J. reine angew. Math. **753** (2019), 89–135
DOI 10.1515/crelle-2016-0072

Journal für die reine und angewandte Mathematik

Uniform congruence counting for Schottky semigroups in $\mathrm{SL}_2(\mathbf{Z})$

By *Michael Magee* at New Haven, *Hee Oh* at New Haven and *Dale Winter* at Princeton

With an appendix by *Jean Bourgain* at Princeton, *Alex Kontorovich* at New Brunswick and *Michael Magee* at New Haven

Abstract. Let Γ be a Schottky semigroup in $\mathrm{SL}_2(\mathbf{Z})$, and for $q \in \mathbf{N}$, let

$$\Gamma(q) := \{\gamma \in \Gamma : \gamma = e \ (\mathrm{mod}\ q)\}$$

be its congruence subsemigroup of level q. Let δ denote the Hausdorff dimension of the limit set of Γ. We prove the following uniform congruence counting theorem with respect to the family of Euclidean norm balls B_R in $M_2(\mathbf{R})$ of radius R: for all positive integer q with no small prime factors,

$$\#(\Gamma(q) \cap B_R) = c_\Gamma \frac{R^{2\delta}}{\#(\mathrm{SL}_2(\mathbf{Z}/q\mathbf{Z}))} + O(q^C R^{2\delta-\epsilon})$$

as $R \to \infty$ for some $c_\Gamma > 0, C > 0, \epsilon > 0$ which are independent of q. Our technique also applies to give a similar counting result for the continued fractions semigroup of $\mathrm{SL}_2(\mathbf{Z})$, which arises in the study of Zaremba's conjecture on continued fractions.

1. Introduction

Let $\mathrm{SL}_2(\mathbf{R})$ act on $\mathbf{R} \cup \{\infty\}$ by Möbius transformations. We say that the collection of elements $g_1, \dots, g_k \in \mathrm{SL}_2(\mathbf{R})$, $k \geq 2$, is a Schottky generating set if there exist mutually disjoint compact intervals $I_1, \dots, I_k, J_1, \dots, J_k$ in $\mathbf{R}$ such that g_i maps the exterior of J_i onto the interior of I_i for each $1 \leq i \leq k$. We call a semigroup $\Gamma \subset \mathrm{SL}_2(\mathbf{R})$ Schottky if it is generated by some Schottky generating set as a semigroup. By the ping-pong argument, Schottky semigroups are necessarily discrete and free. Schottky semigroups are ubiquitous in $\mathrm{SL}_2(\mathbf{R})$; for instance, for any hyperbolic elements $h_1, h_2 \in \mathrm{SL}_2(\mathbf{R})$ with no common fixed points on $\mathbf{R} \cup \{\infty\}$, the pair h_1^m, h_2^m forms a Schottky generating set for all sufficiently large m.

Bourgain is supported in part by NSF grant DMS-1301619. Kontorovich is supported in part by an NSF CAREER grant DMS-1254788 and DMS-1455705, an NSF FRG grant DMS-1463940, and Alfred P. Sloan Research Fellowship, and a BSF grant. Magee was supported in part by NSF Grant DMS-1128155. Oh was supported in part by NSF Grant DMS-1361673.

When Γ is a semigroup in $\mathrm{SL}_2(\mathbf{Z})$ and $q \in \mathbf{N}$, the congruence subsemigroup of Γ of level q is defined by

$$\Gamma(q) := \{\gamma \in \Gamma : \gamma = e \bmod q\}.$$

The main aim of this paper is to study a congruence lattice point counting problem for $\Gamma(q)$ in a Schottky semigroup $\Gamma \subset \mathrm{SL}_2(\mathbf{Z})$ with a uniform power-savings error term. For $R > 0$, consider the ball of radius R with respect to the Frobenius norm:

$$B_R := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbf{R}) : \sqrt{a^2 + b^2 + c^2 + d^2} < R \right\}.$$

The following is a simplified version of our main theorem (see Theorem 11 for a more refined version):

Theorem 1. *If Γ is a Schottky semigroup of* $\mathrm{SL}_2(\mathbf{Z})$, *there exist $Q_0 \in \mathbf{N}$, $c_\Gamma > 0$, $C > 0$ and $\epsilon > 0$ such that for all $q \in \mathbf{N}$ with $(Q_0, q) = 1$,*

$$\#\Gamma(q) \cap B_R = c_\Gamma \frac{R^{2\delta}}{\#\mathrm{SL}_2(\mathbf{Z}/q\mathbf{Z})} + O\big(q^C R^{2\delta - \epsilon}\big),$$

where $\delta > 0$ is the Hausdorff dimension of the limit set of Γ.

The limit set of Γ is the set of all accumulation points of an orbit $\Gamma.o$ in $\mathbf{R} \cup \{\infty\}$.

Remark. (1) When Γ is a Schottky *subgroup* of $\mathrm{SL}_2(\mathbf{Z})$, the analogous result to Theorem 1 was proved by Gamburd [9] for $\delta > 5/6$, by Bourgain–Gamburd–Sarnak [4] for $\delta > 1/2$ and by Oh–Winter [15] for any $\delta > 0$. The last two results are restricted to the moduli condition of q square-free. The counting result of Oh–Winter is deduced from [13] based on the uniform exponential mixing of the geodesic flow for the congruence covers of a Schottky surface, and hence does not apply to the semigroup counting.

(2) So the novelty of Theorem 1 lies in the treatment of a Schottky *semigroup* and the uniformity of the power-savings error term for *all* moduli q (with no small prime factors). The extension to the arbitrary moduli q case relies on the new technology that appears in the Appendix by Bourgain, Kontorovich and Magee.

(3) We also remark that for fixed q, Theorem 1 follows from the work of Naud [14] in this generality. We refer to [4] for more backgrounds on earlier related works.

Our methods also apply to a congruence family of semigroups related to continued fractions and Diophantine approximation. Let $\mathcal{A}$ be a finite set of at least two positive integers. Define $\mathcal{G}_A$ to be the subsemigroup of $\mathrm{GL}_2(\mathbf{Z})$ generated by

$$g_a := \begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix}, \quad a \in \mathcal{A}.$$

We define the *continued fractions semigroup* $\Gamma_{\mathcal{A}}$ as follows:

$$\Gamma_{\mathcal{A}} := \mathcal{G}_{\mathcal{A}} \cap \mathrm{SL}_2(\mathbf{Z}),$$

in other words, $\Gamma_{\mathcal{A}}$ is a semigroup generated by $\{g_a g_{a'} : a, a' \in \mathcal{A}\}$. The continued fractions semigroup $\Gamma_{\mathcal{A}}$ is not a Schottky semigroup; however the methods of proof of Theorem 1 apply as well.

Theorem 2. *Theorem* 1 *also holds for the continued fractions semigroup* $\Gamma_{\mathcal{A}}$.

In order to explain the relation of $\Gamma_{\mathcal{A}}$ with continued fractions, we set

$$[a_1, \dots, a_l, \dots] := \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots \cfrac{1}{a_l + \ddots}}}$$

for any sequence of $a_i \in \mathbf{N}$.

Write

$$\mathfrak{R}_{\mathcal{A}} := \{[a_1, \dots, a_k] : k \in \mathbf{N},\, a_i \in \mathcal{A} \text{ for all } i\}$$

for the set of approximants to $\mathfrak{C}_{\mathcal{A}}$, and $\mathfrak{D}_{\mathcal{A}}$ for the set of denominators of reduced elements of $\mathfrak{R}_{\mathcal{A}}$, that is,

$$\mathfrak{D}_{\mathcal{A}} := \left\{ d : \frac{b}{d} \in \mathfrak{R}_{\mathcal{A}} \text{ for some } b \text{ coprime to } d \right\}.$$

For an integer $A \in \mathbf{N}$, we write $\mathfrak{D}_{[A]} = \mathfrak{D}_{\{1,2,\dots,A\}}$. In [19], Zaremba made the following conjecture, motivated by applications to numerical analysis.

Conjecture 3 (Zaremba). There is some absolute $A \in \mathbf{N}$ such that $\mathfrak{D}_{[A]} = \mathbf{N}$.

Observe that

$$\frac{b}{d} = [a_1, \dots, , a_k]$$

if and only if

$$\begin{pmatrix} 0 & 1 \\ 1 & a_1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & a_2 \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & a_k \end{pmatrix} = \begin{pmatrix} \star & b \\ \star & d \end{pmatrix}.$$

This yields the relation

$$\mathfrak{D}_{\mathcal{A}} = \{\langle \gamma(0,1)^t, (0,1)^t \rangle : \gamma \in \mathcal{G}_{\mathcal{A}}\}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbf{R}^2$, thus enters the semigroup $\mathcal{G}_A$ in the study of continued fractions.

Bourgain and Kontorovich [5, Theorem 1.2] proved that Zaremba's conjecture is true after replacing $\mathbf{N}$ by a density one subset. That is, there is some A such that

$$\#\mathfrak{D}_{[A]} \cap \{1, \dots, N\} = N + o(N). \tag{1.1}$$

Furthermore, they showed that the $o(N)$ term can be taken to be $O(N^{1-c/\log\log N})$ for suitable $c > 0$ (this relies on the Appendix) and $A = 50$ will suffice. The size of A has since been improved to $A = 5$ by Huang [10], following previous innovations by Frolenkov and Kan [8] on the necessary $\delta_{\mathcal{A}}$.

Theorem 2 provides the precise missing ingredient in Bourgain and Kontorovich's work [5], to replacing the $o(N)$ bound for the size of the exceptional set in (1.1) with a power savings error $O(N^{1-\epsilon})$. Indeed, combining Bourgain and Kontorovich's method from [5], Huang's refinement, and with the counting estimate of Theorem 2 and its technical form Theorem 11 in place of [5, Theorem 8.1], one can derive the following improvement of (1.1): for $\mathcal{A} = \{1, 2, 3, 4, 5\}$ and for some $\epsilon > 0$,

$$\#\mathfrak{D}_{\mathcal{A}} \cap \{1, \dots, N\} = N + O(N^{1-\epsilon}). \tag{1.2}$$

The key point is that the uniform lattice point count enables us to replace the parameter $\mathcal{Q} = N^{\alpha_0/\log\log N}$ in [5] and [10] with a power of N.

We remark that a short alternative argument for (1.2) was recently proposed by Bourgain in [2]. His argument deviates from the approach of [5] and hence does not require orbital counting estimates.

We draw the reader's attention to the survey article [3] where other applications to continued fractions are discussed. The reader can also see the survey of Kontorovich [11] that situates Zaremba's conjecture amongst other problems in the 'thin (semi)groups' setting.

Overview of the proofs of Theorems 1 and 2. The basic strategy is to regard our Schottky semigroup setup as an expanding map and to apply transfer operator techniques. Necessary spectral bounds are then deduced by synthesizing work of Bourgain–Varjú, Bourgain–Gamburd–Sarnak, Dolgopyat, and Naud. For now we focus on the arguments for Theorem 1; those for Theorem 2 are similar.

We consider the map $T : I := \bigcup_{i=1}^{k} I_i \to \mathbf{R}$ defined by

$$T|_{I_i} = (g_i)^{-1}$$

and the distortion function $\tau : I \to \mathbf{R}$ given by $\tau(x) = \log|T'(x)|$, which is eventually positive in our setting. The transfer operator $\mathcal{L}_s$ is defined for all $s \in \mathbf{C}$ by

$$\mathcal{L}_s(f)(x) = \sum_{Ty=x} e^{-s\tau(y)} f(y)$$

as a bounded linear operator on $C^1(I)$. Lalley's renewal equation [12] provides a link between the counting problem for Γ and spectral bounds for $\mathcal{L}_s$. Such spectral bounds were obtained by Naud [14], who provided a C^1-operator norm estimate on $\mathcal{L}_s^m$ valid on a strip $|\Re(s) - \delta| < \epsilon$ and so deduced[1] the case $q = 1$ of Theorem 1.

To provide a counting result that is uniformly accurate over congruence semigroups we must actually deal with the congruence transfer operators. More precisely, let

$$c_q : I \to \mathrm{SL}_2(\mathbf{Z}/q\mathbf{Z})$$

be the cocycle given by

$$c_q|_{I_i} = g_i \bmod q,$$

and define the congruence transfer operator

$$\mathcal{L}_{s,q}[F](x) = \sum_{Ty=x} e^{-s\tau y} c_q(y).F(y)$$

[1] Naud uses Ruelle zeta function techniques as in [17], in contrast to our use of the renewal equation.

on the space of $\mathbf{C}^{\Gamma_q}$-valued functions for $\Gamma_q := \mathrm{SL}_2(\mathbf{Z}/q\mathbf{Z})$. The composition $c_q(y).F(y)$ is the result of applying $c_q(y) \in \Gamma_q$ to the vector $F(y) \in \mathbf{C}^{\Gamma_q}$ by the right regular representation of Γ_q. It is also useful throughout the paper to think of F as a function on $I \times \mathbf{C}^{\Gamma_q}$. We fix the standard Hermitian form on $\mathbf{C}^{\Gamma_q}$ that comes from the identification of Γ_q with the standard basis of $\mathbf{C}_q^{\Gamma}$ and defining $\langle g_1, g_2 \rangle = \delta_{g_1, g_2}$. The space $\mathbf{C}^{\Gamma_q} \ominus 1$ is defined to be the space of functions that are orthogonal to constants with respect to the fixed Hermitian form. The vector space $\mathbf{C}^{\Gamma_q} \ominus 1$ inherits a Hermitian form from that of $\mathbf{C}^{\Gamma_q}$. It is with respect to this form that we define the Banach spaces $C^1(I; \mathbf{C}^{\Gamma_q} \ominus 1)$ that play a central role in this paper.

The following is the main technical result:

Theorem 4 (Bounds for congruence transfer operators). *Write $s = a + ib$. There is $Q_0 \in \mathbf{N}$ such that for any $\eta > 0$, there are $\epsilon = \epsilon(\eta) > 0$, $b_0 > 0$, $0 < \rho_\eta < 1$, $C_\eta > 0$, $r > 0$, $0 < \rho_0 < 1$ and $C > 0$ such that the following holds for all $a \in \mathbf{R}$ with $|a - \delta| < \epsilon$ and $b \in \mathbf{R}$:*

(1) *When $|b| \leq b_0$ and $f \in C^1(I; \mathbf{C}^{\Gamma_q} \ominus 1)$,*

$$\|\mathcal{L}_{s,q}^m f\|_{C^1} \leq C q^C \rho_0^m \|f\|_{C^1}$$

when $(q, Q_0) = 1$. Here $\mathbf{C}^{\Gamma_q} \ominus 1$ is the orthogonal complement to the constant functions in the right regular representation of Γ_q.

(2) *When $|b| > b_0$,*

$$\|\mathcal{L}_{s,q}^m\|_{C^1} \leq C_\eta |b|^{1+\eta} \rho_\eta^m$$

uniformly with respect to $q \in \mathbf{N}$.

The transfer operators have two parameters s, the Laplace transform-dual/frequency version of the counting parameter, and q, the modular parameter. Since inverting the Laplace transform that was taken involves an infinite vertical contour, one must obtain spectral bounds that are uniform in s with $\Re(s)$ within some fixed small window of δ. The bounds should also be uniform with respect to the currently considered family of moduli q. These bounds rely on two different inputs that both involve deep ideas.

To address large imaginary part of s considerations, we will use the method of Dolgopyat from [7], and its further development by Naud from [14]. We follow Naud's analysis from [14] up to the point of departure from Naud's work in Lemma 29 where we extend [14, Lemma 5.10] to vector-valued functions. Here, an important point that prevents the cocycle c_q from interfering with the non-stationary phase is that it is locally constant. We mention that this observation was first due to [15] where they consider the congruence transfer operator associated to the Markov partition given by the geodesic flow.

For bounded $\Im(s)$ and varying q we follow the work of Bourgain, Gamburd and Sarnak from [4] and the work of Bourgain, Kontorovich and Magee in the Appendix, which allows us to relate the norm $\|\mathcal{L}_{s,q}^m\|_{C^1}$ to the expander result on the Cayley graphs of the Γ_q with respect to a fixed generating set of g_i's. The main reason behind our successful treatment of arbitrary moduli q case is the work of Bourgain-Varjú establishing the expander result for $\mathrm{SL}_2(\mathbf{Z}/q\mathbf{Z})$ for arbitrary q, as explained in the Appendix.

Acknowledgement. We would like to thank Peter Sarnak for his encouragement and support throughout this project. We thank Jean Bourgain, Alex Kontorovich and Curt McMullen for helpful comments on an earlier version of this paper.

2. Dynamics and Thermodynamics on the boundary

2.1. The dynamical system T. We construct a dynamical system $T : I \to \mathbf{R}$ on a disjoint union of intervals I that plays a central role in the counting estimates of our main Theorems 1 and 2, and set up the notations and the assumptions which will be used throughout the paper.

I: Schottky semigroup case. Let $g_1, \dots, g_{k'}$ $(k' \geq 2)$ be the Schottky generating set in $\mathrm{SL}_2(\mathbf{Z})$. We let $\{\tilde{I}_i, \tilde{J}_i : i = 1, \dots, k'\}$ be the intervals such that g_i maps the exterior of $\tilde{J}_i$ onto the interior of $\tilde{I}_i$ as in the definition of the Schottky generators. Set $g_{k'+\ell} = g_\ell^{-1}$ and $\tilde{I}_{k'+\ell} = \tilde{J}_\ell$ for $1 \leq \ell \leq k'$.

For any $0 \leq \ell \leq k'$, let Γ be the semigroup generated by $g_1, \dots, g_{k'}, g_{k'+1}, \dots, g_{k'+\ell}$; we will call Γ a Schottky semigroup. This is slightly more general than the definition we gave in the Introduction, and the main reason of this extension is to include Schottky groups in our discussion of Schottky semigroups. Note that when $\ell = k'$, Γ coincides with the Schottky *subgroup* generated by $g_1, \dots, g_{k'}$.

Set $p = k' + \ell$. We now define a map $B : \tilde{I} \to \mathbf{R} \cup \{\infty\}$ for $\tilde{I} := \bigcup_{i=1}^{p} \tilde{I}_i$ by the piecewise Möbius action

$$B|_{\tilde{I}_i} = g_i^{-1}.$$

Since $g_i(\infty) \in \tilde{I}_i$, the image of B contains ∞.

The *cylinders* of length n are by definition the sets of the form

$$\tilde{I}_{i_1} \cap B^{-1}(\tilde{I}_{i_2}) \cap \dots \cap B^{-(n-1)}(\tilde{I}_{i_n}),$$

where each $1 \leq i_j \leq p$. Let I be the union of the cylinders of length 2 and define

$$T : I \to \mathbf{R}$$

to be the restriction of B to I. Note that $g_i(\infty) \notin I$ and hence the image of T does not contain ∞; it is for this reason that we replaced $\tilde{I}$ with I. Finally, we say that a sequence $g_{i_1}, g_{i_2}, g_{i_3}, \dots$ of the Schottky generators is *admissible* if no g_{i_j} is followed by its inverse. This means all the words obtained by concatenating consecutive subsequences are reduced. We now let k denote the number of cylinders of length 2.

II: Continued fraction semigroup case. Let $\mathcal{A}$ be a finite subset of $\mathbf{N}$ with at least two elements. For $a \in \mathcal{A}$, set

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

Let Γ be the continued fractions semigroup $\Gamma_{\mathcal{A}}$ generated by $g_a g_{a'}$, $a, a' \in \mathcal{A}$. Since $a, a' \geq 1$, it follows that the trace of any element of Γ is strictly bigger than 2 and hence every element of Γ is hyperbolic.

