive at

$$\sum_{\theta\in P_{Q,\beta}}|S_N(\theta)|\ll\sum_{g_1\in\Omega^{(1)}}(\sharp\Omega^{(3)})^{\frac{1}{2}}\cdot M\cdot(\sharp\mathcal{S}')^{\frac{1}{2}},$$

for

$$\mathcal{S}'=\{(\theta,\theta',g,g')\in P_{Q,\beta}\times P_{Q,\beta}\times[\Omega^{(2)}]^\top\times[\Omega^{(2)}]^\top:\|(g\theta-g'\theta')\eta\|\ll\frac{1}{M}\}$$

with $\eta=g_1^\top e_2$. It is by now standard to the condition in $\mathcal{S}'$ to deduce $q=q'$ and

$$a(g\eta)\equiv a'(g'\eta)\bmod q.$$

One obtains

$$\sharp\mathcal{S}'\ll\sharp\Omega^{(2)}\cdot\frac{N_j^2}{Q^2}.$$

After inserting this above it is easy to reach the desired estimate. □

Corollary 9.17. *Suppose* $1\ll KQ<N^{\frac{5}{52}}$*, then*

$$\int_{W_{K,Q}}|S_N(\theta)|^2d\theta\ll\frac{(\sharp\Omega_N)^2}{N}\cdot\frac{(KQ)^{(1-\delta)\frac{104}{5}}e^{c\log\log(KQ)^2}}{Q^{\frac{1}{2}}}.$$

Proof. This follows directly after combining the two previous propositions as earlier. □

With these estimates at hand we can complete the analysis of the minor arcs.

Theorem 9.18. *Assume $\delta > \delta_0$. Then for $c > 0$ we have*

$$\sum_{n\in\mathbb{Z}} |\mathcal{E}_N(n)|^2 \ll \frac{(\sharp\Omega_N)^2}{N}\mathfrak{Q}^{-c}.$$

Proof. By Parseval, we have

$$\sum_{n\in\mathbb{N}} |\mathcal{E}_N(n)|^2 = \int_0^1 |1-\Psi_{\mathfrak{Q},N}(\theta)|^2 |S_N(\theta)|^2 d\theta.$$

First, we estimate the contribution that overlaps with the major arcs. Note that $1-\psi(x) = |x|$ on $[-1,1]$. For $K \asymp N|\beta|$ we have the estimate

$$\int_{\mathfrak{M}_{\mathfrak{Q}}} |1-\Psi_{\mathfrak{Q},N}(\theta)|^2 |S_N(\theta)|^2 d\theta \ll \sum_{q<\mathfrak{Q}} \sum_{(a,q)=1} \int_{|\beta|<\frac{\mathfrak{Q}}{N}} |\frac{N}{\mathfrak{Q}}\beta|^2 \left(\frac{\sharp\Omega_N}{(N|\beta|\mathfrak{Q})^{1-c}}\right) d\beta$$
$$\ll \frac{(\sharp\Omega_N)^2}{N\mathfrak{Q}^{1-4c}},$$

where we used Proposition 9.14. Note that here $0 < c < (1-\delta)\frac{52}{5} < \frac{1}{4}$.

To treat the pure minor arcs we decompose it inn dyadic regions:

$$\int_{\mathfrak{m}} |S_N(\theta)|^2 d\theta \ll \sum_{Q<\sqrt{N}} \sum_{K<\frac{\sqrt{N}}{Q}} \mathcal{I}_{Q,K},$$

where

$$\mathcal{I}_{Q,K} = \int_{W_{Q,K}} |S_N(\theta)|^2 d\theta.$$

Suppose $Q = N^\alpha$ and $K = N^\kappa$ where $0 \le \alpha < \frac{1}{2}$ and $0 \le \kappa < \frac{1}{2} - \alpha$. We define $\eta = (1-\delta)\frac{104}{5}$. The different estimates we proved will be applied in different regions of the plane (α, κ). Indeed

$$\mathcal{R}_1 = \{(\alpha,\kappa)\colon \kappa > 2(1-\delta) + 4\mathfrak{b}\},$$
$$\mathcal{R}_2 = \{(\alpha,\kappa)\colon \alpha > 4(1-\delta) + 8\mathfrak{b}\},$$
$$\mathcal{R}_3 = \{(\alpha,\kappa)\colon \eta(\alpha+\kappa) < \kappa \text{ and } \alpha+\kappa < \frac{5}{52}\}, \text{ and}$$
$$\mathcal{R}_4 = \{(\alpha,\kappa)\colon \eta(\alpha+\kappa) < \frac{1}{2}\alpha \text{ and } \alpha+\kappa < \frac{5}{52}\}$$

The only remaining job is to collect together the appropriate estimates. □

Finally we can complete the promised proof.

Proof of Theorem 9.1. For the major arcs we have established

$$\mathcal{M}_N(n) \gg \frac{\sharp\Omega_N}{N\log\log(N)} \gg N^{2\delta-1-\frac{1}{1000}}$$

as long as $n \asymp N$. Suppose n is not represented, then the minor arcs must be as big as the major arcs. Indeed we would have

$$|\mathcal{E}_N(n)| = |\mathcal{N}(n) - \mathcal{M}_N(n)| \gg \frac{\sharp\Omega_N}{N\log\log(N)}.$$

Let $\mathfrak{E}(N)$ denote the set of $n \asymp N$ which have a small representation number $R_N(n)$. More precisely

$$\mathfrak{E}(N) = \left\{ \frac{N}{20} \leq n < \frac{N}{10} \colon R_N(n) < \frac{1}{2}\mathcal{M}_N(n) \right\}.$$

The standard argument now proceeds as follows

$$\begin{aligned}
\sharp\mathfrak{E}(N) &\ll \sum_{n\asymp N} \mathbb{1}_{|\mathcal{E}_N(n)| \gg \frac{\sharp\Omega_N}{N\log\log(N)}} \\
&\ll \frac{N^2\log\log(N)^2}{(\sharp\Omega_N)^2} \cdot \sum_{n\asymp N} |\mathcal{E}_N(n)| \\
&\ll \frac{N^2\log\log(N)^2}{(\sharp\Omega_N)^2} \cdot \frac{(\sharp\Omega_N)^2}{N} \cdot \mathfrak{Q}^{-c} \ll N e^{-c\sqrt{\log(N)}}.
\end{aligned}$$

This completes the proof. □

10. Proof of Theorem 1.12

We closely follow [6]. We start with some preliminary estimates. For these we identify $M_{2\times 2}(\mathbb{Z})$ with $\mathbb{Z}^4$ and observe that $\mathrm{Tr}(A^\top B) = A \cdot B$.

Lemma 10.1. *For any square-free $q \geq 1$ and any vector $\mathbf{s} \in \mathbb{Z}^4$ with $(\mathbf{s}, q) = 1$ we have*

$$\sum_{\gamma \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} e_q(\gamma \cdot s) \ll q^{\frac{3}{2}+\epsilon}$$

for any $\epsilon > 0$.

Proof. It suffices to consider the case $q = p$ (by multiplicativity and square-freenes). Write $\mathbf{s} = (x, y, z, w)$ and $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. We treat the case $p \nmid y$ (the other cases are similar). The degenerate case $c = 0$ contributes nothing since

$$\sum_{a\in(\mathbb{Z}/p\mathbb{Z})^\times} \sum_{b\in\mathbb{Z}/p\mathbb{Z}} e_p(ax + by + \overline{a}w) = \sum_{a\in(\mathbb{Z}/p\mathbb{Z})^\times} e_p(ax + \overline{a}w) \sum_{b\in\mathbb{Z}/p\mathbb{Z}} e_p(by) = 0$$

by character orthogonality.

Matrices with $c \neq 0$ contribute

$$\sum_{c\in(\mathbb{Z}/p\mathbb{Z})^\times}\sum_{a,d\in\mathbb{Z}/p\mathbb{Z}} e_p(ax+\bar{c}(ad-1)y+cz+dw)$$
$$= \sum_{c\in(\mathbb{Z}/p\mathbb{Z})^\times} e_p(cz-\bar{c}y)\sum_{a\in\mathbb{Z}/p\mathbb{Z}} e_p(ax)\sum_{d\in\mathbb{Z}/p\mathbb{Z}} e_p(d(\bar{c}ay+w)).$$

The d-sum vanishes unless $a \equiv -c\bar{y}w \bmod p$, in which case it contributes p. The c-sum is a Kloosterman sum which can be bounded by $2\sqrt{p}$. (This is Weil's bound.) □

This together with Lemma 5.4 implies the following.

Proposition 10.2. *Let φ_X be as above. For any square-free $q \geq 1$ and any vector $\mathbf{s} \in \mathbb{Z}^4$ with $(\mathbf{s}, q)$ we have*

$$\sum_{\xi\in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\xi)e_q(\xi\cdot s) \ll q^{-\frac{3}{2}+\epsilon}X^2 + q^3X^{\frac{3}{2}}.$$

Proof. The idea is simply to split the some into congruence classes and use the previous estimates:

$$\sum_{\xi\in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\xi)e_q(\xi\cdot s) = \sum_{\gamma\in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} e_q(\gamma\cdot\mathbf{s}) \sum_{\substack{\xi\in \mathrm{SL}_2(\mathbb{Z}),\\ \xi\equiv\gamma \bmod q}} \varphi_X(\xi)$$
$$\ll \Big|\sum_{\gamma\in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} e_q(\gamma\cdot\mathbf{s})\Big| \frac{X^2}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} + O(q^3X^{\frac{3}{2}}).$$

□

Next we proceed as in the proof of Theorem 1.4 and construct a convenient set $\aleph$.

Proposition 10.3. *Given any $Y \gg 1$ there is a non-empty subset $\aleph = \aleph(Y) \subset \Gamma_2 \cap B_Y$ with the following property. For all square-free q and all $\mathfrak{a}_0 \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ we have*

$$\Big|\frac{\sharp\{\mathfrak{a}\in\aleph\colon \mathfrak{a}\equiv\mathfrak{a}_0 \bmod q\}}{\sharp\aleph} - \frac{1}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}\Big| \ll \mathfrak{G}(Y;q),$$

where $\mathfrak{G}$ is as in Theorem 6.1.