Note that the g_a acts as Möbius transformations on $\mathbf{R} \cup \{\infty\}$ by

$$g_a(z) = \frac{1}{z + a}.$$

Let A denote the largest member of $\mathcal{A}$ and consider the interval $I_A := [\frac{1}{A+1}, 1]$.

For $a \in \mathcal{A}$, let $I_a := g_a I_A$, which can be computed to be

$$I_a = \left[\frac{1}{a+1}, \frac{1}{a+(A+1)^{-1}}\right] \subset \left[\frac{1}{a+1}, \frac{1}{a}\right].$$

The I_a are clearly disjoint as $A \geq 1$. It follows that the g_a generate a free semigroup by the ping-pong argument. We also record for later use that the derivative of the Möbius action has

$$g_a'(z) = \frac{1}{|z+a|^2} \leq (a+(1+A)^{-1})^{-2} \leq (1+(1+A)^{-1})^{-2} \tag{2.1}$$

for all $z \in I_A$. We now set

$$I_{a,a'} := g_a g_{a'} I_A \subset I_a$$

obtaining a disjoint collection of $\#\mathcal{A}^2$ number of closed intervals. Rename these intervals $I_{a,a'}$ and corresponding elements $g_a g_a'$ as I_i's and g_i's, respectively.

Define

$$T : I \to \mathbf{R}, \quad T|_{I_i} = (g_i)^{-1}.$$

Note that $g_a g_{a'} I \subset I_{a,a'}$, in other words, $g_i I \subset I_i$ for each $1 \leq i \leq \#\mathcal{A}^2$. Again, we let $k = \#\mathcal{A}^2$ denote the number of intervals obtained.

Set-up. In the rest of this paper, let Γ be a Schottky semigroup or the continued fractions semigroup, with the associated locally analytic map

$$T : I = \bigcup_i I_i \to \mathbf{R} \quad \text{given by } T|_{I_i} = g_i^{-1}$$

constructed above.

It follows easily from the construction that we have the

Markov property. If $T(I_i) \cap I_j \neq 0$, then $T(I_i) \supset I_j$.

Proposition 5. *The map T is eventually expanding, that is, there are $D > 0$, $\gamma > 1$ such that for all $N \geq 1$ and $x \in T^{-N+1}(I)$,*

$$|(T^N)'(x)| \geq D^{-1}\gamma^N,$$

wherever the derivative exists[2] *in $T^{-N+1}(I)$.*

Proof. For the Schottky semigroup case, this can be proved exactly as in the proof of [1, Proposition 15.4]. For the continued fraction case it follows from (2.1) and the chain rule that for any $z \in I$,

$$|T'(z)| \geq (1+(1+A)^{-1})^4 > 1$$

and hence the claim follows. □

[2] The derivative may have poles.

We also must introduce the following *distortion function* on I.

Definition 6 (Distortion function). The distortion function $\hat{\tau} : I \to \mathbf{R}$ is defined by

$$\hat{\tau}(x) = \log|T'(x)|.$$

This definition is very natural for our purposes. For certain technical calculations, however, it is easier to work with a slightly different version. We consider the Cayley map J from the upper half plane to the unit disc sending i to the center 0 of the disc. We can therefore think of T as acting on the subset $J(I)$ of the unit circle. This gives an alternative distortion function.

Definition 7 (Distortion function II). The distortion function $\tau : I \to \mathbf{R}$ is defined by

$$\tau(x) = \log|(J \circ T \circ J^{-1})'(Jx)|.$$

The two distortion functions mentioned here are cohomologous (that is, there is a function $h = -\log(J')$ such that $\hat{\tau}(x) = \tau(x) + h(x) - h(T(x))$), so are equivalent for many purposes. Sometimes it is convenient to work with one, sometimes the other.

Since T is real analytic, and it is easy to see that T' is never zero on I, it follows that τ is real analytic on I. The iterated version

$$\tau^N(x) := \sum_{i=0}^{N-1} \tau(T^i x)$$

measures the distortion along a trajectory of T. It follows from the eventually expanding property of T that there is an N_0 such that for all $N \geq N_0$, $\hat{\tau}^N > 0$ on the cylinders of length N, that is, $\hat{\tau}$ is *eventually positive.* Since τ is cohomologous to $\hat{\tau}$ we conclude that τ is also eventually positive.

Let d_E denote Euclidean distance in the upper half plane. Fix the basepoint $o = i \in \mathbb{H}$. The following lemma links the lattice point count with the dynamical system we have defined.

Lemma 8. *There exist $C, r > 0$ and $\kappa < 1$ such that if k_0 is a point in I, then for $L \in \mathbf{N}$ and admissible sequence of g_{i_j},*

$$d_E(g_{i_1 \dots} g_{i_L} o, g_{i_1 \dots} g_{i_L} k_0) \leq C\kappa^L. \tag{2.2}$$

If in the general Schottky semigroup case, we also require that $k_0 \notin \tilde{I}_i$, where

$$i = i_L + k' \bmod 2k'.$$

Proof. Inequality (2.2) follows from the fact that Möbius transformations preserve (generalized) circles orthogonal to the boundary of $\mathbb{H}$, together with the eventually expanding property of T. □

We denote by K the limit set of the semigroup Γ, i.e., the set of all accumulation points in $\partial(\mathbb{H}) = \mathbf{R} \cup \{\infty\}$ of the orbit $\Gamma.o$. It follows from Lemma 8 that the limit set K is also given by the T-invariant set

$$K = \bigcap_{i=1}^{\infty} T^{-i}(I).$$

In order to perform counting in congruence classes, we need to twist our dynamical system by a family of locally constant maps. Let $\Gamma_q = \mathrm{SL}_2(\mathbf{Z}/q\mathbf{Z})$.

Definition 9 (Modular cocycle). For every modulus $q \in \mathbf{N}$, define $c_q : I \to \Gamma_q$ by

$$c_q|_{I_i} = g_i \bmod q.$$

This quantity will appear again naturally in Section 3 when we perform the lattice point count.

2.2. Thermodynamics. For a T-invariant probability measure μ on K, let $h_\mu(T)$ denote the measure-theoretic entropy of T with respect to μ. Let $\mathcal{M}(K)^T$ denote the set of all T-invariant probability measures on K.

The *pressure functional* is defined on $f \in L(K)$ by

$$P(f) := \sup_{\mu \in \mathcal{M}(K)^T} \left(h_\mu(T) - \int_K f \, d\mu \right).$$

It follows from the variational principle that $P(-s\tau)$ is strictly decreasing in a real parameter s and has a unique positive zero denoted by s_0. Moreover, it is known that in the current setting $s_0 = \delta$, where δ is the Hausdorff dimension of K.

Let $L(K)$ denote the Banach space of Lipschitz functions on K. For any real-valued $f \in L(K)$, the transfer operator $\mathcal{L}_f$ on $L(K)$ is given by

$$\mathcal{L}_f[G](x) = \sum_{Ty=x} e^{f(y)} G(y).$$

The basic spectral theory of transfer operators is given by the Ruelle–Perron–Frobenius Theorem. We state this following Naud [14], the result can also be found in [16].

Theorem 10 (Ruelle–Perron–Frobenius). *The following statements hold.*

1. *There is a unique probability measure ν_f on K such that $\mathcal{L}_f^*(\nu_f) = e^{P(f)}\nu_f$.*
2. *The maximal eigenvalue of $\mathcal{L}_f$ is $e^{P(f)}$ which belongs to a unique positive eigenfunction $h_f \in L(K)$ with $\nu_f(h_f) = 1$.*
3. *The remainder of the spectrum of $\mathcal{L}_f$ is contained in a disc of radius strictly less than $e^{P(f)}$.*

Our functional analysis takes place for the most part on the Banach space $C^1(I)$ with the norm

$$\|f\|_{C^1(I)} = \|f\|_\infty + \|f'\|_\infty, \tag{2.3}$$

or closely related spaces of vector-valued functions. As in [14] we need to note that Theorem 10 extends reasonably to $\mathcal{L}_f$ acting on $C^1(I)$ given $f \in C^1(I)$. In particular, $\mathcal{L}_f$ acting on $C^1(I)$ has the same spectral properties relative to a positive eigenfunction $h_f \in C^1(I)$ such that $\mathcal{L}_f h_f = e^{P(f)} h_f$. We also view ν_f as a measure on I with support in K.

We will write simply $\mathcal{L}_s = \mathcal{L}_{-s\tau}$ in the sequel.

3. Counting

3.1. From the lattice point count to the boundary dynamics. We now show how one can adapt the work of Lalley [12] to get counting estimates in our setting. Let $\Gamma_q = \mathrm{SL}_2(\mathbf{Z}/q\mathbf{Z})$. We convert questions about the lattice point count in congruence classes into questions about the $\mathbf{R}^{\Gamma_q}$-valued function

$$N_q^*(a, \gamma_0, \varphi) := \sum_{\substack{\gamma \in \Gamma \cup \{e\} \\ d(o,\gamma\gamma_0 o) - d(o,\gamma_0 o) \leq a}} G(\gamma\gamma_0 o)\rho(\pi_q(\gamma)).\varphi,$$

where

- G is a non-negative function on $\mathbb{H} \cup \mathbf{R}$ with the property that there exist an integer M and neighborhood J_M of the length M cylinders in I such that G is locally constant on J_M. We write g for the restriction of G to $\mathbf{R}$.
- $\varphi \in \mathbf{R}^{\Gamma_q}$, $\pi_q : \Gamma \to \Gamma_q$ is reduction mod q and ρ is the right regular representation of Γ_q.
- $o = i \in \mathbb{H}$ is our fixed origin and $\gamma_0 \in \Gamma \cup \mathrm{id}$.

While this might seem mysterious, we explain as follows.

Firstly, and most importantly, the main Theorem 1 *stated in our Introduction is directly analogous to certain estimates for* $N_q^*(a, \mathrm{id}, \varphi)$ *for suitable test* φ.

The distance d vs the matrix norm $\|\gamma\|$. One has the identity

$$\|\gamma\|^2 = 2\cosh(d(i, \gamma i)).$$

With this in hand and our choice $o = i$ of basepoint, the condition $d(i, \gamma\gamma_0 i) - d(i, \gamma_0 i) \leq a$ becomes

$$\frac{\|\gamma\gamma_0\|}{\|\gamma_0\|} \leq R,$$

where $R = \sqrt{2\cosh(a)} = e^{a/2}$.[3)]

The parameter γ_0. Our main Theorem 1 of the Introduction is obtained by setting $\gamma_0 = \mathrm{id}$. However, even to obtain this simplified version, consideration of general γ_0 is necessary in order to set up the forthcoming recursion over the tree-like Γ. This recursive formula leads to the renewal equation.

[3)] More precisely, the condition $\frac{\|\gamma\gamma_0\|}{\|\gamma\|} \leq R$ corresponds to an inequality

$$(*) \quad d(i, \gamma\gamma_0 i) - d(i, \gamma_0 i) \leq 2\log R + \log(1 + e^{-2d(i,\gamma_0 i)}) + \log\left(1 + \sqrt{1 - \frac{1}{R^4 \cosh^2 d(i, \gamma_0 i)}}\right)$$
$$= a + \log(1 + e^{-2d(i,\gamma_0 i)}) + O(e^{-2a}).$$

The difference is only important insofar as it changes the leading constant in our main theorem.

The function G. This function allows one to perform sector estimates by only counting lattice points that fall close to a prescribed part of the boundary $\partial(\mathbb{H})$ of hyperbolic space.

Modular twisting. Let us now explain the modular twisting in the simple case that $G := 1$. Recall that we are supposed to be counting in a given congruence class $\xi \in \Gamma_q$. One can decompose the characteristic function of the singleton set ξ according to its constant coefficient and a part orthogonal to constants, and look at $N(a, \gamma_0, \varphi)$ with φ set in turn to these different components. Since the estimate is additive one can estimate the corresponding quantities separately. The key calculation is that

$$N_q^*(a, \mathrm{id}, \mathbf{1}_\xi) = \sum_{\substack{\gamma \in \Gamma \cup \{e\} \\ d(o,\gamma o) \leq a}} \rho(\pi_q(\gamma)).\mathbf{1}_\xi = \sum_{\substack{\gamma \in \Gamma \cup \{e\} \\ d(o,\gamma o) \leq a}} \mathbf{1}_{\xi \pi_q(\gamma)},$$

so one obtains the congruence lattice point count from reading off a coordinate of the vector-valued $N_q^*(a, \mathrm{id}, \mathbf{1}_\xi)$.

Remark. Whenever we sum over semigroup elements, we have the implied constraint that any concatenation in the summation condition be admissible; we will use the notation $\sum^*$ to emphasize this. For example, we will write

$$\sum_{\substack{\frac{\|\gamma\gamma_0\|}{\|\gamma_0\|} \leq R \\ \gamma \equiv \xi \bmod q}}^{*} G(\gamma\gamma_0 o) := \sum_{\substack{\frac{\|\gamma\gamma_0\|}{\|\gamma_0\|} \leq R \\ \gamma \equiv \xi \bmod q \\ \gamma \cdot \gamma_0 \text{ admissible}}} G(\gamma\gamma_0 o).$$

The most general lattice point count that the upcoming estimates for $N(a, \gamma_0, \varphi)$ will allow us to obtain is the following.

Theorem 11 (Main Theorem, elaborated). *There exist $Q_0 \in \mathbf{N}$, $C > 0$ and $\epsilon > 0$ such that for all $\gamma_0 \in \Gamma$, $\xi \in \mathrm{SL}_2(\mathbf{Z}/q\mathbf{Z})$ and q with $(Q_0, q) = 1$,*

$$\sum_{\substack{\frac{\|\gamma\gamma_0\|}{\|\gamma_0\|} \leq R \\ \gamma \equiv \xi \bmod q}}^{*} G(\gamma\gamma_0 o) = \frac{R^{2\delta}}{|\Gamma_q|} \hat{C}_*(\gamma_0, G|_{\mathbf{R}}) + O\big((\|G\|_\infty + \|[G|_{\mathbf{R}}]'\|_\infty) q^C R^{2(\delta - \epsilon)}\big).$$

Here G is any function in $C^1(\mathbb{H} \cup \mathbf{R})$ which is locally constant on some neighborhood of the cylinders of length M in I for some $M > 0$. The constant $\hat{C}_*(\gamma_0, G|_{\mathbf{R}}) > 0$ is related to C_* from (3.12) *but modified in light of* $(*)$. *The implied constant depends on M.*

We now show how to relate the quantities N_q^* and the dynamics on the boundary. As before, write d_E for Euclidean distance in the upper half plane. Let $\Gamma^{(n)}$ denote those $\gamma \in \Gamma$ which can be written as a reduced word in at least n generators. If $\gamma = g_{i_1} g_{i_2} \dots g_{i_n}$ is written in reduced form, then we define the *shift*

$$\sigma : \Gamma^{(n)} \to \Gamma^{(n-1)}, \quad \sigma(\gamma) = g_{i_2} \dots g_{i_n}.$$

We use the convention that $\Gamma^{(0)} = \Gamma \cup \{e\}$ and $\sigma(g_i) = e$ for all $1 \leq i \leq k$. Throughout the rest of this section we always assume semigroup elements are written in their reduced form.

Define for $\gamma \in \Gamma$

$$\tau_*(\gamma) = d(o, \gamma o) - d(o, (\sigma\gamma)o).$$

and define for $n \geq N$ and $\gamma \in \Gamma^{(n)}$,

$$\tau_*^N(\gamma) = \sum_{j=0}^{N-1} \tau_*(\sigma^j \gamma) = d(o, \gamma o) - d(o, (\sigma^N \gamma)o).$$

We can now recast N_q^* as

$$N_q^*(a, \gamma_0, \varphi) = \sum_{n=0}^{\infty} \sum_{\substack{\gamma \in \Gamma \\ \sigma^n \gamma = \gamma_0}} G(\gamma o)\rho(\pi_q(\gamma\gamma_0^{-1})) \cdot \varphi \mathbf{1}\{\tau_*^n(\gamma) \leq a\}.$$

One obtains by this elementary argument a recursive formula called the *renewal equation*:

$$\text{(3.1)}\quad N_q^*(a, \gamma_0, \varphi) = \sum_{\substack{\gamma \\ \sigma\gamma = \gamma_0}} N_q^*(a - \tau_*(\gamma), \gamma, [\rho(\pi_q(\gamma\gamma_0^{-1}))\varphi]) + G(\gamma_0 o)\varphi \mathbf{1}\{a \geq 0\}.$$

We will now 'push to the boundary', replacing quantities with boundary counterparts under the following Dictionary.

Inside $\mathbb{H}$ (lattice point count)	The boundary $\partial(\mathbb{H})$
σ	T
τ_*	$\tau(x)$ (see Definition 7)
τ_*^N	$\tau^N(x) = \sum_{i=0}^{N-1} \tau(T^i x)$
G	$g = G\vert_I$
ρ	the cocycle c_q (see Definition 9)
$\rho(\pi_q(\gamma\gamma_0^{-1}))$	$c_q^N(x) := c_q(T^{N-1}x)c_q(T^{N-2}x)\dots c_q(Tx)c_q(x)$
$N_q^*(a, \gamma_0, \varphi)$	$N_q(a, x, \varphi) = \sum_{n=0}^{\infty} \sum_{y : T^n y = x} g(y)\rho(c_q^n(y))\varphi \mathbf{1}\{\tau^n(y) \leq a\}$

Table 1

These new quantities play a central role in the remainder of the paper, in place of their old counterparts. We take this opportunity to outline the rest of this section.

- We would like to understand the quantity $N_q^*(a, \gamma_0, \varphi)$. It is not clear how to do this directly, so we compare it to $N_q(a, \gamma_0 k_0, \varphi)$. Unfortunately that comparison is only valid when γ_0 is a "large" group element (see Lemmas 12 and 13), but we can arrange that by repeated application of the finite renewal equation (see (3.1)) so we obtain Lemma 14.
- Next we relate $N_q(a, \gamma k_0, \varphi)$ to the transfer operators. This is done by means of the boundary renewal equation (3.4) and a Laplace transform: we obtain (3.5).
- Spectral bounds for transfer operators (see Theorem 4) together with equation (3.5) and the Laplace inversion formula give us good control on $N_q(a, \gamma k_0, \varphi)$: see Proposition 17.
- Finally, we use control of $N_q(a, \gamma k_0, \varphi)$ to gain control of $N_q^*(a, \gamma_0, \varphi)$.

We will now put this outline into practice.

Assume that $\gamma_0 \neq 1$ (the case $\gamma_0 = 1$ follows from this consideration[4)]). In the Schottky setup, we say that $k_0 \in \tilde{I}_i$ is admissible for γ_0 if

$$\gamma_0 = g_{i_1} \dots g_{i_N}$$

is reduced and $i_N \neq i + k' \bmod 2k'$. We fix such an admissible $k_0 \in K$ now – if in the continued fractions setup this can be chosen arbitrarily in K.