Proof. The proof of this existence result is essentially as before only using Theorem 6.1 (see Proposition 9.5) and we omit the details. □

Our main parameters will be

$$X = N^x,\ Y = N^y \text{ and } Z = N^z \text{ for } x+y+z = 1.$$

We will take $x > 1 - \eta$ close to one and $y, z > 0$ to be small. Let $\aleph = \aleph(Y) \subset \Gamma_2$ be the set constructed in Proposition 10.3. Further put

$$\Xi_0 = \{\xi \in \Gamma \colon \|\xi\| < X\} \text{ and } \Omega_0 = \{\omega \in \Gamma \colon \|\omega\| < Z\}.$$

We don't have good control on the size of $\aleph$ but we know that

$$\sharp\Xi_o \asymp X^{2\delta} \text{ and } \sharp\Omega_0 \asymp Z^{2\delta}.$$

By the pigeon hole principle we find that there is some $l_X \asymp \log(X)$ so that

$$\Xi = \{\gamma \in \Xi_0 \colon l(\gamma) = l_X\}$$

satisfies

$$\sharp\Xi \gg X^{2\delta} \log(X)^{-1}.$$

The same argument applied to Ω_0 yields $l_Z \asymp \log(Z)$ and a set

$$\Omega = \{\omega \in \Omega_0 \colon l(\omega) = l_Z\}$$

with

$$\sharp\Omega \gg Z^{2\delta} \log(Z)^{-1}.$$

Define the product $\Pi = \Xi \cdot \aleph \cdot \Omega$. Since Γ is a free and finitely generated semigroup the representation $\varpi = \xi \cdot \mathfrak{a} \cdot \omega$ of $\varpi \in \Pi$ as a product of $\xi \in \Xi$, $\mathfrak{a} \in \aleph$ and $\omega \in \Omega$ is unique. Furthermore we have

$$\Pi \subset \Gamma \cap B_{100N}.$$

We now turn towards the sifting procedure. Define the sequence $\mathfrak{A} = \{a_N\}$ by

$$a_N(n) = \sum_{\varpi \in \Pi} \mathbb{1}_{\mathrm{Tr}(\varpi)^2 - 4 = n}.$$

The sequence is supported in $\{n \ll T\}$ for $T = N^2$. for a parameter $\mathfrak{Q}$ and any square-free $\mathfrak{q} < \mathfrak{Q}$ we set

$$|\mathfrak{A}_\mathfrak{q}| = \sum_{n \equiv 0 \bmod \mathfrak{q}} a_N(n).$$

We decompose $|\mathfrak{A}_\mathfrak{q}|$ as follows

$$|\mathfrak{A}_\mathfrak{q}| = \sum_{\varpi \in \Pi} \mathbb{1}_{\mathrm{Tr}(\varpi)^2 - 4 \equiv 0 \bmod \mathfrak{q}} = \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q}, \\ \mathfrak{t}^2 \equiv 4 \bmod \mathfrak{q}}} \sum_{\varpi \in \Pi} \mathbb{1}_{\mathrm{Tr}(\varpi) \equiv \mathfrak{t} \bmod \mathfrak{q}}.$$

The congruence condition will now be detected using character orthogonality of $e_\mathfrak{q}(\cdot)$. This looks as follows

$$|\mathfrak{A}_\mathfrak{q}| = \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q}, \\ \mathfrak{t}^2 \equiv 4 \bmod \mathfrak{q}}} \sum_{\varpi \in \Pi} \frac{1}{\mathfrak{q}} \sum_{q \mid \mathfrak{q}} \sum_{r \in (\mathbb{Z}/q\mathbb{Z})^\times} e_q(r(\mathrm{Tr}(\varpi) - \mathfrak{t})).$$

Define

$$\mathfrak{M}_q = \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q}, \\ \mathfrak{t}^2 \equiv 4 \bmod \mathfrak{q}}} \sum_{\varpi \in \Pi} \frac{1}{\mathfrak{q}} \sum_{\substack{q \mid \mathfrak{q}, \\ q < Q_0}} \sum_{r \in (\mathbb{Z}/q\mathbb{Z})^\times} e_q(r(\mathrm{Tr}(\varpi) - \mathfrak{t})).$$

This gives the decomposition

$$|\mathfrak{A}_\mathfrak{q}| = \mathfrak{M}_\mathfrak{q} + \mathfrak{E}_\mathfrak{q}. \tag{34}$$

We will now treat the main term $\mathfrak{M}_\mathfrak{q}$. But first we recall the following estimate (**Exercise**):

$$\sharp\{\mathfrak{t} \in \mathbb{Z}/\mathfrak{q}\mathbb{Z} \colon \mathfrak{t}^2 \equiv 4 \bmod \mathfrak{q}\} = 2^{\nu(\mathfrak{q}) - \mathbb{1}_{2 \mid \mathfrak{q}}}.$$

Proposition 10.4. *Let β be the multiplicative function given at primes by*

$$\beta(p) = \frac{1 + \mathbf{1}_{p \neq 2}}{p} \left(1 + \frac{1}{p^2 - 1}\right).$$

There is a decomposition

$$\mathfrak{M}_\mathfrak{q} = \beta(\mathfrak{q}) \cdot \sharp\Pi + r_1(\mathfrak{q}) + r_2(\mathfrak{q}),$$

where

$$\sum_{\mathfrak{q} < \mathfrak{Q}} |r_1(\mathfrak{q})| \ll \sharp\Pi \cdot \log(\mathfrak{Q})^2 \left(\frac{1}{e^c \sqrt{\log(Y)}} + Q_0^C Y^{-\Theta}\right)$$

and

$$\sum_{\mathfrak{q} < \mathfrak{Q}} |r_2(\mathfrak{q})| \ll \sharp\Pi \cdot \frac{\mathfrak{Q}^\epsilon}{Q_0}.$$

Proof. We start by using the construction of Π to write

$$\mathfrak{M}_q = \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q}, \\ \mathfrak{t}^2 \equiv 4 \bmod \mathfrak{q}}} \sum_{\substack{\xi \in \Xi, \\ \omega \in \Omega}} \frac{1}{\mathfrak{q}} \sum_{\substack{q \mid \mathfrak{q}, \\ q < Q_0}} \sum_{r \in (\mathbb{Z}/q\mathbb{Z})^\times} \sum_{\mathfrak{a}_0 \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} e_q(r(\mathrm{Tr}(\xi \mathfrak{a}_0 \omega) - \mathfrak{t})) \cdot \sharp\{\mathfrak{a} \in \aleph \colon \mathfrak{a} \equiv \mathfrak{a}_0 \bmod q\}.$$

Using Proposition 10.3 we can write

$$\mathfrak{M}_\mathfrak{q} = \mathfrak{M}_\mathfrak{q}^{(1)} + r_1(\mathfrak{q}),$$

for

$$\mathfrak{M}_\mathfrak{q}^{(1)} = \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q}, \\ \mathfrak{t}^2 \equiv 4 \bmod \mathfrak{q}}} \frac{\sharp\Pi}{\mathfrak{q}} \sum_{\substack{q \mid \mathfrak{q}, \\ q < Q_0}} \sum_{r \in (\mathbb{Z}/q\mathbb{Z})^\times} \frac{1}{\sharp \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} \sum_{\mathfrak{a}_0 \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} e_q(r(\mathrm{Tr}(\mathfrak{a}_0) - \mathfrak{t}))$$

and

$$|r_1(\mathfrak{q})| \ll \tau(\mathfrak{q}) \cdot \sharp\Pi \cdot \frac{1}{\mathfrak{q}} \sum_{\substack{q \mid \mathfrak{q}, \\ q < Q_0}} q^4 \mathfrak{B}(Y; q).$$

An l^1-estimate for this error is easy to deduce:

$$\sum_{\mathfrak{q}<\mathfrak{Q}}|r_1(\mathfrak{q})| \ll \sharp\Pi\cdot\sum_{q<Q_0} q^4\mathfrak{B}(Y;q)\sum_{\mathfrak{q}<\mathfrak{Q}}\frac{\tau(\mathfrak{q})}{\mathfrak{q}} \ll \sharp\Pi\cdot\log(\mathfrak{Q})^2\left[\log(Y)^C e^{-c\sqrt{\log(Y)}}+Q_0^C Y^{-\Theta}\right].$$

We now define the real main term as

$$\mathfrak{M}_{\mathfrak{q}}^{(2)}=\mathfrak{M}_{\mathfrak{q}}^{(1)}=\sum_{\substack{\mathfrak{t} \bmod \mathfrak{q},\\ \mathfrak{t}^2\equiv 4 \bmod \mathfrak{q}}}\frac{\sharp\Pi}{\mathfrak{q}}\sum_{q|\mathfrak{q}}\sum_{r\in(\mathbb{Z}/q\mathbb{Z})^\times}\frac{1}{\sharp\,\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})}\sum_{\mathfrak{a}_0\in\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})} e_q(r(\mathrm{Tr}(\mathfrak{a}_0)-\mathfrak{t})).$$

Here we have lifted the restriction $q\le Q_0$. We get

$$\mathfrak{M}_{\mathfrak{q}}^{(1)}=\mathfrak{M}_{\mathfrak{q}}^{(2)}+r_2(\mathfrak{q}).$$

Observe that

$$r_2(\mathfrak{q})\ll\tau(\mathfrak{q})\cdot\sharp\Pi\cdot\frac{1}{\mathfrak{q}}\sum_{\substack{q|\mathfrak{q},\\ q\ge Q_0}}\frac{1}{q}\ll\sharp\Pi\cdot\frac{\mathfrak{q}^\epsilon}{\mathfrak{q}Q_0}.$$

This implies the desired average bound.

It remains to evaluate $\mathfrak{M}_{\mathfrak{q}}^{(2)}$. To do so we define

$$\rho_{\mathfrak{t}}(p)=\frac{1}{\sharp\,\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})}\sum_{\gamma\in\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})}\sum_{r\in(\mathbb{Z}/p\mathbb{Z})^\times} e_p(r(\mathrm{Tr}(\gamma)-\mathfrak{t}))$$

on the primes. We get

$$\mathfrak{M}_{\mathfrak{q}}^{(2)}=\sharp\Pi\cdot\sum_{\substack{\mathfrak{t} \bmod \mathfrak{q},\\ \mathfrak{t}^2\equiv 4 \bmod \mathfrak{q}}}\frac{1}{\mathfrak{q}}\prod_{p|\mathfrak{q}}(1+\rho_{\mathfrak{t}}(p)).$$

Note that $\mathfrak{t}\equiv\pm 2 \bmod p$ for $p\mid\mathfrak{q}$. An elementary computation shows that

$$\rho_{\mathfrak{t}}(p)=\frac{1}{p^2-1}.$$

This directly implies $\mathfrak{M}_{\mathfrak{q}}^{(2)}=\beta(\mathfrak{q})\cdot\sharp\Pi$ as desired. □