Lemma 12. *There is $\kappa < 1$ such that when*

$$\gamma = g_{j_0} \dots g_{j_{n+N}}$$

is a reduced word in Γ, and $\gamma\gamma_0$ is also reduced/admissible then

$$\tau_*^n(\gamma\gamma_0) = \tau^n(\gamma\gamma_0 k_0) + O(\kappa^N).$$

Proof. Let C and κ be the constants from Lemma 8. Then

$$d_E(\gamma\gamma_0 o, \gamma\gamma_0 k_0) \leq C\kappa^{n+N}. \tag{3.2}$$

We also have

$$\begin{aligned}\tau_*(g_{j_0} g_{j_1} g_{j_2} \dots g_{j_{n-1}} \dots g_{j_{n+N}} \gamma_0) = -\log|g'_{j_0}(g_{j_1} g_{j_2} \dots g_{j_{n+N}} \gamma_0 o)| \\ + o(d_E(g_{j_1} g_{j_2} \dots g_{j_{n+N}} \gamma_0 o, \mathbf{R})).\end{aligned}$$

The derivative here is for the action of Γ on the unit disc model obtained via J; a similar estimate is given in [12, p. 41]. Note that the error term can be measured either in the unit disc model or the upper half plane model, as the two are bi-Lipschitz near K. It follows then that

$$\tau_*(g_{j_0} g_{j_1} g_{j_2} \dots g_{j_{n+N}} \gamma_0) = -\log|g'_{j_0}(g_{j_1} g_{j_2} \dots g_{j_{n+N}} \gamma_0 o)| + O(\kappa^{n+N-1}).$$

Since there is some uniform bound for the derivative of $\log|g'_i|$ close to the part of I where g_i is an inverse branch of T, this together with (3.2) implies

$$\tau_*(g_{j_0} g_{j_1} g_{j_2} \dots g_{j_{n+N}} \gamma_0) = -\log|g'_{j_0}(g_{j_1} g_{j_2} \dots g_{j_{n+N}} \gamma_0 k_0)| + O(\kappa^{n+N-1}).$$

By iterating for n steps and summing the geometric series it follows that

$$\tau_*^n(g_{j_0} g_{j_1} g_{j_2} \dots g_{j_{n+N}} \gamma_0) = -\log|(g_{j_0} g_{j_1} g_{j_2} \dots g_{j_{n-1}})'(g_{j_n} \dots g_{j_{n+N}} \gamma_0 k_0)| + O(\kappa^N)$$

or what is the same

$$\tau_*^n(\gamma\gamma_0) = \tau^n(\gamma\gamma_0 k_0) + O(\kappa^N),$$

proving the lemma. □

[4)] By applying the renewal equation (3.1), the quantity $N_q^*(a, 1, \varphi)$ is converted to a constant plus a finite sum of quantities of the form $N_q^*(\cdot, \gamma_0, \cdot)$, where $\gamma_0 \neq 1$.

Lemma 13. *Suppose φ is non-negative. There are N_0, $\kappa < 1$ and C depending on G such that if $N > N_0$ and*

$$\gamma_1 = g_{r_1} \dots g_{r_N} \gamma_0$$

is an admissible concatenation (hence k_0 is admissible for γ_1), then

$$N_q(a - C\kappa^N, \gamma_1 k_0, \varphi) \le N_q^*(a, \gamma_1, \varphi) \le N_q(a + C\kappa^N, \gamma_1 k_0, \varphi).$$

The inequalities are understood between functions on $\mathbf{R}^{\Gamma_q}$.

Proof. We will use the fact that the map $\gamma\gamma_1 \mapsto \gamma\gamma_1 k_0$ on admissible concatenations intertwines the shift σ and the map T. One has

$$N_q^*(a, \gamma_1, \varphi) = \sum_{n=0}^{\infty} \sideset{}{^*}\sum_{\gamma = g_{i_1} \cdots g_{i_n}} G(\gamma\gamma_1 o)\rho(\pi_q(\gamma))\varphi \mathbf{1}\{\tau_*^n(\gamma\gamma_1) \le a\} \tag{3.3}$$

and

$$N_q(a, \gamma_1 k_0, \varphi) = \sum_{n=0}^{\infty} \sideset{}{^*}\sum_{\gamma = g_{i_1} \cdots g_{i_n}} g(\gamma\gamma_1 k_0)\rho(\pi_q(\gamma))\varphi \mathbf{1}\{\tau^n(\gamma\gamma_1 k_0) \le a\}.$$

These can now be compared term by term. If N is large enough, depending on G, then $G(\gamma\gamma_1 o) = g(\gamma\gamma_1 k_0)$ for all terms as all the $\gamma\gamma_1 o$ will lie in the neighborhood J_M, where G is locally constant. On the other hand, we have from Lemma 12 that

$$\mathbf{1}\{\tau_*^n(\gamma\gamma_1) \le a\} \le \mathbf{1}\{\tau^n(\gamma\gamma_1 k_0) \le a + C\kappa^N\}$$

for some C. Also, in the other direction,

$$\mathbf{1}\{\tau^n(\gamma\gamma_1 k_0) \le a - C\kappa^N\} \le \mathbf{1}\{\tau_*^n(\gamma\gamma_1) \le a\}.$$

Given that φ and hence $\rho(\pi_q(\gamma))\varphi$ are positive functions, inserting these inequalities into (3.3) gives the result after suitably choosing N_0. □

Following Lalley [12, p. 22], we iterate the finite renewal equation (3.1) to obtain

$$\begin{aligned} N_q^*(a, \gamma_0, \varphi) = &\sum_{\substack{\gamma \\ \sigma^n \gamma = \gamma_0}} N_q^*(a - \tau_*^n(\gamma), \gamma, \rho[\pi_q(\gamma\gamma_0^{-1})]\varphi) \\ &+ \sum_{m=1}^{n-1} \sum_{\substack{\gamma \\ \sigma^m \gamma = \gamma_0}} G(\gamma o)\rho[\pi_q(\gamma\gamma_0^{-1})]\varphi \mathbf{1}\{a - \tau_*^m(\gamma) \ge 0\} \\ &+ G(\gamma_0 o)\varphi \mathbf{1}\{a \ge 0\}. \end{aligned}$$

We want to increase n so we note that the second line is bounded by (recall k is the number of intervals)

$$\sum_{m=0}^{n-1} k^m \|G\|_\infty \|\varphi\| \ll \|G\|_\infty \|\varphi\| k^n.$$

We will eventually set

$$n = \lfloor ca \rfloor$$

for small enough c. This gives

$$N_q^*(a, \gamma_0, \varphi) = \sum_{\substack{\gamma \\ \sigma^n\gamma=\gamma_0}} N_q^*(a - \tau_*^n(\gamma), \gamma, \rho[\pi_q(\gamma\gamma_0^{-1})]\varphi) + O(\|G\|_\infty\|\varphi\|e^{(\log k)ca}).$$

We can now use Lemma 13 to get:

Lemma 14. *Up to an error of* $O(\|G\|_\infty\|\varphi\|e^{(\log k)ca})$, $N_q^*(a, \gamma_0, \varphi)$ *is sandwiched between*

$$\sum_{\substack{\gamma \\ \sigma^n\gamma=\gamma_0}} N_q(a - \tau_*^n(\gamma) - C\kappa^n, \gamma k_0, \rho[\pi_q(\gamma\gamma_0^{-1})]\varphi)$$

and

$$\sum_{\substack{\gamma \\ \sigma^n\gamma=\gamma_0}} N_q(a - \tau_*^n(\gamma) + C\kappa^n, \gamma k_0, \rho[\pi_q(\gamma\gamma_0^{-1})]\varphi).$$

This sandwiching allows us to convert questions about N_q^*, and hence our main theorem, to questions about N_q. We leave the relation for now since going any further in the comparison requires results from later in the paper. Hopefully by now we have motivated the study of N_q and the dynamical system of Section 2.1.

3.2. The renewal equation: Boundary version. The quantity N_q also satisfies a version of the renewal equation: we first describe a simple version without any congruence aspect. Let $g \in C^1(I)$ as before.

We define

$$N(a, x) = \sum_{n=0}^{\infty} \sum_{\substack{y \\ T^n y=x}} g(y)\mathbf{1}\{\tau^n(y) \le a\},$$

where $\mathbf{1}\{\tau^n(y) \le a\}$ is the characteristic function of $\{\tau^n(y) \le a\}$. Only finitely many of the n give a contribution to the sum, since τ is eventually positive. The *renewal equation* states

$$N(a, x) = \sum_{\substack{y \\ Ty=x}} N(a - \tau(y), y) + g(x)\mathbf{1}\{a \ge 0\}. \tag{3.4}$$

This is related to the transfer operator $\mathcal{L}_{-s\tau}$ by taking a Laplace transform in the a variable. If one defines

$$n(s, x) = \int_{-\infty}^{\infty} e^{-sa} N(a, x)\, da,$$

then (3.4) is transformed into

$$n(s, x) = [\mathcal{L}_s n(s, \cdot)](x) + \frac{g(x)}{s},$$

where $\mathcal{L}_s[f]$ is the transfer operator defined in Section 2.2. The former equation can be recast to

$$s.n(s, \cdot) = (1 - \mathcal{L}_s)^{-1} g.$$

We now adapt our formulae to take account of the congruence aspect. The congruence version of the renewal equation at level q concerns the quantity

$$N_q(a,x,\varphi) := \sum_{n=0}^{\infty} \sum_{\substack{y \\ T^n y = x}} g(y)\rho(c_q^n(y))\varphi \mathbf{1}\{\tau^n(y) \le a\} \in \mathbf{C}^{\Gamma_q}$$

from before. This congruence renewal equation reads

$$N_q(a,x,\varphi) = \sum_{\substack{y \\ Ty = x}} \rho(c_q(y)) N_q(a - \tau(y), y, \varphi) + g(x)\varphi \mathbf{1}\{0 \le a\}.$$

Consider the *congruence transfer operator* $\mathcal{L}_{s,q}$ on $\mathbf{C}^{\Gamma_q}$-valued functions defined as follows:

$$\mathcal{L}_{s,q}[F](x) := \sum_{Ty=x} e^{-s\tau(y)} c_q(y).F(y),$$

where c_q is the modular cocycle given in Definition 9. Then parallel arguments to before give for

$$n_q(s,x,\varphi) = \int_{-\infty}^{\infty} e^{-sa} N_q(a,x,\varphi)\, da$$

the formula

$$sn_q(s,x,\varphi) = [(1 - \mathcal{L}_{s,q})^{-1} g \otimes \varphi](x), \tag{3.5}$$

where $g \otimes \varphi$ is the vector-valued function taking $x \mapsto g(x)\varphi$.

3.3. Spectral theory of transfer operators. Recall that we work with the Banach space $C^1(I)$ with norm as in (2.3) and the similar Banach spaces $C^1(I;\mathbf{C}^{\Gamma_q})$ of $\mathbf{C}^{\Gamma_q}$-valued functions. In Theorem 4 we summarized the spectral properties of $\mathcal{L}_{s,q}$ that we prove in this paper, and that will be used to estimate equation (3.5). The proof of Theorem 4 is deferred to Sections 4 and 5. We now continue with our counting estimates using Theorem 4 as a given.

3.4. Continuing the count. Notice that N_q and hence n_q are linear in φ. We split into two cases as we can write

$$\varphi = \varphi_0 + \varphi',$$

where φ_0 is constant and φ' is orthogonal to constants. The analysis of $N_q(a,x,\varphi_0)$ boils down to that of $N(a,x)$, which is in principle understood without any of the results of this paper. We take up the analysis in the case that

$$\varphi' \in \mathbf{C}^{\Gamma_q} \ominus 1,$$

that is, orthogonal to constants. Assume this is the case from now on.

One obtains from (3.5) and Theorem 4 that for any $\eta > 0$,

$$|s| \|n_q(s,\cdot,\varphi')\|_{C^1} \le \begin{cases} Cq^C(1-\rho_0)^{-1}\|g \otimes \varphi\|_{C^1} & \text{if } |b| \le b_0, \\ C_\eta |b|^{1+\eta}(1-\rho_\eta)^{-1}\|g \otimes \varphi\|_{C^1} & \text{if } |b| > b_0, \end{cases} \tag{3.6}$$

with the same quantifiers and constants as in Theorem 4. Consolidating constants, for any $\eta > 0$ there is $C' = C'(\eta)$ such that

$$|s| \|n_q(s,\cdot,\varphi')\|_{C^1} \le C' \max(q^C, |b|^{1+\eta}) \|g \otimes \varphi\|_{C^1} \tag{3.7}$$

whenever $|a - \delta| < \epsilon$ for some sufficiently small ϵ.

We also note that given the bounds in Theorem 4, it follows that the correspondence

$$s \mapsto (1 - \mathcal{L}_{s,q})^{-1} g \otimes \varphi'$$

gives a holomorphic family of C^1 functions in the region $|a - \delta| < \epsilon$ for fixed g and φ', hence $n_q(s, x, \varphi')$ is holomorphic for s in this region. This is essential for the contour shifting argument to follow. Now we follow technical work of Bourgain, Gamburd and Sarnak [4, pp. 25–26] to extract information about $N_q(a, x, \varphi')$.

Following [4, equation (9.4)], let k be a smooth nonnegative function on $\mathbf{R}$ such that $\int k = 1$, $\text{support}(k) \subset [1, 1]$ and[5]

$$|\hat{k}(\xi)| \leq B \exp(-|\xi|^{1/2}) \quad \text{for some } B,$$

where

$$\hat{k}(\xi) := \int_{\mathbf{R}} e^{-\xi t} k(t)\, dt.$$

Then let for small $\lambda > 0$,

$$k_\lambda(t) = \lambda^{-1} k(t\lambda^{-1}),$$

this has the effect that

$$\hat{k_\lambda}(\xi) = \hat{k}(\lambda\xi), \quad |\hat{k_\lambda}(\xi)| \leq B \exp(-|\lambda\xi|^{1/2}). \tag{3.8}$$

Consider the smoothed quantity of interest

$$\int_{-\infty}^{\infty} k_\lambda(t) N_q(a + t, x, \varphi')\, dt = \frac{1}{2\pi i} \int_{s \in \delta + i\mathbf{R}} e^{as} n_q(s, x, \varphi') \hat{k_\lambda}(s)\, ds$$

by inverting the Laplace transform and interchanging the order of integration. From (3.7), n_q is well enough behaved that this is possible. For technical reasons let $\epsilon' = \min(\delta/2, \epsilon/2)$. We can shift the contour to $\Re(s) = \delta - \epsilon'$ to get that the above is the same as

$$\begin{aligned} &\frac{1}{2\pi i} \int_{s \in \delta - \epsilon' + i\mathbf{R}} e^{as} n_q(s, x, \varphi') \hat{k_\lambda}(s)\, ds \\ &\quad = \frac{1}{2\pi} e^{a(\delta - \epsilon')} \int_{\theta \in \mathbf{R}} e^{ai\theta} n_q(\delta - \epsilon' + i\theta, x, \varphi') \hat{k_\lambda}(\delta - \epsilon' + i\theta)\, d\theta, \end{aligned}$$

where $s = \delta - \epsilon' + i\theta$. Putting in the bound (3.6) for n_q together with (3.8) gives the new bound

$$\begin{aligned} &\frac{BC'}{2\pi} e^{a(\delta - \epsilon')} \|g \otimes \varphi'\|_{C^1} \biggl(q^C \int_{|\theta| \leq b_0} |\delta - \epsilon' + i\theta|^{-1} e^{-|\lambda(\delta - \epsilon' + i\theta)|^{1/2}}\, d\theta \\ &\qquad\qquad + \int_{|\theta| > b_0} |\delta - \epsilon' + i\theta|^{-1} |\theta|^{1+\eta} e^{-|\lambda(\delta - \epsilon' + i\theta)|^{1/2}}\, d\theta \biggr) \\ &\quad \leq \frac{BC'}{2\pi} e^{a(\delta - \epsilon')} \|g \otimes \varphi'\|_{C^1} \left(\frac{4q^C b_0}{\delta - \epsilon'} + C'' |\lambda|^{-1-\eta} \right) \end{aligned}$$

for some new absolute constants C', C''. Putting this together (choosing $\eta = 1$ is enough) gives the following.

[5] The assumption that $\hat{k}$ has stretched exponential decay is overly strong here: it would be sufficient for example that $\hat{k}$ be uniformly bounded and in L^1 of any vertical line in $\mathbf{C}$ with real part sufficiently close to δ.

Lemma 15. *There is $Q_0 > 0$ provided by Theorem* 4 *and positive constants ϵ', C, κ_1, κ_2 such that for q with $(Q_0, q) = 1$ and any $g \in C^1(I)$, $\varphi' \in \mathbf{C}^{\Gamma_q} \ominus 1$ we have*

$$\left\| \int_{-\lambda}^{\lambda} k_\lambda(t) N_q(a+t, x, \varphi')\, dt \right\| < e^{a(\delta-\epsilon')} \|g \otimes \varphi'\|_{C^1} \big(\kappa_1 q^C + \kappa_2 |\lambda|^{-2}\big),$$

where the norm on the left-hand side is the one in $\mathbf{C}^{\Gamma_q}$.

We now describe $N_q(a, x, \varphi_0)$ with φ_0 a constant function. In this case the counting reduces to the non-congruence setting. The following is a straightforward adaptation of [4, Proposition 10.2] to our setting. This effectivizes work of Lalley [12], using the work of Naud [14] as input to get a power saving error term. Let $\underline{1}$ be the constant function in $\mathbf{C}^{\Gamma_q}$ taking on the value 1.

Lemma 16. *There exists $\epsilon'' > 0$ such that for any q, $g \in C^1(I)$ we have*

$$\int_{-\lambda}^{\lambda} k_\lambda(t) N_q(a+t, x, \underline{1})\, dt = C(x,g) e^{\delta a} \underline{1} + O\big(\|g\|_{C^1} |\Gamma_q| \lambda^{-3} e^{(\delta-\epsilon'')a}\big),$$

where

$$C(x,g) = \left(\frac{\int g \, d\nu_{-\delta\tau}}{\delta \int \tau \, d\nu_0} \right) h_{-\delta\tau}(x)$$

is a C^1 function of x and the error is estimated in C^1 norm, and ν, h are the measures and functions we defined in Theorem 10.

We remark that the $|\Gamma_q| \|g\|_{C^1}$ in the error term above comes from $\|g \otimes \varphi_0\|_{C^1}$. We can now put these lemmas together to get

Proposition 17. *There is $Q_0 > 0$ provided by Theorem* 4 *such that when $(Q_0, q) = 1$, the following holds. There is $\epsilon > 0$ such that for any non-negative $\varphi \in \mathbf{R}^{\Gamma_q} \subset \mathbf{C}^{\Gamma_q}$,*

$$N_q(a, x, \varphi) = \frac{C(x,g) e^{\delta a} \langle \varphi, \underline{1} \rangle \underline{1}}{|\Gamma_q|} + O\big(e^{(\delta-\epsilon)a} q^C \|g\|_{C^1} \|\varphi\|\big),$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product.

Proof. Decompose φ as

$$\varphi = \frac{\langle \varphi, \underline{1} \rangle \underline{1}}{|\Gamma_q|} + \varphi'.$$

Then Lemmas 15 and 16 give that

$$\int_{-\lambda}^{\lambda} k_\lambda(t) N_q(a+t, x, \varphi)\, dt = \frac{C(x,g) e^{\delta a} \langle \varphi, \underline{1} \rangle \underline{1}}{|\Gamma_q|} + e^{a(\delta-\epsilon)} O\big(\|g\|_{C^1} \|\varphi\| (\kappa_1 q^C + \kappa_2 \lambda^{-2} + \lambda^{-3})\big)$$

by using that

$$\|g \otimes \varphi'\|_{C^1} \le \|\varphi'\| \|g\|_{C^1}$$

and replacing ϵ', ϵ'' with a new small enough ϵ. Now taking $\lambda = e^{-a\epsilon/6}$, we have that the error term is

$$e^{a(\delta-\epsilon/2)}O(q^C \|g\|_{C^1}\|\varphi\|).$$

Since φ is non-negative, $N_q(a,x,\varphi)$ is increasing in a and hence

$$N_q(a-\lambda,x,\varphi) \le \int_{-\lambda}^{\lambda} k_\lambda(t)N_q(a+t,x,\varphi)\,dt \le N_q(a+\lambda,x,\varphi)$$

which is enough to get the result given the exponentially shrinking λ, by replacing ϵ with some smaller value. □

With the precise asymptotics of Proposition 17 at hand, we return to estimating $N_q^*(a,\gamma_0,\varphi)$. Using Lemma 14 along with Proposition 17 gives

$$\begin{aligned} N_q^*(a,\gamma_0,\varphi) &= \big(1+O(\delta C\kappa^n)\big)\frac{e^{\delta a}}{|\Gamma_q|}\langle\varphi,\underline{1}\rangle\underline{1}\sum_{\substack{\gamma\\ \sigma^n\gamma=\gamma_0}} C(\gamma k_0,g)e^{-\delta\tau_*^n(\gamma)} \\ &\quad + O\Bigg(q^C\|g\|_{C^1}\|\varphi\|e^{(\delta-\epsilon)a}\sum_{\substack{\gamma\\ \sigma^n\gamma=\gamma_0}} e^{-(\delta-\epsilon)\tau_*^n(\gamma)}\Bigg) \\ &\quad + O(\|G\|_\infty\|\varphi\|e^{(\log k)ca}). \end{aligned}$$

Given that $n = \lfloor ca \rfloor$ for some small c yet to be chosen, the κ^n term will not be significant. We do however have to describe the terms

$$\sum_{\substack{\gamma\\ \sigma^n\gamma=\gamma_0}} C(\gamma k_0,g)e^{-\delta\tau_*^n(\gamma)}$$

and

$$\sum_{\substack{\gamma\\ \sigma^n\gamma=\gamma_0}} e^{-(\delta-\epsilon)\tau_*^n(\gamma)}.$$

The latter can be bounded using Lemma 12 with $N=0$ to give $\tau_*^n(\gamma) = \tau^n(\gamma k_0) + O(1)$ and hence

$$\sum_{\substack{\gamma\\ \sigma^n\gamma=\gamma_0}} e^{-(\delta-\epsilon)\tau_*^n(\gamma)} \ll \sum_{\substack{k\\ T^n k=\gamma_0 k_0}} e^{-(\delta-\epsilon)\tau^n(k)} = [\mathscr{L}^n_{-(\delta-\epsilon)}1](\gamma_0 k_0). \tag{3.9}$$

We know that $\mathscr{L}_{-(\delta-\epsilon)\tau}$ is bounded by $\exp(P(-(\delta-\epsilon)\tau))$ by the Ruelle–Perron–Frobenius Theorem. We now therefore require $n < \frac{a\epsilon}{2P(-(\delta-\epsilon)\tau)}$ so that

$$[\mathscr{L}^n_{-(\delta-\epsilon)}1](\gamma_0 k_0) \ll \exp(nP(-(\delta-\epsilon)\tau)) \ll \exp\bigg(\frac{a\epsilon}{2}\bigg).$$

To describe the main term

$$\frac{e^{\delta a}}{|\Gamma_q|}\langle\varphi,\underline{1}\rangle\underline{1}\sum_{\substack{\gamma\\ \sigma^n\gamma=\gamma_0}} C(\gamma k_0,g)e^{-\delta\tau_*^n(\gamma)}, \tag{3.10}$$

we require the following result of Lalley (cf. [12, Theorem 4]). It says that there is a version of the maximal eigenfunction $h_{-\delta\tau}$ on Γ, as opposed to K.

Lemma 18. *Fix $k_0 \in K$. There exist a unique positive function $h_* : \Gamma \to \mathbf{R}$ and $\theta > 1$ so that if $\gamma \in \Gamma^{(n)}$,*

$$h_*(\gamma) = h_{-\delta\tau}(\gamma k_0) + O(\theta^{-n}).$$

Also, for all $\gamma \in \Gamma$,

$$h_*(\gamma) = \sum_{\substack{\gamma' \\ \sigma(\gamma')=\gamma}} e^{-\delta\tau_*(\gamma')} h_*(\gamma'). \tag{3.11}$$

Now recall the definition of $C(\cdot, g)$ from Lemma 16. If we define the corresponding function on Γ according to the pairing of h_* with $h_{-\delta\tau}$,

$$C_*(\gamma, g) = \left(\frac{\int g \, d\nu_{-\delta\tau}}{\delta \int \tau \, d\nu_0}\right) h_*(\gamma), \tag{3.12}$$

we get from Lemma 18 that

$$C_*(\gamma, g) = C(\gamma k_0, g) + O(\|g\|_{C^1} \theta^{-n})$$

when $\gamma \in \Gamma^{(n)}$. This means that the main term contribution (3.10) to $N_q^*(a, \gamma_0, \varphi)$ is

$$\begin{aligned} &\frac{e^{\delta a}}{|\Gamma_q|} \langle \varphi, \underline{1} \rangle \underline{1} \Bigg(\sum_{\substack{\gamma \\ \sigma^n \gamma = \gamma_0}} C_*(\gamma, g) e^{-\delta \tau_*^n(\gamma)} + O\Big(\|g\|_{C^1} \theta^{-n} \sum_{\substack{\gamma \\ \sigma^n \gamma = \gamma_0}} e^{-\delta\tau_*^n(\gamma)}\Big)\Bigg) \\ &\quad = \frac{e^{\delta a}}{|\Gamma_q|} C_*(\gamma_0, g) \langle \varphi, \underline{1} \rangle \underline{1} + e^{\delta a} O(\theta^{-n} \|\varphi\| \|g\|_{C^1}) \end{aligned}$$

by using (3.11) and a calculation similar to that in (3.9) to give

$$\sum_{\substack{\gamma \\ \sigma^n \gamma = \gamma_0}} e^{-\delta\tau_*^n(\gamma)} \ll [\mathscr{L}_{-\delta}^n 1](\gamma_0 k_0) \ll 1.$$

We now let $n = \lfloor ca \rfloor$ with

$$c = \min\left(\frac{\delta - \epsilon}{4 \log k}, \frac{\epsilon}{2P(-(\delta - \epsilon)\tau)}\right).$$

Then the result of the preceding discussion is that

$$N_q^*(a, \gamma_0, \varphi) = \frac{e^{\delta a}}{|\Gamma_q|} C_*(\gamma_0, g) \langle \varphi, \underline{1} \rangle \underline{1} + O\big((\|\varphi\| (\|g\|_{C^1} + \|G\|_\infty) q^C e^{(\delta - \epsilon')a}\big)$$

for some $\epsilon' = \epsilon'(\kappa, \theta, \epsilon, \mathscr{A})$. When $\varphi(\gamma) = \mathbf{1}\{\gamma = \xi\}$ we have that

$$\langle \varphi, \underline{1} \rangle = 1$$

and hence evaluating $N_q^*(a, \gamma_0, \mathbf{1}\{\gamma = \xi\})$ gives

$$\sum_{\substack{\gamma \in \Gamma \\ d(o, \gamma\gamma_0 o) - d(o, \gamma_0 o) \le a \\ \pi_q(\gamma) = \xi}}^{*} G(\gamma\gamma_0 o) = \frac{e^{\delta a}}{|\Gamma_q|} C_*(\gamma_0, g) + O\big((\|g\|_{C^1} + \|G\|_\infty) q^C e^{(\delta - \epsilon')a}\big).$$

This proves our main Theorem 11, given Theorem 4.

4. Bounds for transfer operators: Large imaginary part

In this section we will prove part (2) of Theorem 4.

4.1. Non-local integrability. Recall from Section 2 the set I, K, the map $T : I \to \mathbf{R}$, the cocycles c_q and Γ. We need to introduce symbolic dynamics. We write A for the $k \times k$ matrix with (i, j) entry equal to 1 if $T(I_i) \supset I_j$ and 0 otherwise. Such a matrix A is called the *transition matrix*. We say that a sequence (i_j) with entries in $1, \dots, k$ is admissible if $T(i_j) \supset i_{j+1}$ for all j in the index set of the sequence. When $T(I_i) \supset I_j$, we define T_i^{-1} on I_j to be the unique locally defined branch of T^{-1} that maps I_j to I_i.

Let Σ_A^+ (resp. Σ_A^-) be the space of positively (resp. negatively) indexed admissible sequences on $\{1, \dots, k\}$. We define for $\xi \in \Sigma_A^-$ the function

$$\Delta_\xi(u, v) = \sum_{i=0}^{\infty} \tau(T_{\xi_{-i}}^{-1} \circ \cdots \circ T_{\xi_0}^{-1} u) - \tau(T_{\xi_{-i}}^{-1} \circ \cdots \circ T_{\xi_0}^{-1} v)$$

on $I_j \times I_j$ such that $T(I_{\xi_0}) \supset I_j$. It follows from the expanding property of T that Δ_ξ is C^1 where it is defined. Naud (following others) defines a temporal distance function

$$\varphi_{\xi,\eta}(u, v) = \Delta_\xi(u, v) - \Delta_\eta(u, v)$$

which is defined for each $\xi, \eta \in \Sigma_A^-$ and $u, v \in I_j$

Definition 19 (Non-local integrability (NLI)). An eventually positive function τ is said to have property (NLI) if there are $j_0 \in \{1, \dots, k\}$, $\xi, \eta \in \Sigma_A^-$ with $T(I_{\xi_0}) \cap T(I_{\eta_0}) \supset I_{j_0}$ and $u_0, v_0 \in K \cap I_{j_0}$ such that

$$\frac{\partial \varphi_{\xi,\eta}}{\partial u}(u_0, v_0) \neq 0.$$

Proposition 20. *The distortion functions τ and $\hat{\tau}$ have the non-local integrability property.*

Proof. In the two cases of Schottky semigroups and the continued fractions semigroups we are considering, we always have two hyperbolic elements $h_i := g_i^{-1}$, $h_j := g_j^{-1}$ (with g_i, g_j from the generating set) satisfying (1) $T|_{I_i} = h_i$ and $T|_{I_j} = h_j$, (2) the h_i and h_j have distinct repelling (resp. attractive) fixed points on $\mathbf{R} \cup \{\infty\}$ and (3) the semigroup generated by h_i and h_j consists of hyperbolic elements. Given such elements, Naud's argument in [14, Proof of Lemma 4.4] shows the non-local integrability properties of $\hat{\tau}(x) = \log |T'(x)|$ and $\tau(x)$. □

4.2. Beginning Dolgopyat's argument. One novelty of this paper is the following version of [14, Theorem 2.3] that is uniform in the congruence aspect.

Proposition 21. *There is $b_0 > 0$ such that part* (2) *of Theorem* 4 *holds. That is, for any $\eta > 0$, there is $0 < \rho_\eta < 1$ such that*

$$\|\mathcal{L}_{s,q}^m\|_{C^1} \ll_\eta |b|^{1+\eta} \rho_\eta^m$$

when $|b| > b_0$ and $q \in \mathbf{N}$, as in Theorem 4.

We now show how to relate this proposition to the construction of certain Dolgopyat operators. Recall the Ruelle–Perron–Frobenius Theorem (Theorem 10) and its notation. Let h_a be the normalized positive eigenfunction of $\mathcal{L}_{-a\tau}$ corresponding to the maximal eigenvalue $\exp(P(-a\tau))$. We set

$$\tau_a = -a\tau - P(-a\tau) - \log(h_a \circ T) + \log(h_a).$$

We now renormalize our transfer operators by defining

$$L_{s,q} := \mathcal{L}_{\tau_a - ib\tau,q}.$$

This is the same as

$$L_{s,q} = \exp(-P(-a\tau))M_{h_a}^{-1}\mathcal{L}_{s,q}M_{h_a}, \tag{4.1}$$

where M_{h_a} is multiplication by h_a. It now follows by arguments as in Naud [14, p. 132] that it is enough to prove Proposition 21 and Theorem 4 with $L_{s,q}$ in place of $\mathcal{L}_{s,q}$. We also note here that the maximal eigenfunction of L_a is the constant function, with eigenvalue 1, that is, $L_a 1 = 1$ for $a \in \mathbf{R}$.

The rest of the passage to the estimates in the next section is routine but we give some of the details for completeness. One shows that in order to prove Proposition 21 it is enough to prove the following lemma.

Lemma 22. *With the same conditions as Theorem* 4, *there are $N > 0$ and $\rho \in (0,1)$ such that when $|a-\delta|$ is sufficiently small and $|b|$ is sufficiently large we have*

$$\int_K |L_{s,q}^{nN} W|^2 \, d\nu_0 \leq \rho^n,$$

where $W \in C^1(I; \mathbf{C}^{\Gamma_q})$, $d\nu_0 = h_{-\delta\tau}\nu_{-\delta\tau}$ is the Gibbs measure on K, and $\|W\|_{(b)} \leq 1$, which stands for the warped Sobolev norm

$$\|W\|_{(b)} = \|W\|_\infty + |b|^{-1}\|W'\|_\infty.$$

These estimates are uniform in q.

This corresponds to [15, Theorem 3.1] in the work of Oh and Winter and is the uniform version of [14, Proposition 5.3].

Lemma 22 implies Proposition 21 by the use of a priori estimates for the transfer operators that allow one to convert an L^2 estimate into a C^1 bound. These estimates are given in [14, Lemma 5.2] for complex-valued functions. They are however easily proved for vector-valued functions giving

Lemma 23. *There are $\kappa_1, \kappa_2, a_0, b_0 > 0$ and $R < 1$ such that for $|a-\delta| < a_0$ and $|b| > b_0$ we have for all $f \in C^1(I; \mathbf{C}^{\Gamma_q})$,*

$$\|[L_{s,q}^n f]'\|_\infty \leq \kappa_1 |b| \|L_a^n f\|_\infty + R^n \|L_a^n |f'|\|_\infty$$

and

$$\|L_{\delta,q}^n f\|_\infty \leq \int_K |f| \, d\nu_0 + \kappa_2 R^n \|f\|_{L(K)}.$$

Lemma 23 together with Lemma 22 imply Proposition 21 by arguments appearing in [14, pp. 133–134]. Roughly speaking the ingredients are Cauchy–Schwarz to access Lemma 22, remarks regarding the behavior of τ_a^m for a close to δ that appear elsewhere in this paper, and splitting up exponents in the form $m = nN + r$.

The proof of Lemma 22 proceeds through the construction of certain Dolgopyat operators that we give in the next subsection.

4.3. Construction of uniform Dolgopyat operators.

We follow the notation of Naud (cf. [14]). For $A > 0$ we consider the cone

$$\mathcal{C}_A := \{H \in C^1(I) : H > 0 \text{ and } |H'(x)| \leq AH(x) \text{ for all } x \in I\}.$$

In this subsection we establish a uniform version of the key lemma of Naud [14, Lemma 5.4]. This is also analogous to [15, Theorem 3.3].

Lemma 24 (Construction of uniform Dolgopyat operators). *Suppose τ has the* (NLI) *property. There exist $N > 0$, $A > 1$ and $\rho \in (0, 1)$ such that for all $s = a + ib$ with $|a - \delta|$ small and $|b| > b_0$ large, there exists a finite set of operators $(\mathcal{N}_s^J)_{J \in \mathcal{E}_s}$ that are bounded on $C^1(I)$ and satisfy the following three conditions:*

(1) *The cone $\mathcal{C}_{A|b|}$ is stable by $\mathcal{N}_s^J$ for all $J \in \mathcal{E}_s$.*

(2) *For all $H \in \mathcal{C}_{A|b|}$ and all $J \in \mathcal{E}_s$,*

$$\int_K |\mathcal{N}_s^J H|^2 \, d\nu_0 \leq \rho \int_K |H|^2 \, d\nu_0.$$

(3) *Given $H \in \mathcal{C}_{A|b|}$ and $f \in C^1(I; \mathbf{C}^{\Gamma_q})$ such that $|f| \leq H$ and $|f'| \leq A|b|H$, there is $J \in \mathcal{E}_s$ with*

$$|L_{s,q}^N f| \leq \mathcal{N}_s^J H \quad \text{and} \quad |(L_{s,q}^N f)'| \leq A|b|\mathcal{N}_s^J H.$$

When we write $|f|$ for $f \in C^1(I; \mathbf{C}^{\Gamma_q})$, we refer to the function obtained by taking pointwise Euclidean (l^2) norms. We now show that the existence of these operators implies Lemma 22.

Proof that Lemma 24 *implies Lemma* 22. Given this construction (that is, Lemma 24), Lemma 22 is proved following the argument of [15, p. 21] or one in [14, p. 135]. Indeed, given non-zero $f \in C^1(I; \mathbf{C}^{\Gamma_q})$ with $\|f\|_{(b)} \leq 1$ (cf. Lemma 22 for the definition of $\|\|_{(b)}$), we define

$$H = \|f\|_{(b)} 1.$$

One sees that H and f are as in Lemma 24, that is, $H \in \mathcal{C}_{A|b|}$, $|f| \leq H$, and $|f'| \leq A|b|H$ as $A > 1$. One gets then by part (3) of Lemma 24 that

$$|L_{s,q}^N f| \leq \mathcal{N}_s^J H, \quad |(L_{s,q}^N f)'| \leq A|b|\mathcal{N}_s^J H$$

for some $J \in \mathcal{E}_s$. Since $\mathcal{C}_{A|b|}$ is stable under the N_s^J, one can repeat this to get for some sequence $J_1, \ldots, J_n \in \mathcal{E}_s$ that

$$\int_K |L_s^{nN} f|^2 \, d\nu_0 \leq \int_K |\mathcal{N}_s^{J_N} \ldots \mathcal{N}_s^{J_1} H|^2 \, d\nu_0 \leq \rho^n \int_K |H|^2 \, d\nu_0 \leq \rho^n$$

by using part (2) of Lemma 24. □

The first two properties of Lemma 24 were proved by Naud in [14]; we follow closely Naud's construction of the operators in the following.

4.4. Consequences of non-local integrability (NLI). Naud notes the following consequence of (NLI) that we will use later.

Lemma 25 (Proposition 5.5 of [14]). *If τ has property* (NLI)*, there are $m, m', N_0 > 0$ such that for all $N > N_0$, there are two branches α_1^N, α_2^N of T^{-N} with*

$$m' \geq \left|\frac{d}{du}[\tau^N \circ \alpha_1^N - \tau^N \circ \alpha_2^N](u)\right| \geq m > 0 \quad \textit{for all } u \in I.$$

We remark here that the lower bound is the harder one. The upper bound follows from the expanding property of T and regularity of τ.

Now suppose we deal with τ with property (NLI). Let ξ, η, u_0, v_0 and j_0 be as in Definition 19.

Throughout the rest of this paper, the assignments $N \to \alpha_1^N$ and $N \to \alpha_2^N$ are fixed as those given by Lemma 25*.*

We do however need to know some of the details about how the α_i^N have been constructed, which we give now.

As in the proof of [14, Proposition 5.5] there are $\epsilon > 0$ and an open interval $\mathcal{U}$ with

$$I_{j_0} \supset \mathcal{U} \ni u_0$$

such that

$$\left|\frac{\partial \varphi_{\xi,\eta}}{\partial u}(u', v_0)\right| > \epsilon$$

for all $u' \in \mathcal{U}$. We define for any n,

$$\beta_1^n = T_{\xi_{-n+1}}^{-1} \circ \cdots \circ T_{\xi_0}^{-1} \quad \text{and} \quad \beta_2^n = T_{\eta_{-n+1}}^{-1} \circ \cdots \circ T_{\eta_0}^{-1},$$

two branches of T^{-n} on I_{j_0}. In the proof of [14, Proposition 5.5], Naud also constructs

$$\psi : I \to \mathcal{U}$$

which is a branch of $T^{-\hat{p}}$ for some $\hat{p}$ a fixed positive integer related to the mixing and expanding properties of T. The image of ψ is a disjoint union of k closed intervals each of which is diffeomorphic to some I_j by ψ. We denote by U_0 the image of ψ. We will use the parameterization

$$N = \tilde{N} + \hat{p}.$$

Then the α_i^N are defined by

$$\alpha_i^N = \beta_i^{\tilde{N}} \circ \psi.$$

As $\tilde{p}$ is fixed, $\tilde{N}$ and N are coupled. They are to be chosen, depending on b and other demands in the following.