We still need to control the error $\mathfrak{E}_{\mathfrak{q}}$. It suffices to do so on average. Define

$$\mathcal{E}=\sum_{\mathfrak{q}<\mathfrak{Q}}|\mathfrak{E}_{\mathfrak{q}}|=\sum_{\mathfrak{q}<\mathfrak{Q}}\Big|\sum_{\substack{\mathfrak{t} \bmod \mathfrak{q},\\ \mathfrak{t}^2\equiv 4 \bmod \mathfrak{q}}}\sum_{\varpi\in\Pi}\frac{1}{\mathfrak{q}}\sum_{\substack{q|\mathfrak{q},\\ q\ge Q_0}}\sum_{r\in(\mathbb{Z}/q\mathbb{Z})^\times} e_q(r(\mathrm{Tr}(\varpi)-\mathfrak{t}))\Big|.$$

Theorem 10.5. *For any $\epsilon>0$ and any $1\gg Q_0<\mathfrak{Q}<N$ we have*

$$\mathcal{E}\ll N^\epsilon\cdot\sharp\Pi\cdot(XZ)^{1-\delta}\left[\frac{1}{Q_0^{\frac{1}{4}}}+\frac{1}{Z^{\frac{1}{4}}}+\frac{\mathfrak{Q}^4}{X^{\frac{1}{4}}}\right].$$

Proof. We define

$$\zeta(\mathfrak{q}) = \frac{|\mathfrak{E}_{\mathfrak{q}}|}{\mathfrak{E}_{\mathfrak{q}}}.$$

This gives us

$$\begin{aligned}\mathcal{E} &= \sum_{\mathfrak{q}<\mathfrak{Q}} \zeta(\mathfrak{q}) \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q},\\ \mathfrak{t}^2\equiv 4 \bmod \mathfrak{q}}} \sum_{\varpi\in\Pi} \frac{1}{\mathfrak{q}} \sum_{\substack{q|\mathfrak{q},\\ q\geq Q_0}} \sum_{r\in(\mathbb{Z}/q\mathbb{Z})^\times} e_q(r(\operatorname{Tr}(\varpi)-\mathfrak{t}))\\ &= \sum_{Q_0\leq q<\mathfrak{Q}} \frac{1}{q} \sum_{r\in(\mathbb{Z}/q\mathbb{Z})^\times} \sum_{\varpi\in\Pi} e_q(r\operatorname{Tr}(\varpi))\cdot\zeta_1(q,r),\end{aligned}$$

for

$$\zeta_1(q,r) = q \sum_{\substack{\mathfrak{q}<\mathfrak{Q},\\ \mathfrak{q}\equiv 0 \bmod q}} \frac{\zeta(\mathfrak{q})}{\mathfrak{q}} \sum_{\mathfrak{t}^2\equiv 4 \bmod \mathfrak{q}} e_q(-r\mathfrak{t}).$$

Now we decompose $\Pi = \Xi\cdot\aleph\cdot\Omega$ and break the q-sum into dyadic pieces. This gives

$$\mathcal{E} \ll \sum_{\mathfrak{a}\in\aleph} \sum_{\substack{Q_0<Q<\mathfrak{Q},\\ \text{dyadic}}} \frac{1}{Q}\mathcal{E}_1(Q;\mathfrak{a}),$$

where

$$\mathcal{E}_1(Q;\mathfrak{a}) = \sum_{q\asymp Q}\Big|\sum_{r\in(\mathbb{Z}/q\mathbb{Z})^\times} \zeta_1(q,r)\sum_{\xi\in\Xi}\sum_{\omega\in\Omega} e_q(r\operatorname{Tr}(\xi\mathfrak{a}\omega)\Big|.$$

We are done as soon as we can show

$$\mathcal{E}_1(Q;\mathfrak{a}) \ll N^\epsilon Q\cdot\sharp\Xi\cdot\sharp\Omega\cdot(XZ)^{1-\delta}\left[\frac{1}{Q^{\frac{1}{4}}}+\frac{1}{Z^{\frac{1}{4}}}+\frac{Q^4}{X^{\frac{1}{4}}}\right].$$

To temporary remove the absolute value in $\mathcal{E}_1(Q;\mathfrak{a})$ we artificially introduce another number $\zeta(q)\in S^1$. Taking out the ξ-sum and applying Cauchy-Schwarz we obtain

$$\mathcal{E}_1(Q;\mathfrak{a})^2 \ll \sharp\Xi\cdot\sum_{\xi\in\mathrm{SL}_2(\mathbb{Z})}\varphi_X(\xi)\left|\sum_{q\asymp Q}\zeta_2(q)\sum_{r\in(\mathbb{Z}/q\mathbb{Z})^\times}\zeta_1(q,r)\sum_{\omega\in\Omega} e_q(r\operatorname{Tr}(\xi\mathfrak{a}\omega))\right|^2.$$

Note that we artificially inserted the weight function φ_X and extended the ξ-sum to all of $\mathrm{SL}_2(\mathbb{Z})$. Note that we have the easy estimate $\zeta_1(q,r)\ll\mathfrak{Q}^\epsilon$. The trace is linear so that we can open the square to get

$$\mathcal{E}_1(Q;\mathfrak{a})^2 \ll \mathfrak{Q}^\epsilon\cdot\sharp\Xi\cdot\sum_{q,q'\asymp Q}\sum_{\omega,\omega'\in\Omega}\sum_{\substack{r\in(\mathbb{Z}/q\mathbb{Z})^\times\\ r'\in(\mathbb{Z}/q'\mathbb{Z})^\times}}\left|\sum_{\xi\in\mathrm{SL}_2(\mathbb{Z})}\varphi_X(\xi)e\left(\xi\cdot\left[\frac{r}{q}\mathfrak{a}\omega-\frac{r'}{q'}\mathfrak{a}\omega'\right]\right)\right|^2.$$