4.5. Construction of Dolgopyat operators. The following result is proved by Naud [14, Proposition 5.6].

Proposition 26 (Triadic partition). *There are $A_1, A_1' > 0$ and $A_2 > 0$ such that when $\epsilon > 0$ is small enough, there is a finite collection $(V_i)_{1 \le i \le Q}$ of closed intervals ordered along U_0 such that:*

(1) $\mathcal{U} \supset \bigcup_{i=1}^{Q} V_i \supset U_0$, $V_i \cap \operatorname{Int} U_0 \neq \emptyset$ *for all i and* $\operatorname{Int} V_i \cap \operatorname{Int} V_j = \emptyset$ *when $i \neq j$.*

(2) *For all $1 \le i \le Q$, $\epsilon A_1' \le |V_i| \le \epsilon A_1$.*

(3) *For all $1 \le j \le Q$ with $V_j \cap K \neq \emptyset$, one has either $V_{j-1} \cap K \neq \emptyset$ and $V_{j+1} \cap K \neq \emptyset$ or $V_{j-2} \cap K \neq \emptyset$ and $V_{j-1} \cap K \neq \emptyset$ or $V_{j+1} \cap K \neq \emptyset$ and $V_{j+2} \cap K \neq \emptyset$. In other words, intervals that intersect K come at least in triads.*

(4) *For all $1 \le i \le Q$ with $V_i \cap K \neq \emptyset$, $V_i \cap K \subset U_0$ and* $\operatorname{dist}(\partial V_i, K) \ge A_2 |V_i|$.

Now following Naud, we can construct the Dolgopyat operators. Suppose that we are working at frequency $s = a + ib$. Then for fixed ϵ' to be chosen, we construct a triadic partition $(V_i)_{i=1}^{Q}$ of U_0 with $\epsilon = \epsilon'/|b|$ as in Proposition 26. Then for all $i \in \{1, 2\}$ and $j \in \{1, \dots, Q\}$ we set

$$Z_j^i = \beta_i^{\tilde{N}}(V_j \cap U_0).$$

We will write

$$X_j = \{x \in I : \psi(x) \in V_j\}, \quad 1 \le j \le Q.$$

Properties (4) and (2) of Proposition 26 imply that

$$\operatorname{dist}(K \cap V_j, \partial V_j) \ge A_2 |V_j| \ge \frac{A_2 A_1' \epsilon'}{|b|} \tag{4.2}$$

whenever $K \cap V_j \neq 0$. For such j we can find a C^1 cutoff χ_j on I that is $\equiv 1$ on the convex hull of $K \cap V_j$ and $\equiv 0$ outside V_j. Due to (4.2) we can ensure that

$$|\chi_j'| \le A_3 \frac{|b|}{\epsilon'}, \quad A_3 = A_3(A_2, A_1').$$

Then the index set $\mathcal{I}_s$ is defined to be

$$\mathcal{I}_s := \{(i, j) : 1 \le i \le 2,\ 1 \le j \le Q,\ V_j \cap K \neq \emptyset\}.$$

Allow $0 < \theta < 1$ to be fixed shortly. For all $J \subset \mathcal{I}_s$ we define $\chi_J \in C^1(I)$ by

$$\chi_J(x) = \begin{cases} 1 - \theta \chi_j(\psi(T^N x)), & \text{if } x \in Z_i^j \text{ for } (i, j) \in J, \\ 1, & \text{else.} \end{cases}$$

Then the Dolgopyat operators on $C^1(I)$ are defined by

$$\mathcal{N}_s^J(f) = L_a^N(\chi_J f).$$

Recall that L_a is the transfer operator at $s = a$.

Let us return to our Lemma 24 so that we can complete our definitions.

Definition 27. We say that $J \subset I_s$ is dense if for all $1 \le j \le Q$ with $V_j \cap K \neq \emptyset$ there is some $1 \le j' \le Q$ with $(i, j') \in J$ for some $i \in \{1, 2\}$ and with $|j - j'| \le 2$.

We define $\mathcal{E}_s$ of Lemma 24 to be the set of $J \subset \mathcal{J}_s$ such that J is dense.

The following is proved in [14] – we have tried to contain everything that we use as a black box here.

Proposition 28 (Naud). *There are constants a_0, b_0, A, N_0 such that for each sufficiently small ϵ' there is $\theta_0(\epsilon')$ and $\rho(\epsilon')$ such that when $N > N_0$, $\theta < \theta_0(\epsilon')$, $|a - \delta| < a_0$ and $|b| > b_0$, properties* (1) *and* (2) *of Lemma* 24 *hold for our $(N, |b|, \theta, \epsilon')$ parameterized and $\mathcal{E}_s$-indexed Dolgopyat operators with respect to this ρ.*

Furthermore, there is positive C_0 such that when $|a - \delta| < a_0$ we have for arbitrary N,

$$|(\tau_a^N \circ \alpha^N)'(x)| \leq C_0, \tag{4.3}$$

and when $N > N_0$, $b > b_0$, we have

$$|([\tau_a^N + ib\tau^N] \circ \alpha^N)'(x)| \leq \frac{A|b|}{4}. \tag{4.4}$$

This was a factor in how A was chosen.

The proof of the inequalities above are discussed in [14, p. 137].

This fully completes the definition of the Dolgopyat operators modulo choice of ϵ', θ and N – the A and ρ required for Lemma 24 *are that specified by Proposition* 28 *given these parameters.*

4.6. Proof of Lemma 24, property (3). Our remaining task in this subsection is to prove property (3) of Lemma 24. This is proved for complex-valued functions by Naud in [14, pp. 140–144]. Naud makes some use of taking quotients of values of functions that we will have to work around.

We give the details now. Recall that ϵ', θ are still undetermined. The following technical lemma is the vector-valued version of [14, Lemma 5.10]. Recall that $c_q : I \to U(\mathbf{C}^{\Gamma_q})$ is our twisting unitary-valued map at level q. We will need to consider the quantity $c_q^N(\alpha_i^N x)$, defined as in the Dictionary of Table 1, where α_i^N, $i = 1, 2$, are the two particular branches of T^{-N} that are given by Lemma 25. We record the key fact here that since c_q is locally constant, so too is c_q^N for any N.

Lemma 29 (Key technical fact towards non-stationary phase). *Let $H \in \mathcal{C}_{A|b|}$ and let $f \in C^1(I; \mathbf{C}^{\Gamma_q})$ such that $|f| \leq H$ and $|f'| \leq A|b|H$. For $i = 1, 2$, define for θ a small real parameter and for any q,*

$$\Theta_1(x) := \frac{|e^{[\tau_a^N + ib\tau^N](\alpha_1^N x)} c_q^N(\alpha_1^N x) f(\alpha_1^N x) + e^{[\tau_a^N + ib\tau^N](\alpha_2^N x)} c_q^N(\alpha_2^N x) f(\alpha_2^N x)|}{(1 - 2\theta) e^{\tau_a^N(\alpha_1^N x)} H(\alpha_1^N x) + e^{\tau_a^N(\alpha_2^N x)} H(\alpha_2^N x)},$$

$$\Theta_2(x) := \frac{|e^{[\tau_a^N + ib\tau^N](\alpha_1^N x)} c_q^N(\alpha_1^N x) f(\alpha_1^N x) + e^{[\tau_a^N + ib\tau^N](\alpha_2^N x)} c_q^N(\alpha_2^N x) f(\alpha_2^N x)|}{e^{\tau_a^N(\alpha_1^N x)} H(\alpha_1^N x) + (1 - 2\theta) e^{\tau_a^N(\alpha_2^N x)} H(\alpha_2^N x)}.$$

Then for N large enough, one can choose θ, ϵ' small enough such that for j with $X_j \cap K \neq \emptyset$, there are j' with $|j - j'| \leq 2$, $X_{j'} \cap K \neq \emptyset$ and $i \in \{1, 2\}$ such that

$$\Theta_i(x) \leq 1 \quad \textit{for all } x \in X_{j'}.$$

Before giving the proof we must state a simple lemma from [14]. The proof goes through easily in our vector-valued setting. This is also covered in [15, Lemma 3.29].

Lemma 30 ([14, Lemma 5.11]). *Let $Z \subset I$ be an interval with $|Z| \leq \frac{c}{|b|}$. Let $H \in \mathcal{C}_{A|b|}$ and $f \in C^1(I; \mathbf{C}^{\Gamma_q})$ with $|f| \leq H$ and $|f'| \leq A|b|H$. Then for c small enough, we have either*

$$|f(u)| \leq \frac{3}{4}H(u) \quad \textit{for all } u \in Z$$

or

$$|f(u)| \geq \frac{1}{4}H(u) \quad \textit{for all } u \in Z.$$

We also need the following piece of trigonometry from [14, Lemma 5.12].

Lemma 31 (Sharp triangle inequality). *Let V be a finite-dimensional complex vector space with Hermitian inner product $\langle \cdot, \cdot \rangle$. For non-zero vectors z_1, z_2 with $|z_1|/|z_2| \leq L$ and*

$$\Re\langle z_1, z_2 \rangle \leq (1-\eta)|z_1||z_2|, \tag{4.5}$$

there is $\delta = \delta(L, \eta)$ such that

$$|z_1 + z_2| \leq (1-\delta)|z_1| + |z_2|.$$

We remark that while Lemma 31 is elementary, the fact that there is no dependence on the dimension of V is one of the crucial points in our arguments.

Proof of Lemma 29. Choose ϵ' small enough so that Lemma 30 holds for all $Z = Z_j^i$ (with $c = \epsilon'$). As in [14] by choosing N large enough it is possible to assume $|Z_j^i| \leq |V_j|$ for all j, i. We also enforce $\theta < 1/8$ so that $1 - 2\theta \geq 3/4$.

Now let V_j, V_{j+1}, V_{j+2} all have non-empty intersection with K. One of the $j, j+1, j+2$ will be the j' of the lemma. Set $\widehat{X}_j = X_j \cup X_{j+1} \cup X_{j+2}$ and assume as in Naud that $\widehat{X}_j$ is contained in one connected component of I; note that the set $\widehat{X}_j$ is connected.

Following from our choice of θ, if there are $j' \in \{j, j+1, j+2\}$ and $i \in \{1, 2\}$ with $|f(u)| \leq \frac{3}{4}H(u)$ when $u \in Z_{j'}^i$ then $\Theta_i(u) \leq 1$ on $Z_{j'}^i$ and we are done. So we can assume that $|f(u)| > \frac{3}{4}H(u)$ for some u in each $Z_{j'}^i$. Hence by Lemma 30, for all i, j' we have

$$|f(u)| \geq \frac{1}{4}H(u) > 0 \quad \text{for all } u \in Z_{j'}^i. \tag{4.6}$$

We make the definition

$$z_i(x) := \exp\big([\tau_a^N + ib\tau^N](\alpha_i^N x)\big) c_q^N(\alpha_i^N x) f(\alpha_i^N x), \quad z_i : \widehat{X}_j \to \mathbf{C}^{\Gamma_q}, \quad i = 1, 2.$$

The result follows from Lemma 31 after establishing bounds on the relative size and angle of z_1, z_2 uniformly in appropriate $X_{j'}$.

Control of relative size. Firstly we wish to control the relative size of z_1, z_2. This is done by Naud and his estimates go through directly in our case, after making all substitutions of the form

$$\left|\frac{z_1(x)}{z_2(x)}\right| \to \frac{|z_1(x)|}{|z_2(x)|}$$

and bearing in mind that c_q^N is a unitary-valued function. This caters to our inability to divide non-zero vectors. The output of Naud's argument in [14, pp. 141–142] is that given an index $j' \in \{j, j+1, j+2\}$, either $|z_1(x)| \le M|z_2(x)|$ for all $x \in X_{j'}$ or $|z_2(x)| \le M|z_1(x)|$ for all $x \in X_{j'}$, where

$$M = 4\exp(2NB_a)\exp(2A\epsilon' A_1)$$

and

$$B_a = a\|\tau\|_\infty + |P(-a\tau)| + 2\|\log h_a\|_\infty$$

is a locally bounded function that arises in the estimation of τ_a^N (cf. [14, p. 139]). Returning to the overall argument, this means that we are done when we can establish (4.5) with some η uniformly on some $X_{j'}$.

Control of relative angle. The key argument here is to very carefully control the angles between the functions z_1 and z_2. One sets

$$\Phi(x) := \frac{\langle z_1(x), z_2(x)\rangle}{|z_1(x)||z_2(x)|},$$

which is the same as

$$\Phi(x) = \exp(ib(\tau^N(\alpha_1^N x) - \tau^N(\alpha_2^N x)))\frac{\langle c_q^N(\alpha_1^N x) f(\alpha_1^N x), c_q^N(\alpha_2^N x) f(\alpha_2^N x)\rangle}{|f(\alpha_1^N x)||f(\alpha_2^N x)|}.$$

Define

$$u_i(x) = c_q^N(\alpha_i^N x)\frac{f(\alpha_i^N x)}{|f(\alpha_i^N x)|}, \quad x \in \widehat{X}_j, \quad i = 1, 2.$$

Then the u_i are C^1 as f is non-vanishing through (4.6). We have

$$(c_q^N . f) \circ \alpha_i^N = |f \circ \alpha_i^N| . u_i,$$

so that, differentiating on both sides and using $(c_q^N)' \equiv 0$,

$$(c_q^N \circ \alpha_i^N).(f \circ \alpha_i^N)' = |f \circ \alpha_i^N|' u_i + |f \circ \alpha_i^N| u_i'.$$

As u_i has constant length 1 it follows that u_i and u_i' are orthogonal (in $\mathbf{R}^{2|\Gamma_p|}$). Therefore

$$|[f \circ \alpha_i^N]'|^2 = (|f \circ \alpha_i^N|')^2 + |f \circ \alpha_i^N|^2 |u_i'|^2.$$

It now follows that

$$|u_i'(x)| \le \frac{|[f \circ \alpha_i^N]'(x)|}{|f(\alpha_i^N x)|}.$$

We estimate the right-hand side by a direct calculation using the chain rule with the expanding property of T and our assumptions on H from (4.6) and the hypotheses of Lemma 29. Indeed, Naud performs a similar calculation [14, p. 142] which yields

$$|u_i'(x)| \le 8A|b|\frac{D}{\gamma^N}. \tag{4.7}$$

Note that we can rewrite the central quantity Φ as

$$\Phi(x) = \exp(ib(\tau^N(\alpha_1^N x) - \tau^N(\alpha_2^N x)))\langle u_1(x), u_2(x)\rangle.$$

We can use (4.7) and Cauchy–Schwarz to get

$$\left|\frac{d}{dx}\langle u_1, u_2\rangle\right| = |\langle u_1', u_2\rangle + \langle u_1, u_2'\rangle| \le 16A|b|\frac{D}{\gamma^N}. \tag{4.8}$$

Note that we have the diameter bound

$$\operatorname{diam}(\widehat{X}_j) \le 3A_1\frac{\epsilon'}{|b|}\|(\psi^{-1})'\|_\infty$$

so that using (4.8) we have

$$|\langle u_1(x_1), u_2(x_1)\rangle - \langle u_1(x_2), u_2(x_2)\rangle| \le 3\cdot 16\cdot AA_1\|(\psi^{-1})'\|_\infty\epsilon'\frac{D}{\gamma^N}$$

for any $x_1, x_2 \in \widehat{X}_j$; note here that the cocycles $c_q^N(\alpha_i^N x)$ are constant on $\hat{X}_j$. We now enforce $\epsilon' < \frac{1}{10}$ and N large enough so that

$$48\cdot AA_1\|(\psi^{-1})'\|_\infty\frac{D}{\gamma^N} < 1.$$

Let us cut off one branch of reasoning. Suppose that there is $x_0 \in \widehat{X}_j$ with

$$|\langle u_1(x_0), u_2(x_0)\rangle| < \frac{1}{10}.$$

Then for all $x \in \widehat{X}_j$ we have

$$|\langle u_1(x), u_2(x)\rangle| < \frac{1}{5}.$$

It would follow that $|\Re\Phi(x)| < \frac{1}{5}$ for all $x \in \widehat{X}_j$ and the Lemma would be proved by our argument with trigonometry.

Therefore we can now assume

$$|\langle u_1(x), u_2(x)\rangle| \ge \frac{1}{10}$$

for all $x \in \widehat{X}_j$. Then the new function

$$U(x) = \frac{\langle u_1(x), u_2(x)\rangle}{|\langle u_1(x), u_2(x)\rangle|} \in \mathbf{C}$$

is C^1 on $\widehat{X}_j$ of constant length 1 and by an argument we have made before

$$|U'(x)| \le \frac{|\langle u_1, u_2\rangle'(x)|}{|\langle u_1(x), u_2(x)\rangle|} \le 10\cdot 16\cdot A|b|\frac{D}{\gamma^N}, \tag{4.9}$$

using (4.8). We can write

$$U(x) = \exp(i\phi(x))$$

for some C^1 real-valued $\phi : \widehat{X}_j \to \mathbf{R}$. Then (4.9) reads

$$|\phi'(x)| \le 160A|b|\frac{D}{\gamma^N}. \tag{4.10}$$

As we assume $\Phi \ne 0$ on $\widehat{X}_j$, we can find a C^1 function that we will denote

$$\arg\Phi : \widehat{X}_j \to S^1 = \mathbf{R}/2\pi\mathbf{Z}, \quad \Phi(x) = \exp(i\arg\Phi(x))\cdot|\Phi(x)|.$$

Now define

$$F(x) = (\tau^N(\alpha_1^N x) - \tau^N(\alpha_2^N x)), \quad x \in \widehat{X}_j.$$

The critical output of the (NLI) property for τ, Lemma 25, tells us that

$$0 < m \leq |F'(x)| \leq m' \tag{4.11}$$

when we choose $N > N_0$, which we do. As

$$\arg\Phi = bF + \phi,$$

we now have, incorporating (4.11) and (4.10),

$$|b|\left(m - 10 \cdot 16A\frac{D}{\gamma^N}\right) \leq |(\arg\Phi)'| \leq |b|\left(m' + 10 \cdot 16A\frac{D}{\gamma^N}\right).$$

We fix, finally, N large enough so that we gain $C_2 > C_1 > 0$ (depending only on N, m, m', A, D, and γ) with

$$|b|C_1 \leq |(\arg\Phi)'| \leq |b|C_2.$$

By estimating diameters of X_{j+1} and $\widehat{X}_j$ from Proposition 26 together with the mean value theorem, the total cumulative change of argument of Φ between $x_j \in X_j$ and $x_{j+2} \in X_{j+2}$, written Δ, is between

$$C_3\epsilon' \leq \Delta \leq C_4\epsilon',$$

where

$$C_3 = C_1 A_1' \inf_{U_0} |(\psi^{-1})'| > 0, \quad C_4 = C_2 3A_1 \|(\psi^{-1})'\|_\infty.$$

We now enforce $\epsilon' < \pi/(2C_4)$ so that we no longer need to worry about $\arg\Phi$ winding around the circle. We are about to conclude. Now ϵ' is fixed. By our trigonometric strategy, we are done with

$$\theta = \delta\left(M, \left(\frac{C_3\epsilon'}{100}\right)^2\right)$$

unless there exist $x_j \in X_j$ and $x_{j+2} \in X_{j+2}$ with

$$\Re\Phi(x_k) > 1 - \left(\frac{C_3\epsilon'}{100}\right)^2, \quad k = j, j+2.$$

In this case, by the Schwarz inequality we know

$$|\Phi(x_k)| \leq 1, \quad k = j, j+2,$$

so it follows that now using the principal branch for arg and, e.g., $|\sin x| \leq 2|x|$

$$|\arg\Phi(x_k)| \leq \frac{C_3\epsilon'}{50}, \quad k = j, j+2.$$

Given that the argument of Φ moves at least by $C_3\epsilon'$ in one direction between x_j and x_{j+2} and does not move more than $\pi/2$ (hence does not wind), this is a contradiction. □

We can now conclude this section with the following proof.