At this point we write

$$\frac{r}{q}\mathfrak{a}\omega - \frac{r'}{q'}\mathfrak{a}\omega' = \frac{\mathbf{s}}{q_0}.$$

We need some information on $\mathbf{s}$ and q_0. Both these depend on $q, q', r, r', \omega, \omega'$ and $\mathfrak{a}$. Let us introduce some more notation

$$\tilde{q} = (q, q'),\ q = q_1\tilde{q},\ q' = q_1'\tilde{q} \text{ and } \hat{q} = [q, q'] = q_1q_1'\tilde{q}.$$

All q's are square-free. Observe that $q_1q_1' \mid q_0$ and $q_0 \mid \hat{q}$. Further set $\tilde{q}_0 = (q_0, \tilde{q})$ and $\hat{q} = q_0\hat{q}_0 = q_1q_1'\tilde{q}_0\hat{q}_0$. We must have $q_0 = q_1q_1'\tilde{q}_0$ and $Q \ll \hat{q} \ll Q^2$.

We also obtain the congruence

$$q_1'r\omega \equiv q_1r'\omega' \bmod \hat{q}_0.$$

Looking at the determinant yields

$$(q_1'r)^2 \equiv (q_1r')^2 \bmod \hat{q}_0.$$

Since $1 = (q_1r', \hat{q}_0) = (q_1'r, \hat{q}_0)$ we find u with $u^2 \equiv 1 \bmod \hat{q}_0$ so that

$$q_1'r \equiv uq_1r' \bmod \hat{q}_0.$$

The number of such u can be estimates by $2^{\nu(\hat{q}_0)} \ll N^\epsilon$. We deduce that

$$\omega \equiv u\omega' \bmod \hat{q}_0.$$

With this at hand we obtain

$$\mathcal{E}_1(Q;\mathfrak{a})^2 \ll \mathfrak{Q}^\epsilon \cdot \sharp\Xi \cdot \sum_{Q\ll\hat{q}\ll Q^2} \sum_{\substack{q_1q_1'\tilde{q}_0\hat{q}_0=\hat{q},\\ q=q_1\tilde{q}_0\hat{q}_0\asymp Q,\\ q'=q_1'\tilde{q}_0\hat{q}_0\asymp Q,\\ q_0=q_1q_1'\tilde{q}_0}} \sum_{\substack{u \bmod \hat{q}_0\\ q^2\equiv 1 \bmod \hat{q}_0}} \sum_{r\in(\mathbb{Z}/q\mathbb{Z})^\times} \sum_{\substack{r'\in(\mathbb{Z}/q'\mathbb{Z})^\times,\\ q_1'r\equiv uq_1r' \bmod \hat{q}_0}} \sum_{\omega'\in\Omega}$$
$$\sum_{\substack{\omega\in \mathrm{SL}_2(\mathbb{Z}),\\ \omega\equiv u\omega' \bmod \hat{q}_0,\\ \mathbf{s}=q_0(\frac{r}{q}\mathfrak{a}\omega-\frac{r'}{q'}\mathfrak{a}\omega')}} \varphi_Z(\omega)\left|\sum_{\xi\in\mathrm{SL}_2(\mathbb{Z})}\varphi_X(\xi)e_{q_0}(\xi\cdot\mathbf{s})\right|^2.$$

We work from inside out. First, the ξ-sum can be estimated using Proposition 10.2 (contributing $(q_0^{-\frac{3}{2}}X^2 + q_0^3X^{\frac{3}{2}})$). Similarly we treat the ω-sum (contributing $(\hat{q}_0^{-3}Z^2 + Z^{\frac{3}{2}})$) and the ω'-sum is treated trivially (contributing $\sharp\Omega$). The r'-sum can be estimated by $\frac{q'}{\hat{q}_0}$ and the r-sum by q. Together this makes

$$\frac{qq'}{\hat{q}_0} = \frac{qq'q_0}{\hat{q}} \ll \frac{Q^2q_0}{\hat{q}}.$$

As noted earlier we can bound the u-sum by N^ϵ. We are left with

$$\mathcal{E}_1(Q;\mathfrak{a})^2 \ll (\mathfrak{Q}N)^\epsilon \cdot \sharp\Xi \cdot \sum_{Q\ll\hat{q}\ll Q^2}\sum_{q_0\hat{q}_0=\hat{q}} \frac{Q^2 q_0}{\hat{q}}\sharp\Omega\cdot(\hat{q}_0^{-3}Z^2+Z^{\frac{3}{2}})(q_0^{-\frac{3}{2}}X^2+q_0^3X^{\frac{3}{2}})$$

$$\ll N^\epsilon Q^2(\sharp\Xi\cdot\sharp\Omega)^2(XZ)^{2(1-\delta)}\sum_{Q\ll\hat{q}\ll Q^2}\frac{1}{\hat{q}}\left[\frac{1}{\hat{q}^{\frac{1}{2}}}+\frac{1}{Z^{\frac{1}{2}}}+\frac{Q^8}{X^{\frac{1}{2}}}\right].$$

This directly implies the desired estimate. □

We are now ready to put main term and error together.

Theorem 10.6. *For any sufficiently small $\eta>0$ there is $A=A(\eta)$ sufficiently large so that the sequence $\mathfrak{A}$ has level of distribution*

$$\mathfrak{Q}=T^{\frac{1}{32}-\eta}.$$

More precisely there is a multiplicative function $\beta\colon\mathbb{N}\to\mathbb{R}$ so that

$$\prod_{w\le p<z}(1-\beta(p))^{-1}\le C\cdot\left(\frac{\log(z)}{\log(w)}\right)^2,$$

for some $C>1$ and any $2\le w<z$. (This is a quadratic sieve condition.) Furthermore there is a decomposition

$$|\mathfrak{A}_q|=\beta(\mathfrak{q})\cdot\sharp\Pi+r(\mathfrak{q}),$$

so that for all K

$$\sum_{\substack{\mathfrak{q}<\mathfrak{Q},\\ \text{square-free}}}\ll_K\frac{\sharp\Pi}{\log(N)^K}.$$

Finally, we can choose $X=N^{1-\eta}$ (in the construction of Π) so that $\sharp\Pi\gg N^{2\delta-\eta}$.

Proof. From Proposition 10.4 and Theorem 10.5 we get

$$|\mathfrak{A}|=\beta(\mathfrak{q})\cdot\sharp\Pi+\underbrace{r_1(\mathfrak{q})+r_2(\mathfrak{q})+\mathfrak{E}_{\mathfrak{q}}}_{r(\mathfrak{q})}.$$

Recall that we had $X=N^x$, $Y=N^y$ and $Z=N^z$ with $x+y+z=1$. Further set

$$\mathfrak{Q}=T^\alpha=N^{2\alpha}\text{ and }Q_0=N^{\alpha_0}.$$

Note that we can estimate

$$\sum_{\mathfrak{q}<\mathfrak{Q}}|r(\mathfrak{q})|\ll\sharp\Pi\cdot\log(\mathfrak{Q})^2\left(\frac{1}{e^c\sqrt{\log(Y)}}+Q_0^CY^{-\Theta}\right)+\sharp\Pi\cdot\frac{\mathfrak{Q}^\epsilon}{Q_0}+\mathcal{E}$$

by Proposition 10.4. The first two contributions are controlled as long as $y > 0$ and $0 < \alpha_0 < \frac{y\Theta}{C}$. Finally $\mathcal{E}$ can be estimated by Theorem 10.5. The so obtained bound suffices for

$$\begin{aligned} \frac{\alpha_0}{4} &> (x+z)(1-\delta), \\ \frac{z}{4} &> (x+z)(1-\delta) \text{ and} \\ \frac{x}{4} &> 8\alpha + (x+z)(1-\delta). \end{aligned} \tag{35}$$

What is left is an exercise in maximizing α for which there is such a triple (x, y, z). Let η be sufficiently small and put $\alpha = \frac{1}{32} - \eta$. Further assume $\delta > 1 - \eta$ and set $x = 1 - \eta$. It is elementary to show that (35) is satisfied.

We set

$$z = \frac{\eta}{1 + C/\Theta},\ y = z \cdot \frac{C}{\Theta} \text{ and } \alpha_0 = \frac{5}{6}z.$$

Further assume that $\delta > 1 - \frac{\eta}{5(1+C/\Theta)}$. We can guarantee this condition by making A sufficiently large. (Note that $\aleph \in \Gamma_2$ is independent of A!)

We have $y\frac{\Theta}{C} > \alpha_0$ as desired. Finally we observe

$$\frac{z}{4} > \frac{\alpha_0}{4} = \frac{5}{24}z < \frac{1}{5}z > 1 - \delta > (x+z)(1-\delta).$$

Thus all requirements are met and the proof is complete. □

This *level distribution theorem* implies the following key *sieving theorem.*

Theorem 10.7. *Let Π_{AP} denote the set of $\varpi \in \Pi$ for which $\mathrm{Tr}(\varpi)^2 - 4$ is almost prime. That is*

$$\Pi_{AP} = \{\varpi \in \Pi \colon p \mid (\mathrm{Tr}(\varpi)^2 - 4) \implies p > N^{\frac{1}{350}}\}.$$

Then for any sufficiently small η there is an $A = A(\eta)$ sufficiently large and a choice of X, Y, Z so that

$$\sharp\Pi_{AP} > N^{2\delta - \eta}.$$

Proof. Put $\alpha = \frac{1}{34}$. An application of Brun's sieve (**Exercise**) shows that

$$\sum_{\substack{n, \\ (n, P_z)=1}} a_N(n) \gg \frac{\sharp\Pi}{\log(N)^2},$$

for $P_z = \prod_{p<z} p$ and $z \leq T^{\alpha/(9\kappa+1)} = T^{1/646} = N^{1/323}$. We can take $z = N^{\frac{1}{350}}$. Of course any $n = \mathrm{Tr}(\varpi)^2 - 4$ co-prime to P_z has no prime factors below z. The rest of the theorem follows directly. □

It is an easy **Exercise** to show that

$$\sharp\{\gamma \in \Gamma_A \colon \mathrm{Tr}(\gamma) = t\} \ll t^{1+\epsilon}$$

for all $A < \infty$ and $t \geq 1$.