Proof of Lemma 24*, property* (3). Choose N, θ and ϵ' so that Proposition 28 holds as well as Lemma 29. Increasing N if necessary we may also assume that

$$\frac{D}{\gamma^N} \leq \frac{1}{4}.$$

Suppose we are given $H \in \mathcal{C}_{A|b|}$ and a function $f \in C^1(I; \mathbf{C}^{\Gamma_q})$ such that $|f| \leq H$ and $|f'| \leq A|b|H$. The second inequality stated in property (3) is softer so we prove this first. The complex scalar version of this inequality is proved in [14, p. 138].

We calculate

$$[L_{s,q}^N f](x) = \sum_{\alpha^N} \exp([\tau_a^N + ib\tau^N](\alpha^N x)) c_q^N(\alpha^N x) f(\alpha^N x),$$

where

$$c_q^N(y) = c_q(T^{N-1}y) \dots c_q(Ty).c_q(y)$$

and the sum is over branches of T^{-N}. Therefore

$$\begin{aligned}[L_{s,q}^N f]'(x) = &\sum_{\alpha^N} ([\tau_a^N + ib\tau^N] \circ \alpha^N)'(x) \exp([\tau_a^N + ib\tau^N](\alpha^N x)) c_q^N(\alpha^N x) f(\alpha^N x) \\ &+ \sum_{\alpha^N} \exp([\tau_a^N + ib\tau^N](\alpha^N x)) c_q^N(\alpha^N x)(f \circ \alpha^N)'(x),\end{aligned}$$

c_q^N being locally constant. Using that c_q^N is unitary and bounding derivatives of α^N with the eventually expanding property and chain rule gives

$$\begin{aligned}|[L_{s,q}^N f]'(x)| \leq &\sum_{\alpha^N} |([\tau_a^N + ib\tau^N] \circ \alpha^N)'(x)| \exp([\tau_a^N](\alpha^N x)) H(\alpha^N x) \\ &+ \frac{D}{\gamma^N} \sum_{\alpha^N} \exp([\tau_a^N](\alpha^N x)) A|b| H(\alpha^N x).\end{aligned}$$

Using inequality (4.4) in Proposition 28 and our choice of N, we get

$$|[L_{s,q}^N f]'(x)| \leq \frac{1}{2} A|b|[L_a^N H](x) \leq A|b|[\mathcal{N}_s^J H](x)$$

given the very mild assumption $\theta < 1/2$.

Now we turn to the more difficult first inequality of Lemma 24, property (3). Given that we have established Lemma 29 in the vector-valued setting, the proof follows by the same argument as in [14, p. 143]. We give the details here for completeness.

Let J be the set of indices (i, j), where $\Theta_i(x) \leq 1$ when $x \in X_j$. The statement of Lemma 29 is precisely that this set of indices is dense (recall Definition 27) and hence $J \in \mathcal{E}_s$ as required. We will prove

$$|L_{s,q}^N f| \leq \mathcal{N}_s^J H = L_a(\chi_J H).$$

Fix x. Notice that if $x \notin \operatorname{Int} X_j$ for any j, then for all branches α^N of T^{-N}, $\alpha^N x \notin Z_j^i$ and so $\chi_J(\alpha^N x) = 1$ for any J. More generally, if $x \notin \operatorname{Int} X_j$ for any j appearing as a coordinate in J, then $\chi_J(\alpha^N x) = 1$. Therefore

$$|[L_{s,q}^N f](x)| \leq \sum_{\alpha^N} \exp(\tau_a^N(\alpha^N x)) H(\alpha^N x) = \mathcal{N}_s^J[H](x).$$

We are left to consider x, J such that $x \in \operatorname{Int}(X_j)$ and J contains (i, j) for some i.

Suppose that $(i,j)=(1,j)$ and $(2,j)\notin J$. Then for $\alpha^N\neq\alpha_1^N$ a branch of T^{-N}, $\chi_J(\alpha^N x)=1$ (the only other possibility would have been $\alpha^N=\alpha_2^N$). Then using $\Theta_1(x)\leq 1$ gives

$$\begin{aligned}|L_{s,q}^N[f](x)| &\leq \sum_{\alpha^N\neq\alpha_1^N,\alpha_2^N} \exp(\tau_a^N(\alpha^N(x))H(\alpha^N(x))\\ &\qquad + (1-2\theta)\exp(\tau_a^N(\alpha_1^N(x))H(\alpha_1^N(x)) + \exp(\tau_a^N(\alpha_2^N(x))H(\alpha_2^N(x))\\ &\leq \mathcal{N}_s^J[H](x).\end{aligned}$$

The case $(i,j)=(2,j)$ and $(1,j)\notin J$ is treated the same way. Finally, if $(1,j)$ and $(2,j)$ are in J, then $\Theta_1(x),\Theta_2(x)\leq 1$ from which one can estimate

$$\begin{aligned}&|\exp([\tau_a^N+ib\tau^N](\alpha_1^N x))f(\alpha_1^N x)+\exp([\tau_a^N+ib\tau^N](\alpha_2^N x))f(\alpha_2^N x)|\\ &\quad\leq (1-\theta)\exp(\tau_a^N(\alpha_1^N(x))H(\alpha_1^N(x))+(1-\theta)\exp(\tau_a^N(\alpha_2^N(x))H(\alpha_2^N(x))\\ &\quad\leq \exp(\tau_a^N(\alpha_1^N(x))\chi_J(\alpha_1^N x)H(\alpha_1^N(x))+\exp(\tau_a^N(\alpha_2^N(x))\chi_J(\alpha_2^N x)H(\alpha_2^N(x)).\end{aligned}$$

Also noting that $\chi_J(\alpha^N x)=1$ when $\alpha^N\neq\alpha_i^N$, $i=1,2$, the previous inequality shows

$$|L_{s,q}^N[f](x)|\leq \mathcal{N}_s^J[H](x)$$

in our final remaining case. The proof is complete. □

5. Bounds for transfer operators: Small imaginary part

In this section we will prove the first part of Theorem 4. The key point is to think of $W\in C^1(I,\mathbf{C}^{\Gamma_q})$ as a function on $I\times\Gamma_q$ and decouple the variables. This allows us to relate the transfer operator to a convolution operator on $\mathbf{C}^{\Gamma_q}$. The relevant convolution operators have good spectral radius bounds that stem from the expander theory of Γ_q as described in the Appendix – the expansion technology requires that we restrict q to be coprime to a finite bad modulus Q_0, we make this restriction throughout. We now begin decoupling arguments in order to relate part (1) of Theorem 4 to the main result of the Appendix that we state as Theorem 33 below.

5.1. Accessing the convolution. We define E_q to be the space of functions of the group $\Gamma_q=\mathrm{SL}_2(\mathbf{Z}/q\mathbf{Z})$ that are orthogonal to all functions lifted from $\Gamma_{q'}$ for $q'|q$. We set out to show that when we iterate $L_{s,q}^n$ we suitably contract the C^1 norm.

We have calculated already that for $W\in C^1(I,\mathbf{C}^{\Gamma_q})$ with $\|W\|_{C^1}<\infty$ and taking on values only in the orthocomplement to constant functions

$$[L_{s,q}^N W](x)=\sum_{\alpha^N}\exp([\tau_a^N+ib\tau^N](\alpha^N x))c_q^N(\alpha^N x)W(\alpha^N x), \tag{5.1}$$

where the sum is over branches of T^{-N} on the interval containing x. We can write these branches in a special form. They are given precisely by sequences

$$\alpha^N=g_{i_1}g_{i_2}\dots g_{i_N},$$

where the g_{i_j} form an admissible sequence. If in the general Schottky semigroup setting, we also require that if $x\in\tilde{I}_i$ for $1\leq i\leq 2k'$, then $i_N\neq i+k' \bmod 2k'$, recalling the notation of Section 2.1.

It will be convenient to make the parametrization

$$N = M + R, \quad M, R > 0.$$

We then write

$$\alpha^N = \alpha^M \alpha^R,$$

where

$$\alpha^M = g_{i_1} \dots g_{i_M}, \quad \alpha^R = g_{i_{M+1}} \dots g_{i_N}, \tag{5.2}$$

and view these as globally defined maps on I. Then α^N is uniquely parameterized by the two sequences appearing in (5.2). When we write α^M and α^R, henceforth we always mean compositions of these forms. Notice that the choice of α^R is restricted depending on x and α^M is restricted depending on $g_{i_{M+1}}$.

For each of the intervals I_i we pick a point $x_0(i) \in I_i$. For each α^M we pick $i_0 = i_0(\alpha^M)$ such that α^M gives a well-defined branch on I_{i_0}. Then

$$d(\alpha^N x, \alpha^M x_0(i_0)) = d(\alpha^M(\alpha^R x), \alpha^M x_0(i_0)) \leq \frac{D}{\gamma^M} \operatorname{diam}(I)$$

by the eventually expanding property of T. Then

$$\|W(\alpha^N x) - W(\alpha^M x_0(i_0))\| \leq \frac{D}{\gamma^M} \operatorname{diam}(I) \|W\|_{C^1}.$$

It follows then that

$$\begin{aligned}[L_{s,q}^N W](x) = &\sum_{\alpha^M} \sum_{\alpha^R}{}^{*} \exp([\tau_a^N + ib\tau^N](\alpha^N x)) c_q^N(\alpha^N x) W(\alpha^M x_0(i_0)) \\ &+ O\bigg(\|W\|_{C^1} \frac{D}{\gamma^M} \operatorname{diam}(I) \sum_{\alpha^N} \exp(\tau_a^N(\alpha^N x))\bigg),\end{aligned}$$

where the star on summation means that we restrict to those α^R with necessary restriction on g_{i_N} coming from x and $g_{i_{M+1}}$ coming from α^M. Note that i_0 depends on α^M. We will assume that $D\gamma^{-M}$ is small, say $< 1/(100 \operatorname{diam}(I))$ and note that the sum in the error term is

$$\sum_{\alpha^N} \exp(\tau_a^N(\alpha^N x)) = L_a^N[1](x) = 1(x) = 1$$

as the operator has been normalized. So then

$$\begin{aligned}[L_{s,q}^N W](x) = &\sum_{\alpha^M} \sum_{\alpha^R}{}^{*} \exp([\tau_a^N + ib\tau^N](\alpha^N x)) c_q^N(\alpha^N x) W(\alpha^M x_0(i_0)) \\ &+ O(\|W\|_{C^1} \gamma^{-M}).\end{aligned} \tag{5.3}$$

This is an important estimate as it allows us access the expansion properties coming from c_q by decoupling M and N.

Recall that c_q was obtained by reducing the matrices g_i modulo q to obtain a locally constant mapping $c_q : I \to \Gamma_q$. This mapping can be reinterpreted as a unitary-valued map $c_q : I \to U(\mathbf{C}^{\Gamma_q})$ via the right regular representation of Γ_q.

For any specified α^M as in (5.2) and $x \in I$ we construct the complex-valued measure on Γ_q,

$$\mu_{s,x,\alpha^M} = \sum_{\alpha^R}{}^{*} \exp([\tau_a^N + ib\tau^N](\alpha^M\alpha^R x))\delta_{c_q^R(\alpha^R x)^{-1}}, \tag{5.4}$$

where δ_g gives mass one to $g \in \Gamma_q$. We note for the reader's convenience that one can calculate from the definition in the Dictionary of Table 1, Section 3,

$$c_q^R(\alpha^R x) = g_{i_N} g_{i_{N-1}} \cdots g_{i_{M+1}} \bmod q, \quad c_q^M(\alpha^M x_0(i_0)) = g_{i_M} g_{i_{M-1}} \cdots g_{i_1} \bmod q.$$

For any $f \in C^1(I; \mathbf{C}^{\Gamma_q})$ and α^M as in (5.2) we construct a complex-valued measure φ_{f,α^M} by

$$\varphi_{f,\alpha^M} = \sum_{g \in \Gamma_q} f(\alpha^M x_0(i_0))|_g \delta_{g c_q^M(\alpha^M x_0(i_0))^{-1}},$$

where c_q is thought of as Γ_q-valued and $f(\alpha^M x_0(i_0))$ thought of as a $\mathbf{C}$-valued function on Γ_q, with $|_g$ standing for evaluation at g. Also recall $i_0 = i_0(\alpha^M)$. Then

$$\begin{aligned}[\varphi_{f,\alpha^M} \star \mu_{s,x,\alpha^M}] &= \sum_{g \in \Gamma_q} \sum_{\alpha^R}{}^{*} \exp([\tau_a^N + ib\tau^N](\alpha^M\alpha^R x)) f(\alpha^M x_0(i_0))|_g \\ &\qquad\qquad \times \delta_{g c_q^M(\alpha^M x_0(i_0))^{-1}} \star \delta_{c_q^R(\alpha^R x)^{-1}} \\ &= \sum_{g \in \Gamma_q} \sum_{\alpha^R}{}^{*} \exp([\tau_a^N + ib\tau^N](\alpha^M\alpha^R x)) f(\alpha^M x_0(i_0))|_g \delta_{g c_q^N(\alpha^M\alpha^R x)^{-1}}.\end{aligned}$$

This means that, now viewed as a function on $\mathrm{SL}_2(\mathbf{Z}/q\mathbf{Z})$,

$$[\varphi_{f,\alpha^M} \star \mu_{s,x,\alpha^M}] = \sum_{\alpha^R}{}^{*} \exp([\tau_a^N + ib\tau^N](\alpha^N x)) c_q^N(\alpha^M\alpha^R x) f(\alpha^M x_0(i_0)).$$

The reader should compare this with (5.3).

5.2. Bounds for μ_{s,x,α^M}.

We need a bound for $\|\mu_{s,x,\alpha^M}\|_1$ to use the result of the Appendix. Firstly we write

$$|\mu_{s,x,\alpha^M}| \le \sum_{\alpha^R}{}^{*} \exp(\tau_a^N(\alpha^N x))\delta_{c_q^R(\alpha^R x)}.$$

Notice that

$$\tau_a^N(\alpha^M\alpha^R x) = \tau_a^M(\alpha^M\alpha^R x) + \tau_a^R(\alpha^R x).$$

Then

$$\|\mu_{s,x,\alpha^M}\|_1 \le \sum_{\alpha^R}{}^{*} \exp(\tau_a^M(\alpha^M\alpha^R x)) \exp(\tau_a^R(\alpha^R x)).$$

We now decouple: let α_0^R be any arbitrary choice of α^R (a sequence of the g_i that is compatible with α^M and x). Then

$$\tau_a^M(\alpha^M\alpha^R x) - \tau_a^M(\alpha^M\alpha_0^R x) = \sum_{n=0}^{M-1} \tau_a(T^n\alpha^N x) - \tau_a(T^n\alpha_0^N x)$$

and noting that $T^n\alpha^M\alpha^R x$ and $T^n\alpha^M\alpha_0^R x$ are within

$$\frac{D}{\gamma^{M-n}} \operatorname{diam}(I)$$

of one another, we have

$$\begin{aligned}\tau_a^M(\alpha^M\alpha^R x) &\le \tau_a^M(\alpha^M\alpha_0^R x) + D.\operatorname{diam}(I)\sup_{y\in I}|[\tau_a]'(y)|\sum_{n=0}^{M-1}\frac{1}{\gamma^{M-n}} \\ &\le \tau_a^M(\alpha^M\alpha_0^R x) + \kappa_1(D,\gamma,I,\tau,a_0)\end{aligned} \tag{5.5}$$

for $|a-\delta| < a_0$ (as τ_a is roughly constant in a close to δ). Therefore

$$\begin{aligned}\|\mu_{s,x,\alpha^M}\|_1 &\le \exp(\kappa_1 + \tau_a^M(\alpha^M\alpha_0^R x))\sum_{\alpha^R}{}^{*}\exp(\tau_a^R(\alpha^R x)) \\ &\le \exp(\kappa_1 + \tau_a^M(\alpha^M\alpha_0^R x))[L_a^R 1](x) = \exp(\kappa_1 + \tau_a^M(\alpha^M\alpha_0^R x))\end{aligned}$$

by the normalization of L_a. We record this bound in the following.

Lemma 32. *Given a_0 small enough, there is $\kappa_1 = \kappa_1(a_0)$ such that for all x and α^M,*

$$\|\mu_{s,x,\alpha^M}\|_1 \le \exp(\kappa_1 + \tau_a^M(\alpha^M\alpha_0^R x))$$

for $|a-\delta| < a_0$. Here α_0^R is any admissible choice of α^R as in (5.2) *compatible with α^M and x.*

We are now in a position to use the main result of the Appendix, which for the convenience of the reader we also state here.

Theorem 33 (Bourgain–Kontorovich–Magee, Appendix). *There is a finite modulus Q_0 and $c > 0$ such that when $R \approx c\log q$, $(q, Q_0) = 1$, $|a-\delta| < a_0$ and $\varphi \in E_q$, we have*

$$\|\varphi \star \mu_{s,x,\alpha^M}\|_2 \le Cq^{-1/4}B\|\varphi\|_2,$$

given that

$$\|\mu\|_1 < B.$$

Using Lemma 32, Theorem 33 now implies that when $R \approx c\log q$ for suitable c and $|a-\delta| < a_0$, for any $\varphi \in E_q$ we have

$$\|\varphi \star \mu_{s,x,\alpha^M}\|_2 \le Cq^{-1/4}\exp(\kappa_1 + \tau_a^M(\alpha^M\alpha_0^R x))\|\varphi\|_2. \tag{5.6}$$

Then if we use (5.3) we obtain

$$\begin{aligned}\|[L_{s,q}^N W](x)\|_{l^2(\Gamma_q)} &\le \sum_{\alpha^M}\|[\varphi_{W,\alpha^M} \star \mu_{s,x,\alpha^M}]\|_{l^2(\Gamma_q)} + O(\|W\|_{C^1}\gamma^{-M}) \\ &\le Cq^{-1/4}\exp(\kappa_1)\sum_{\alpha^M}\exp(\tau_a^M(\alpha^M\alpha_0^R x))\|\varphi_{W,\alpha^M}\|_{l^2(\Gamma_q)} \\ &\quad + O(\|W\|_{C^1}\gamma^{-M}).\end{aligned}$$

We have now chosen some α_0^R and we are assuming the previous conditions on $|a-\delta| < a_0$. Since trivially

$$\|\varphi_{W,\alpha^M}\|_{l^2(\Gamma_q)} \le \|W\|_\infty$$

we can continue to bound $\|[L_{s,q}^N W](x)\|$ up to $O(\|W\|_{C^1}\gamma^{-M})$ by

$$\begin{aligned} &Cq^{-1/4}\exp(\kappa_1)\|W\|_\infty \sum_{\alpha^M} \exp(\tau_a^M(\alpha^M\alpha_0^R x)) \\ &\leq Cq^{-1/4}\exp(\kappa_1)\|W\|_\infty L_a^N[1](T^M\alpha^M\alpha_0^R x) \\ &= Cq^{-1/4}\exp(\kappa_1)\|W\|_\infty. \end{aligned}$$

We have now proved, by choosing $N > \kappa_{10}\log q$ so that there is room for the requisite R and big enough M the following lemma.

Lemma 34. *Let $(q, Q_0) = 1$. There are $a_0, q_0, \kappa_{10}, \epsilon > 0$ and $\gamma' > 1$ such that when $|a - \delta| < a_0$, we have*

$$\|L_{s,q}^N W\|_\infty \leq q^{-\epsilon}\|W\|_\infty + \gamma'^{-N}\|W\|_{C^1}$$

when $N > \kappa_{10}\log q$, $q > q_0$, and $W \in E_q$ with $\|W\|_{C^1} < \infty$.