Proof of Theorem 1.12. Let $\alpha = \frac{1}{350}$ and $\eta > 0$ sufficiently small. We have already fund A sufficiently large and a set $\Pi \subset \Gamma_A \cap B_N$ with $\sharp\Pi_{AP} > N^{2\delta-\eta}$. Recall that

$$\Pi_{AP} = \{\varpi \in \Pi \colon p \mid (\mathrm{Tr}(\varpi)^2 - 4) \implies p > N^{\alpha}\}.$$

Also define

$$\Pi_{AP}^{\square} = \{\varpi \in \Pi_{AP} \colon \mathrm{Tr}(\varpi)^2 - 4 \text{ not square-free}\}.$$

We can estimate

$$\begin{aligned}\sharp\{\gamma \in \Gamma_A \cap B_N \colon \mathrm{Tr}(\gamma)^2 - 4 \text{ square-free}\} &\geq \sharp\{\gamma \in \Pi_{AP} \colon \mathrm{Tr}(\gamma)^2 - 4 \text{ square-free}\} \\ &> N^{2\delta-\eta} - \sharp\Pi_{AP}^{\square}.\end{aligned}$$

Suppose $\gamma \in \Pi_{AP}^{\square}$. Then there is a prime $p > N^{\alpha}$ so that $p^2 \mid (\mathrm{Tr}(\gamma)^2 - 4)$. However, this implies that there is $\epsilon \in \{\pm q\}$ so that

$$(p^2 + \epsilon 2) \mid \mathrm{Tr}(\gamma).$$

This implies $p \ll \sqrt{N}$. With these observations made we can trivially estimate

$$\begin{aligned}\sharp\Pi_{AP}^{\square} &\leq \sum_{N^{\alpha} < p \ll N^{\frac{1}{2}}} \sum_{\substack{t < N, \\ t^2 - 4 \equiv 0 \bmod p^2}} \sharp\{\gamma \in \Gamma_A \cap B_N \colon \mathrm{tr}(\gamma) = t\} \\ &\ll \sum_{N^{\alpha} < p \ll N^{\frac{1}{2}}} \frac{N}{p^2} \cdot N^{1+\epsilon} \ll N^{2-\alpha+\epsilon}.\end{aligned}$$

We are done as soon as $2\delta > 2 - \alpha$. This can be arranged by assuming $\delta - \frac{\eta}{2} > 1 - 1/700$, which will fold for A sufficiently large. □

References

1. David Borthwick, *Spectral theory of infinite-area hyperbolic surfaces*, Progress in Mathematics, vol. 256, Birkhäuser Boston, Inc., Boston, MA, 2007. MR 2344504
2. Jean Bourgain, Alex Gamburd, and Peter Sarnak, *Affine linear sieve, expanders, and sum-product*, Invent. Math. **179** (2010), no. 3, 559–644. MR 2587341
3. ———, *Generalization of Selberg's $\frac{3}{16}$ theorem and affine sieve*, Acta Math. **207** (2011), no. 2, 255–290. MR 2892611
4. Jean Bourgain and Alex Kontorovich, *On representations of integers in thin subgroups of* $\mathrm{SL}_2(\mathbb{Z})$, Geom. Funct. Anal. **20** (2010), no. 5, 1144–1174. MR 2746949
5. ———, *On Zaremba's conjecture*, Ann. of Math. (2) **180** (2014), no. 1, 137–196. MR 3194813
6. ———, *Beyond expansion II: low-lying fundamental geodesics*, J. Eur. Math. Soc. (JEMS) **19** (2017), no. 5, 1331–1359. MR 3635355
7. Jean Bourgain, Alex Kontorovich, and Peter Sarnak, *Sector estimates for hyperbolic isometries*, Geom. Funct. Anal. **20** (2010), no. 5, 1175–1200. MR 2746950
8. Jens Carsten Jantzen, *Character formulae from Hermann Weyl to the present*, Groups and analysis, London Math. Soc. Lecture Note Ser., vol. 354, Cambridge Univ. Press, Cambridge, 2008, pp. 232–270. MR 2528469
9. Svetlana Katok, *Fuchsian groups*, Chicago Lectures in Mathematics, University of Chicago Press, Chicago, IL, 1992. MR 1177168

10. Alex Kontorovich, *Applications of thin orbits*, Dynamics and analytic number theory, London Math. Soc. Lecture Note Ser., vol. 437, Cambridge Univ. Press, Cambridge, 2016, pp. 289–317. MR 3618792
11. Alex Kontorovich, D. Darren Long, Alexander Lubotzky, and Alan W. Reid, *What is ... a thin group?*, Notices Amer. Math. Soc. **66** (2019), no. 6, 905–910. MR 3929581
12. George Lusztig, *Some problems in the representation theory of finite Chevalley groups*, The Santa Cruz Conference on Finite Groups (Univ. California, Santa Cruz, Calif., 1979), Proc. Sympos. Pure Math., vol. 37, Amer. Math. Soc., Providence, R.I., 1980, pp. 313–317. MR 604598
13. Geordie Williamson, *Schubert calculus and torsion explosion*, J. Amer. Math. Soc. **30** (2017), no. 4, 1023–1046, With a joint appendix with Alex Kontorovich and Peter J. McNamara. MR 3671935