5.3. Bounds for Lipschitz norms. In order to iterate Lemma 34 (this is our aim) we also need bounds for

$$\|L_{s,q}^N W\|_{C^1}$$

under the same conditions as in Lemma 34. This amounts to estimating

$$\sup_I |[L_{s,q}^N W]'|$$

and so we can proceed along similar lines as before. Indeed one calculates from (5.1) that

$$\begin{aligned} [L_{s,q}^N W]'(x) = &\sum_{\alpha^N} ([\tau_a^N + ib\tau^N]\circ\alpha^N)'(x)\exp([\tau_a^N + ib\tau^N](\alpha^N x))c_q^N(\alpha^N x)W(\alpha^N x) \\ &+ \sum_{\alpha^N} \exp([\tau_a^N + ib\tau^N](\alpha^N x))c_q^N(\alpha^N x)[W\circ\alpha^N]'(x) \end{aligned}$$

using that c_q^N is locally constant. The second set of terms are bounded by

$$\frac{D}{\gamma^N}\sum_{\alpha^N}\exp(\tau_a^N(\alpha^N x))\|W\|_{C^1}$$

which can be bounded by

$$\frac{D}{\gamma^N}\|W\|_{C^1}L_a^N[1](x) = \frac{D}{\gamma^N}\|W\|_{C^1}.$$

So we have

$$[L_{s,q}^N W]'(x) = \Sigma + O\left(\frac{D}{\gamma^N}\|W\|_{C^1}\right), \tag{5.7}$$

where

$$\Sigma := \sum_{\alpha^N}([\tau_a^N + ib\tau^N]\circ\alpha^N)'(x)\exp([\tau_a^N + ib\tau^N](\alpha^N x))c_q^N(\alpha^N x)W(\alpha^N x).$$

We can go through the same decoupling argument as before to get

$$\Sigma = \sum_{\alpha^N}([\tau_a^N + ib\tau^N] \circ \alpha^N)'(x) \exp([\tau_a^N + ib\tau^N](\alpha^N x)) c_q^N(\alpha^N x) W(\alpha^M x_0(i_0))$$
$$+ O\biggl(\|W\|_{C^1} \frac{D}{\gamma^M} \operatorname{diam}(I) \sum_{\alpha^N} |([\tau_a^N + ib\tau^N] \circ \alpha^N)'(x)| \exp(\tau_a^N(\alpha^N x))\biggr),$$

recalling $i_0 = i_0(\alpha^M)$. Note that since there are constants C_1 and a_0 such that when $|a - \delta| < a_0$ we have

$$|[\tau_a^N \circ \alpha^N]'(x)| \le C_1$$

for $x \in I$ (see for example [14, p. 138]), we have

$$|[[\tau_a^N + ib\tau^N] \circ \alpha^N]'(x)| \le C_1 + |b| \sup_I |\tau'| \sum_{i=0}^{N-1} \frac{D}{\gamma^i} \le \kappa_{11} \tag{5.8}$$

for some $\kappa_{11} = \kappa_{11}(a_0, b_0)$ when $|b| \le b_0$. Therefore we have the decoupled equation

$$\Sigma = \sum_{\alpha^N}([\tau_a^N + ib\tau^N] \circ \alpha^N)'(x) \exp([\tau_a^N + ib\tau^N](\alpha^N x)) c_q^N(\alpha^N x) W(\alpha^M x_0(i_0))$$
$$+ O_{b_0}(\|W\|_{C^1}\gamma^{-M})$$

valid when $|b| < b_0$ and $|a - \delta| < a_0$ for some fixed a_0. We denote the first of these two terms by Σ'. Now similarly to before we define complex-valued measures

$$\mu'_{s,x,\alpha^M} = \sum_{\alpha^R}{}^{*}([\tau_a^N + ib\tau^N] \circ \alpha^M\alpha^R)'(x) \exp([\tau_a^N + ib\tau^N](\alpha^M\alpha^R x))\delta_{c_q^R(\alpha^R x)^{-1}},$$
$$\varphi_{f,\alpha^M} = \sum_{g \in \Gamma_q} f(\alpha^M x_0(i_0))|_g \delta_{g c_q^M(\alpha^M x_0(i_0))^{-1}}$$

for $f \in C^1(I; \Gamma^q)$, α^M as in (5.2). Then the key observation is that

$$\|\Sigma'\| = \biggl\| \sum_{\alpha^M} \varphi_{W,\alpha^M} \star \mu'_{s,x,\alpha^M} \biggr\|_{l^2(\Gamma_q)}.$$

5.4. Bounds for μ'_{s,x,α^M}.

We have

$$\|\mu'_{s,x,\alpha^M}\|_1 \le \sup_I |[\tau_a^N + ib\tau^N] \circ \alpha^M\alpha^R)'(x)| \sum_{\alpha^R}{}^{*} \exp(\tau_a^N(\alpha^M\alpha^R x)).$$

By (5.8), μ'_{s,x,α^M} is dominated (in absolute value) by $\kappa_{11}(b_0)\mu_{s,x,\alpha^M}$ when $|b| < b_0$. Thus we can use our previous bound (5.6) to deduce that for the same choice of $R = c\log q$ and a as before,

$$\begin{aligned}\|\Sigma'\| &\le Cq^{-1/4} \exp(\kappa_1)\kappa_{11}\|W\|_\infty \sum_{\alpha^M} \exp(\tau_a^M(\alpha^M\alpha_0^R x))\\ &\le Cq^{-1/4} \exp(\kappa_1)\kappa_{11}\|W\|_\infty L_a^N[1](\alpha_0^R x)\\ &\le Cq^{-1/4} \exp(\kappa_1)\kappa_{11}\|W\|_\infty\end{aligned}$$

whenever $|a-\delta| < a_0$, $|b| < b_0$ are the ranges specified by previous lemmas and $N > \kappa_{12} \log q$. It now follows from (5.7) that with these conditions on N, q, a, b we have in light of Lemma 34 and the prior bound (5.7)

$$\|L_{s,q}^N W\|_{C^1} \leq \kappa_{13} q^{-\epsilon} \|W\|_\infty + \kappa_{14} \gamma^{-N} \|W\|_{C^1} + \gamma'^{-N} \|W\|_{C^1} \tag{5.9}$$

for some $\epsilon > 0$ when $W \in E_q$.

By iterating the estimate (5.9) one obtains:

Lemma 35. *For $b_0 > 0$ given, there are a_0, q_0, κ and $\epsilon > 0$ such that when $|a-\delta| < a_0$, $|b| < b_0$ and $N = \lceil \kappa \log q \rceil$ with $q > q_0$ and $(q, Q_0) = 1$ we have*

$$\|L_{s,q}^{nN} W\|_{C^1} \leq q^{-n\epsilon}$$

for all E_q-valued $W \in C^1(I; \mathbf{C}^{\Gamma_q})$ with $\|W\|_{C^1} = 1$.

5.5. The new subspace structure and the proof of Part (1) of Theorem 4. We note first the following consequence of Lemma 35.

Lemma 36. *For all $b_0 > 0$ there are $0 < \rho < 1$, a_0, q_0 and C such that when $|a-\delta| < a_0$, $|b| \leq b_0$ and $q_0 < q$, $(q, Q_0) = 1$, we have for all $m > 0$,*

$$\|L_{s,q}^m f\|_{C^1} \leq C q^C \rho^m \|f\|_{C^1} \quad \text{when } f \in E_q.$$

This is an easy exercise and the reader can get the details from [15, proof of Theorem 4.3].

Recall the new subspace structure of Γ_q. For any $q'|q$ there is a projection $\Gamma_q \to \Gamma_{q'}$. The kernel of this projection will be denoted $\Gamma_q(q')$, the congruence subgroup of level q' in Γ_q. These have the property that if $q''|q'$ then $\Gamma_q(q') \leq \Gamma_q(q'')$. This groups give an orthogonal decomposition of the right regular representation

$$\mathbf{C}^{\Gamma_q} = \bigoplus_{q'|q} E_{q'}^q, \tag{5.10}$$

where $E_{q'}^q$ consists of functions invariant under $\Gamma_q(q')$ but not invariant under $\Gamma_q(q'')$ for any q'' such that $q''|q'$, $q'' \neq q'$. Then the E_q from before matches E_q^q as defined here.

The decomposition (5.10) gives rise to a corresponding direct sum decomposition

$$C^1(I; \mathbf{C}^{\Gamma_q}) = C^1(I) \oplus \bigoplus_{1 \neq q'|q} C^1(I; E_{q'}^q).$$

It is clear that the subspaces $E_{q'}^q$ are invariant under the transfer operator $L_{s,q}$ and taking derivatives.

Note that if $f \in E_{q'}^q$, then f descends to a well-defined function F on $\Gamma_q / \Gamma_q(q') \cong \Gamma_{q'}$ which is not invariant under any congruence subgroup of $\Gamma_{q'}$, hence in $E_{q'}^{q'}$. Also, if G is a function in $E_{q'}^{q'}$, then G lifts through the previous isomorphism to a function g in $E_{q'}^q$ for any q' such that $q'|q$. This gives rise to a map of Banach spaces

$$\Phi_{q,q'} : C^1(I; E_{q'}^{q'}) \to C^1(I; E_{q'}^q)$$

for any $q'|q$ with the property that

$$\|\Phi_{q,q'}(f)\|_{C^1} = \sqrt{|\Gamma_q(q')|} \|f\|_{C^1}.$$

This map is equivariant under the transfer operators in the sense that

$$\Phi_{q,q'}[L_{s,q'}f] = L_{s,q}\Phi_{q,q'}[f]$$

for any $f \in E^{q'}_{q'}$. In other words, the action of $L_{s,q}$ on a summand in (5.10) is determined by the action of the corresponding transfer operator on $E^{q'}_{q'}$ for some $q'|q$. We decompose $f \in C^1(I;\mathbf{C}^{\Gamma_q})$ as

$$f = f_1 + \sum_{1 \neq q'|q} f_{q'}$$

with $f_{q'} \in E^{q}_{q'}$. If we assume that q has no proper divisors $\leq q_0$ from Lemma 36, then for any m, with all norms C^1 norms,

$$\begin{aligned} \|L^m_{s,q}f - L^m_{s,q}f_1\| &\leq \sum_{q_0<q'|q} \|L^m_{s,q}f_{q'}\| \\ &= \sum_{q_0<q'|q} \sqrt{\#\Gamma_q(q')}\|L^m_{s,q'}\Phi^{-1}_{q,q'}f_{q'}\| \\ &\leq C \sum_{q_0<q'|q} \sqrt{\#\Gamma_q(q')}(q')^C\rho^m\|\Phi^{-1}_{q,q'}f_{q'}\| \\ &\leq Cq^C\rho^m \sum_{1\neq q'|q} \|f_{q'}\|. \end{aligned}$$

This bound can be changed to

$$\|L^m_{s,q}f - L^m_{s,q}f_1\| \leq C'q^{C'}\rho^m\|f\|$$

for some $C' = C'(\Gamma, b_0)$ by noting that individually

$$\|f_{q'}\| \leq \|f\|$$

and that any number q has $\ll_\epsilon q^\epsilon$ divisors for any $\epsilon > 0$. The analogous estimates hold for the unnormalized $\mathcal{L}_{s,q}$ (by perturbation theory and (4.1)). That is, by possibly adjusting constants slightly and decreasing a_0,

$$\|L^m_{s,q}f - L^m_{s,q}f_1\| \leq C'q^{C'}\rho^m\|f\|.$$

In particular, part (1) of Theorem 4 now follows from the special case that $f_1 = 0$ so that $f \in C^1(I;\mathbf{C}^{\Gamma_q} \ominus 1)$.

A. Thermodynamic expansion to arbitrary moduli

By *Jean Bourgain* at Princeton, *Alex Kontorovich* at New Brunswick
and *Michael Magee* at New Haven

A.1. Statements. We import all the notation from the rest of the paper. We are led to study the measure μ on $G = \mathrm{SL}_2(q)$ given by

$$\mu = \sum_{\alpha^R}{}^{*} \exp([\tau^N_a + ib\tau^N](\alpha^M\alpha^R x))\delta_{c^R_q(\alpha^R x)}, \tag{A.1}$$

this differs from the μ_{s,x,α^M} of equation (5.4) by taking inverses of group elements. This makes spectral bounds for the right action of μ_{s,x,α^M} and those for the left action of μ equivalent. Here $N = M + R$, $x \in I$,

$$\alpha^M = g_{i_1} g_{i_2} \dots g_{i_M}$$

is fixed, and the starred summation means that it is restricted to those

$$\alpha^R = g_{i_{M+1}} g_{i_{M+2}} \dots g_{i_N}$$

where the sequence $g_{i_1}, \dots, g_{i_N}$ is admissible and α^R is a well-defined local branch of T^{-R} near x. In practice this may rule out one possible value for i_N. See Section 5.1 for more details. Also recall the "new subspace" $E_q \subset l^2(G)$ defined in Section 5.1 and the constant a_0 coming from Proposition 28.

Our goal in this Appendix is to prove the following:

Theorem A.1. *There are a finite modulus Q_0 and a constant $c > 0$ such that when $R \approx c \log q$, $(q, Q_0) = 1$, $|a - \delta| < a_0$ and $\varphi \in E_q$, we have*

$$\|\mu * \varphi\|_2 \le C q^{-1/4} B \|\varphi\|_2, \tag{A.2}$$

given that

$$\|\mu\|_1 < B.$$

Recall that in Section 5.1 we chose for each α^M an $i_0 = i_0(\alpha^M)$ such that α^M is a well-defined local branch of T^{-M} on I_{i_0}. We also chose for each i an $x(i)$ in I_i. More generally, for each admissible composition $\alpha = g_{i_1} \dots g_{i_j}$ of semigroup elements we now choose an $i(\alpha)$ such that α is a well-defined branch of T^{-j} on $I_{i(\alpha)}$. This choice depends only on i_j. Let $o = x(i(\alpha^R))$.

To begin, we define a measure ν by

$$\nu \equiv \exp(\tau_a^M(\alpha^M x(i_0)))\mu_1, \tag{A.3}$$

where μ_1 is the measure given by

$$\mu_1 \equiv \sum_{\alpha^R}^{*} \exp(\tau_a^R(\alpha^R o))\delta_{c_q^R(\alpha^R o)}.$$

Lemma A.2. *We have*

$$|\mu| \le C\,\nu. \tag{A.4}$$

Proof. Use the "contraction property" in inequality (5.5) and argue as in the proof of Lemma 32. □

We will now manipulate μ_1. We assume that R can be decomposed further as

$$R = R'L, \tag{A.5}$$

with L to be chosen later (a sufficiently large constant independent of R' and q). Now split α^R as

$$\alpha^R = \alpha_{R'}^L \alpha_{R'-1}^L \dots \alpha_2^L \alpha_1^L,$$

where the α_k^L are branches of T^{-L} given by

$$\alpha_{R'}^L = g_{i_{M+1}} \dots g_{i_{M+L}}, \quad \alpha_{R'-1}^L = g_{i_{M+L+1}} \dots g_{i_{M+2L}}$$

and so on. For each $0 \leq p \leq R'-1$ we also split

$$\alpha_{R'-p}^L = \alpha_{R'-p}^{L-2} \alpha_{R'-p}^{(2)},$$

where $\alpha_{R'-p}^{L-2} = g_{i_{M+pL+1}} \dots g_{i_{M+(p+1)L-2}}$ and $\alpha_{R'-p}^{(2)} = g_{i_{M+(p+1)L-1}} g_{i_{M+(p+1)L}}$. The reason for isolating two indices will become clear later.

Write out

$$\begin{aligned} \tau_a^R(\alpha^R o) &= \sum_{i=0}^{R-1} \tau_a(T^i \alpha^R o) \\ &= \sum_{i=0}^{R'-1} \sum_{\ell=0}^{L-1} \tau_a(T^{iL+\ell} \alpha^R o) \\ &= \sum_{i=0}^{R'-1} \sum_{\ell=0}^{L-1} \tau_a(T^{iL+\ell} \alpha_{R'-i}^L \alpha_{R'-i-1}^L \dots \alpha_1^L o) \\ &= \sum_{i=0}^{R'-1} \tau_a^L(\alpha_{R'-i}^L \alpha_{R'-i-1}^L \dots \alpha_1^L(o)). \end{aligned} \tag{A.6}$$

We now perform decoupling term by term in the above. We will use the shorthand

$$\alpha^{Lj} \equiv \alpha_j^L \alpha_{j-1}^L \dots \alpha_1^L.$$

For $j \geq 2$, we compare each term in (A.6) of the form

$$\tau_a^L(\alpha^{Lj}(o))$$

to

$$\tau_a^L(\alpha_j^L \alpha_{j-1}^{L-2} x(i(\alpha_{j-1}^{L-2}))).$$

This gives

$$\begin{aligned} \tau_a^L(\alpha^{Lj}(o)) &= \tau_a^L(\alpha_j^L \alpha_{j-1}^{L-2} x(i(\alpha_{j-1}^{L-2}))) \\ &\quad + O\big(\sup |[\tau_a^L \circ \alpha_j^L]'| d(\alpha_{j-1}^{L-2} x(i(\alpha_{j-1}^{L-2})), \alpha_{j-1}^{L-2} \alpha_{j-1}^{(2)} \dots \alpha_1^L o)\big) \\ &= \tau_a^L(\alpha_j^L \alpha_{j-1}^{L-2} x(i(\alpha_{j-1}^{L-2}))) + O(\gamma^{-(L-2)}), \end{aligned} \tag{A.7}$$

where we used the bound (4.3) of Proposition 28, valid when a is within a_0 of δ.

We will also use the formula

$$\delta_{c_q^R(\alpha^R o)} = \delta_{c_q^L(\alpha^L o)} * \delta_{c_q^L(\alpha^{2L} o)} * \delta_{c_q^L(\alpha^{3L} o)} * \dots * \delta_{c_q^L(\alpha^{R'L} o)}. \tag{A.8}$$

Then combining (A.6) and (A.8), we write

$$\begin{aligned} \mu_1 &= \sum^*_{\alpha_1^L, \alpha_2^{L-2}, \dots, \alpha_{R'}^{L-2}} \; \sum^*_{\alpha_2^{(2)}, \dots, \alpha_{R'}^{(2)}} \exp(\tau_a^R(\alpha^R o))) \delta_{c_q^R(\alpha^R o)} \\ &= \sum^*_{\alpha_1^L, \alpha_2^{L-2}, \dots, \alpha_{R'}^{L-2}} \; \sum^*_{\alpha_2^{(2)}, \dots, \alpha_{R'}^{(2)}} \exp\left(\sum_{j=1}^{R'} \tau_a^L(\alpha^{jL}(o)) \right) \\ &\qquad \times \delta_{c_q^L(\alpha^L o)} * \delta_{c_q^L(\alpha^{2L} o)} * \delta_{c_q^L(\alpha^{3L} o)} * \dots * \delta_{c_q^L(\alpha^{R'L} o)}. \end{aligned} \tag{A.9}$$

Starred summation means that the outer sum is restricted to be compatible with α^M and x, and given the collection of α_k^{L-2} from the outer sum, we then restrict to those $\alpha_k^{(2)}$ that form admissible compositions overall. We now decouple, replacing each term of the form

$$e^{\tau_a^L(\alpha^{jL}(o))} \mapsto e^{\tau_a^L(\alpha_j^L \alpha_{j-1}^{L-2} x(i(\alpha_{j-1}^{L-2})))} \equiv \beta_j$$

with $j \geq 2$, at a cost of a multiplicative factor of $\exp(c\gamma^{-L})$; here c is proportional to the implied constant of (A.7). When $j = 1$, no replacement is performed, and we set

$$\beta_1 \equiv e^{\tau_a^L(\alpha_1^L o)}.$$

Inserting this into (A.9) gives

$$\text{(A.10)} \qquad \mu_1 \leq \sideset{}{^*}\sum_{\alpha_1^{L-2}, \alpha_2^{L-2}, \dots, \alpha_{R'}^{L-2}} \sideset{}{^*}\sum_{\alpha_1^{(2)}} \beta_1 \delta_{c_q^L(\alpha^L o)} * \exp(c\gamma^{-L})^{R'-1} \times \Bigg(\sideset{}{^*}\sum_{\alpha_2^{(2)}, \dots, \alpha_{R'}^{(2)}} \prod_{j=2}^{R'} \beta_j \delta_{c_q^L(\alpha^{2L} o)} * \delta_{c_q^L(\alpha^{3L} o)} * \dots * \delta_{c_q^L(\alpha^{R'L} o)} \Bigg).$$

Note that, although β_j depends on all of the indices in $\alpha_j^L \alpha_{j-1}^{L-2}$, because α_j^{L-2} and α_{j-1}^{L-2} are fixed in the outermost sum, we can and will treat β_j as a function of $\alpha_j^{(2)}$.

We claim that each term $c_q^L(\alpha^{jL} o)$ also only depends on one $\alpha_j^{(2)}$. This is because we have

$$\alpha^{jL} = g_{k_1} \dots g_{k_L} \alpha^{(j-1)L}$$

for some choice of g_{k_m}, and hence for whatever o is chosen, we have

$$c_q^L(\alpha^{jL} o) = c_q(g_{k_L} \alpha^{(j-1)L} o) c_q(g_{k_{L-1}} g_{k_L} \alpha^{(j-1)L} o) \dots c_q(g_{k_1} \dots g_{k_L} \alpha^{(j-1)L} o),$$

see the Dictionary of Table 1, Section 3. From the definition of c_q we have

$$c_q(g_{k_m} o') = g_{k_m} \bmod q$$

for any $o' \in I$, where g_{k_m} is a local inverse branch of T near o'. Thus

$$\text{(A.11)} \qquad c_q^L(\alpha^{jL} o) = g_{k_L} \dots g_{k_1} \bmod q.$$

Here

$$\text{(A.12)} \qquad g_{k_{L-1}} g_{k_L} = \alpha_j^{(2)}.$$

This means we may distribute the convolution and product over the sum, writing (A.10) as

$$\text{(A.13)} \qquad \mu_1 \leq \exp(c\gamma^{-L})^{R'-1} \sideset{}{^*}\sum_{\alpha_1^{L-2}, \alpha_2^{L-2}, \dots, \alpha_{R'}^{L-2}} \Bigg(\sideset{}{^*}\sum_{\alpha_1^{(2)}} \beta_1 \delta_{c_q^L(\alpha^L o)} \Bigg) * \Bigg(\sideset{}{^*}\sum_{\alpha_2^{(2)}} \beta_2 \delta_{c_q^L(\alpha^{2L} o)} \Bigg) * \dots * \Bigg(\sideset{}{^*}\sum_{\alpha_{R'}^{(2)}} \beta_{R'} \delta_{c_q^L(\alpha^{R'L} o)} \Bigg).$$

We give each convolved term in (A.13) a name, defining, for each $j \geq 1$, the measure

$$\text{(A.14)} \qquad \eta_j = \eta_j^{(\alpha_j^{L-2}, \alpha_{j-1}^{L-2})} \equiv \sideset{}{^*}\sum_{\alpha_j^{(2)}} \beta_j \delta_{c_q^L(\alpha^{jL} o)}.$$

Note this parameterization makes sense since the admissibility of $\alpha_j^{(2)}$ depends only on α_j^{L-2} and α_{j-1}^{L-2}. We have thus proved the following:

Proposition A.3. *We have*

$$\mu_1 \le \exp(c\gamma^{-L})^{R'-1} \sideset{}{^*}\sum_{\alpha_1^{L-2},\alpha_2^{L-2},\dots,\alpha_{R'}^{L-2}} \eta_1 * \eta_2 * \cdots * \eta_{R'}. \tag{A.15}$$

Next we observe that each of the measures η_j is nearly flat, in that their coefficients in (A.14) differ by constants:

Lemma A.4. *There is some $c' > 0$ such that for any $L > 0$, for each $j \ge 1$ and any $\alpha_j^{(2)}$ and $\alpha_j^{(2)\prime}$, we have*

$$\frac{\beta_j'}{\beta_j} \le c'.$$

Proof. The first $L-2$ terms of β_j and β_j' agree, so we again use the "contraction property" from (5.5). □

Since the measures η_j are nearly flat, we may now apply the expansion result in [6].

Theorem A.5. *Assume that L is sufficiently large (depending only on Γ). Then for $\varphi \in L_0^2(G)$, we have*

$$\|\eta_j * \varphi\|_2 \le (1 - C_1)\|\eta_j\|_1\|\varphi\|_2. \tag{A.16}$$

Here $C_1 > 0$ depends on Γ but not on q.

Proof of Theorem A.5. Recalling (A.14), we can write

$$\|\eta_j * \varphi\|_2^2 = \langle \widetilde{A}\varphi, \varphi\rangle, \tag{A.17}$$

where $\widetilde{A}$ acts by convolution with the measure

$$A \equiv \sideset{}{^*}\sum_{\alpha_j^{(2)},\alpha_j^{(2)\prime}} \beta_j\, \beta_j' \delta_{c_q^L(\alpha^{jL}o)c_q^L((\alpha^{jL})'o)^{-1}}.$$

Using the notation of (A.11) and (A.12), note that

$$\begin{aligned} c_q^L(\alpha^{jL}o)c_q^L((\alpha^{jL})'o)^{-1} &= \alpha_j^{(2)} \cdot g_{k_{L-1}} \cdots g_{k_1}(\alpha_j^{(2)\prime} \cdot g_{k_{L-1}} \cdots g_{k_1})^{-1} \\ &= \alpha_j^{(2)}(\alpha_j^{(2)\prime})^{-1}. \end{aligned}$$

We will now appeal to the following spectral gap modulo q for the group generated by the coefficients $\alpha_j^{(2)}(\alpha_j^{(2)\prime})^{-1}$.

Proposition A.6 (Spectral gap). *There are some modulus Q_0 and some $\epsilon > 0$ such that for all indices j, for all q coprime to Q_0 and for all $\phi \in \ell_0^2(G)$ with $\|\phi\|_2 = 1$ there is some pair $\alpha_j^{(2)}, \alpha_j^{(2)\prime}$ such that*

$$\|\alpha_j^{(2)}(\alpha_j^{(2)\prime})^{-1} * \phi - \phi\|_2 > \epsilon. \tag{A.18}$$

The statement of Proposition A.6 is well known to be equivalent to other uniform spectral gap properties. The uniform spectral gap is known to exist in the current setting for the following reasons.

Continued fractions setting. Here we need the products $\alpha_j^{(2)}(\alpha_j^{(2)'})^{-1}$ to generate a group with Zariski closure SL_2. Since all sequences of g_{i_j} are admissible, the $\alpha_j^{(2)}$ appearing in (A.14) do not depend on j. Recall that in the continued fractions setting, each g_i is already a product of two generators $\left(\begin{smallmatrix}0&1\\1&a\end{smallmatrix}\right)\left(\begin{smallmatrix}0&1\\1&b\end{smallmatrix}\right)$. It is easy to see then that the $\alpha_j^{(2)}$ generate a Zariski dense subgroup whenever the alphabet $\mathcal{A}$ of $\Gamma_{\mathcal{A}}$ has at least two letters, in fact, it would have been enough to take for the $\alpha^{(2)}$ blocks of length 1. On the other hand, we do need sufficiently many of the $\left(\begin{smallmatrix}0&1\\1&a\end{smallmatrix}\right)$ to be involved as the products $\left(\begin{smallmatrix}0&1\\1&a\end{smallmatrix}\right)\left(\begin{smallmatrix}0&1\\1&b\end{smallmatrix}\right)^{-1} = \left(\begin{smallmatrix}1&0\\a-b&1\end{smallmatrix}\right)$ are lower-triangular. Proposition A.6 then follows from the expansion result of Bourgain and Varjú [6]. In the cases that the $\alpha_j^{(2)}(\alpha_j^{(2)'})^{-1}$ generate all of $\mathrm{SL}_2(\mathbf{Z})$, Proposition A.6 is a well-known consequence of Selberg's "3/16 Theorem" from [18].

Schottky semigroup/group setting. Note that this setting contains the case that Γ is a Schottky *group* as in [4]. Again, it will be enough to show that the $\alpha_j^{(2)}(\alpha_j^{(2)'})^{-1}$ generate a Zariski dense sub*group* of $\mathrm{SL}_2(\mathbf{Z})$. This is the reason why we needed to make $\alpha_j^{(2)}$ a block of length 2. Indeed, suppose that the Schottky semigroup is generated by at least two Schottky generators and let g, h be two of these generators. For example, if α_j^{L-2} ends in g while α_{j-1}^{L-2} starts with g^{-1}, then the summation in (A.14) contains $\alpha_j^{(2)}$ of the form

$$gh, gh^{-1}, hg^{-1}, hh, h^{-1}g^{-1}, h^{-1}h^{-1}.$$

It is then easy to see that the $\alpha_j^{(2)}(\alpha_j^{(2)'})^{-1}$ generate a Zariski dense group (if Γ has more than two generators, this is also clear). We may then apply the Bourgain–Varjú expansion result [6] to obtain a spectral gap for the group generated by $\alpha_j^{(2)}(\alpha_j^{(2)'})^{-1}$. Now, this group and its generator set (and hence also its expansion constant ϵ as in (A.18)) depend on α_j^{L-2} and α_{j-1}^{L-2} (or rather just their starting/ending letters). But as Γ is finitely generated, only a finite number of groups/generators arise in this way, and we simply take ϵ to be the worst one, yielding Proposition A.6.

We now resume our proof of Theorem A.5. To this end, assume without loss of generality that $\|\varphi\|_2 = 1$ and let $\alpha_j^{(2)}, \alpha_j^{(2)'}$ be the pair provided by Proposition A.6 applied to φ, and ϵ the provided constant. Since there is a uniform bound on the size of the support of A, Lemma A.4 gives

$$\beta_j \beta_j' \gg \|A\|_1 \tag{A.19}$$

with an uniform positive implied constant (here $\beta_j\beta_j'$ is the coefficient of $\alpha_j^{(2)}(\alpha_j^{(2)'})^{-1}$ in A). It follows by routine arguments from (A.19) together with (A.18) for φ, with the associated ϵ, that the operator norm of $\widetilde{A}$ acting on $\ell_0^2(G)$ is $\|\widetilde{A}\|_{\mathrm{op}} \le (1-\epsilon')\|A\|_1$ for some ϵ' depending on ϵ. The resulting bound on (A.17) establishes Theorem A.5, since $\|A\|_1 = \|\eta_j\|_1^2$. □

Corollary 37. *Assume that L is sufficiently large (depending only on Γ). Then there is some $C_2 > 0$ also depending only on Γ so that, for any $\varphi \in L_0^2(G)$, we have*

$$\|\mu_1 * \varphi\|_2 \le (1-C_2)^R \|\mu_1\|_1 \|\varphi\|_2. \tag{A.20}$$

Proof. Beginning with (A.15), apply (A.16) R' times to get

$$\|\mu_1 * \varphi\|_2 \le \exp(c\gamma^{-L})^{R'-1} \sum^*_{\alpha_1^{L-1},\dots,\alpha_{R'}^{L-1}} (1-C_1)^{R'} \prod_{j=1}^{R'} \|\eta_j\|_1 \|\varphi\|_2.$$

Applying contraction yet again gives

$$\sum^*_{\alpha_1^{L-1},\dots,\alpha_{R'}^{L-1}} \prod_{j=1}^{R'} \|\eta_j\|_1 \le \exp(c\gamma^{-L})^{R'-1} \|\mu_1\|_1,$$

whence (A.20) follows on taking L large enough and recalling (A.5). □

Returning to the measure ν in (A.3), we have from (A.20) that

$$\|\nu * \varphi\|_2 \le (1-C_2)^R \|\nu\|_1 \|\varphi\|_2. \tag{A.21}$$

To conclude Theorem A.1, we need the following:

Lemma A.7. *Let μ be a complex distribution on the group $G = \mathrm{SL}_2(q)$ and assume that $|\mu| \le C\nu$. Let $E_q \subset L^2_0(G)$ be the subspace defined in Section* 5.1*, and let $A : E_q \to E_q$ be the operator acting by convolution with μ. Then*

$$\|A\| \le C' \left[\frac{|G| \|\widetilde{\nu} * \nu\|_2^2}{q} \right]^{1/4}.$$

Here $\widetilde{\mu}(g) = \overline{\mu(g^{-1})}$.

Proof. Note that the operator A^*A is self-adjoint, positive, and acts by convolution with $\widetilde{\mu} * \mu$. Let λ be an eigenvalue of A^*A. Since A acts on E_q, Frobenius gives that λ has multiplicity $\mathrm{mult}(\lambda)$ at least Cq. We then have that

$$\begin{aligned} \lambda^2 \mathrm{mult}(\lambda) \le \mathrm{tr}[(A^*A)^2] &= \sum_{g \in G} \langle (A^*A)^2 \delta_g, \delta_g \rangle = \sum_{g \in G} \|\widetilde{\mu} * \mu * \delta_g\|_2^2 \\ &= |G| \|\widetilde{\mu} * \mu\|_2^2 \le C^4 |G| \|\widetilde{\nu} * \nu\|_2^2. \end{aligned}$$

The claim follows, as $\|A\| = \max_\lambda \lambda^{1/2}$. □

We apply the lemma to μ in (A.1) using (A.4), giving

$$\|\mu * \varphi\|_2 \le Cq^{1/2} \|\widetilde{\nu} * \nu\|_2^{1/2}. \tag{A.22}$$

It remains to estimate the ν convolution.

Proposition A.8. *Choosing R to be of size $C \log q$ for suitable C, we have that*

$$\|\widetilde{\nu} * \nu\|_2 \le 2 \frac{\|\nu\|_1^2}{|G|^{1/2}}. \tag{A.23}$$

Proof. Let

$$\psi \equiv \delta_e - \frac{1}{|G|}\mathbf{1}_G \in L_0^2(G),$$

and note that $\|\psi\|_2 < 1$. Then

$$\begin{aligned}\|\widetilde{\nu} * \nu\|_2 = \|\widetilde{\nu} * \nu * \delta_e\|_2 &\leq \|\widetilde{\nu} * \nu * \left(\frac{1}{|G|}\mathbf{1}_G\right)\|_2 + \|\widetilde{\nu} * \nu * \psi\|_2 \\ &\leq \frac{\|\nu\|_1^2}{|G|^{1/2}} + \|\nu\|_1 \|\nu * \psi\|_2,\end{aligned}$$

where we used the triangle inequality and Cauchy–Schwarz. Since $\psi \in L_0^2(G)$, we apply inequality (A.21), giving

$$\|\nu * \psi\|_2 < (1 - C_2)^R \|\nu\|_1 < \frac{\|\nu\|_1}{|G|^{1/2}}$$

by a suitable choice of $R = C \log q$. The claim follows immediately. □

Finally, we give a proof of Theorem A.1.

Proof of Theorem A.1. Insert (A.23) into (A.22) and use (A.4) and $|G| > Cq^3$. Clearly (A.2) holds with $B = C\|\nu\|_1$. □

References

[1] *D. Borthwick*, Spectral theory of infinite-area hyperbolic surfaces, Progr. Math. **256**, Birkhäuser, Boston 2007.

[2] *J. Bourgain*, Partial quotients and representation of rational numbers, C. R. Math. Acad. Sci. Paris **350** (2012), no. 15–16, 727–730.

[3] *J. Bourgain*, Some Diophantine applications of the theory of group expansion, in: Thin groups and superstrong approximation, Math. Sci. Res. Inst. Publ. **61**, Cambridge University Press, Cambridge (2014), 1–22.

[4] *J. Bourgain*, *A. Gamburd* and *P. Sarnak*, Generalization of Selberg's $\frac{3}{16}$ theorem and affine sieve, Acta Math. **207** (2011), no. 2, 255–290.

[5] *J. Bourgain* and *A. Kontorovich*, On Zaremba's conjecture, Ann. of Math. (2) **180** (2014), no. 1, 137–196.

[6] *J. Bourgain* and *P. P. Varjú*, Expansion in $\mathrm{SL}_d(\mathbf{Z}/q\mathbf{Z}), q$ arbitrary, Invent. Math. **188** (2012), no. 1, 151–173.

[7] *D. Dolgopyat*, On decay of correlations in Anosov flows, Ann. of Math. (2) **147** (1998), no. 2, 357–390.

[8] *D. Frolenkov* and *I. D. Kan*, A strengthening of a theorem of Bourgain–Kontorovich II, Mosc. J. Comb. Number Theory **4** (2014), no. 1, 78–117.

[9] *A. Gamburd*, On the spectral gap for infinite index "congruence" subgroups of $\mathrm{SL}_2(\mathbf{Z})$, Israel J. Math. **127** (2002), 157–200.

[10] *S. Huang*, An improvement to Zaremba's conjecture, Geom. Funct. Anal. **25** (2015), no. 3, 860–914.

[11] *A. Kontorovich*, From Apollonius to Zaremba: Local-global phenomena in thin orbits, Bull. Amer. Math. Soc. (N.S.) **50** (2013), 187–228.

[12] *S. P. Lalley*, Renewal theorems in symbolic dynamics, with applications to geodesic flows, non-Euclidean tessellations and their fractal limits, Acta Math. **163** (1989), no. 1–2, 1–55.

[13] *A. Mohammadi* and *H. Oh*, Matrix coefficients, counting and primes for orbits of geometrically finite groups, J. Eur. Math. Soc. (JEMS) **17** (2015), no. 4, 837–897.

[14] *F. Naud*, Expanding maps on Cantor sets and analytic continuation of zeta functions, Ann. Sci. Éc. Norm. Supér. (4) **38** (2005), no. 1, 116–153.

[15] *H. Oh* and *D. Winter*, Uniform exponential mixing and resonance free regions for convex cocompact congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$, J. Amer. Math. Soc. **29** (2016), 1069–1115.

[16] *W. Parry* and *M. Pollicott*, Zeta functions and the periodic orbit structure of hyperbolic dynamics, Astérisque **187–188**, Société Mathématique de France, Paris 1990.

[17] *D. Ruelle*, An extension of the theory of Fredholm determinants, Publ. Math. Inst. Hautes Études Sci. **72** (1990), 175–193.

[18] *A. Selberg*, On the estimation of Fourier coefficients of modular forms, Proc. Sympos. Pure Math. **8** (1965), 1–15.

[19] *S. K. Zaremba*, La méthode des "bons treillis" pour le calcul des intégrales multiples, in: Applications of number theory to numerical analysis (Montreal 1971), Academic Press, New York (1972), 39–119.

Michael Magee, Mathematics Department, Yale University, New Haven, CT 06511, USA
e-mail: michael.magee@yale.edu

Hee Oh, Mathematics Department, Yale University, New Haven, CT 06511, USA;
and Korea Institute for Advanced Study, Seoul, Korea
e-mail: hee.oh@yale.edu

Dale Winter, Institute for Advanced Study, Princeton, NJ 08540, USA
e-mail: dale.alan.winter@gmail.com

Jean Bourgain, Institute for Advanced Study, Princeton, NJ 08540, USA
e-mail: bourgain@math.ias.edu

Alex Kontorovich, Rutgers University, New Brunswick, NJ, USA
e-mail: alex.kontorovich@rutgers.edu

Eingegangen 1. Februar 2016, in revidierter Fassung 28. November 2016